Textbook of Clinical Trials

Second Edition

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

Textbook of Clinical Trials

Second Edition

Edited by

David Machin

Division of Clinical Trials and Epidemiological Sciences,
National Cancer Centre, Singapore, UK Children's Cancer Study Group,
University of Leicester, UK and Medical Statistics Group, School of Health and Related Sciences,
University of Sheffield, UK

Simon Day

Medicines and Healthcare products Regulatory Agency,
London, UK

Sylvan Green

Arizona Cancer Centre, Tucson, Arizona, USA

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, ONT, L5R 4J3, Canada

Wiley also publishes its books in a variety of electronic formats. Some content that appears
in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data:

Textbook of clinical trials / edited by David Machin, Simon Day, Sylvan
Green. – 2nd ed.
p. ; cm.
Includes bibliographical references and index.
ISBN-13: 978-0-470-01014-3 (hardback : alk. paper)
ISBN-10: 0-470-01014-2 (hardback : alk. paper)
1. Clinical trials. I. Machin, David. II. Day, Simon. III. Green,
Sylvan B.
[DNLM: 1. Clinical Trials. 2. Research Design. QV 771 T3443
2006]
R853.C55T49 2006
610.72′4 – dc22
2006017012

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-470-01014-3
ISBN-10: 0-470-01014-2

Typeset in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry
in which at least two trees are planted for each one used for paper production.

Contents

INTRODUCTION 1

CANCER 47

Contributors

EDGARDO ABALOS
Centro Rosarino de Estudios Perinatales (CREP), Pueyrredón
985 – 2000, Rosario, Argentina

VATHSALA ANANTHARAMAN
Renal Medicine, Singapore General Hospital, Outram Road,
Singapore 169608, Singapore, Email: grmvat@sgh.com.sg

ERIK ANG
Erik Ang Plastic Surgery, Suite 1208, Mount Elizabeth
Medical Centre, Singapore 228510, Singapore, Email:
erik@angplasticsurgery.com

ANTONY J. ARTHUR
School of Nursing, Faculty of Medicine and Health Sciences,
University of Nottingham, Queen's Medical Centre,
Nottingham NG7 2UH, UK, Email:
tony.arthur@nottingham.ac.uk

VIJAYARAMAN ARUTCHELVAM
97 Darras Road, Darras Hall, Ponteland, Newcastle upon Tyne
NE20 9PQ, UK, Email: riarut@aol.com

PHILIP BATH
Division of Stroke Medicine, Institute of Neuroscience,
University of Nottingham, D Floor South Block, Queen's
Medical Centre, Nottingham NG7 2UH, UK, Email
philip.bath@nottingham.ac.uk

DONALD A. BERRY
Department of Biostatistics, University of Texas, MD
Anderson Cancer Center, 1515 Holcombe Blvd,
Box 0447, Houston, TX 77030, USA, Email:
dberry@odin.mdacc.tmc.edu

SILADITYA BHATTACHARYA
Department of Obstetrics and Gynaecology, University of
Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK, Email:
s.bhattacharya@abdn.ac.uk

AMAN U. BUZDAR

Department of Breast Medical Oncology, University of Texas,
MD Anderson Cancer Center, 1515 Holcombe Blvd,
Box 1354, Houston, TX 77030, USA

LIANA CAMPÓDONICO

Centro Rosarino de Estudios Perinatales (CREP), Pueyrredón
985 – 2000, Rosario, Argentina

EDITH D. CANBY-HAGINO

Division of Urology, Department of Surgery, University of
Texas Health Sciences Center, 7703 Floyd Curl Drive, San
Antonio, TX 78229, USA

DAVID CASARETT

University of Pennsylvania, Philadelphia, PA USA

RANDALL D. CEBUL

Center for Health Care Research and Policy, Case Western
Reserve University School of Medicine at MetroHealth
Medical Center, Cleveland, Ohio, USA, Email: rdc@case.edu

EDWIN CHAN

Clinical Trials and Epidemiology Research Unit, SingHealth
Health Services, Blk B #02-02, 226 Outram Road, Singapore
169039, Singapore, Email: edwin@cteru.com.sg

JONATHAN C.K. CLARKE

Moorfields Eye Hospital and Institute of Ophthalmology,
MEM & 100, University College London, 11–43 Bath
Street, London EC1V 9EL, UK, Email: j.clarke@ucl.ac.uk

MIKE CLARKE

UK Cochrane Centre, NHS R&D Programme, Summertown
Pavilion, Middle Way, Oxford OX2 7LG, UK, Email:
mclarke@cochrane.co.uk

NICOLA CRICHTON

RCN Development Centre, Faculty of Health, London South
Bank University, 103 Borough Road, London SE1 0AA, UK,
Email: crichtnj@lsbu.ac.uk

NEAL V. DAWSON

Center for Health Care Research and Policy, Case Western
Reserve University School of Medicine at MetroHealth
Medical Center, Cleveland, Ohio, USA, Email: nvd@case.edu

SIMON DAY

Licensing Division, Medicines and Healthcare products
Regulatory Agency, Room 13–205 Market Towers, 1 Nine
Elms Lane, London SW8 5NQ, UK, Email:
simon.day@mhra.gsi.gov.uk

ALLAN DONNER

Department of Epidemiology and Biostatistics, Faculty of
Medicine and Dentistry, The University of Western Ontario,
Kresge Building, Room K201, Ontario, Canada N6A 5C1,
Email: allan.donner@schulich.uwo.ca

GRAHAM DUNN

Biostatistics Group, Division of Epidemiology and Health
Sciences, University of Manchester, Stopford Building, Oxford
Road, Manchester M13 9PT, UK, Email:
graham.dunn@man.ac.uk

BATYA ELUL

Mailman School of Public Health, Columbia University, New
York, NY 10034, USA, Email: be2124@columbia.edu

TIMOTHY M.M. FARLEY

Department of Reproductive Health and Research, World
Health Organization, 1211 Geneva 27, Switzerland, Email:
farleyt@who.int

STEPHEN L. GEORGE

Department of Biostatistics and Bioinformatics, CALGB
Statistical Center, Duke University Medical Center,
Box 2714, Durham, NC 27710, USA, Email:
Stephen.George@duke.edu

RICHARD GOLDBERG

Division of Hematology/Oncology, The University of North
Carolina at Chapel Hill, CB# 7305, 3009 Old Clinic Bldg,
Chapel Hill, NC 27599-7305, USA

MICHAEL D. HUGHES

Center for Biostatistics in AIDS Research, Harvard School of
Public Health, 677 Huntington Avenue, Boston, MA 02115,
USA, Email: mhughes@sdac.harvard.edu

CAROL JAGGER

Trent Institute for HSR, Department of Epidemiology and
Public Health, University of Leicester, 22–28 Princess Road
West, Leicester LE1 6TP, UK, Email: cxj@leicester.ac.uk

DR OLGA JONASSON*

University of Illinois College of Medicine, Chicago, USA

ANDERS KÄLLÉN

Jullovsvagen 13, S-224 67 Lund, Sweden, Email:
anders.kallen@astrazeneca.com

JOHAN P.E. KARLBERG

Clinical Trials Centre, Faculty of Medicine, The University of
Hong Kong, Hong Kong SAR, PR China, Email:
jpekarl@hkucc.hku.hk

PENG-TEE KHAW

Moorfields Eye Hospital and Institute of Ophthalmology,
University College London, 11–43 Bath Street, London EC1V
9EL, UK, Email: p.khaw@ucl.ac.uk

KYUNGMANN KIM

Department of Biostatistics and Medical Informatics,
University of Wisconsin–Madison, K6/446 Clinical Sciences
Center, 600 Highland Avenue, Madison, WI 53792-4675, USA

*Dr Olga Jonasson died in August 2006 whilst this book was in production.

INGELA KRANTZ
Skaraborg Institute, Stationsgatan 12, SE-541 30 Skövde,
Sweden, Email: ingela.krantz@skaraborg-institute.se

ROBERT S. KROUSE
Surgical Care Line, 2–112, Southern Arizona Veterans Affairs
Health Care System, 3601 S. 6th Avenue, Tucson, AZ
85723, USA, Email: robert.krouse@med.va.gov

PHILIP W. LAVORI
VA Medical Centre, 795 Willow Road, Menlo Park, CA
94025-2539, USA, Email: philip.lavori@med.va.gov

PING-CHUNG LEUNG
Institute of Chinese Medicine, Room 74026, 5/F, Clinical
Sciences Building, Prince of Wales Hospital, Shatin, Hong
Kong, Email: pingcleung@cuhk.edu.hk

PAUL LIMBERG
Gastroenterology and Hepatology, Mayo Clinic, 200 1st Street
SW, Rochester, MN 55905, USA, Email:
limburg.paul@mayo.edu

PING-YU LIU
Fred Hutchinson Cancer Research Centre, Seattle, WA
98109-1024, USA Email: pliu@fhcrc.org

EDWARD C.M. LO
Dental Public Health, Faculty of Dentistry, The University
of Hong Kong, 34 Hospital Road, Sai Ying Pun,
Hong Kong

THOMAS E. LOVE
Center for Health Care Research and Policy, Case Western
Reserve University School of Medicine at MetroHealth
Medical Center, Cleveland, Ohio, USA, Email: tel3@case.edu

DAVID MACHIN
Division of Clinical Trials and Epidemiological Sciences,
National Cancer Centre Singapore, 11 Hospital Drive,
Singapore 16910, Singapore, Email: ctedav@nccs.com.sg,
david.machin1@onetel.net

ISHA MALIK
Specialist Registrar in Endocrinology, 154 Hampstead Road,
Newcastle upon Tyne NE4 8AB, UK, Email:
ishamalik@doctors.org.uk

COLMAN McGRATH
Dental Public Health, Faculty of Dentistry, The University of
Hong Kong, 34 Hospital Road, Sai Ying Pun, Hong Kong

COSETTA MINELLI
Centre for Biostatistics and Genetic Epidemiology, Department
of Health Sciences, University of Leicester, 22–28 Princess
Road West, Leicester LE1 6TP, UK, Email: cm109@le.ac.uk

JILL MOLLISON

Centre for Statistics in Medicine, Wolfson College Annexe,
Oxford OX2 6UD, UK

SHARON B. MURPHY

Children's Cancer Research Institute, Mail Code 7784, The
University of Texas Health Science Center at San Antonio,
7703 Floyd Curl Drive, San Antonio, TX 78229-3900, USA,
Email: murphysb@uthscsa.edu

PAUL S. MYLES

Department of Anaesthesia and Pain Management, Alfred
Hospital, Commercial Road, Melbourne, Victoria
3004, Australia, Email: p.myles@alfred.org.au

LUIGI NALDI

Clinica Dermatologica, Ospedali Riuniti di Bergamo, L. go
Barozzi 1, 24100 Bergamo, Italy, Email: luigi.naldi@gised.it

PETROS PERROS

Endocrine Unit, Freeman Hospital, Freeman Road, Newcastle
upon Tyne NE7 7DN, UK, Email: petros.perros@ncl.ac.uk

GILDA PIAGGIO

Department of Reproductive Health and Research, World
Health Organization, 1211 Geneva 27, Switzerland, Email:
piaggiog@who.int

KRISHNA PRASAD

Licensing Division, Medicines and Healthcare products
Regulatory Agency, 1 Market Towers, Nine Elms Lane,
London SW8 5NQ, UK, Email:
krishna.prasad@mhra.gsi.gov.uk

DOMENIC REDA

Cooperative Studies Program Coordinating Center (151K),
Hines VA Hospital, PO Box 5000-151K, Hines, IL
60141, USA, Email: Domenic.Reda@va.gov

JOHN RIGG

Department of Anaesthesia and Clinical Perioperative Services,
Armadale Health Service, PO Box 460, Armadale, Western
Australia 6992, Australia, Email: jrarigg@cyllene.uwa.edu.au

DANIEL SARGENT

Mayo Clinic, 200 First Street, SW, Kahler 1A, Rochester, MN
55905, USA, Email: sargent.daniel@mayo.edu

CHARLES A. SCHIFFER

Division of Hematology/Oncology, Karmanos Cancer Institute,
Wayne State University School of Medicine, Detroit, MI
48201, USA, Email: schiffer@karmanos.org

JOAN H. SCHILLER

Division of Hematology/Oncology, University of Texas
Southwestern Medical Center, Dallas, Texas, USA

M. KATHERINE SHEAR
Panic, Anxiety and Traumatic Grief Program, Western Psychiatric Institute and Clinic, 3811 O'Hara Street, Pittsburgh, PA 15213, USA, Email: shearmk@msx.upmc.edu

JONATHAN J. SHUSTER
College of Medicine, University of Florida, PO Box 100212, Gainesville, FL 32610-0212, USA, Email: jshuster@biostat.ufl.edu

TERRY L. SMITH
University of Texas, MD Anderson Cancer Center, 1515 Holcombe Blvd, Box 0447, Houston, TX 77030, USA, Email: tsmith@odin.mdacc.tmc.edu

VERNON K. SONDAK
H. Lee Moffitt Cancer Center, The University of South Florida, 12902 Magnolia Dr., Tampa, FL 33612, USA, Email: sondakvk@moffitt.usf.edu

COLIN SONG
Department of Plastic, Reconstructive and Aesthetic Surgery, Singapore General Hospital, Outram Road, Singapore 169608, Singapore, Email: gplson@sgh.com.sg

CATHERINE M. TANGEN
The Southwest Oncology Group, 1730 Minor Ave., Suite 1900, Seattle, WA 98101-1468, USA; and the Fred Hutchinson Cancer Research Center, PO Box 19024, Seattle, WA 98109-1024, USA

NICHOLAS TARRIER
University of Manchester, Academic Division of Clinical Psychology, Education and Research Building, Wythenshawe Hospital, Southmoor Road, Manchester M23 9LT, UK, Email: ntarrier@man.ac.uk

IAN M. THOMPSON
Division of Urology, Department of Surgery, University of Texas Health Sciences Center, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA, Email: thompsoni@uthscsa.edu

JOSÉ VILLAR
Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford OX3 9DU, UK, Email: villarj@who.int

BEVERLY WINIKOFF
Gynuity Health Projects, New York, NY 10010, USA, Email: bwinikoff@gynuity.org

MAY C.M. WONG
Dental Public Health, Faculty of Dentistry, The University of Hong Kong, 34 Hospital Road, Sai Ying Pun, Hong Kong, Email: mcmwong@hkucc.hku.hk

WOFFLES WU

Aesthetic and Laser Centre, Suite 09–02, Camden Medical
Centre, 1 Orchard Boulevard, Singapore 249615, Singapore,
Email: woffles@magix.com.sg

TIL WYKES

Institute of Psychiatry, De Crespigny Park, London SE5 8AF,
UK, Email: t.wykes@iop.kcl.ac.uk

Preface to Second Edition

Although only three years have passed since the first edition of the *Textbook of Clinical Trials*, there has been sufficient interest in the book that a revised version has been made possible. Thus, while the objectives remain the same, this very much expanded volume describes an even wider range of medical areas in which clinical trials have had impact. These additional areas include: several cancers, primary care and health services research, maternal and perinatal health and early termination of pregnancy, anaesthesia, pain and surgery, transplantation, infectious diseases and HIV, wound healing, additional respiratory diseases, stroke, ophthalmology, palliative care and trials in nursing care, as well as considerations for evaluating therapies in very rare diseases.

We are grateful to those authors of the first edition who have refreshed their chapters, and also to the new authors who have enabled this expanded edition to be produced.

Clinical trials continue to be essential to society in the evaluation of therapies for existing and new diseases. As we went to press for the first edition, the SARS epidemic was emerging. As we go to press with this second edition, the threat of avian flu is high on the agenda of health care providers, governments and the media. The need to evaluate efficiently and reliably the benefits and risks of potential medicines for a wide range of diseases remains with us.

David Machin, Simon Day and Sylvan Green

Preface to the First Edition

This *Textbook of Clinical Trials* is not a textbook of clinical trials in the traditional sense. Rather, it catalogues in part the impact of clinical trials – particularly the randomised controlled trial – on both the practice of medicine and allied fields and on the developments and practice of medical statistics. The latter has evolved in many ways through the direct needs of clinical trials and the consequent interaction of statistical and clinical disciplines. The impact of the results from clinical trials, particularly the randomised controlled trial, on the practice of clinical medicine and other areas of health care has been profound. In particular, they have provided the essential under-pinning to evidence-based practice in many disciplines and are one of the key components for regulatory approval of new therapeutic approaches throughout the world.

Probably the single most important contribution to the science of comparative clinical trials was the recognition, more than 50 years ago, that patients should be allocated to the options under consideration at random. This was the foundation for the science of clinical trial research and placed the medical statistician at the centre of the process. Although the medical statistician may be at the centre, he or she is by no means alone. Indeed the very nature of clinical trial research is multidisciplinary so that a 'team' effort is always needed from the concept stage, through design, conduct, monitoring and reporting.

Some of the developments impacting on clinical trials have been truly statistical in nature, for example Cox's proportional hazards model, while others such as the intention-to-treat (ITT) principle are – in some sense – based more on experience. Other important statistical developments have not depended on technical advancement, but rather on conceptual advancement, such as the now standard practice of reporting confidence intervals rather then relying solely on p-values at the interpretation stage. Of major importance over this same time period has been the expansion in data processing capabilities and the range of analytical possibilities only made possible by the tremendous development in computing power. However, despite many advances, the majority of randomised controlled trials remain simple in design – most often a comparison between two randomised groups.

On the medical side there have been many changes including new diseases that raise new issues. Thus, as we write, SARS has emerged: the final extent of the epidemic is unknown, diagnosis is problematical and no specific treatment is available. In more established diseases there have been major advances in the types of treatment available, be they in surgical technique, cancer chemotherapy or psychotropic drugs. Advances in medical and associated technologies are not confined to curative treatments but extend, for example, to diagnostic methods useful in screening for disease, vaccines for disease prevention, drugs and devices for female and male contraception, and pain relief and psychological support strategies in palliative care.

Clinical trials imply some intervention affecting the subjects who are ultimately recruited into them. These subjects will range from the very healthy, perhaps women of a relatively young age recruited to a contraceptive development trial, to those (perhaps elderly) patients in terminal decline from a long-standing illness. Each group studied in a clinical trial, from unborn child to aged adult, brings its own constraint on the ultimate design of the trial in mind. So too does the relative efficacy of the current standard. If the outcome is death and the prognosis poor, then bolder steps may be taken in the choice of treatments to test. If the disease is self-limiting or the outcome cosmetic then a more conservative approach to treatment options would be justified.

In all this activity the choice of clinical trial design and its ultimate conduct are governed by essential ethical constraints, the willingness of subjects to consent to the trial in question and their right to withdraw from the trial should they wish.

Thus the *Textbook of Clinical Trials* addresses some of these and many other issues as they impact on patients with cancer, cardiovascular disease, dermatological, dental, mental and ophthalmic health, gynaecology and respiratory diseases. In addition, chapters deal with issues relating to complementary medicine, contraception and special issues in children and special issues in older patients. A brief history of clinical trials and a summary of some pertinent statistical issues are included.

David Machin, Simon Day and Sylvan Green

INTRODUCTION

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

1

The Development of Clinical Trials

SIMON DAY
Licensing Division, Medicines and Healthcare products Regulatory Agency, London, UK

INTRODUCTION

This chapter traces some of the development of clinical trials – from very early anecdotal reports of informal evaluations of medicines (some not necessarily considered 'medicines' by today's standards), medical practices, and so on – through to modern, well established principles, which include blinding, randomisation, clear protocols and analysis plans, etc. Some events have been milestones, whilst others have contributed in more modest ways to, what is now often considered as, the 'gold standard' of evidence for evaluating therapies.

EARLIEST STORIES

The modern-day birth of clinical trials is usually considered to be the publication by the UK Medical Research Council in 1948 of a trial for the treatment of pulmonary tuberculosis with streptomycin, and we will return to this example later in the chapter. However, earlier but less well-documented examples do exist. The comparative concept of assessing therapeutic efficacy has been known from ancient times. Lilienfeld[1] cites a description of a nutritional experiment involving a control group in the Book of Daniel from the Old Testament:

> Then Daniel said to the guard whom the master of the eunuchs had put in charge of Hananiah, Mishael, Azariah and himself, 'Submit to us this test for ten days. Give us only vegetables to eat and water to drink; then compare our looks with those of the young men who have lived on the food assigned by the king, and be guided in your treatment of us by what you see.' The guard listened to what they said and tested them for ten days. At the end of ten days they looked healthier and were better nourished than all the young men who had lived on the food assigned them by the king. So the guard took away the assignment of food and the wine they were to drink, and gave them only the vegetables.

Daniel lived around the period 800 BC and although it may not be possible to confirm the accuracy of the account, what is clear is that when this passage was written – around 150 BC – the ideas certainly existed.

The passage from Daniel describes not just a control group, but a *concurrent* control group.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

This fundamental element of comparative, experimental (and in this case clinical) research did not begin to be widely practised until the latter half of the twentieth century.

Much later than the book of Daniel, but still very early, is an example from the fourteenth century: it is a letter from Petrarch to Boccaceto cited by Witkosky:[2]

> I solemnly affirm and believe, if a hundred or a thousand of men of the same age, same temperament and habits, together with the same surroundings, were attacked at the same time by the same disease, that if one followed the prescriptions of the doctors of the variety of those practicing at the present day, and that the other half took no medicine but relied on Nature's instincts, I have no doubt as to which half would escape.

During the fourteenth to sixteenth centuries, the Renaissance period was a time of great development in many forms ranging from art to science. This period provides other examples including an unplanned comparison of treatment of battlefield wounds. Packard[3] describes how, during a battle to capture the castle of Villaine in 1537, the surgeon Ambroise Paré was using the standard treatment (*sic*) of pouring boiling oil over soldiers' wounds. During the battle, he ran out of oil so he resorted to using a mixture of egg yolks, oil of roses and turpentine. The reason for this particular concoction seems unknown. The superiority of the new 'treatment' became evident the next day:

> I raised myself very early to visit them, when beyond my hope I found those to whom I applied the digestive medicament feeling but little pain, their wounds neither swollen nor inflamed, and having slept through the night. The others to whom I had applied the boiling oil were feverish with much pain and swelling about their wounds. Then I determined never again to burn thus so cruelly by arquebusses.

Today, we might (at best) call such an experience a 'natural experiment'; at worst we would simply consider it an anecdotal experience, completely confounded with time and so possibly also type and severity of wounds, weather conditions and a host of other unknown factors.

Perhaps the most famous historical example of a planned, prospective controlled, comparative, clinical trial is from the eighteenth century: that where Lind[4] found oranges and lemons to be the most effective of six dietary treatments for scurvy on board ships. His account (reproduced from Anderson)[5] reads thus:

> On 20th of May, 1747, I took 12 patients in the scurvy, on board the *Salisbury* at sea. The cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude with weakness of their knees. They laid together in one place being the proper treatment for the sick in the forehold; and had one diet common to all, water gruel, sweetened with sugar in the morning; fresh mutton broth often for dinner; at other times pudding, boiled biscuit with sugar and for supper, barley and raisins, rice and currants, sago and wine, or the like. Two of these were ordered each a quart of cider a day. Two others took 25 drops of elixir vitriol three times a day upon an empty stomach; using a gargle strongly acidulated with it for their mouths. Two others took two spoonfuls of vinegar three times a day upon an empty stomach, having the gruel and other food well acidulated with it, and also the gargle for their mouths. Two of the worse patients were put on a course of sea water. Of this they drank half a pint a day and sometimes more or less as it operated by way of gentle physic. Two others had each two oranges and one lemon given them every day. These they eat with greediness at different times, upon an empty stomach. They continued for six days under this course, having consumed the quantity that could be spared. The two remaining patients took the bigness of a nutmeg three times a day of an electuary recommended by a hospital surgeon, made of garlic, mustard seed, *Rad. Raphan.*, balsam of Peru, and gum Myrrh, using for drink barley water well acidulated with tamarinds; by a concoction of which with the addition of cream of tartar they were greatly purged three or four times during the course.
>
> The consequence was that the most sudden and good effects were perceived from the use of oranges and lemons, one of those who had taken them being

> at the end of six days fit for duty. The spots were not indeed at that time quite off his body, nor his gums sound; but without any other medicine, than a gargarism of elixir vitriol, he became quite healthy before we came into Plymouth, which was on the 16th June. The other was the best recovered of any in his condition, and being now deemed pretty well, was appointed nurse to the rest of the sick.

Pierre-Charles-Alexandre Louis,[6] a nineteenth-century clinician and pathologist, introduced the numerical aspect to comparing treatments. His idea was to compare the results of treatments on groups of patients with similar degrees of disease (which is not quite the case with Lind), and so to truly compare 'like with like':

> I come now to therapeutics, and suppose that you have some doubt as to the efficacy of a particular remedy: How are you to proceed? . . . You would take as many cases as possible, of as similar a description as you could find, and would count how many recovered under one mode of treatment, and how many under another; in how short a time they did so, and if the cases were in all respects alike, except in the treatment, you would have some confidence in your conclusions; and if you were fortunate enough to have a sufficient number of facts from which to deduce any general law, it would lead to your employment in practice of the method which you had seen oftenest successful.

'Like with like' was an important step forward from Lind's investigation of the treatment of scurvy. Note, although early in Lind's passage he says that 'Their cases were as similar as I could have them', later he acknowledges (partly through a clear and detailed description of the study) that the two worst cases both received the same treatment: 'Two of the worst patients were put on a course of sea water.' His use of the verb 'put' (rather than, perhaps, 'received') implies an intention on Lind's part. Perhaps he expected that the sea water might be the best treatment. It was more than a century later when Bradford Hill used a formal randomisation procedure for creating groups of cases that were 'in all respects alike, except in the treatment'.

RANDOMISATION

The use of randomisation was a major contribution to experimental design, put forward by the statistician and geneticist R.A. Fisher in agricultural trials (see, for example, Fisher,[7] Fisher and McKenzie).[8] Fisher randomised plots of crops to receive different treatments. In clinical trials, there had been early schemes to use 'group randomisation' whereby patients were divided into two groups and then the treatment for each group was randomly assigned. The Belgian medicinal chemist van Helmont[9] described an early example of this:

> Let us take out of the hospitals, out of the Camps, or from elsewhere, 200, or 500 poor People that have Fevers, Pleurisies, &c, Let us divide them into halves, let us cast lots, that one half of them may fall to my share, and the others to yours . . . we shall see how many funerals both of us shall have: *But* let the reward of the contention or wager, be 300 florens, deposited on both sides.

Considering modern-day standards of trials it is interesting to compare and contrast features such as:

- a description of some sort of inclusion criteria;
- a pre-specified, clinically relevant, endpoint (although today we might use the more politically correct term 'all-cause mortality'); and
- some indication of sample size (although not very definitively chosen).

More recently, Amberson and McMahon[10] used group randomisation in a trial of sanocrysin for the treatment of pulmonary tuberculosis. Today, the more common term to describe such trials is 'cluster' randomised trials; a good review is contained in an issue of the review journal *Statistical Methods in Medical Research* (see Donner and Klar).[11]

Systematic assignment was used by Fibiger,[12] who alternately assigned diphtheria patients to serum treatment or an untreated control group. Alternate assignment is frowned upon today, partly because knowledge of the future treatment allocations may selectively bias the admission of patients into the treatment group,[13] also because any unknown patterns of patient presentation may turn out to be correlated with the treatment assignment. 'Proper' randomisation[14] will avoid this possibility. Diehl *et al.*[15] reported a common cold vaccine study with University of Minnesota students as subjects where proper random assignment and blinding of patients to treatments appears to have been used:

> At the beginning of each year . . . students were assigned at random . . . to a control group or an experimental group. The students in the control groups . . . received placebos . . . All students thought they were receiving vaccines . . . Even the physicians who saw the students . . . had no information as to which group they represented.

However, Gail[16] points out that although this appears to be a randomised clinical trial, a further unpublished report by Diehl clarifies that this is another instance of systematic assignment:

> At the beginning of the study, students who volunteered to take these treatments were assigned alternately and without selection to control groups and experimental groups.

Bradford Hill, in the study of streptomycin in pulmonary tuberculosis,[17] used random sampling numbers in assigning treatments to subjects, so that the subject was the unit of randomisation. This study is now generally acknowledged to be the 'first properly randomised clinical trial' – although it was not fully blinded, as discussed below.

Later, Bradford Hill and the British Medical Research Council continued with further randomised trials: chemotherapy of pulmonary tuberculosis in young adults,[18] antihistaminic drugs in the prevention and treatment of the common cold,[19] cortisone and aspirin in the treatment of early cases of rheumatoid arthritis,[20,21] and long-term anticoagulant therapy in cerebrovascular disease.[22]

BLINDING

The common cold vaccine study published by Diehl *et al.*[15] cited earlier, in which University of Minnesota students were alternately assigned to vaccine or placebo, was a masked (or blinded) clinical trial:

> All students thought they were receiving vaccines . . . Even the physicians who saw the students . . . had no information as to which group they represented.

Partial blinding was used in the early Medical Research Council trials in which Bradford Hill was involved. Thus, in the first of those trials, the study of streptomycin in tuberculosis,[17] although patients and their treating physicians were not blinded to the treatment assignment, the X-ray films were viewed by two radiologists and a clinician, each reading the films independently and not knowing if the films were of C (control, bed-rest alone) or S (streptomycin and bed-rest) cases.

Bradford Hill[23] noted in respect of using such blinding and randomisation:

> If [the clinical assessment of the patient's progress and of the severity of the illness] is to be used effectively, without fear and without reproach, the judgements must be made without any possibility of bias, without any overcompensation for any possible bias, and without any possible accusation of bias.

Simply overcoming bias may not be sufficient: overcoming any possible *accusation of bias* is an important justification for blinding and randomisation. It is not clear if Bradford Hill considered the blind assessment of the X-rays (hence, the outcome measure) was adequate, or whether blinding of patients and treating physicians was necessary. Today, blinding (including treatment allocation concealment) and randomisation are considered the two most important (although not necessarily completely adequate) aspects of a good, well-controlled clinical trial.

In the second MRC trial, the antihistamine common cold study,[19] placebos, indistinguishable from the drug under test, were used. Here, Bradford Hill noted:

> in [this] trial . . . feelings may well run high . . . either of the recipient of the drug or the clinical observer, or indeed of both. If either were allowed to know the treatment that had been given, I believe that few of us would without qualms accept that the drug was of value – if such a result came out of the trial.

In the United States, the National Institutes of Health started their first randomised trial in 1951. It was a National Heart Institute study of adrenocorticotropic hormone (ACTH), cortisone and aspirin in the treatment of rheumatic heart disease.[24] This was followed in 1954 by a randomised trial of retrolental fibroplasia (now known as retinopathy of prematurity), sponsored by the National Institute of Neurological Diseases and Blindness.[25] During the four decades following the pioneering trials of the 1940s and 1950s, there was a large growth in the number of randomised trials not only in Britain and the United States, but also in Canada and mainland Europe.

ETHICS

Experimentation in medicine is as old as medicine itself and there is nothing necessarily wrong with that. Some experiments on humans have, however, been conducted without concern for the welfare of the subjects, who may have been prisoners or disadvantaged people. Katz[26] provides examples of nineteenth-century studies in Russia and Ireland of the consequences of intentionally infecting people with syphilis and gonorrhoea. McNeill[27] describes how, during the same period in the United States, physicians put slaves into pit ovens to study heat stroke, and poured scalding water over them as an experimental cure for typhoid fever. He even describes how one slave had two fingers amputated in a 'controlled trial', one finger with anaesthetic and one without! The benefits of the strength of causal evidence obtained from a well-controlled trial hardly outweigh the ethical unacceptability.

Unethical experiments on human beings have continued into the twentieth century and have been described by, for example, Beecher,[28] Freedman[29] and McNeil.[27] In 1932 the US Public Health Service began a study in Tuskegee, Alabama, of the natural progression of untreated syphilis in 400 black men. The intentional withholding of treatment may be the first point of unacceptable ethics; the fact that the experiment was restricted to black men only (while whites received treatment) is yet further concern. It is quite remarkable that the study continued right up until 1972 when a newspaper reported that the subjects were uninformed or misinformed about the purpose of the study.[29] Shirer,[30] amongst others, describes how during the Nazi regime from 1933 to 1945, German doctors conducted experiments, mainly on Jews, but also on Gypsies, mentally disabled persons, Russian prisoners of war and Polish concentration camp inmates. The Nazi doctors were later tried and found guilty of these atrocities in 1946–7 at Nuremberg and this led to the writing, by three of the trial judges, of the Nuremberg Code (see US Government Printing Office).[31] This was the first international effort to lay down ethical principles of clinical research. Principle 1 of the Code states:

> The voluntary consent of the human subject is absolutely essential. This means that the person involved should have legal capacity to give consent; should be so situated as to be able to exercise free power of choice, without the intervention of any element of force, fraud, deceit, duress, overreaching, or other ulterior form of constraint or coercion; and should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision.

Other principles of the Code are that experiments should yield results for the good of society, that unnecessary suffering and injury should be avoided, and that the subject should be free to withdraw from the experiment at any time and for any reason. Modern-day standards of acceptable ethics go beyond simply obtaining the patient's consent: a study that is unethical from the physician's (or society's) point of view cannot be considered acceptable simply by a patient giving his or her consent.

Other early advocates of informed consent were Charles Francis Withington and William Osler. Withington[32] realised the 'possible conflict between the interests of medical science and those of the individual patient', and concluded in favour of 'the latter's indefensible rights'. Osler[33] insisted on informed consent in medical experiments. Despite this early advocacy, and the 1946–7 Nuremberg Code, the application of informed consent to medical experiments did not take hold until the 1960s. Bradford Hill,[34] based on his experience in a number of early randomised clinical trials sponsored by the MRC, believed that it was not feasible to draw up a detailed code of ethics for clinical trials that would cover the variety of ethical issues that came up in these studies. He also considered that the patient's consent was not warranted in all clinical trials – a view that would not be generally supported today. Gradually the medical community has come to recognise the need to protect the reputation and integrity of medical research (as well as protecting patients and research subjects) and in 1955 a human experimentation code was adopted by the Public Health Council in the Netherlands.[35] Later, in 1964, the World Medical Assembly issued the Declaration of Helsinki[36] essentially adopting the ethical principles of the Nuremberg Code, with consent being 'a central requirement of ethical research' (see Faden *et al.*).[37] The Declaration of Helsinki has been updated and amended several times: Tokyo 1975, Venice 1983, Hong Kong 1989, Cape Town 1996 and Edinburgh 2000.

THE PHARMACEUTICAL INDUSTRY

The impact of advances in clinical trial thinking has had a major impact on the pharmaceutical industry, which relies heavily on trials for providing evidence of efficacy and safety of the products that it develops. However, in return, the industry has had a major influence on standards of clinical trials for exactly similar reasons – namely that it carries out so many of them.

Until around about the time of the thalidomide disaster (see, for example, Shah),[38] new medicines were licensed largely upon evidence that they were safe. Efficacy was less of an issue. Changes were introduced following thalidomide (although it should be noted that this was not the only product that prompted changes). It is interesting to observe that in Britain, the body that advises the Licensing Authority has, until recently, been called the Committee on Safety of Medicines (see, for example, Day).[39] In October 2005, along with other changes to the procedural aspects of licensing medicines (although not the scientific aspects), a new body, the Commission on Human Medicines, was established. It has a similar remit to the former Committee, although its change of name – clearly to encompass more than only safety – is a better descriptor.

Setting aside the semantics of the naming of advisory committees, licensing of new medicines today requires (amongst other things) clear and convincing evidence of efficacy. This, of course, best comes from high-quality clinical trials. Numerous guidelines (to guide industry as well to set common standards for assessment) have been developed covering many aspects of drug development and the demonstration of safety and efficacy in clinical trials. Amongst these – and being the most over arching – are those of the International Conference on Harmonisation (www.ich.org) and of the guidelines produced by that body, 'E9' covers statistical principles for clinical trials.[40,41] That document has served as an excellent state of the art for most aspects of clinical trial design, conduct, analysis and reporting. However, with ever-increasing commercial pressures to bring new products to the marketplace more quickly, statisticians and other scientists working in the pharmaceutical industry have a keen interest to use – and often contribute to – new developments in clinical trial design and analysis.

DATA MONITORING

In the modern randomised clinical trial, particularly for trials of life-threatening conditions, the accumulating data are often monitored for safety and efficacy by an independent data monitoring committee (see, for example, Ellenberg *et al.*)[42] One of the earliest examples of this was in 1968 when such a committee was established to serve the Coronary Drug Project, a large multicentre trial sponsored in the United States by the National Heart Institute of the National Institutes of Health.[43,44] In 1967, after a presentation of interim outcome data by the study coordinators to all participating investigators, Thomas Chalmers, clearly with great insight, wrote to the policy board chairman expressing his concern:

> that knowledge by the investigators of early nonstatistically significant trends in mortality, morbidity, or incidence of side effects might result in some investigators – desirous of treating their patients in the best possible manner, i.e., with the drug that is ahead – pulling out of the study or unblinding the treatment groups prematurely.

We can note here the distinction between collective ethics and individual ethics – what is best for the trial, as opposed to what might be best for the individual patients. Following this letter, a more formal data and safety monitoring committee was established for the Coronary Drug Project consisting of scientists who were not contributing data to the study. Thereafter, the practice of sharing accumulating outcome data with the study's investigators, and others closely connected with the study, was discontinued. The data safety and monitoring committee assumed responsibility for deciding when the accumulating data warranted changing the study protocol or terminating the study.

The first formal recognition of the need for interim analyses, and the recognition that such analyses affect the probability of the Type I error, came with the publication in the 1950s of papers on sequential clinical trials by Bross[45] and Armitage.[46] The principal advantage of a sequential trial over a fixed sample size trial is that, when the length of time needed to reach an endpoint is short, e.g. weeks or months, the sample size required to detect a substantial benefit from one of the treatments is reduced from what it would be in a more traditional 'fixed sample size' design.

In the 1970s and 1980s solutions to interim analysis problems came about in the form of group sequential methods and stochastic curtailment.[47–49] In the group sequential trial, the frequency of interim analyses is usually limited to a small number, say between three and six. The boundaries proposed by Pocock[50] use constant nominal significance levels; those proposed by Haybittle[51] and Peto[52] use stringent significance levels for all except the final test; in the O'Brien–Fleming[53] method, stringency gradually decreases; in the method by Lan and DeMets,[54] the total Type I error probability is gradually spent in a manner that does not require the timing of analyses to be prespecified. More details of these newer methods in the development of clinical trials are given in the next chapter.

RECENT YEARS . . .

In recent years we have seen a huge increase in the number of trials carried out and published, and in the advancement of methodological aspects relating to trials. Whilst many see the birth of clinical trials (certainly in their modern-day guise) as being the MRC streptomycin trial,[17] there remains some controversy (see, for example, D'Arcy Hart,[55,56] Gill[57] and Clarke[58]). However, it is interesting to note that one of the most substantial reviews of historical aspects of trials is based on Bull's work for a 1951 MD thesis.[59] He cites 135 historical examples and other supporting references – but no mention of Bradford Hill and the MRC. The modern-day story of clinical trials perhaps begins where Bull ended.

Today, there are many academic papers devoted to the methodology of clinical trials; there are many books on the general methods of trials, as well as others on specific technical points of trials and those in specific therapeutic areas. There are journals specifically devoted to clinical trials (*Clinical Trials: Journal of the Society for Clinical Trials* and *Contemporary Clinical Trials*) and there is a professional society – the Society for Clinical Trials (www.sctweb.org).

ACKNOWLEDGEMENTS

An earlier version of this chapter was based on work by F. Ederer, in: Armitage P and Colton T (Eds) *Encyclopedia of Biostatistics*, 1998: John Wiley & Sons, Ltd, Chichester.

REFERENCES

1. Lilienfeld AM. Ceteris paribus: the evolution of the clinical trial. *Bull Hist Med* (1982) **56**: 1–18.
2. Witkosky SJ. *The Evil That Has Been Said of Doctors: Extracts From Early Writers*, trans with annotations by TC Minor, The Cincinnati Lancet-Clinic, Vol. 4l, New Series Vol. 22, 447–448 (l889).
3. Packard FR. *The Life and Times of Ambroise Paré*, 2nd ed. New York: Paul B. Hoeber (l92l) 27, 163.
4. Lind J. *A Treatise of the Scurvy*. Edinburgh: Sands Murray Cochran, (1753) 191–3.
5. Anderson M. *An Introduction to Epidemiology*. London: Macmillan, (1980) 163–4.
6. Louis PCA. The applicability of statistics to the practice of medicine. *London Med Gaz* (1837) **20**: 488–491.
7. Fisher RA. The arrangement of field experiments. *J Min Agric* (1926) **33**: 503–13.
8. Fisher RA, and McKenzie WA, Studies in crop variation: II. The manurial response of different potato varieties. *J Agric Sci* (1923) **13**: 315.
9. van Helmont JB. *Oriatrike or Physik Refined*, trans J Chandler. London: Lodowick Loyd (l662).
10. Amberson B, McMahon PM. A clinical trial of sanocrysin in pulmonary tuberculosis. *Am Rev Tuberc* (1931) **24**: 401.
11. Donner A, Klar N. Cluster randomization trials (editorial). *Stati Methods Med Res* (2000) **9**: 79–80.
12. Fibiger I. Om Serum Behandlung of Difteri. *Hospitalstidende* (1898) **6**: 309–25, 337–50.
13. Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *Br Med J* (1999) **318**: 1209.
14. Day S. Randomisation. In: Everitt B and Palmer C. *Encyclopaedic Companion to Medical Statistics*, London: Edward Arnold, (2005) 291–4.
15. Diehl HS, Baker AB, Cowan DW. Cold vaccines: an evaluation based on a controlled study. *J Am Med Assoc* (1938) **111**: 1168–73.
16. Gail MH. Statistics in Action. *J Am Stat Assoc* (1996) **91**: 1–13.
17. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J* (l948) **ii**: 769–82.
18. Medical Research Council. Chemotherapy of pulmonary tuberculosis in young adults. *Br Med J* (1952) **i**: 1162–8.
19. Medical Research Council. Clinical trials of antihistaminic drugs in the prevention and treatment of the common cold. *Br Med J* (1950) **ii**: 425–31.
20. Medical Research Council. A comparison of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis – I. *Br Med J* (1954) **i**: 1223–7.
21. Medical Research Council. A comparison of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis – II. *Br Med J* (1955) **ii**: 695–700.
22. Hill AB, Marshall J, Shaw DA. A controlled clinical trial of long-term anticoagulant therapy in cerebrovascular disease. *Q J Med* (1960) **29** (NS): 597–608.
23. Hill AB. *Statistical Methods in Clinical and Preventative Medicine*. Edinburgh: E and S Livingstone (1962).
24. Rheumatic Fever Working Party. The evolution of rheumatic heart disease in children: five-year report of a co-operative clinical trial of ACTH, cortisone, and aspirin. *Circulation* (l960) **22**: 505–15.
25. Kinsey VE. Retrolental fibroplasia. *Am Med Assoc Arch Ophthalmol* (1956) **56**: 481–543.
26. Katz J. The Nuremberg Code and the Nuremberg Trial. A reappraisal. *J Am Med Assoc* (l996) **276**: 1662–6.
27. McNeill PM. *The Ethics and Politics of Human Experimentation*. Cambridge: Press Syndicate of the University of Cambridge (1993).
28. Beecher HK. Ethics and clinical research. *New Engl J Med* (1966) **274**: 1354–1360.
29. Freedman B. Research, unethical. In: Reich WT, ed, *Encyclopedia of Bioethics*. New York: Free Press, (1995) 2258–61.
30. Shirer WL. *The Rise and Fall of the Third Reich*. New York: Simon and Schuster (1960).
31. US Government Printing Office. *Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No. 10*, Vol. 2. Washington DC: US Government Printing Office (l949) 181–2.

32. Withington CF. *Time Relation of Hospitals to Medical Education*. Boston: Cupples Uphman (1886).
33. Osler W. The evolution of the idea of experiment. *Trans Congr Am Physicians Surg* (l907) **7**: 1–8.
34. Hill AB. Medical ethics and controlled trials. *Br Med J* (1963) **i**: 1043–9.
35. Netherlands Minister of Social Affairs and Health. *4 World Med J* (1957) 299–300.
36. World Medical Assembly. *Declaration of Helsinki: recommendations guiding physicians in biomedical research involving human subjects*. Helsinki: World Medical Assembly (1964).
37. Faden RR, Beauchamp T, King NMP. *A History of Informed Consent*, New York: Oxford University Press, (1986).
38. Shah RR. Thalidomide, drug safety and early drug regulation in the UK. *Adverse Drug React Toxicol Rev* (2001) **20**: 199–255.
39. Day S. Medicines and Healthcare products Regulatory Agency (MHRA). In: Everitt B and Palmer C. *Encyclopaedic Companion to Medical Statistics*. London: Edward Arnold, (2005) 218–219.
40. Lewis JA. Statistical Principles for Clinical Trials (ICH E9): an introductory note on an international guideline. *Stat Med* (1999) **18**: 1903–4.
41. ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Stat Med* (1999) **18**: 1905–42.
42. Ellenberg SS, Fleming TR, DeMets DL. *Data Monitoring Committees in Clinical Trials*. Chichester: John Wiley & Sons Ltd (2002).
43. Canner P. Monitoring of the data for adverse or beneficial treatment effects. *Controlled Clin Trials* (1983) **4**: 467–83.
44. Friedman L. The NHLBI model: a 25-year history. *Stat Med* (1993) **12**: 425–31.
45. Bross I. Sequential medical plans, *Biometrics* (1952) **8**: 188–295.
46. Armitage P. Sequential tests in prophylactic and therapeutic trials. *Q J Med* (1954) **23**: 255–74.
47. Armitage P. Interim analysis in clinical trials, *Stat Med* (1991) **10**: 925–37.
48. Friedman LM, Furberg CD, and DeMets DL. *Fundamentals of Clinical Trials*, 2nd edn. Boston: Wright (1985).
49. Pocock SJ. *Clinical Trials: A Practical Approach*. Chichester: John Wiley & Sons, Ltd (1983).
50. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* (1977) **64**: 191–9.
51. Haybittle JL. Repeated assessment of results of cancer treatment. *Br J Radiol* (1971) **44**: 793–7.
52. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith P. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. 1. Introduction and design. *Bri J Cancer* (1976) **34**: 585–612.
53. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* (1979) **35**: 549–56.
54. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* (1983) **70**: 659–63.
55. D'Arcy Hart P. History of randomised controlled trials (letter to the Editor). *Lancet* (1972) **i**: 965.
56. D'Arcy Hart P. Early controlled clinical trials (letter to the Editor). *Br Med J* (1996) **312**: 378–9.
57. Gill DBEC. Early controlled trials (letter to the Editor). *Br Med J* (1996) **312**: 1298.
58. Clarke M. Early controlled trials (letter to the Editor). *Br Med J* (1996) **312**: 1298.
59. Bull JP. The historical development of clinical therapeutic trials. *J Chron Dis* (1959) **10**: 218–248.

2

General Statistical Issues

DAVID MACHIN
National Cancer Centre & Clinical Trials and Epidemiological Research Unit, Singapore and United Kingdom
Children's Cancer Study Group, University of Leicester, UK

INTRODUCTION

Just as in any other field of scientific and medical research, the choice of an appropriate design for a clinical trial is a vital element. In many circumstances, and for many of the trials described in this text, these designs may not be overly complicated. For example, a large majority will compare two therapeutic or other options in a parallel group trial. In this case the analytical methods used for description and analysis too may not be over complicated. The vast majority of these are described in basic medical statistics textbooks and implemented in standard software packages. Nevertheless there are circumstances in which more complex designs, such as sequential trials, are utilised and for which specialist methods are required. There are also often rather complex statistical problems associated with monitoring the progress of clinical trials, their interim analysis, stopping rules for early closure and the possibility of extending recruitment beyond that initially envisaged.

Although the clinical trial itself may not be of complex design in the statistical sense, the associated trial protocol should carefully describe (and in some detail) the elements essential for its conduct. Thus the protocol will describe the rationale for the trial, the eligible group of patients or subjects, the therapeutic options and their modification should the need arise. It will also describe the method of patient allocation to these options, the specific clinical assessments to be made and their frequency, and the major endpoints to be used for evaluation. It will also include a justification of the sample size chosen, an indication of the analytical techniques to be used for summary and comparisons, and the proforma for data collection.

Of major concern in all aspects of clinical trial development and conduct is the ethical necessity which is written into the Declaration of Helsinki of 1964[1] to ensure the well-being of the patients or subjects under study. This in itself requires that clinical trials are well-planned and conducted with due concern for the patient's welfare and safety. It also requires that the trial is addressing an important question, the answer to which will bring eventual benefit to the patients themselves or at least to future patients.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

EVIDENCE-BASED MEDICINE

We have indicated that in many circumstances trials have been conducted that are unrealistically small, some unnecessarily replicated, while others have not been published as their results have not been considered of interest. It has now been recognised that to obtain the best current evidence with respect to a particular therapy, all pertinent clinical trial information needs to be obtained, and if circumstances permit, the overview is completed by a meta-analysis of the trial results. This recognition has led to the Cochrane Collaboration and a worldwide network of overview groups addressing numerous therapeutic questions.[2] In certain situations this has brought definitive statements with respect to a particular therapy. For others it has lead to the launch of a large-scale confirmatory trial.

Although it is not appropriate to review the methodology here, it is clear that the 'overview' process has led to many changes to the way in which clinical trial programmes have developed. They have provided the basic information required in planning new trials, impacted on an appropriate trial size and publication policy, and very importantly raised reporting standards. They are impacting directly on decisions that affect patient care and questioning conventional wisdom in many areas.

TYPES OF CLINICAL TRIALS

In broad terms, the types of trials conducted in human subjects may be divided into four phases. These phases represent the stages in, for example, the development of a new drug and which requires early dose finding and toxicity data in man, indications of potential activity, comparisons with a standard to determine efficacy and then (in certain circumstances) post-marketing trials. The nomenclature of Phase I, II, III and IV has been developed for drug development purposes and there may or may not be exact parallels in other applications. For example, a trial to assess the value of a health educational programme will be a Phase III study as will a trial comparing two surgical procedures for closing a cleft palate. In both these examples, any one of the Phase I, II or IV trials would not necessarily be conducted. The objectives of each phase in a typical development programme for a drug are summarised in Table 2.1.

Table 2.1. Objectives of the trials of different phases in the development of drug (after Day[3])

Phase	Objective
I	The earliest types of studies that are carried out in humans. They are typically done using small numbers of healthy subjects and are to investigate pharmacodynamics, pharmacokinetics and toxicity
II	Carried out in patients, usually to find the best dose of drug and to investigate safety
III	Generally major trials aimed at conclusively demonstrating efficacy. They are sometimes called confirmatory trials and, in the context of pharmaceuticals, typically are the studies on which registration of a new product will be based
IV	Studies carried out after registration of a product. They are often for marketing purposes as well as to gain broader experience with using the new product

Source: Reproduced by permission of John Wiley & Sons Ltd.

Without detracting from the importance of Phase I, II and IV clinical trials, the main focus of this text is on Phase III comparative trials. In this context, reference will often be made to the 'gold standard' randomised controlled trial (RCT). This does not imply that this is the only type of trial worthy of conduct, but rather that it provides a benchmark against which other trial designs are measured.

PHASE I AND II TRIALS

The traditional outline of a series of clinical trials moving sequentially through Phases I to IV is useful to consider in an idealistic setting, although in practice the sequential manner is not always followed (for reasons that will become clear).

Pocock[4] (pp. 2–3) also gives a convenient summary of the four phases while Temple[5] gives a discussion of them with emphasis from a regulatory perspective. Whether the sequential nature of the four phases is adhered to or not, the objectives of each phase are usually quite clearly defined.

As we have indicated in Table 2.1, Phase I studies aim to investigate the metabolism of a drug and its pharmacodynamics and pharmacokinetics. Typical pharmacokinetic data would allow, for example, investigation of peak drug concentrations in the blood, the half-life and the time to complete clearance. Such studies will assist in defining what doses should be used and the dosing frequency (once daily, twice daily, hourly) for future studies. These Phase I studies (certainly the very first ones) are almost always undertaken in healthy volunteers and would naturally be the very first studies undertaken in humans. However, later in a drug development programme it may be necessary to study its effects in patients with specific diseases, in those taking other medications or patients from special groups (infants, elderly, ethnic groups, pregnant women).

Most of the objectives of a Phase I study can often be met with relatively few subjects – many studies have fewer than 20 subjects. In essence, they are much more like closely controlled laboratory experiments than population-based clinical trials.

Broadly speaking, Phase II trials aim to set the scene for subsequent confirmatory Phase III trials. Typically, although exceptions may occur, these will be the first 'trials' in patients. They are also the first to investigate the existence of possible clinical benefits to those patients. However, although efficacy is important in Phase II it may often be in the form of surrogate; for example, tumour response rather then survival time in a patient with cancer. Along with efficacy, these studies will also be the first to give some detailed data on side effects.

Although conducted in patients, Phase II trials are typically still highly controlled and use highly defined (often narrow) patient groups so that extraneous variation is kept to a minimum. These are very much exploratory trials aimed at discovering *if* a compound *can* show useful clinical results. Although it is not common, some of these trials may have a randomised comparison group.

NON-RANDOMISED EFFICACY STUDIES

HISTORICAL CONTROLS

In certain circumstances when a new treatment has been proposed investigators have recruited patients in single-arm studies. The results from these patients are then compared with information on similar patients having (usually in the past) received a relevant standard therapy for the disease in question. However, such comparisons may well be biased in many possible ways such that it may not be reasonable to ascribe the difference (if any) observed to the treatments themselves. Nevertheless it has been argued that using regression models to account for possible confounding variables may correct such biases,[6] but this is at best a very uncertain procedure and is not often advocated. Similar problems arise if all patients are recruited prospectively but allocation to treatment is not made at random. Of course, there will be situations in which randomisation is not feasible and there is no alternative to the use of historical controls or non-randomised prospective studies. One clear example of this is the early evaluation of the Stanford Heart Transplant Program in which patients could not be randomised to receive or not a donor heart. Many careful analyses and reviews of this unique data set have been undertaken and these have established the value of the Program, but progress would have been quicker (and less controversial) had randomisation been possible.

In the era of evidence-based medicine, information from non-randomised but comparative studies is categorised as providing weaker evidence than randomised trials (see Altman,[7] p. 3279).

PHASE III CONTROLLED TRIALS

EQUIPOISE AND UNCERTAINTY

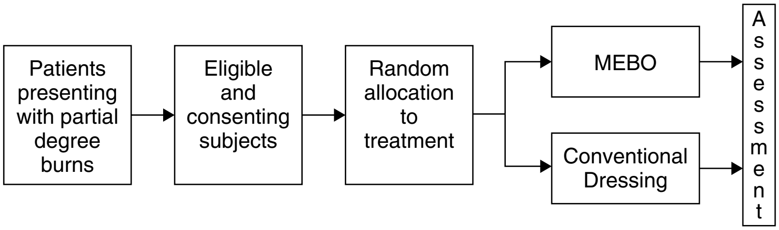
As indicated, the randomised controlled trial is the standard against which other trial designs may be compared. One such trial, and there are many other examples described in subsequent chapters, compared conventional treatment, *C*, with a complementary medicine alternative in patients with severe burns[8]. The complementary medicine was termed Moist Exposed Burns Ointment (*MEBO*). One essential difference between the two treatments was that *C* covered the wounds (dressed) whilst *MEBO* left them exposed (not dressed). See Figure 2.1.

In this trial patients with severe burns were emergency admissions requiring immediate treatment, so that once eligibility was confirmed and consent obtained, randomisation immediately followed and treatment was then commenced. Such a trial is termed a two-treatment parallel group design. This is the most common design for comparative clinical trials. In these trials subjects are independently allocated to receive one of several treatment options. No subject receives more than one of these treatments.

In addition there is genuine uncertainty as to which of the options is best for the patient. It is this uncertainty which provides the necessary equipoise, as described by Freedman[9] and Weijer *et al.*[10] to justify random allocation to treatment after due consent is given. Enkin,[11] in a debate with Weijer *et al.*,[10] provides a counter view.

There are at least two aspects of the eligibility requirements that are important. The first is that the patient indeed has the condition (here severe burns) and satisfies all the other requirements. There must be no specific reasons why the patient should not be included. For example, in some circumstances pregnant or lactating women (otherwise eligible) may be excluded for fear of impacting adversely either on the foetus or the newborn child. The second is that all (here two) the therapeutic options are equally appropriate for this particular patient. Only if both these aspects are satisfied should the patient be invited to consent to participate in the trial. There will be circumstances in which a patient may be eligible for the trial but the attending physician feels (for whatever reason) that one of the trial options is 'best' for the patient. In this case the patient should receive that option, no consent for the trial is then required and the randomisation would not take place. In such circumstances, the clinician should not randomise the patient in the hope that the patient will receive the 'best' option; then, if he or she did not, withdraw the patient from the trial.

The consent procedure itself will vary from trial to trial and will, at least to some extent, depend on local ethical regulations in the country in which the trial is being conducted. The ideal is fully informed and written consent by the patient


Patients presenting with partial degree burns
Eligible and consenting subjects
Random allocation to treatment
MEBO
Conventional Dressing
Assessment

Source: Reproduced from Ang *et al.*,[8]

Figure 2.1. Randomised controlled trial to compare conventional treatment and Most Exposed Burns Ointment (MEBO) for the treatment of patients with severe burns (after Ang *et al.*[8]).

him- or herself. However, departures from this may be appropriate. For example, such departures may concern patients with severe burns who may be unconscious at admission, very young children for whom a proxy must be used to obtain the consent for them, or patients with hand burns that are so severe that they affect their ability to write their signature.

Clearly, during the period in which patients are being assessed for eligibility and their consent obtained, both the attending physician and the patient will be fully aware of the potential options being compared in the RCT. However, neither must be aware, at this stage, of the eventual treatment allocation. It is important therefore that the randomisation list, for the current as well as for future patients, is held by a neutral third party. In most circumstances, this should be an appropriate trial office that is contacted by the responsible clinician once eligibility and consent are obtained. This contact may be made by telephone, fax, direct access by modem into a trial database, email or the web – whichever is convenient in the particular circumstance. It is then important that therapy is instituted as soon as practicable after the randomisation is obtained.

In specific cases, the randomisation can be concealed within opaque and sealed envelopes which are distributed to the centres in advance of patient recruitment. Once a patient is deemed eligible, the envelope is taken in the order specified in a prescribed list, opened and the treatment thereby revealed. Intrinsically, there is nothing wrong with this process but, because of the potential for abuse, it is not regarded as entirely satisfactory. However, in some circumstances it will be unavoidable; perhaps a trial is being conducted in a remote area with poor communications. In such cases, every precaution should be taken to ensure that the process is not compromised.

The therapeutic options should be well described within the trial protocol and details of what to do, if treatment requires modification or stopping for an individual patient, should be given. Stopping may arise either when patients merely refuse to take further part in the trial or from safety concerns with a therapy under test.

STANDARD OR CONTROL THERAPY

In the early stages of the development of a new therapy it is important to compare this with the current standard for the disease in question. In certain circumstances, the 'new' therapy may be compared against a 'no treatment' control. For example, in patients receiving surgery for the primary treatment of head and neck cancer followed by best supportive care, the randomised controlled trial may be assessing the value of adding post-operative chemotherapy. In this case the 'control' group are those who receive no adjuvant treatment, whilst the 'test' group receive chemotherapy. In certain circumstances, patients may receive a placebo control. For example, in the randomised controlled trial conducted by Chow *et al.*[12] in those with advanced liver cancer, patients are randomised to receive either placebo or tamoxifen. In this trial both patients and the attending physicians are 'blinded' to the actual treatment given to individual patients. Such a 'double-blind' or 'double-masked' trial is a design that reduces any potential bias to a minimum. Such designs are not possible, however, in many circumstances and neither are those with a 'no treatment' control. In many situations, the 'control' will be the current best practice against which the new treatment will be compared. Should this turn out to be better than current practice then this, in its turn, may become standard practice against which future developments will be compared.

In general there will be both baseline and follow-up information collected on all patients. The baseline (pre-randomisation) information will be that required to determine eligibility together with other information required to describe the patients recruited to the trial together with those which are thought likely to influence prognosis. The key follow-up information will be that which is necessary to determine the major endpoint(s) of the randomised controlled trial. Thus in the example of the burns patients these may when the unhealed body surface area finally closes or the size and severity of the resulting scars. To establish the first of these, the burns areas may have to be monitored on a daily basis to determine exactly when the endpoint is achieved, whereas the latter may be a single assessment at the anniversary of the initial burn accident. Pre-trial information on these endpoints, possibly from clinical experience or other studies, will usually form the basis of the anticipated difference between treatments (the effect size), help determine the trial size and be the variables for the statistical comparisons.

LARGE SIMPLE TRIALS

It has become recognised over time, particularly in the fields of cardiovascular disease and cancer, that there are circumstances where small therapeutic advantages may be worthwhile demonstrating. In terms of trial size, the smaller the potential benefit, essentially the effect size, then the larger the trial must be in order to be reasonably certain that the small benefit envisaged really exists at all.

Trials of this size, often involving many thousands of patients, are a major undertaking. To be practical, they must be in common diseases in order to recruit the required numbers of patients in a reasonable time frame. They must be testing a treatment that has wide applicability and can be easily administered by the clinicians responsible or the patients themselves. The treatments must be relatively non-toxic, else the small benefit will be outweighed by the side effects. The trials must be simple in structure and restricted as to the number of variables recorded, so that the recruiting clinicians are not overburdened by the workload attached to large numbers of trial patients going through the clinic. They also need to be simple in this respect, for the responsible trial centre to cope with the large amounts of patient data collected.

One example of such a trial tests the value of aspirin in patients with cardiovascular disease.[13] This trial concerned a very common disease using a very simple and low-cost treatment taken as tablets with very few side effects. The resulting estimates of absolute survival gain were (as expected) small but the benefits in public health terms enormous. Similar types of trials have been conducted in patients with breast cancer; one in particular compared the three adjuvant treatment possibilities: tamoxifen, anastrozole or their combination.[14]

INTERVENTION AND PREVENTION TRIALS

The focus so far has been on randomised controlled trials in patients with medical conditions requiring treatment or a medical intervention of some sort. Such designs do apply to situations such as trials in normal healthy women in which alternative forms of contraception are being tested.[15] However, there are quite different situations in which the object of a trial is to evaluate alternative strategies for preventing disease or for detecting its presence earlier than is routine. For example, intervention trials to encourage 'safe sex' to prevent the spread of AIDS or breast screening trials to assess the value of early detection of disease on subsequent treatment and prognosis. In such cases, it may be impossible to randomise on an individual subject basis. Thus an investigator may have to randomise communities to test out different types of health promotion or different types of vaccines, when problems of contamination or logistics, respectively, mean that it is better to randomise a group rather than an individual. This is the approach adopted by Donner *et al.*[16] in a trial to compare a reduced antenatal care model with a standard care model. For this trial, because of the clustered randomisation of the alternatives on a clinic-to-clinic basis, the Zelen[17] single consent design was utilised.

PHASE IV TRIALS – POST-MARKETING SURVEILLANCE

Within a regulatory framework, Phase IV trials are generally considered as 'post-registration' trials: that is, trials of products that already have a marketing authorisation. However, within this post-registration period studies may be carried out for a variety of purposes, some within their existing licence and others outside of that licence. Studies may also be undertaken in countries where a marketing authorisation has not been approved, in which case they are regulatory or Phase-III-type studies, at least for that country. Studies may be undertaken to expand the indications listed on a marketing authorisation for either a different disease or a different severity of the indicated condition. They may be undertaken to gain more safety data for newly registered products: this later situation is more usually what is considered as a Phase IV study.

Historically, pharmaceutical companies used to carry out studies that were solely for marketing purposes and answered very few (if any) research questions. These were termed ‘seeding studies’, although with tighter ethical constraints such studies are now very rare, if indeed they take place at all. Certainly the ‘hidden’ objective of many Phase IV studies carried out by pharmaceutical companies may be to increase sales, but if the means of doing so is via answering a useful scientific or medical question then this should be of benefit both to society and to the company.

Many trials organised by academic departments should also be considered as Phase IV studies. Classic examples such as the RISC Group[13] trials looking at the cardiovascular benefits of aspirin are studies of licensed products to expand their use.

ALTERNATIVE DESIGNS

For illustrative purposes we have used the two-arm parallel group RCT but these designs can be generalised to compare three or more groups as appropriate. In addition, there are other designs in common use which include 2 × 2 factorial and crossover designs.[18]

MORE THAN TWO GROUPS

Although not strictly a different design, a parallel group trial with more than two treatments to compare does pose some difficulties. For example, how is the size of a trial comparing three treatments, A, B and C, determined, since there are now three possible anticipated effect sizes that one can use for planning purposes? These correspond to the treatment comparisons $A-B$, $A-C$ and $B-C$. The number of patients required for each of these comparisons may give three very different sample sizes. It is then necessary to decide which of these will form the basis for the final trial recruitment target, N. The trial will then randomise the patients equally into the three treatment groups. In many circumstances, a three-arm trial will tend to require some 50% more patients than a two-arm trial comparing two of the three treatments under consideration.

Once the trial has been completed, the resulting analysis (and reporting) is somewhat more complex than for a simple two-group comparison. However, it is the importance of the questions posed, rather than the ease of analysis, which should determine the design chosen. A good example of the use of such a design is the previously mentioned trial in post-menopausal patients with breast cancer in which three options are compared.[14]

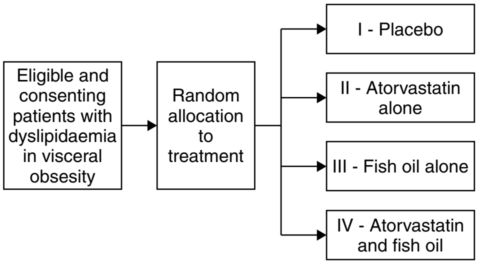
Nevertheless practical considerations may rule out this choice of design. The design poses particular problems in data monitoring. For example, if an early advantage appears to favour one particular treatment, this suggests the trial might be stopped early as a consequence. Then it may not be clear whether the randomisation between the other treatment groups should or should not continue. Were the trial to stop early, then the questions relating to the other comparisons would be unlikely to have been resolved at this stage. Should the (reduced) trial continue, then there may be very complex issues associated with its analysis and reporting.

In addition, a potentially serious problem of bias can arise. At the onset of the trial, the clinician has to assess whether or not all the three options (A, B or C) available for treatment are suitable for the particular patient under examination. If any one of these were not thought to be appropriate (for whatever reason), then the patients’ consent would not be sought to enter the trial and to be randomised. Suppose, later in the trial, an interim analysis suggests recruitment to A is no longer necessary and that arm is closed. For future patients, the clinician now has to assess whether or not only the two options (B or C) are suitable for the particular patient under examination. As a consequence, the patients now going into the trial are no longer potentials for A and hence may be somewhat different than those entering at the earlier stages. Although this will not bias the final comparison between B and C, it does implies that there will be a bias if the patients entering at this stage are compared with those receiving A.

In the above, we have assumed that the three treatments A, B and C are, in a sense, unrelated, albeit all suitable for the patients in mind. However, if they were related, e.g. perhaps three doses of the same drug, then note of this structure may change the approach to design from that outlined here.

FACTORIAL DESIGNS

In a trial conducted by Chan *et al.*[19] patients with dyslipidaemia in visceral obesity were randomised to either atorvastatin alone, fish oil alone, both or neither (placebo) to investigate their influence on lipid levels. This trial may take the form of no treatment (**1**), atorvastatin alone (***a***), fish oil alone (***f***) or both atorvastatin and fish oil (***af***). These alternative options are summarised in Figure 2.2.


Eligible and consenting patients with dyslipidaemia in visceral obsesity
Random allocation to treatment
I - Placebo
II - Atorvastatin alone
III - Fish oil alone
IV - Atorvastatin and fish oil

Source: Reproduced by permission of Blackwell Publishing Ltd.

Figure 2.2. Randomised 2 × 2 factorial trial to determine the value of Atorvastatin, fish oil or both in patients with dyslipidaemia in visceral obesity[19]..

In contrast to the two-group parallel design of Figure 2.1, where *MEBO* is compared with conventional treatment C, a factorial design poses two questions simultaneously. Those patients assigned to groups I and II are compared with those receiving III and IV, to assess the value of fish oil. This estimates the effect of fish oil and is termed the ‘main effect’ of that treatment. Those assigned to I and III are compared with those receiving II and IV to assess the main effect of atorvastatin. In most situations the final trial size will require fewer patients than would be necessary if the two questions were posed in two distinct parallel group trials of the format of Figure 2.2.

In addition, the factorial design allows the comparison of groups I and IV with II and III, and this estimates the so-called fish oil by atorvastatin interaction. For example, suppose both the main effect of fish oil alone and the main effect of atorvastatin alone prove to be beneficial in this context; then an interaction would arise if the combination treatment fish oil–atorvastatin gives a benefit greater (or lesser) than the sum of these constituent parts. As is the situation here, factorial designs can be of a double placebo type, where subjects of group I of Figure 2.2 receive a double placebo, one representing each treatment factor.

In principle, the 2×2 or 2^2 design can be extended to the 2^k $(k > 2)$ one. Circumstances where this kind of design may be useful are if perhaps the first two factors are major therapeutic or curative options and the third is a factor for a secondary question not associated directly with efficacy but (say) to relieve pain in such subjects. However, the presence of a third factor of whatever type increase the complexity of the trial design (not itself a particularly difficult statistical problem) which may have implications on the patient consent procedures and the timing of randomisation(s). Nevertheless, a 2^3 design in a trial of falls prevention in the elderly has been successfully conducted by Day *et al.*[20] Piantadosi[21] describes several examples of the use of these designs.

However, Green[22] has issued a cautionary note that some of the assumptions behind the use of factorial designs may not be entirely justified and so any proposals for the use of such designs should be considered carefully.

CROSSOVER TRIALS

In contrast to the design of Figure 2.1, in the crossover trial each patient will receive both treatment options, one followed by the other in two periods of time. The two treatments will be given either in the order A followed by B (AB) or the reverse (BA). The essential features of a crossover design are summarised in Figure 2.3 for the trial conducted by Hong *et al.*[23] in patients with erectile dysfunction.

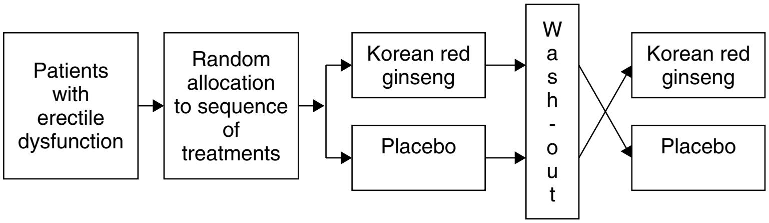
Typically, in the two-period, two-treatment crossover trial, for eligible patients there is a run-in stage in which the subject receives neither treatment. At the end of this, randomisation to either AB or BA takes place. Following active treatment in Period I (in effect either A or B depending on the randomisation), there is a wash-out interval in which again no treatment is given, after which Period II commences and the (other) treatment of the sequence initiated. The characteristics of this design, e.g. the possible run-in and the wash-out period, imply that only certain types of patients in which active treatment can be withheld in this way are suitable to be recruited. Further, there is an implication that the patient returns to essentially the same state at the beginning of Period II as he or she was at the commencement of Period I. This ensures that the between-treatment comparison ($A \vee B$) within the patient remains unaffected by anything other than the change in treatment itself and random variation. These considerations tend to restrict the applicability of the design to patients with chronic conditions such as, for example, arthritis, asthma or migraine. Senn[24] describes in careful and comprehensive detail the role of crossover designs in clinical studies.

A clear advantage of the design is that the patient receives both options and so the analysis includes within-patient comparisons which are more sensitive than between-patient comparisons, implying that such trials would require fewer subjects than a parallel group design comparing the same treatment options.

EQUIVALENCE TRIALS

In certain situations, a new therapy may bring certain advantages over the current standard, possibly in a reduced side-effects profile, easier administration or cost. Nevertheless, it may not be anticipated to be better with respect to the primary efficacy variable. Under such conditions, the new approach may be expected to be at least ‘equivalent’ to the standard in relation to efficacy.

Trials to show that two (or more) treatments are ‘equivalent’ to each other pose special problems in design, management and analysis.[25] ‘Proving the null hypothesis’ in a significance testing scenario is never possible: the strict interpretation when a statistically significant difference has *not* been found is that ‘there is


Patients with erectile dysfunction
Random allocation to sequence of treatments
Korean red ginseng
Placebo
Wash-out
Korean red ginseng
Placebo

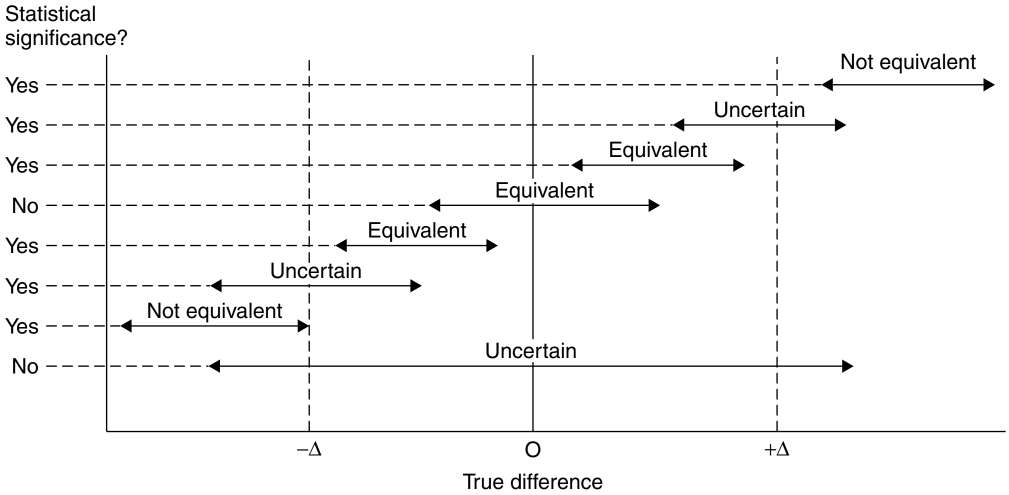
Source: Reproduced from Hong *et al.*,[23] with permission © (2002), reprinted with permission from the American Urological Association.

Figure 2.3. Randomised placebo-controlled, two-period crossover trial of Korean red ginseng in patients with erectile dysfunction[23].

insufficient evidence to demonstrate a difference'. Small trials typically fail to detect differences between treatment groups but not necessarily because no difference exists. Indeed it is unlikely that two different treatments will ever exert truly identical effects.

A level of 'therapeutic equivalence' should be defined and this is a medical question, not a statistical one. For example, in a study of weight gain in pre-term infants, if two treatments show mean increases in weight to within 25 g per week then they may be considered as therapeutically equivalent. Note that in this example 25 g is *not* the mean weight gain that is expected per week – we would hope that would be much more. But if infants receiving one feeding regimen had a mean increase of 150 g per week then we would consider an alternative treatment to be equivalent if the mean weight gain were between 125 and 175 g per week.

Conventionally, to show a treatment difference, we would state the null hypothesis as being that there is no difference between the treatments and then look for evidence to refute that null hypothesis. In the case of equivalence we specify the range of equivalence, Δ (25 g per week in the above example), and then test *two* null hypotheses. We test that the observed difference is statistically significantly *greater* than $-\Delta$; and that the observed difference is statistically significantly *less* than $+\Delta$. In practice it is much easier to consider a confidence interval for the difference between the treatment means and draw this on a graph with the agreed limits of equivalence. Figure 2.4 shows various scenarios. Some cases show equivalence, some fail to show equivalence; some cases show a statistically significant difference, others fail to show a difference. Note that it is quite possible to show a statistically significant difference between two treatments yet also demonstrate therapeutic equivalence. These are not contradictory statements but simply a realisation that although there is evidence that one treatment works better than another, the size of the benefit is so small that it has little or no practical advantage.



Examples of possible results of using the confidence interval approach: –Δ to +Δ is the prespecified range of equivalence; the horizontal lines correspond to possible trial outcomes expressed as confidence intervals, with the associated significance test result shown on the left; above each line is the decision concerning equivalence.

Source: Reproduced from Jones *et al.*,[26] with permission from the BMJ Publishing Group.

Figure 2.4. Schematic diagram to illustrate the concept of equivalence (from Jones *et al.*[26]).

The analysis and interpretation can be quite straightforward but the design and management of equivalence trials is often much more complex. In general, careless or inaccurate measurement, poor follow-up of patients, poor compliance with study procedures and medication all tend to bias results towards no difference. Since we are trying to offer evidence of equivalence, poor study design and procedures may therefore actually help to hide treatment differences. In general, therefore, the quality of equivalence trials should be demonstrably high.

One special subset of equivalence trials is termed 'non-inferiority' trials. Here we only wish to be sure that one treatment is 'not worse than' or is 'at least as good as' another treatment: if it is better, that is fine (even though superiority would not be required to bring it into common use). All we need is to get convincing evidence that the new treatment is not worse than the standard. We still have the same design and management considerations but here, looking at Figure 2.4, we would only be concerned with whether or not the *lower* limit of the confidence interval is *greater* than the non-inferiority margin ($-\Delta$).

SEQUENTIAL TRIALS

There has been an implicit assumption in the trial designs discussed above that the total (and final) sample size is determined at the design stage and before recruitment commences to the trial in question. This fixed sample size approach essentially implies that the data collected during the conduct of the trial will only be examined for efficacy once the trial has closed to patient accrual. However, the vast majority of Phase III trials will tend to recruit patients over perhaps an extended interval of time and so information on efficacy will be accumulated over this period. A sequential trial is one designed to utilise this accumulating knowledge to better effect – perhaps to decrease the final trial size if the data are indicating an advantage to one of the treatments and this can be firmly established at an early stage, or to extend the trial size in other circumstances.

In fact, Donaldson *et al.*[27] give examples where trials that had been conducted using a fixed sample size approach might have been curtailed earlier has a sequential design been utilised. Fayers *et al.*[28] describe the issues faced when designing a sequential trial using α-interferon in patients with renal carcinoma. The accumulating patient data from this trial crossed an early termination boundary which inferred an advantage to α-interferon.[29]

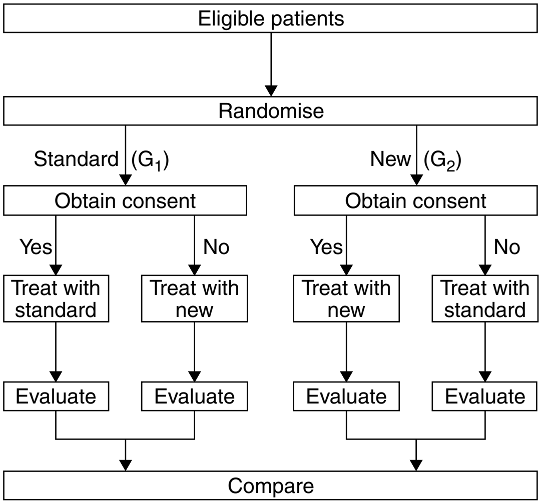
A fully sequential design will monitor the trial patient by patient as the information on the trial endpoint is observed from them. Alternatively, a 'group' sequential trial will utilise information from successive groups of patients. Computer programs to assist in the implementation of these designs are available[30] and a review of some of the issues is given by Whitehead.[31]

There are, however, several problems associated with the use of sequential designs. These problems range from difficulties of financing a trial of uncertain size, making sure the data are fully up to date as the trial progresses, to the more technical concerns associated with the calculation of the appropriate confidence intervals. However, Whitehead,[32] see also Jennison and Turnbull,[33] has argued very persuasively that all these objections can be resolved. Nevertheless, in relative terms the use of sequential designs is still somewhat limited.

ZELEN'S DESIGNS

Although not strictly an alternative 'design' in the sense of those of this section, Zelen's randomised consent design combines aspects of design with problems associated with obtaining consent from patients to participate in clinical trials. They were motivated by the difficulties expressed by clinicians in obtaining consent from women whom they wished to recruit to trials with breast cancer.[17,34] Essentially subjects are randomised to one of two treatment groups. Those who were randomised to the standard treatment (conventional dressing in Figure 2.1) are all treated with it. For these patients no consent to take part in the trial is sought. On the other hand, those who are randomised to the experimental treatment (*MEBO* dressing in Figure 2.1) are asked for their consent: if they agree they are treated with the experimental treatment; if they disagree they are treated with the standard treatment. This is known as Zelen's single consent design. An alternative is that those randomised to the standard treatment may also be asked if they accept that treatment; again, they are actually given their treatment of choice. This latter double consent design is described in Figure 2.5. In either case, the analysis must be by intention-to-treat: that is, it is based on the treatment to which patients were randomised, not the treatment they actually received, although one would hope that only a minority would not have the treatment allocated.

The properties of these designs have been examined in some detail by Altman *et al.*[35] who concluded that: 'There are serious statistical arguments against the use of randomised consent designs, which should discourage their use.' In any event, they have rarely been used in practice although they continue to be advocated.[36]



Source: After Altman *et al.*[35] (Figure 3).

Figure 2.5. The sequence of events to follow in Zelen's double randomised consent design, seeking consent in conjunction with randomisation (after Altman *et al.*,[35] Figure 3).

Nevertheless, in the large antenatal care trial described by Donner *et al.*[16] the Zelen single consent design is utilised. However, this is a cluster randomised trial and the issues are somewhat different.

BAYESIAN METHODS

The essence of Bayesian methodology in the context of the design of clinical trials is to incorporate relevant information into the design process. At a later stage, Bayesian methods may assist in data monitoring (as trial data accumulate) and with the final analysis and interpretation of the (now complete) trial data. In theory the information available and which is pertinent to the trial in question can be summarised into a prior distribution. This may include (hard) information from the published literature and/or elicited clinical opinion. The mean of such a distribution would correspond to the possible effect size which can then be assessed by the design team as clinically worthwhile in their context and hence used for sample size estimation purposes. The same prior, or that prior updated from new external evidence accumulated during the course of the trial, may be used by the trial Data Monitoring Committee (DMC) to help form the basis of recommendations with respect to future conduct of the trial – perhaps to close the trial as efficacy has been clearly established or increase the planned size of the trial as appropriate. Finally, once the trial data are complete, these can be combined with the prior (or updated prior) to obtain the posterior distribution, from which Bayesian estimates of treatment effect and corresponding credibility intervals can be calculated.

However, despite the feasibility of the above approach few trials to date have implemented a full Bayesian approach. A review of articles in the *British Medical Journal* from 1996 to November 1999 found no examples.[35] Nevertheless, Spiegelhalter *et al.*[37] show convincingly how the concept of optimistic and sceptical prior distributions obtained from the clinical teams may be of assistance in interpreting the results of a trial, while Tan *et al.*[38] describe how the methodology may be useful in the context of trials in rare diseases when few patients will be available. Despite some technical difficulties, it is fairly certain that Bayesian methods will become more prominent in Phase II trial methodology. For example, Tan Tai *et al.*[39] suggest how such ideas may be useful in a Phase II programme.

RANDOMISATION AND ALLOCATION TO TREATMENT

We have indicated above that randomisation of patients to the treatment they receive is an important part of the 'gold standard'. In fact it is the key element. The object of randomisation is to help ensure that the final comparison of treatment options is as unbiased as possible; that is, that any difference or lack of difference observed between treatments in efficacy is not due to the method by which patients are chosen for the options under study. For example, if the attending clinician chose which of *MEBO* or *C* should be given to each patient, then any differences observed may be due, at least in part, to the selection process itself rather than to a true difference in efficacy.

Apart from the possible effect of the allocated treatments themselves, observed differences may arise through the play of chance alone or possibly an imbalance of patients with differing prognoses in the treatment groups, or both. The object of the statistical analysis will be to take account of any imbalance and assess the role of chance. Some imbalance in the major prognostic variables may be avoided by stratifying the randomisation by prognostic group and ensuring that an equal number of patients are allocated within each stratum to each of the options. This may be achieved by arranging the randomisation to be balanced within predetermined blocks of patients within each of the strata. Blocks are usually chosen as neither too small nor too large, sizes four or six often being used. Sometimes the block size, perhaps between these options, is chosen at random for successive sequences of patients within a stratum of patient types.

Alternatively, the balance of treatments between therapeutic options can be made using a dynamic allocation procedure such as those described by Taves[40] and Pocock and Simon[41]. In such schemes, during the randomisation procedure a patient is identified to belong to a predetermined category according to certain covariates. This category may, for example, be defined as those of a particular age, gender and tumour stage group. Once the category is determined, then randomisation to the treatment options may proceed as described above. One option, however, is to allocate the next patient to the treatment with the fewest patients already assigned within that category. In this case, the allocation at that stage is deterministic. A better option in such circumstances is to weight the randomisation, perhaps in the ratio of 3:2 in favour of the option with the fewest patients. Clearly, if numbers are equal, the randomisation would revert to 1:1.

We have implicitly assumed that, for two treatments, a 1:1 randomisation will take place. For all practical purposes, this will be statistically the most efficient. However, the particular context may suggest other ratios. For example, if the patient pool is limited for whatever reason, then the clinical team may argue that they should obtain more information within the trial from the test treatment rather than the well-known standard. Perhaps, there is a concern with the toxicity profile rather than just the efficacy per se. In such circumstances, a randomisation ratio of say 2:3 or 1:2 in favour of the test therapy may be decided. However, some loss of statistical power will ensue and this loss should be quantified before a decision on the allocation ratio is finally made.

ENDPOINTS

DEFINING THE ENDPOINT(S)

The protocol for every clinical trial will detail the assessments to be made on the patients recruited. Some of these assessments may focus on aspects of the day-to-day care of the patient whilst others may focus more on those measures which will be necessary in order to determine the trial endpoint(s) for each subject. It is important that these endpoints are unambiguously defined so that they can be determined for each patient recruited to the trial. It is good practice to define which endpoint is the major endpoint of the trial as this will be used to determine trial size and be the main focus for the efficacy evaluation. In many situations, there may be several endpoints of interest, but in this case it is important to order them in order of priority or at least to identify those of primary or secondary importance. If there are too many endpoints defined, then the multiplicity of comparisons then made at the analysis stage may result in spurious statistical significance. This is a major concern if endpoints for health-related quality of life and health economic evaluations are added to the already established more clinical endpoints.

SINGLE MEASURES

In some trials a single measure may be sufficient to determine the endpoint in each patient. For example, the endpoint may be the diastolic blood pressure measured at a particular time, say 28 days, post-randomisation in each patient. In this case the treatment groups will be summarised by the respective means. In some situations the endpoint may be patient response; for example, the patient becomes normo-tensive following a period of treatment. Those who respond are termed successes and those that do not failures. In this case, the treatment groups will be summarised by the proportion of responders. If, on the other hand, the patients are categorised as normo-tensive, still hypotensive but diastolic blood pressure (DBP) nevertheless reduced, or still hypotensive and DBP not improved, then this would correspond to an ordered categorical variable. Alternatively, the endpoint may be defined as the time from randomisation and inception of treatment for the patient to become normo-tensive. In this situation repeated (say daily) measures of DBP will be made until the value recorded is normo-tensive (as defined in the protocol). The interval between the date of randomisation and the date of recording the first occurrence of a normo-tensive recording is the endpoint measure of interest. Such data are usually summarised using survival time methods.

A particular feature of time-to-event studies occurs when the endpoint cannot be determined. For example, in the trial monitoring the DBP it may be that a patient never becomes normo-tensive during the trial observation period. In this case the time from randomisation until the end of the trial observation period represents the time a patient has been under observation but has not yet become normo-tensive. Such a survival time is termed censored and is often denoted by, say, 28+, which here means the patient has been observed for 4 weeks but still remains hypotensive. In contrast, an observation of 28 means the patient has been observed for 4 weeks and became normo-tensive on the last observation day.

REPEATED MEASURES

In the trial taking repeated DBP assessments, these are recorded in order to determine a single outcome – ‘time to becoming normo-tensive’. In other situations, the successive values of DBP themselves may be utilised in making the formal comparisons. If the number of observations made on each subject is the same, then the analysis may be relatively straightforward, perhaps using repeated measures analysis of variance. On the other hand, if the number of observations recorded varies or if the intervals between successive observations vary from patient to patient or if there is occasional missing data, then the summary and analysis of such data may be quite complex. One option is to calculate the area under the curve (AUC) and use this as a single measure for each patient, thus avoiding the use of more complex analytical methods.[42]

However, the AUC method is now being superseded somewhat in Phase III trials by the use of general estimating equations and multilevel modelling. The technical details are beyond the scope of this book but most good statistics packages[43] now include facilities for these types of analyses. Nevertheless, these methods have not yet had much impact on the reporting of clinical trials, although a good example of their use has been provided by Brown *et al.*[44]

QUALITY OF LIFE

In many trials, endpoints such as the percentage of patients responding to treatment, survival time or direct measures such as DBP have been used. In other situations, more psychosocial measures have been utilised such as pain scores, perhaps measured using a visual analogue scale, and emotional functioning scores perhaps assessed by patients completing a questionnaire themselves. Such self-completed questionnaires have also been developed to measure aspects of quality of life (QoL) in patients undergoing treatment for their disease. One such instrument is the SF-36 of Ware and Sherbourne,[45] part of which is reproduced in Figure 2.6.

The QoL domains measured by these instruments may then be used as the definitive endpoints for clinical trials in certain circumstances. For example, in patients with terminal cancer the main thrust of therapy may be for palliation (rather than cure) so that aspects of QoL may be the primary concerns for any comparison of alternative approaches to management and care of such patients. If a single aspect of this QoL measured at one time point is to be used for comparison purposes then no new principles are required for either trial design purposes or analysis. On the other hand, and more usually, there may be several aspects of the QoL instrument that may need to be compared between treatment groups and these features will usually be assessed over time. This is further complicated by often unequal numbers of assessments available from each patient caused either by missing assessments in the series for a variety of reasons related or unrelated to their health status, or perhaps in terminal patients by their death. Fayers and Machin[46] and Fairclough[47] discuss these features of QoL data in some detail.

As we have discussed previously, there is also a problem associated with the numerous statistical tests of significance of the multiple QoL outcomes. These pose problems of interpretation which have also been addressed by Fayers and Machin[46] (Chapter 11). In short, a cautious approach is needed to ensure apparently ‘statistically significant’ differences are truly those. One way to overcome this problem is for the clinical protocol to rank the domains of QoL to be measured in terms of their relative importance and to confine the formal statistical tests and confidence intervals to these only.

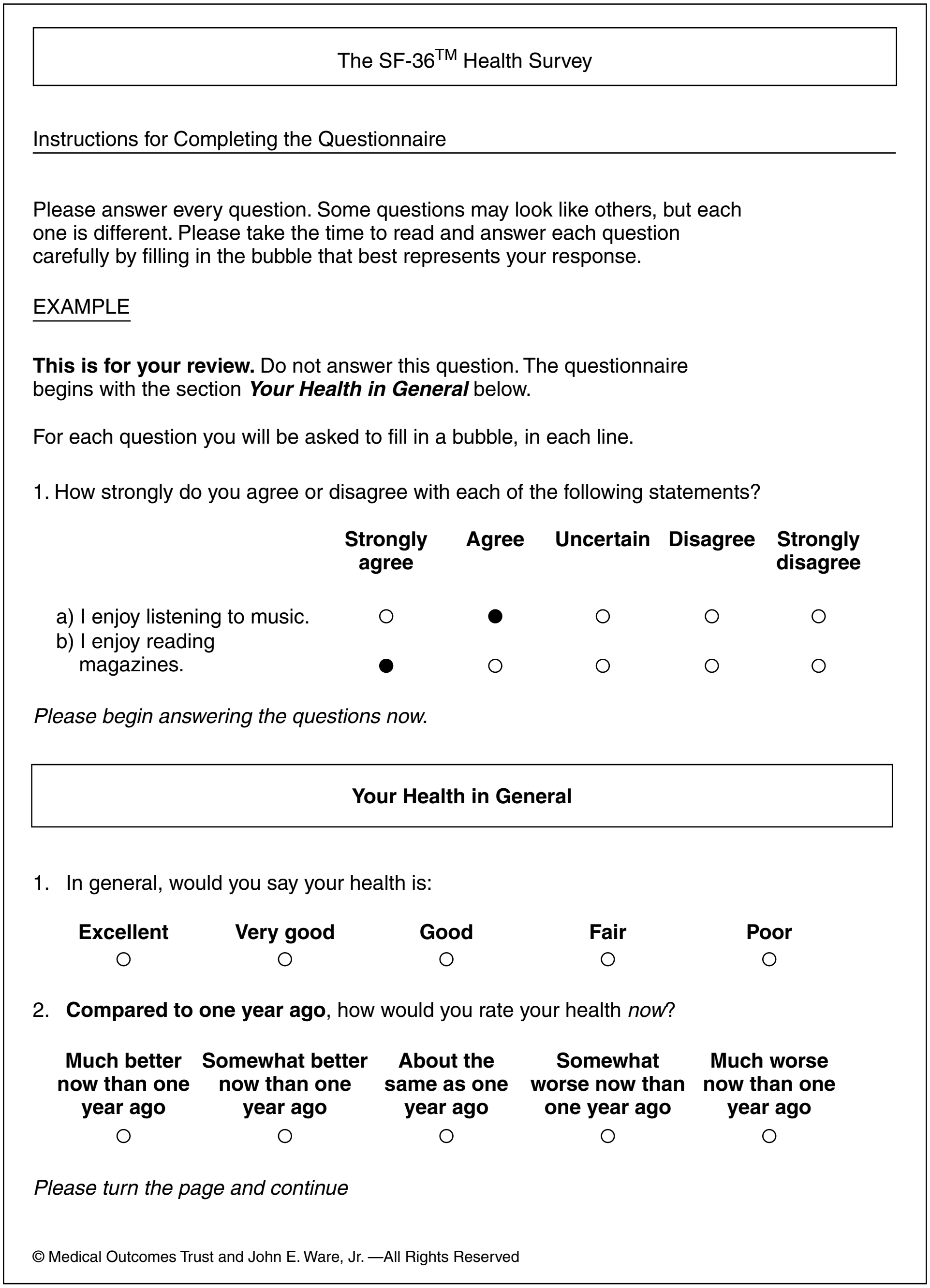
ECONOMIC EVALUATION

Most trials are intended primarily to address questions of efficacy. Safety is frequently an important (though secondary) objective. Health economics is increasingly often now evaluated as part of a randomised controlled trial.

There are four main types of cost analyses that are usually considered:

- *cost minimisation,* simply to determine the best treatment to minimise the total cost of treating the disease;
- *cost effectiveness,* a trade-off between the cost of caring for a patient and the level of efficacy offered by a treatment;
- *cost benefit,* a trade-off between the cost of caring for a patient and the overall benefit (not restricted to efficacy).
- *cost utility,* the trade-off between costs and all measures of ‘utility’ which may include efficacy, Q.L, greater life expectancy or increased productivity.

One of the big difficulties with pharmacoeconomic evaluations is determining what *indirect costs* should be considered. *Direct costs* are usually easier: costs of medication, costs of those giving the care (doctors, nurses, health visitors) and the basic costs of occupation of a hospital bed. Indirect costs include loss of earnings and productivity, loss of earnings and productivity of spouses or other family members who may care for a sick relative, and contribution to



The SF-36^{TM} Health Survey

Instructions for Completing the Questionnaire

Please answer every question. Some questions may look like others, but each one is different. Please take the time to read and answer each question carefully by filling in the bubble that best represents your response.

EXAMPLE

This is for your review. Do not answer this question. The questionnaire begins with the section ***Your Health in General*** below.

For each question you will be asked to fill in a bubble, in each line.

1. How strongly do you agree or disagree with each of the following statements?

	Strongly agree	**Agree**	**Uncertain**	**Disagree**	**Strongly disagree**
a) I enjoy listening to music.	○	●	○	○	○
b) I enjoy reading magazines.	●	○	○	○	○

Please begin answering the questions now.

Your Health in General

1. In general, would you say your health is:

Excellent	**Very good**	**Good**	**Fair**	**Poor**
○	○	○	○	○

2. **Compared to one year ago**, how would you rate your health *now*?

Much better now than one year ago	**Somewhat better now than one year ago**	**About the same as one year ago**	**Somewhat worse now than one year ago**	**Much worse now than one year ago**
○	○	○	○	○

Please turn the page and continue

For permission to use contact: Dr John Ware, Medical Outcomes Trust, 20 Park Plaza Suite 1014, Boston, MA 02116-4313, USA

Source: Reproduced from Ware and Sherbourne,[45] with permission.

Figure 2.6. Item Short Form Health Survey (SF-36) (from Ware and Sherbourne[45]).

hospital/pharmacy overhead costs. Because of the ambiguity associated with these indirect costs, most pharmacoeconomic evaluations performed as part of a clinical trial tend to focus solely on direct costs.

If we were to design a trial primarily to compare costs associated with different treatments we would follow the basic ideas of blinding and randomisation and then record subsequent costs incurred by the patient and the health provider. A very careful protocol would be necessary to define which costs are being considered so that this is measured consistently for all patients. A treatment that is not very effective might, for example, result in the patient needing more frequent consultations. The increased physician, nurse and other paramedical personnel contact time would then be recorded as a cost but it needs to be clear whether patient travel costs, for example (still direct costs, but not to the health service), are included, or not. However, most trials are aimed primarily at assessing efficiency and a limitation of investigating costs in a clinical trial is that the schedule of, and frequency of, visits by the patient to the physician may be very different to what it would be in routine clinical practice. Typically patients are monitored more frequently and more intensely in a trial setting than in routine clinical practice. The costs recorded, therefore, in a clinical trial may well be different (probably greater but possibly less) than in clinical practice. This is sometimes put forward as a major objection to pharmacoeconomic analyses carried out in conjunction with clinical trials. The same limitation does, of course, apply to efficacy evaluations: the overall level of efficacy seen in clinical trials is often not realised in clinical practice. However, if we keep in perspective that in a clinical trial, it is the *relative* efficacy of one treatment over another (even if one of them is a placebo) then this limitation, whilst still important, can be considered less of an overall objection. The same argument should be applied in pharmacoeconomic evaluations and the relative increase/decrease in costs of one treatment over another can be reported.

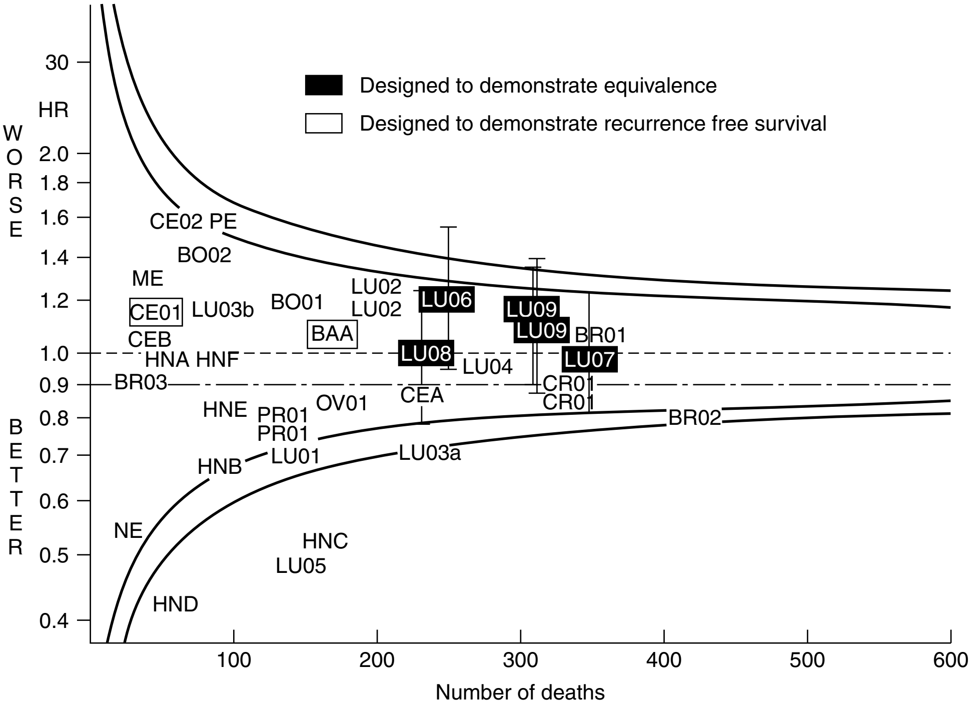
Recommendations of how trials incorporating health economics assessment have been given by the BMJ Economic Evaluation Working Party.[48] Neymark *et al.*[49] discuss some of the methodological issues as they relate to cancer trials.

TRIAL SIZE

When designing a new trial, a realistic assessment of the potential benefit (the anticipated effect size) of the proposed test therapy must be made at the onset. The history of clinical trial research suggests that, in certain circumstances, rather ambitious or over-optimistic views of potential benefit have been claimed at the design stage. This has led to trials of inadequate size for the questions posed.

The retrospective review by Machin *et al.*[50] of the published trials of the UK Medical Research Council in solid tumour cancers is summarised in the funnel plot of Figure 2.7. The benefit observed, as expressed by the hazard ratio (HR) for the new treatment, is plotted against the number of deaths reported in the trial publication. Those trials within the left-hand section of the funnel have relatively few deaths observed and so will be of correspondingly low power. Of these trials, some have an observed HR that is below the horizontal line at HR = 0.9. This line has been drawn at a level that is thought to represent a clinically worthwhile advantage to the test treatment. These specific points would have been outside the funnel had they been estimated from more observed deaths. Thus we might conclude from Figure 2.7 that, had these trials been larger, the corresponding treatments might have been observed to bring worthwhile benefit, rather than being dismissed as 'not statistically' significant.

However, it is a common error to assume that the lack of statistical significance following a test of hypothesis implies no difference between groups. Conversely a statistically significant result does not necessarily imply a clinically significant (important) result. Nevertheless, the



Source: After Machin *et al.*[50]

Figure 2.7. Retrospective review of UK Medical Research Council trials in solid tumours published prior to 1996 (after Machin *et al.*[50]).

message of Figure 2.7 is that potentially useful therapies may be overlooked if the trials are too small.

ANTICIPATED (PLANNING) EFFECT SIZE

A major factor in determining the size of a RCT is the anticipated effect size or clinically worthwhile difference. In broad terms, if this is not large then it should be of sufficient clinical, scientific or public health importance to warrant the consequentially large trial that will be required to answer the question posed. If the anticipated effect is large, the RCT will be relatively small, in which case the investigators may need to question their own 'optimistic' view of the potential benefit. In either case, a realistic view of the possible effect size is important. In practice, it is usually important to calculate the sample size for a range of values of the effect size. In this way the sensitivity of the resulting sample sizes to this range of values will provide options for the investigating team.

Estimates of the anticipated effect size may be obtained from the available literature, formal meta-analyses of related trials or may be elicited from clinical opinion. In circumstances where there is little prior information available, Cohen[51] has proposed a standardised effect size, Δ. In the case when the difference between two treatments A and B is expressed by the difference between their means $(\mu_A - \mu_B)$ and σ is the standard deviation (SD) of the endpoint variable which is assumed to be a continuous measure, then $\Delta = (\mu_A - \mu_B)/\sigma$. A value of $\Delta \leq 0.1$ is considered a small standardised effect, $\Delta \approx 0.5$ as moderate and $\Delta \geq 1$ as large (see also Day[3]). Experience has suggested that in many clinical areas these can be taken as a good practical guide for design purposes. However, for large simple trials, the equivalent of effects sizes as small as $\Delta = 0.05$ or less may be clinically important.

SAMPLE SIZE

Once the trial has been concluded, then a formal test of the null hypothesis of no difference between treatments is often made. We emphasise later that it is always important to provide an associated confidence interval for the estimate of treatment difference observed. The test of the null hypothesis has an associated false positive rate and the alternative hypothesis a false negative rate. The former is variously known also as the Type I error rate, test size or significance level, α. The latter is the Type II error rate β, and $1 - \beta$ is the power. When designing a clinical trial it is often convenient to think in hypothesis testing terms and so set α and β and a specific effect size Δ for consideration. For determining the size of a trial, α and β are typically taken as small; for example, $\alpha = 0.05$ (5%) and $\beta = 0.1$ (10%) or equivalently the power $1 - \beta = 0.9$ (90%) is large.

If the trial is ultimately to compare the means obtained from the two treatment groups, then with randomisation to each treatment in equal numbers, the total sample size, N, is given by

$$N = \frac{4(z_{1-\alpha/2} + z_{1-b})^2}{\Delta^2}, \qquad (2.1)$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are obtained from tables of the standardised normal distribution for given α and β.

If we set in equation (2.1) a two-sided $\alpha = 0.05$ and a power of $1 - \beta = 0.9$, then $z_{1-\alpha/2} = z_{0.975} = 1.96$ and $z_{1-\beta} = z_{0.9} = 1.2816$, so that $N = 42.028/\Delta^2 \approx 42/\Delta^2$. For large, moderate and small sizes of Δ of 1, 0.5 and 0.1, the corresponding sample sizes are 42, 168 and 4200 respectively. More realistically these may be rounded to 50, 200 and 4500. For a large simple trial with $\Delta = 0.05$, this implies 16 000 patients may be recruited.

This basic equation has to be modified to adapt to the specific trial design (parallel group, factorial, crossover or sequential), the type of randomisation, the allocation ratio, as well as the particular type of endpoint under consideration. Machin *et al.*[52] provide examples for many different situations.

A good clinical trial design is that which will answer the question posed with the minimum number of subjects possible. An excessively large trial not only incurs higher costs but also is unethical. Too small a trial size leads to inconclusive results, since there is a greater chance of missing the clinically important difference, a resulting waste of resources and, this too is unethical.

MONITORING TRIAL PROGRESS

DATA MONITORING COMMITTEES

It is clear that a randomised controlled trial is a major undertaking and which clearly involves human subjects in the process. Thus, as we have stated, it is important that some form of equipoise in respect to the treatments under test is required to justify the randomisation. However, once the trial is in progress, information accumulates and as it does so it may be that the initial equipoise becomes disturbed. Indeed the very point of a clinical trial is to upset the equipoise in favour of the best (if indeed one truly is) treatment.

Clearly there will be circumstances when such early information may be sufficient to answer convincingly the question posed by the trial. In this case the trial should close to further patient entry. One circumstance when this will arise is when the actual benefit far exceeds that which the design team envisaged. For example, Lau *et al.*[53] stopped a trial in patients with respectable hepatocellular carcinoma after early results on 43 patients suggested a substantial benefit to adjuvant intra-arterial iodine-131-labelled lipiodol. Their decision was subsequently criticised by Pocock and White[54] who suggested the result was ‘Too good to be true’ as early stopping may yield biased estimates of the treatment effect. A confirmatory trial is now in progress to substantiate or refute these findings.[55] Essentially, although very promising, the trial results as published provided insufficient evidence for other clinical teams to adopt the test therapy for their patients.

Nevertheless in this, and for the majority of clinical trials, it is clearly important to monitor the accumulating data. It has also been recognised that such monitoring should be reviewed (not by the clinical teams involved in entering patients into the trial themselves) but by an independent Data Monitoring Committee (DMC). The membership and remit of a DMC will usually depend on the particular trial(s) under review. For example, the European Organisation for the Research on the Treatment of Cancer has a standing committee of three clinical and one statistical member, none of whom are involved in any way with the trials under review. This independent DMC reviews reports on trial progress prepared by the data centre teams and makes specific recommendations to the relevant trial coordinating group. Early thoughts on the structure of DMCs for the UK Medical Research Council Cancer Therapy Committee are provided by Parmar and Machin.[56] To emphasise the importance of 'Independence' such committees sometimes choose the acronym IDMC.

SAFETY

Although an IDMC will be concerned with the relative efficacy of the treatments under test, issues of safety will also be paramount in many circumstances. In many cases, safety issues may dominate the early stages of a trial when relatively new and untested treatment modalities are first put into wider use, whereas in the later stages detailed review of safety may not be required as no untoward experiences have been observed in the early stages. In contrast, serious safety issues may force a recommendation for early closure of the trial even in situations where early indications of benefit in terms of efficacy are present. Clearly the role of the IDMC or (in view of the 'Safety' aspects) the IDMSC is to provide a balanced judgement on these possibly conflicting aspects when making their report. This judgement will derive from the current evidence from the trial itself, external evidence perhaps on new information since the trial was inaugurated, and their own collective experience.

INTERIM ANALYSIS AND EARLY STOPPING RULES

At the planning stage of a clinical trial the design team will be aware of the need to monitor the progress of the trial by reports to an IDMSC. On these occasions the data centre responsible for the conduct of the trial will expect to prepare reports on many aspects of trial progress including especially safety and efficacy. This requirement is often detailed in the trial protocol. The detail may specify those aspects that are likely to be of major concern and also the timing (often expressed in terms of patient numbers or events observed) of such reports.

An interim report may include a formal (statistical) comparison of treatment efficacy. This comparison will then be repeated on the accumulating data for each IDMSC and finally following the close of the trial once the relevant data are to hand. These repeated statistical tests raise the possibility of an increased chance of falsely declaring a difference in efficacy between treatments. To compensate for this, methods of adjusting for the multiple looks at the accumulating data have been devised. Many of these are reviewed by Piantadosi[21] (Chapter 10).

Several of these methods of interim analysis also include 'stopping rules'; that is, they incorporate procedures or boundaries which once crossed by the data under review imply that the trial should terminate. However, all these methods are predicated on obtaining timely and complete data, very rapid analysis and report writing and immediate review by the IDMSC.[57] They also focus on only one aspect (usually efficacy) and so do not provide a comprehensive view of the whole situation.

The nature of the essential balance required between a formal statistical approach to interim looks at the data and the less structured nature of IDMSC decision making is provided by Ashby and Machin,[58] Machin[59] and Piantadosi[21]

(Chapter 10.8). Parmar *et al.*[60] describe an approach for monitoring large trials using Bayesian methods.

REPORTING CLINICAL TRIALS

The first rule after completing a clinical trial is to report the results – whether they are positive, negative or equivocal. Selective reporting whereby results of positive studies tend to be published and negative studies tend not to be published presents a distorted view of the true situation. This approach to reporting is particularly important for clinical trial overviews and meta-analysis where it is clearly important to be able to include *all* relevant studies (not just the published ones) in the overall synthesis.

The second aspect of reporting is the standard of reporting, particularly the amount of necessary detail given in any trial report. The most basic feature that has repeatedly been emphasised is to give estimates (with confidence intervals) of treatment effects and not just p-values. Guidelines for referees (useful also for authors) have been published in several journals including those of the *British Medical Journal*.[35] The Consolidation of the Standards of Reporting Trials (CONSORT) statement is an international recommendation adopted by many leading medical journals.[61]

One particular feature of the CONSORT statement is that the outcomes of ‘all’ patients randomised to a clinical trial are to be reported. Thus a full note has to be provided on those, for example, who post-randomisation refuse the allocation and perhaps then insist on the competitor treatment. Two examples of how the patient flow through a trial is summarised are given in Figure 2.8.

It is of some interest to note that the writing team for Lau *et al.*[53] were encouraged by the journal to include information on late (post-interim analysis) randomisations into their report. It is clear that no such stipulation was required of the MRC Renal Cancer Collaborators.[29]

As indicated, the statistical guidelines referred to, and the associated checklists for statistical review of papers for international journals,[62] require confidence intervals (CIs) to be given

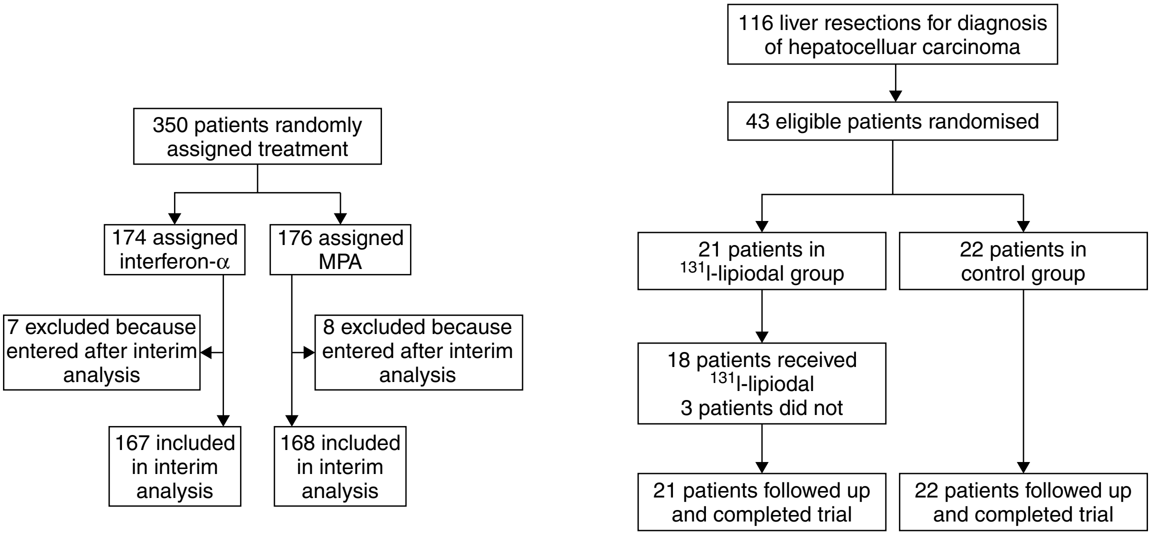

350 patients randomly assigned treatment
174 assigned interferon-α
176 assigned MPA
7 excluded because entered after interim analysis
8 excluded because entered after interim analysis
167 included in interim analysis
168 included in interim analysis
116 liver resections for diagnosis of hepatocelluar carcinoma
43 eligible patients randomised
21 patients in ¹³¹I-lipiodal group
22 patients in control group
18 patients received ¹³¹I-lipiodal 3 patients did not
21 patients followed up and completed trial
22 patients followed up and completed trial

Figure 2.8. Trial profiles following the CONSORT guidelines (after MRC Renal Cancer Collaborators,[29] Lau *et al.*[53]) © (1999), reprinted with permission from Elsevier.

for the main results. These are intended as an important prerequisite to be supplemented by the p-value from the associated hypothesis test. Methods for calculating CIs are provided in many standard statistical packages as well as the specialist software of Altman *et al.*[7] (Chapter 17).

INTENTION-TO-TREAT (ITT)

As we have indicated, once patients have been randomised they should start treatment as specified in the protocol as soon as it is practicably possible. For the severely burnt patients either *MEBO* or *C* dressings can be immediately applied. On the other hand, if patients, once randomised, have then to be scheduled for surgery, then there may be considerable delay before therapy is activated. This delay may provide a period in which the patients change their minds about consent or indeed in those with life-threatening illness some may die before the scheduled date of surgery. Thus, the number of patients actually starting the protocol treatment allocated may be less than the number randomised to receive it. The 'intention-to-treat' principle is that once randomised the patient is retained in that group for analysis whatever occurs, even in situations where a patient after consent is randomised to (say) A but then refuses and even insists on being treated by option B. The effect of such a patient is to dilute the estimate of the true difference between A and B. However, if such a patient was analysed as if allocated to treatment B, then the trial is no longer properly randomised and the resulting comparison may be seriously biased.

However, in certain circumstances, the 'intention-to-treat' may be replaced or supplemented by a 'per protocol' summary.[63] For example, if the toxicity and/or side-effects profile of a new agent are to be summarised, any analysis including those patients who were randomised to the drug but then did not receive it (for whatever reason) could seriously underestimate the true levels. If this is indeed appropriate for such endpoints, then the trial protocol should state that such an analysis is intended from the onset.

One procedure that used to be in widespread use was, once the protocol treatment and follow-up were complete and all the trial specific information collected on a patient, to review these data in detail. This review would, for example, check that the patient eligibility criteria were satisfied and that there had been no important protocol deviations while on treatment. Any patients, following this review, then found to be ineligible or protocol violators would then, in principle, be set aside and excluded from the trial results.

One particular problem is one in which patients are recruited to a trial on the basis of clinical examination during which a biopsy specimen is taken and sent for review. In the meantime the patient is randomised and treatment commenced, but once the report is returned the patient is found not to comply with the eligibility criteria. The above review process would automatically exclude this patient whereas Freedman and Machin[64] argue otherwise.

Usually, this review would not be blind to the treatment received (in fact even if the trial is double-blind): there may be clues once the data are examined in this way as to which treatment is which. As a consequence, this process would tend to exclude more patients on the more aggressive treatment. For example, the review conducted by Machin *et al.*[50] of some early randomised trials in patients with cancer conducted by the UK Medical Research Council showed that the earlier publications systematically reported on fewer patients in the more aggressive treatment arm despite a 1:1 randomisation. The exclusion of a larger proportion of patients receiving the more aggressive therapy would tend to bias the results in its favour. Thus any patients who had 'difficulties' with the treatment, perhaps the more sick patients, were not included in the assessment of its efficacy. This type of exclusion was widespread practice, the consequences of which included the development of ITT policies and standards for reporting clinical trials, the latter policy insisting that the progress of all the randomised patients should be reported.

In general, the application of ITT is conservative in the sense that it will tend to dilute between treatment differences, Piaggio and Pinol[65] have pointed out that for equivalence trials ITT will not be conservative but will tend to favour the equivalence hypothesis.

SELECTED FURTHER READING ON CLINICAL TRIALS

GENERAL

Day S. *Dictionary for Clinical Trials*. Chichester: John Wiley & Sons, Ltd (1999).

Redmond C, Colton T, eds. *Biostatistics in Clinical Trials*. Chichester: John Wiley & Sons, Ltd (2000).

Senn SJ. *Statistical Issues in Drug Development*. Chichester: John Wiley & Sons, Ltd (1997).

Williams CJ, ed. *Introducing New Treatments for Cancer: Practical, Ethical and Legal Problems*. Chichester: John Wiley & Sons, Ltd (1992).

DESIGN AND CONDUCT OF CLINICAL TRIALS

Buyse ME, Staquet MJ, Sylvester RJ, eds. *Cancer Clinical Trials: Methods and Practice*. Oxford: Oxford University Press (1984).

Crowley J, ed. *Handbook of Statistics in Clinical Oncology*. New York: Marcel Dekker (2000).

Girling DJ, Parmar MKB, Stenning SP, Stephens RJ, Stewart LA, eds. *Clinical Trials in Cancer: Principles and Practice*. Oxford: Oxford University Press (2003).

Jadad A. *Randomised Controlled Trials: A User's Guide*. London: British Medical Journal Books (1998).

Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC (2000).

Machin D, Campbell MJ. *Design of Studies for Medical Research*. Chichester: John Wiley & Sons, Ltd (2005).

Matthews JNS. *Introduction to Randomised Controlled Clinical Trials*. London: Arnold (2000).

McFadden ET. *Management of Data in Clinical Trials*. New York: John Wiley & Sons, Inc. (1997).

Piantadosi S. *Clinical Trials: A Methodologic Perspective*. New York: John Wiley & Sons, Inc. (1997).

Pocock SJ. *Clinical Trials: A Practical Approach*. Chichester: John Wiley & Sons, Ltd (1983).

Senn SJ. *Cross-over Trials in Clinical Research*. Chichester: John Wiley & Sons, Ltd (1993).

Whitehead J. *Design and Analysis of Sequential Clinical Trials*, Revised 2nd edn. Chichester: John Wiley & Sons, Ltd (1997).

Wooding WM. *Planning Pharmaceutical Clinical Trials*. New York: John Wiley & Sons, Inc. (1994).

QUALITY OF LIFE

Fairclough DL. *Design and Analysis of Quality of Life Studies in Clinical Trials*. Andover: CRC Press (2002).

Fayers PM, Machin D. *Quality of Life: Assessment, Analysis and Interpretation of Patient-Reported Outcomes*. 2nd Edition. Chichester: John Wiley & Sons, Ltd (2007).

Fayers PM and Hays R (Eds). *Assessing Quality of Life in Clinical Trials: Methods and Practice* (2nd Edition). Oxford: Oxford University Press (2006).

SAMPLE SIZE

Machin D, Campbell MJ, Fayers PM, Pinol APY. *Sample Size Tables for Clinical Studies*. Oxford: Blackwell Science (1997).

EVIDENCE-BASED MEDICINE

Altman DG, Chalmers I, eds. *Systematic Reviews*. London: British Medical Journal Books (1995).

Leanndro G. *Meta-analysis in Medical Research*: Oxford: BMJ Books, Blackwell Publishing (2005).

Stevens A, Abrams KR, Brazier J, Fitzpatrick R, Lilford R. *Methods in Evidence Based Healthcare*. London: Sage (2001).

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods of Meta-Analysis in Medical Research*. Chichester: John Wiley & Sons, Ltd (2000).

REFERENCES

1. World Medical Association. World Medical Association Declaration of Helsinki, 1996. Republished in *J Am Med Assoc* (1997) **277**: 925–6.
2. Chalmers I, Sackett D, Silagy C. Cochrane collaboration. In: C Redmond, T Colton, eds, *Biostatistics in Clinical Trials* Chichester: John Wiley & Sons, Ltd (2000) 65–7.
3. Day S. *Dictionary for Clinical Trials*. Chichester: John Wiley & Sons, Ltd (1999).
4. Pocock SJ. *Clinical Trials: A Practical Approach*. Chichester: John Wiley & Sons, Ltd (1983).
5. Temple R. Current definitions of phases of investigation and the role of the FDA in the conduct of clinical trials. *Am Heart J* (2000) **139**: S133–5.
6. Gehan EA, Freireich EJ. Non-randomized controls in cancer clinical trials. *New Eng J Med* (1974) **290**: 198–203.
7. Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with Confidence*, 2nd edn. London: British Medical Journal Books (2000).

8. Ang ES-W, Lee S-T, Gan CS-G, See PG-J, Chan Y-H, Ng L-H, Machin D. Evaluating the role of alternative therapy in burn wound management: randomized trial comparing Moist Exposed Burn Ointment with conventional methods in the management of patients with second-degree burns. *Medscape Gen Med* (2001) **3**(2): 3.
9. Freedman B. Equipoise and the ethics of clinical research. *New Eng J Med* (1987) **317**: 141–5.
10. Weijer C, Shapiro SH, Glass KC. Clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial. *Br Med J* (2000) **321**: 756–8.
11. Enkin MW. Clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial. *Br Med J* (2000) **321**: 756–8.
12. Chow PK-H, Tai B-C, Tan C-K, Machin D, Johnson PJ, Khin M-W, Soo K-C. No role for high-dose tamoxifen in the treatment of inoperable hepatocellular carcinoma: an Asia-Pacific double-blind randomised controlled trial. *Hepatology* (2002) **36**: 1221–6.
13. RISC Group. Risk of myocardial infarction and death during treatment with low dose aspirin and intravenous heparin in men with unstable coronary artery disease. *Lancet* (1990) **336**: 827–30.
14. The ATAC (Arimidex, Tamoxifen Alone or in Combination) Trialists' Group. Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomised trial. *Lancet* (2002) **359**: 2131–39.
15. Piya-Anant M, Koetsawang S, Patrasupapong N, Dinchuen P, d'Arcangues C, Piaggio G, Pinol A. Effectiveness of cyclofem® in the treatment of depo medroxyprogesterone acetate induced amenorrhea. *Contraception* (1998) **57**: 23–8.
16. Donner A, Piaggio G, Villar J, Pinol A, Al-Mazrou Y, Ba'aqeel H, Bakketeig L, Belizán, JM, Berendes H, Carroli G, Farnot U, Lumbiganon P. Methodological considerations in the design of the WHO antenatal care randomised controlled trial. *Paediat Perinatal Epidemiol* (1998) **12** (Suppl 2): 59–74.
17. Zelen M. A new design for randomized clinical trials. *New Eng J Med* (1979) **300**: 1242–5.
18. Wooding WM. *Planning Pharmaceutical Clinical Trials*. New York: John Wiley & Sons, Inc. (1994).
19. Chan DC, Watts GF, Mori TA, Barrett PH, Beilin LJ, Redgrave TG. Factorial study of the effects of atorvastatin and fish oil on dyslipidaemia in visceral obesity. *Eur J Clin Invest* (2002) **32**: 429–36.
20. Day L, Fildes B, Gordon I, Fitzharris M, Flamer H, Lord S. Randomised factorial trial of falls prevention among older people living in their own homes. *Br Med J* (2002) **325**: 128–1.
21. Piantadosi S. Factorial designs. In: Redmond C and Colton T, eds, *Biostatistics in Clinical Trials*. Chichester: John Wiley & Sons, Ltd (2000).
22. Green S. Factorial designs with time-to-event end points. In: J Crowley ed., *Handbook of Statistics in Clinical Oncology*, New York: Marcel Dekker (2000) Chapter 9, 161–71.
23. Hong B, Ji YH, Hong JH, Nam KY, Ahn Ty. A double-blind crossover study evaluating the efficacy of Korean red ginseng in patients with erectile dysfunction: a preliminary report. *J Urol* (2002) **168**: 2070–73.
24. Senn SJ. *Cross-over Trials in Clinical Research*. Chichester: John Wiley & Sons, Ltd (2002).
25. Simon R (2000) Therapeutic equivalence trials. In J Crowley (ed) *Handbook of Statistics in Clinical Oncology*. Chapter 10, pp 173–187. Marcel Dekker, New York.
26. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *Br Med J* (1996) **313**: 36–39.
27. Donaldson AN, Whitehead J, Stephens R, Machin D. A simulated sequential analysis based on data from two MRC trials. *Br J Cancer* (1993) **68**: 1171–78.
28. Fayers PM, Cook PA, Machin D, Donaldson AN, Whitehead J, Ritchie A, Oliver RTD, Yuen P. On the development of the Medical Research Council trial of α-interferon in metastatic renal carcinoma. *Stat Med* (1994) **13**: 2249–60.
29. Medical Research Council Renal Cancer Collaborators. Interferon-α and survival in metastatic renal carcinoma: early results of a randomised controlled trial. *Lancet* (1999) **353**: 14–17.
30. MPS Research Unit. *PEST 4: Operating Manual*. University of Reading (2000).
31. Whitehead J. Use of the triangular test in sequential clinical trials. In: Crowley J ed, *Handbook of Statistics in Clinical Oncology*. New York: Marcel Dekker (2000) Chapter 12, 211–28.
32. Whitehead J. *Design and Analysis of Sequential Clinical Trials*, Revised 2nd edn. Chichester: John Wiley & Sons, Ltd (1997).
33. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman and Hall/CRC (2000).
34. Zelen M. Randomised consent trials. *Lancet* (1992) **340**: 375.
35. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. In: Altman DG, Machin D, Bryant TN, Gardner MJ, *Statistics with Confidence*, 2nd edn. London: British Medical Journal Books, 171–90 (2000).

36. Machin D, Lee ST. The ethics of randomised trials in the context of cleft palate research. *Plastic Reconstruct Surg* (2000) **105**: 1566–68.
37. Spiegelhaleter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. *J R Stat Soc* (1994) **A 157**: 357–416.
38. Tan S-B, Dear KBG, Bruzzi P, Machin D. Strategy for randomized clinical trials in rare cancers. *Br Med J* (2003) **327**: 47–9.
39. Tan S-B, Machin D, Tai B-C, Foo K-F, Tan E-H. A Bayesian re-assessment of two Phase II trials of gemcitabine in metastatic nasopharyngeal cancer. *Br J Cancer* (2002) **86**: 843–50.
40. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *Clin Pharmacol Ther* (1974) **15**: 443–53.
41. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* (1975) **31**: 103–5.
42. Matthews JNS, Altman DG, Campbell MJ, Royston JP. Analysis of serial measurements in medical research. *Br Med J* (1990) **300**: 230–5.
43. StataCorp. *STATA Statistical Software: Release 9.0*. Stata Corporation, College Station, TX (2001).
44. Brown JE, King MT, Butow PN, Dunn SM, Coates AS. Patterns over time in quality of life, coping and psychological adjustment in late stage melanoma patients: an application of multilevel models. *Qual Life Res* (2000) **9**: 75–85.
45. Ware JE, Sherbourne CD. The MOS 36-Item Short Form Health Survey (SF-36) I: conceptual framework and item selection. *Med Care* (1991) **30**: 473–83.
46. Fayers PM, Machin D. *Quality of Life: Assessment, Analysis and Interpretation*. Chichester: John Wiley & Sons, Ltd (2000).
47. Fairclough DL. *Design and Analysis of Quality of Life Studies in Clinical trials*. Andover: CRC Press (2002).
48. BMJ Economic Evaluation Working Party. Guidelines for authors and peer reviewers of economic submissions to the BMJ. *Br Med J* (1996) **313**: 275–83.
49. Neymark N, Kiebert W, Torfs K, Davies L, Fayers P, Hillner B, Gelber R, Guyatt G, Kind P, Machin D, Nord E, Osoba D, Revicki D, Schulman K, Simpson K. Methodological and statistical issues of quality of life (QoL) and economic evaluation in cancer clinical trials: report of a workshop. *Eur J Cancer* (1998) **34**: 1317–33.
50. Machin D, Stenning SP, Parmar MKB, Fayers PM, Girling DJ, Stephens RJ, Stewart LA, Whaley JB. Thirty years of Medical Research Council randomized trials in solid tumours. *Clin Oncol* (1997) **9**: 20–28.
51. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Earlbaum (1988).
52. Machin D, Campbell MJ, Fayers PM, Pinol APY. *Sample Size Tables for Clinical Studies*. Oxford: Blackwell Science (1997).
53. Lau WY, Leung TWT, Ho SKW, Chan M, Machin D, Lau J, Chan ATC, Yeo W, Mok TSK, Yu SCH, Leung NWY, Johnson PJ. Adjuvant intra-arterial iodine-131-labelled lipiodol for respectable hepatocellular carcinoma: a prospective randomised trial. *Lancet* (1999) **353**: 797–801.
54. Pocock SJ, White I. Trials stopped early: too good to be true? *Lancet* (1999) **353**: 797–801.
55. Tan SB, Machin D, Cheung YB, Chung YFA, Tai BC. Following a trial that stopped early: what next for adjuvant intra-arterial iodine-131-labelled lipiodol in respectable hepatocellular carcinoma? *J Clin Oncol* (2002) **20**: 1709.
56. Parmar MKB, Machin D. Monitoring clinical trials: experience of, and proposals under consideration by, the Cancer Therapy Committee of the British Medical Research Council. *Stat Med* (1993) **12**: 497–504.
57. Machin D. Discussion of 'The what, why and how of Bayesian clinical trials monitoring'. *Stat Med* (1994) **13**: 1385–89.
58. Ashby D, Machin D. Stopping rules, interim analysis and data monitoring committees. *Br J Cancer* (1993) **68**: 1047–50.
59. Machin D. Interim analysis and ethical issues in the conduct of trials. In: Williams CJ ed, *Introducing New Treatments for Cancer: Practical, Ethical and Legal Problems*. Chichester: John Wiley & Sons, Ltd (1992) Chapter 15.
60. Parmar MKB, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E. Monitoring of large randomised clinical trials: a new approach with Bayesian methods. *Lancet* (2001) **358**: 375–81.
61. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schultz KF, Simel D, Stroup DF. Improving the quality of reporting randomized controlled trials: the CONSORT statement. *J Am Med Assoc* (1996) **276**: 637–9.
62. Gardner MJ, Machin D, Campbell MJ, Altman DG. Statistical checklists. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds, *Statistics with Confidence*, 2nd edn. London: British Medical Journal Books (2000) 101–10.
63. Lewis JA, Machin D. Intention to treat: who should use ITT. *Br J Cancer* (1993) **68**: 647–50.
64. Freedman LS, Machin D. Pathology review in cancer research. *Br J Cancer* (1993) **68**: 1047–50.
65. Piaggio G, Pinol APY. Use of the equivalence approach in reproductive health. *Stat Med* (2001) **20**: 3571–87.

3

The Cochrane Collaboration

MIKE CLARKE
UK Cochrane Centre, NHS R & D Programme, Oxford OX2 7LG, UK

INTRODUCTION

The Cochrane Collaboration is an international organisation dedicated to helping people make well-informed decisions about health care. It does this through preparing, maintaining and promoting the accessibility of systematic reviews of the effects of health care interventions. Most of these reviews are based on randomised trials but, in some circumstances, other types of study might be brought together, appraised, summarised and combined within the Cochrane review. This chapter begins by setting the scene for the need for systematic reviews, outlines how these reviews are done and then describes The Cochrane Collaboration as of December 2005, 12 years after it was established.

THE NEED FOR SYSTEMATIC REVIEWS

At the end of the nineteenth century, at the founding meeting of the Association of Medical Librarians in Philadelphia, USA, George Gould included a vision of the future of health information in his inaugural address. 'I look forward,' he said, 'to such an organisation of the literary records of medicine that a puzzled worker in any part of the civilized world shall in an hour be able to gain a knowledge pertaining to a subject of the experience of every other man in the world'.[1] That was in 1898. Now, in the early years of the twenty-first century, many might share a modified version of Gould's vision in which *everyone* making a decision about health care, whether it be *their own or someone else's*, would be able to obtain the necessary *knowledge from good-quality research*, which they need to make the best decision possible. Unfortunately, the vision has not been realised. Many challenges, and some solutions that Gould is unlikely to have thought of, have arisen through the intervening decades.

The conduct and acceptance of systematic reviews is one of the solutions but it is not a recent phenomenon. The idea of bringing together evidence in a systematic way was described at another conference slightly earlier in the nineteenth century, and further north on the American continent. Lord Rayleigh, at the 1884 meeting of the British Association for the Advancement of Science in Montreal, said:

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

> If, as is sometimes supposed, science consisted in nothing but the laborious accumulation of facts, it would soon come to a standstill, crushed, as it were, under its own weight. The suggestion of a new idea, or the detection of a law, supersedes much that has previously been a burden on the memory, and by introducing order and coherence facilitates the retention of the remainder in an available form. Two processes are thus at work side by side, the reception of new material and the digestion and assimilation of the old. One remark, however, should be made. The work which deserves, but I am afraid does not always receive, the most credit is that in which discovery and explanation go hand in hand, in which not only are new facts presented, but their relation to old ones is pointed out.[2]

In the late twentieth century, the arrival of electronic publishing and the ability to use this to update summaries of research through the digestion and assimilation of the old, alongside the reception of new material, brought together these concepts of up to date, 'living' summaries of research – research synthesis. The arrival of the internet provided a mechanism for the rapid dissemination of these up to date summaries.

However, the task of doing such reviews also became considerably more difficult during the century as the amount of research and the volume of the health care literature exploded. A meeting of the Medical Library Association, nearly 60 years after the first, and nearly half a century ago, provides ample illustration of this. John Bugher told the meeting: 'The sheer volume of publication in the medical area, as all of you are very much aware, becomes more formidable each year.' He added, 'I am told that this increase in medical literature is exponential and that it tends to double each fifteen or sixteen years.'[3] If Bugher was concerned about this in 1957, one can only wonder what he would have thought of the situation in the 1990s, when there were more than 20 000 health care journals, publishing a total of 2 million articles per year,[4] and little hope that people needing to gather knowledge to help them make a decision about health care would be able to do so in dozens or hundreds of hours, never mind a single one.

THE CONDUCT OF SYSTEMATIC REVIEWS

Systematic reviews represent one attempt to do the work of identifying, assessing and summarising the relevant research in order to help the person make a health care decision. These reviews allow that person to consider the relevant evidence, without having to find it all for themselves. In order for such summaries of evidence to be reliable enough to support well-informed decisions, they need to be robust and to minimise bias. These biases are both those that might exist in the research being summarised and those in the conduct of the review. The former will be dealt with by choosing studies that have used designs most appropriate to the topic being investigated. In the case of the effects of different interventions, this leads to reliance on randomised trials. But there is nothing intrinsic in systematic reviews that makes them only relevant to randomised trials. Systematic reviews are needed, and have been done, of other types of study where these would be the most appropriate to tackle the question of interest. Thus, systematic reviews of prognosis, test accuracy, burden of disease and genetic predisposition have all been done, and should be done, as a means of providing the most reliable summary possible of the existing research. Systematic reviews should be done to inform health care decisions but also to inform the design of new research, to ensure that these new studies are the most appropriate for moving knowledge forward.[5]

The conduct of a systematic review requires that the objectives and eligibility criteria for the review are set out clearly in advance. This should include the decisions, and the rationale behind these decisions, for which types of study, interventions, participants and outcome measures will be included in the review. The reviewers would then seek as many of the relevant studies as possible. To overcome publication bias, they will need to look for both published and unpublished studies. Having identified the studies, they need to appraise them. If the aim of the review is to produce as reliable a summary as possible of the existing research, it needs to include only research that has been done to an adequate standard. After deciding on what is eligible for the review, the reviewers would then compile as complete a data set as they can and, if appropriate and possible, they might combine the results of the individual studies in a meta-analysis. This will provide a more mathematically precise estimate of the effect measured in the independent studies but will be meaningless if the studies are not sufficiently similar for an average of their findings to be sensible. The review needs to be written in a structured way, to enable users to find what they are seeking within it, and, ideally, it should be updated periodically to reflect new research, and other evidence and information about the subject it covers.

SYSTEMATIC REVIEWS AND THE COCHRANE COLLABORATION

The preceding paragraph describes the process for preparing and maintaining systematic reviews in general, and is that followed by Cochrane reviews. And, as noted above, these reviews are predominantly reviews of randomised trials. This is because of the overarching aim of helping people make well-informed decisions about the *effects* of health care interventions. As such, the minimisation of bias that is only possible through random allocation becomes a key feature of the studies to be included.

HISTORY OF THE COCHRANE COLLABORATION

The Cochrane Collaboration is the largest organisation in the world engaged in the production and maintenance of systematic reviews. It was established in 1993, at the first Cochrane Colloquium. A year earlier, Iain Chalmers, Kay Dickersin and Thomas Chalmers wrote an editorial in the British Medical Journal describing the need for such an organisation[6] and drawing inspiration from the following statement by Archie Cochrane, published in 1972: ‘It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, updated periodically, of all relevant randomised controlled trials’.[7]

The editorial coincided with the opening of the first Cochrane Centre in Oxford, UK. There are now 12 Cochrane Centres around the world (see below). A chronology of significant events in the – still short – history of The Cochrane Collaboration is available on its web site: www.cochrane.org.

The Cochrane Collaboration has 10 guiding principles:

- Collaboration, by internally and externally fostering good communications, open decision making and teamwork.
- Building on the enthusiasm of individuals, by involving and supporting people of different skills and backgrounds.
- Avoiding duplication, by good management and coordination to maximise economy of effort.
- Minimising bias, through a variety of approaches such as scientific rigour, ensuring broad participation, and avoiding conflicts of interest.
- Keeping up to date, by a commitment to ensure that Cochrane reviews are maintained through identification and incorporation of new evidence.
- Striving for relevance, by promoting the assessment of health care interventions using outcomes that matter to people making choices in health care.
- Promoting access, by wide dissemination of the outputs of The Cochrane Collaboration, taking advantage of strategic alliances, and by promoting appropriate prices, content and media to meet the needs of users worldwide.
- Ensuring quality, by being open and responsive to criticism, applying advances in methodology, and developing systems for quality improvement.
- Continuity, by ensuring that responsibility for reviews, editorial processes and key functions is maintained and renewed.

- Enabling wide participation in the work of The Cochrane Collaboration by reducing barriers to contributing and by encouraging diversity.

The first Cochrane Colloquium, at which the international Cochrane Collaboration was established, took place in Oxford in October 1993, with 77 people from 19 countries. Since then there have been a series of annual Cochrane Colloquia in different parts of the world, with upwards of 1000 people attending. The most recent colloquia were in Ottawa, Canada, in 2004 and Melbourne, Australia, in 2005.

There are currently more than 13 000 people actively involved in the work of The Cochrane Collaboration, in almost 100 countries. The number of people involved has increased by 10–20% each year through the five years to 2005. During this time, particular emphasis has been placed on increasing the involvement of people from low- and middle-income countries. This is reflected in the even more rapid increase in the number of people actively involved in the preparation and maintenance of Cochrane reviews from these countries. The total has risen from about 300 in the year 2000 to more than 1200 in 2005. The work of all these people and of The Cochrane Collaboration generally is supported by hundreds of different organisations, including health service providers, research funding agencies, governments, international organisations and universities.

STRUCTURE OF THE COCHRANE COLLABORATION

Cochrane reviews are prepared and maintained within Cochrane Review Groups. There are about 50 of these, each responsible for a specific area of health care, such as pregnancy and childbirth, stroke, lung cancer and HIV/AIDS. Each group has an editorial base, where the core staff and a coordinating editor work, along with a more extensive editorial team of editors, advisers and specialists. A full list of the Cochrane Review Groups as of December 2005 is given in Table 3.1.

Table 3.1. Cochrane Review Groups

Acute Respiratory Infection
Airways
Anaesthesia
Back
Bone, Joint and Muscle Trauma
Breast Cancer
Colorectal Cancer
Consumers and Communication
Cystic Fibrosis and Genetic Disorders
Dementia and Cognitive Improvement
Depression, Anxiety and Neurosis
Developmental, Psychosocial & Learning Problems
Drugs and Alcohol
Ear, Nose and Throat Disorders
Effective Practice and Organization of Care
Epilepsy
Eyes and Vision
Fertility Regulation
Gynaecological Cancer
Haematological Malignancy
Heart
Hepato-Biliary
HIV/AIDS
Hypertension
Incontinence
Infectious Diseases
Inflammatory Bowel Disease
Injuries
Lung Cancer
Menstrual Disorders and Infertility
Metabolic and Endocrine Disorders
Methodology Review Group
Movement Disorders
Multiple Sclerosis
Musculoskeletal
Neonatal
Neuromuscular Disease
Oral Health
Pain, Palliative and Supportive Care
Peripheral Vascular Diseases
Pregnancy and Childbirth
Prostatic Diseases and Urological Cancers
Renal
Schizophrenia
Sexually Transmitted Diseases
Skin
Stroke
Tobacco Addiction
Upper Gastrointestinal and Pancreatic Diseases
Wounds

The reviews themselves are prepared and kept up to date by several thousand authors, often working in small teams of 2–6 people. These authors may be researchers who have done trials, but equally may be health care professionals, patients or carers. There is no requirement that the authors are ‘experts’ at the outset, but many find that the process of doing the Cochrane review makes them into an expert in the topic area for the review. In keeping with the principles of The Cochrane Collaboration, the authors of Cochrane reviews are helped to prepare and maintain these reviews and, unlike submission to other journals, once the topic area of a Cochrane review has been agreed with the editorial team of the Cochrane Review Group, there is a commitment to publish it, regardless of the findings, but providing that it is of high enough quality.

The increasing size of The Cochrane Collaboration over the last decade is reflected also in the growth in the total number of authors of Cochrane reviews. In the year 2000, there were nearly 3000, and by 2005 this had increased to more than 7500. Very few of these are paid to work on their reviews. The main motivation is a desire to answer reliably a question about the relative effects of interventions for people with particular conditions.

The work of Cochrane Review Groups and the authors of Cochrane reviews are supported by a variety of other Cochrane entities. Cochrane Methods Groups bring together people with expertise in particular areas of methodology such as information retrieval, statistical analysis, qualitative research and patient-reported outcomes. The groups provide advice to The Cochrane Collaboration and, to varying extents, conduct empirical research to help identify ways in which the quality of systematic reviews and other evaluations or health care can be improved further. Table 3.2 lists the 11 Cochrane Methods Groups.

Table 3.2. Cochrane Methods Groups

Applicability and recommendations
Economics
Individual patient data meta-analysis
Information retrieval
Non-randomised studies
Patient reported outcomes
Prospective meta-analysis
Qualitative
Reporting bias
Screening and diagnostic tests
Statistical methods

Table 3.3. Cochrane Fields and Networks

Cancer Network
Child health
Complementary medicine
Consumer Network
Health care of older people
Health equity
Health promotion and public health
Neurological Network
Occupational health
Prehospital and emergency health
Primary health care
Rehabilitation and related therapies
Vaccines

There are 13 Cochrane Fields or Networks (Table 3.3). These have broad areas of interest and expertise. They span the scope of all or many Cochrane Review Groups, and include a Consumer Network helping to promote the interests of users of health care. The Fields and Networks work with Cochrane Review Groups to identify people to help with the preparation and assessment of Cochrane reviews. They also strive to promote the accessibility of Cochrane reviews to people in the broad areas of health care they cover, such as primary care, cancer and child health.

The work of all Cochrane entities, and their members, is supported by 12 regional Cochrane Centres (Table 3.4). Cochrane Centres have regional, linguistic or geographic responsibilities to support the activities of members of The Cochrane Collaboration. Wherever people live in the world, there is one Cochrane Centre that acts as their reference centre within the Collaboration.

Table 3.4. Cochrane centres

Australasian
Brazilian
Canadian
Chinese
Dutch
German
Iberoamerican
Italian
Nordic
South African
United Kingdom
United States

Some of the centres have branches within the same country (e.g. the US Cochrane Centre is based in Baltimore, with branches in San Francisco and Boston) or in other countries for which they have responsibility (e.g. there is a South-East Asian Cochrane Network, New Zealand Branch and Singapore Branch of the Australasian Cochrane Centre).

Policy within The Cochrane Collaboration is set by the Steering Group, the members of which are elected by, and from within, these Cochrane entities. The Steering Group also acts as the Board of Directors of The Cochrane Collaboration, which is registered as a charity and a company in England.

Some Cochrane Centres have a special responsibility for the collaboration as a whole, above and beyond their regional roles. One such example is the Nordic Cochrane Centre at the Righshospitalet in Copenhagen, Denmark, which develops the collaboration's Information Management System (IMS). This is the software that is used to prepare and maintain Cochrane reviews and to submit them for publication. It includes Review Manager, or RevMan, which is the software tool used to write Cochrane reviews and, where appropriate, to perform and display meta-analyses of the results of the included studies.

Parts of the IMS were initially developed by Update Software Ltd, the original publishing partner of The Cochrane Collaboration, before responsibility was transferred to the Nordic Cochrane Centre. Much further development has been done by this Centre since then, in consultation with Cochrane entities and users of the software. For the collaboration's first decade, the IMS worked mainly as standalone software running on local computers, with reviewers sharing their files by disk or email attachment. As the Collaboration grew, and the number of reviews and the vital task of keeping these up to date got larger, a better way to share these documents and information was needed. This is now being developed as the new IMS. It is based on a central server, which will allow the global membership of the Collaboration and, often, of individual review teams to work together more closely and effectively.

ACCESSIBILITY OF COCHRANE REVIEWS

Cochrane reviews are published in *The Cochrane Database of Systematic Reviews (CDSR)*. As of the end of 2005, this contains the full text of more than 2500 complete Cochrane reviews containing details of the eligible studies, results and conclusions. Each of these reviews will be kept up to date as new evidence and information accumulates. The reviews were preceded by published protocols setting out how the review would be done and providing an explicit description of the methods to be followed. There are currently a further 1600 published protocols for new reviews in progress and hundreds of additional Cochrane reviews at earlier stages of preparation.

The growth in The Cochrane Collaboration is also apparent in the growth in the number of Cochrane reviews. When the first issue of *CDSR* was published, at the beginning of 1995, it included 36 full Cochrane reviews. This had risen to 500 in 1999, 1000 in 2001 and crossed the 2000 barrier in April 2004. Every year, between 300 and 400 protocols reach the stage of full review, a few hundred existing reviews are updated so substantively that they can be considered to be the equivalent of new reviews, and several hundred more are brought up to date in other ways. The Cochrane Collaboration's aim is for all reviews to the updated every two years, but as the number of reviews continues to grow this challenge, which is also a special feature of Cochrane reviews, becomes more difficult to achieve. Cochrane reviews are indexed in MEDLINE and, since 2005, have been included in Science Citation Index, with an official impact factor likely to be available in 2008.

The main way of accessing the *CDSR* is on the internet, where it is published as part of *The Cochrane Library*. It is also available on CD-ROM. *The Cochrane Library* is published by John Wiley and Sons, Ltd and is available on a subscription basis. However, several countries, beginning with Ireland and Northern Ireland in 2002, have national licences which make *The Cochrane Library* free at the point of use to everyone in these countries, including Australia, Denmark, England, Ireland, Northern Ireland, Norway, Spain and Wales.

OTHER OUTPUT FROM THE COCHRANE COLLABORATION

The output of The Cochrane Collaboration also includes a number of other unique resources: the *Cochrane Central Register of Controlled Trials (CENTRAL)*, the *Cochrane Database of Methodology Reviews* and the *Cochrane Methodology Register*.

The first of these, *CENTRAL*, was created in recognition of the difficulties of finding randomised trials. It serves to bring together, in a single place, records for more than 450 000 reports of studies that are, or might be, randomised trials in health care, stretching back over many decades. In contrast, when the Collaboration was established in 1993, fewer than 20 000 reports of randomised trials could be found easily in MEDLINE, even though that database alone contained several tens of thousands more such reports. The Cochrane Collaboration's efforts to identify and make accessible information on reports of trials that might be suitable for inclusion in Cochrane reviews have included extensive programmes of hand searching journals and conference abstracts (in which the journal or conference proceedings is checked from cover to cover to look for relevant reports) and electronic searching of bibliographic databases such as MEDLINE and EMBASE. Suitable records have been added to *CENTRAL*, with coordination by the US Cochrane Center.[8]

The *Cochrane Database of Methodology Reviews* contains the full text for Cochrane methodology reviews. These are systematic reviews of issues relevant to the conduct of reviews of health care interventions, or evaluations of health care more generally. They include, for example, reviews of evidence relating to publication bias, comparisons of the results from randomised trials and non-randomised trials, and peer review. There were 11 full Cochrane methodology reviews and published protocols for several more in December 2005.

The raw material for Cochrane methodology reviews comes to a large extent from the *Cochrane Methodology Register*. This can be thought of as filling the role that *CENTRAL* fills for health care interventions, for studies of the methods used to minimise bias and evaluate health care. It contains more than 7000 records relating to published reports of empirical research, registrations for ongoing research, reviews of such research and useful resources such as criteria to assess the quality of studies.

CONCLUSIONS

Systematic reviews of previous studies are vital to not only the design of new research but also to its interpretation. They bring together the findings from relevant research in as unbiased a way as possible and provide a key component in evidence-informed decision making. The Cochrane Collaboration, as the largest single organisation involved in the preparation and maintenance of systematic reviews of the effects of health care interventions, plays a unique role in the provision of this type of information.

Over the coming years, The Cochrane Collaboration needs to strive to ensure that its work is sustainable, by meeting challenges that have been apparent throughout its first decade, as well as new challenges. For example, even with 4000 Cochrane reviews underway, and results already available from 2500 of these, there is still a large amount of work to be done. In 2002, it was estimated that at least 10 000 systematic reviews would be needed to catch up and cover all health care interventions that had been investigated in randomised trials by that time.[9] The Cochrane Collaboration's recent work on Evidence Aid, highlighting reviews of relevance in natural disasters and other health care emergencies revealed that an up to date systematic review was not available for a similar proportion – three-quarters – of topics identified as priorities after the tsunami of 26 December 2004 the Indian Ocean.[10]

One of the biggest challenges facing The Cochrane Collaboration but also one of its most important features is maintaining, or updating, its reviews. When there are 10 000 Cochrane reviews, these will need to be assessed and, if necessary, updated at the rate of approximately 5000 per year. The Cochrane Collaboration's continued sustainability and growth will depend on cooperation with funders, providers and users of health care and of health care research. In this way, and through the increased accessibility of Cochrane reviews, the updated version of George Gould's vision from the late nineteenth century might be achieved early in the twenty-first.

REFERENCES

1. Gould GM. The work of an association of medical librarians. *Med Libr* (1898) **1**: 15–19.
2. Lord Rayleigh. Address by the Rt. Hon. Lord Rayleigh. In: Report of the fifty-fourth meeting of the British Association for the Advancement of Science, held at Montreal in August and September 1884. London: John Murray (1885) 3–23.
3. Bugher JC. International aspects of public health. *Bull Med Libr Assoc* (1958) **46**: 7–16.
4. Mulrow CD. Rationale for systematic reviews. *Br Med J* (1994) **309**: 597–9.
5. Clarke M. Doing new research? Don't forget the old. *PLoS Med* (2004) **1**: 100–2.
6. Chalmers I, Dickersin K, Chalmers TC. Getting to grips with Archie Cochrane's agenda. *Br Med J* (1992) **305**: 786–8.
7. Cochrane AL. 1931–1971: a critical review, with particular reference to the medical profession. In: *Medicines for the Year 2000*. London: Office of Health Economics (1979) 1–11.
8. Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S. Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials. *Eval Health Prof* (2002) **25**: 38–64.
9. Mallett S, Clarke M. How many Cochrane reviews are needed to cover existing evidence on the effects of healthcare interventions? *Evidence Based Med* (2003) **8**: 100–1.
10. Tharyan P, Clarke M, Green S. How The Cochrane Collaboration is responding to the Asian tsunami. *PLoS Med* (2005) **2**: e169.

CANCER

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

4

Breast Cancer

DONALD A. BERRY, TERRY L. SMITH AND AMAN U. BUZDAR
Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

INTRODUCTION

More than 200 000 women in the United States are diagnosed annually with breast cancer. About 40 000 women die from the disease each year. Among women, it is the most common malignancy, and is exceeded only by lung cancer as the leading cause of cancer death. Although the risk of breast cancer is substantially higher in older women, many cases occur in young women. Of cases diagnosed in the United States in 1998, 5% occurred in women under the age of 35, 30% in women aged 35 to 49, 31% in women aged 50 to 64, and 33% in women aged 65 or older.[1] For US women, the lifetime risk of developing breast cancer is about 13%. About 0.1% of US women carry an inherited mutation of a breast/ovarian cancer susceptibility gene, BRCA1 or BRCA2, and so have a lifetime risk in excess of 50%.

From 1940 to the early 1980s, breast cancer incidence in the United States increased by a fraction of a per cent per year when adjusted for age. Chiefly because of the widespread dissemination of screening mammography beginning in the early 1980s, invasive breast cancer incidence has increased by 3–4% per year into the 1990s. Over the same period, the rate of ductal carcinoma *in situ* (DCIS) increased by about sixfold, from about 5 cases per 100 000 women in 1980 to more than 30 per 100 000 in 1998.

Despite the increasing incidence of breast cancer among US women since the early 1980s, and perhaps indirectly because of this increase, the annual age-adjusted rate of breast cancer mortality has decreased by almost 2% per year in the 1990s. Researchers generally attribute this improvement in breast cancer survival to increased use of screening mammography and to improvements in the treatment of breast cancer.[2] Efforts are underway to better delineate the relative impacts of factors influencing breast cancer survival.

An important development for breast cancer research in the 1980s and 1990s was not directly related to science. These years saw the formation of strong advocacy groups that worked to promote research in breast cancer. Federal funding has increased more than sixfold since 1990, and grass-roots action has resulted in an unprecedented programme of breast cancer research funding administered by the Department of Defense. In addition, patient advocates have become highly educated about research

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

issues and many serve regularly alongside professional scientists on various governmental boards guiding the direction of research expenditures and treatment recommendations. Patient advocates also serve on cooperative group committees that plan clinical trials in breast cancer, institutional review boards, and data safety monitoring boards.

Advocacy groups have worked to increase the number of women who participate in clinical trials. The Clinical Trial Initiative of the National Breast Cancer Coalition Fund (NBCCF) maintains a registry of clinical trials and urges women with breast cancer to participate (see NBCCF[3]). Before a clinical trial can be included in their registry, experts from the NBCCF ascertain that it addresses an important, novel research question related to breast cancer, and that its design is scientifically rigorous and employs appropriate and meaningful outcomes.

STAGING

Breast cancer is staged using a system developed by the American Joint Committee on Cancer, and based on the size and other characteristics of primary tumour (T), the status of ipsilateral lymph nodes (N), and the presence or absence of distant metastases (M) (AJCC[4] and Singletary *et al.*[5]). The stage of disease, ranging from 0 to IV, is based on combinations of these TNM rankings.

Stage 0 consists of ductal and lobular carcinoma *in situ* (DCIS, LCIS), non-invasive and possibly non-malignant forms of the disease. Stages I to III are invasive stages in which the tumour is confined to the breast or its immediate vicinity. Higher stage indicates larger primary tumours or greater locoregional tumour involvement. Patients having evidence of distant metastasis are classified as Stage IV.

Distribution of disease stage at the time of breast cancer diagnosis varies by country, depending on the health care system's approach to diagnosis and reporting. In the United States, the approximate proportions of women diagnosed with Stage 0 through Stage IV disease are 21%, 42%, 29%, 5% and 4%, respectively.[1] An additional tumour classification method based on histopathologic examination has limited discrimination ability because 70–80% of tumours are of a single type: infiltrating ductal carcinoma.

PROGNOSIS

Breast cancer is heterogeneous. Many breast cancers are slowly growing and their carriers survive for many years and die of other causes. Other tumours are very aggressive and may have spread to distant sites by the time the primary tumour is diagnosed. This heterogeneity has implications for research in all phases of the disease, beginning with screening and diagnostic methods through the evaluation of treatments for advanced disease.

Stage is the most widely recognised determinant of patient outcome. Stage IV disease is generally regarded to be incurable, with median survival in the range of 18 to 24 months, although a small fraction of patients with Stage IV disease achieve complete remission following systemic chemotherapy, and survive for many years.[6] On the other hand, patients with Stage I disease, consisting of a small primary tumour and no involved lymph nodes, have at least a 90% probability of being disease-free after five years. Lymph node involvement is associated with a worse prognosis, with five-year disease-free rates ranging from 50 to 75%. Tumour grade, proliferative activity and menopausal status play relatively minor roles.

Although stage is an important prognostic factor, it is of limited use as a determinant of treatment outcome. The relative benefits of treatment are reasonably consistent across stages – although the absolute benefit can be much greater for higher stage disease. Much current research focuses on factors that may predict clinical benefit from certain treatment approaches ('predictive factors') in contrast to the more conventional 'prognostic factors' which are regarded as indicators of general tumour aggressiveness, irrespective of type of therapy.

The best-studied predictive factor is oestrogen-receptor (ER) status, which is an important indicator of whether a tumour will respond to hormonal treatment. Tamoxifen and other selective oestrogen-receptor modulators (SERMs) are highly effective in patients with hormone-sensitive breast cancer, but they have no benefit in patients whose tumours are ER negative and progesterone-receptor negative (Early Breast Cancer Trialists' Collaborative Group, EBCTCG[7]). Patients who benefit from SERMs may also benefit from aromatase inhibitors.[8]

HER-2 (also referred to as HER-2/neu, ErbB2, c-erbB-2) is a member of the epidermal growth factor receptor family that is overexpressed in 20% to 40% of breast tumours, and has been cited in numerous reports as conveying poor prognosis.[9] Studies in early breast cancer have suggested that patients with HER-2 positive tumours are more likely to benefit from anthracycline therapy.[10–12] Trastuzumab, a monoclonal antibody against HER-2, is effective in delaying progression in Stage IV disease that overexpresses HER-2,[13] and is being evaluated for its efficacy in treating HER-2-overexpressing primary tumours.

HISTORICAL PERSPECTIVE ON CLINICAL TRIALS IN BREAST CANCER

SURGERY AND RADIOTHERAPY

Scientific understanding of the biology of breast cancer has changed radically in the past 50 years. Results of large randomised trials have played a major role in this transition. From the nineteenth century and up into the 1970s, breast cancer was understood to be a local/regional disease that spread by direct extension along lymphatic pathways to distant sites. This concept gave rise to the surgical methods promoted by W.S. Halsted[14–16] around the turn of the twentieth century, i.e. extensive resection of the breast, regional lymphatics, lymph nodes and muscle. This surgical technique, known as radical mastectomy, remained the principal approach to treatment of breast cancer throughout the first half of the century, sometimes combined with radiotherapy.

When the concept of large-scale randomised clinical trials to investigate alternative therapies was proposed in the 1960s, controversy arose among breast cancer researchers as well as in other medical fields. In a heated exchange, a prominent breast cancer surgeon denounced such studies as 'a great leap backward in the treatment of breast cancer'.[17] Despite such opposition, pioneers in the field persisted in designing trials to address important therapeutic questions of the time, and, moreover, were able to persuade patients to participate in this novel idea of assigning treatment by randomisation. These early trials compared various surgical and radiotherapy approaches. In a trial of almost 1700 women implemented in 1971, there were no significant survival differences between conventional radical mastectomy, total mastectomy with radiation, and total mastectomy with removal of axillary nodes.[18,19] Results of this and other trials of the era challenged long-held views of the disease and gradually convinced researchers that their concept of breast cancer as a local disease which could best be treated by radical local treatment techniques was incorrect. Rather, breast cancer came to be understood as a systemic disease that could benefit from systemic therapy, and radical local therapies were no longer regarded as essential for prolonging survival.

CHEMOTHERAPY

Cytotoxic agents for treatment of solid tumours were first developed in the 1950s. Breast cancer proved to be highly sensitive to several of these, when used as single agents in small trials. Subsequently, combinations of these cytotoxic agents were evaluated, one of the earliest being the Cooper regimen (cyclophosphamide, methotrexate, 5-fluorouracil, vincristine and prednisone).[20] With the understanding of breast cancer as a systemic disease and the proven sensitivity of breast cancer cells to cytotoxic agents, the stage was set for the rapid development of adjuvant chemotherapy once this concept was introduced in the

1970s. A randomised trial comparing surgery followed by combination chemotherapy to surgery alone demonstrated that disease recurrence could be significantly reduced using this adjuvant therapy approach.[21]

The introduction of doxorubicin for treatment of breast cancer is illustrative of the series of clinical trials typically undertaken for the development of new agents. Small trials conducted in solid tumours in the early 1970s established safety and dosing, and these were quickly followed by Phase II trials of the agent in metastatic breast cancer. Subsequently, doxorubicin was evaluated in combination with other agents, and randomised trials established that higher response rates could be achieved in metastatic disease with combinations that included doxorubicin. These successes prompted the introduction of various doxorubicin and other anthracycline-containing combinations as adjuvant therapy for primary breast cancer. Known by such acronyms as 'FAC' = 'CAF', 'FEC', 'AC', these combinations continue to play a prominent role in the treatment of breast cancer.[22,23] Anthracycline-containing therapies further reduce the risk of recurrence and favourably impact survival in early breast cancer.[24]

HORMONAL THERAPY

Hormonal therapy is a key component of therapy when tumours are hormone-receptor positive. Early trials focused on ovarian ablation by surgery or chemical means. The anti-oestrogen agent tamoxifen was introduced in the 1970s, at a time when there was high regard for the potential of cytotoxic agents, but little interest in hormonal therapies. Early small trials in metastatic breast cancer were equivocal and could have led to abandoning the agent. However, the weight of evidence from laboratory studies and several small trials pointed to superior efficacy with prolonged administration in ER positive disease. After a series of large randomised trials, tamoxifen is now regarded as standard therapy for pre- and post-menopausal women with ER positive tumours.[25] Tamoxifen may be the single most important advance in treating breast cancer. Questions remain about the optimum treatment duration even though a trial conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP) comparing 5 and 10 years of tamoxifen therapy concluded there was little or no advantage to longer therapy.[26]

HIGH-DOSE CHEMOTHERAPY WITH BONE MARROW TRANSPLANT OR STEM CELL SUPPORT

An unresolved question in therapy of breast cancer that has presented an unusual challenge for the conduct of clinical trials is that of high-dose chemotherapy supported by autologous bone marrow transplant or peripheral blood progenitor cells. Ten trials addressing the question of high-dose versus standard-dose chemotherapy have been reported. Two of these were subsequently discredited following an international investigation. Only two of the remaining eight trials entered more than 200 patients. Financial issues, patient and physician acceptance and competing treatment strategies have compromised accrual, and it is unclear if ongoing trials can be completed. The available evidence suggests that high-dose therapy provides little or no benefit for patients regardless of their disease stage.[27,28]

MAMMOGRAPHY

Eight large randomised trials conducted since 1963 assessed the value of screening mammography for reducing breast cancer mortality. These are of particular interest for the scrutiny they have undergone in recent years. The preponderance of evidence from the randomised trials indicates a benefit associated with screening mammography.[29–31] However, a meta-analysis concluded that six of the eight trials were seriously flawed and the remaining two trials showed insignificant breast cancer mortality differences between the screened and non-screened groups.[32] The National Cancer Institute recommends screening mammography every 1 to 2 years for women aged 40 and older, while recognising that there are risks associated with false-positive results.

PREVENTION

Beginning in the 1990s, coinciding with the detection of methods for identifying women at high risk of breast cancer, the first large-scale trials were mounted to determine if the incidence of breast cancer could be reduced in targeted high-risk groups. These trials established that breast cancer incidence could be greatly reduced by daily doses of tamoxifen.[33,34] This reduction was due entirely to a lower incidence of ER positive tumours with no change in the incidence of ER negative tumours. This suggests that prophylactic tamoxifen will not have as great an impact on survival as it does on incidence, although none of the prevention trials address survival as an endpoint.

The STAR trial (Study of Tamoxifen and Raloxifene), which compared the effects of tamoxifen and raloxifene in the treatment of 19 747 postmenopausal women at increased risk of breast cancer, showed similar incidence of breast cancer in the two treatment groups. However, women treated with raloxifene had lower incidence of uterine hyperplasia.[35]

MAJOR TRIAL GROUPS

One of the largest cooperative groups conducting trials in breast cancer in the United States is the NSABP. Trials from this group are often referred to by their 'B' numbers, e.g. B-06, which established the equivalence of lumpectomy to total mastectomy.[36] Other major cooperative groups conducting clinical trials in breast cancer are the Eastern Cooperative Oncology Group (ECOG), Cancer and Leukemia Group B (CALGB), Southwest Oncology Group (SWOG), Breast Cancer International Research Group (BCIRG), European Organisation for Research and Treatment of Cancer (EORTC), North Central Cancer Treatment Group (NCCTG), and the National Cancer Institute of Canada (NCIC).

An important information resource regarding the benefits of treatment for early breast cancer is the Early Breast Cancer Trialists' Collaborative Group (EBCTCG). This group, based at Oxford University, serves as a centre for data synthesis rather than actual conduct of clinical trials. Beginning in 1983, this group has collected data from virtually all major randomised trials conducted in early breast cancer, published or not. Data from more that 200 000 women have been analysed, using statistical techniques for meta-analysis, with results published at the end of each five-year analysis cycle, beginning in 1985. These publications have addressed the role of radiation, ovarian ablation, polychemotherapy, tamoxifen and quality of life; these 'overview' articles are frequently cited in support of treatment approaches.[7,24,25,37–42] The weaknesses of meta-analysis have been widely discussed in the statistical literature, chief among these being the issue of heterogeneity among the trials being combined. For example, the overviewers combine various therapeutic regimens under the single rubric 'polychemotherapy'. However, these overview reports have allowed researchers to reliably assess moderate-size treatment effects which could not have been detected in individual trials. Treatments causing even moderate reductions in mortality, if implemented widely among women with breast cancer, could prevent or delay thousands of deaths due to the disease. The meta-analysis has also addressed questions of treatment efficacy within subsets, e.g. the confirmation of benefit of adjuvant tamoxifen in ER positive pre-menopausal women[7] as well as in post-menopausal women.

TIME-DEPENDENT HAZARDS

In this section we address a methodological issue that arises quite generally in survival analysis. Consider disease-free survival. This is the usual primary outcome measure in evaluating adjuvant therapies, with results presented in the form of Kaplan–Meier survival curves and compared using statistical tests that take into account the entire survival distributions. The simple hazard function, which is in effect the derivative of the survival curve, can serve as an effective graphical aid to understanding treatment and covariate effects. Moreover, it can reveal important effects that are not apparent in the survival curves, themselves.

We will use trial CALGB 8541 as an example.[43] This trial considered three different dose schedules of CAF (cyclophosphamide, doxorubicin and 5-fluorouracil) in node positive breast cancer. The schedules consisted of four cycles of CAF at 600, 60, 600 mg/m^2 (high dose), six cycles at 400, 40, 400 mg/m^2 (moderate dose) or four cycles at 300, 30, 300 mg/m^2 (low dose). The primary endpoint was disease-free survival, which is shown in Figure 4.1 for the three dose groups using Kaplan–Meier plots. Details of the comparison are provided in the original report, and will not be repeated here.

Time-to-event curves such as those in Figure 4.1 do not tell the whole story regarding any benefit of increasing dose and dose intensity. A clearer picture is contained in plots of hazard over time. The hazard in any particular time period is the proportion of events occurring during that time period in comparison with the number of patients who are at risk at the beginning of the period. For example, if there are 100 patients in a group and 10 of these recur in the first year, then the first-year hazard is 10%. Going into the second year, only 90 patients are at risk. If another 10 recur in the second year, then the second-year hazard is $10/90 = 11\%$. When calculating hazards from survival plots such as those in Figure 4.1 (which incorporate censored observations), we subtract the current year's survival proportion from the previous year's survival proportion and divide by the previous year's survival proportion. The resulting yearly values are shown in Figure 4.2.

Some authors like to smooth hazard estimates over time. We prefer to show the raw estimates. The reason is that each time period provides a 'nearly independent' trial of therapeutic comparisons. Depending on what assumptions are made about the underlying survival distribution, these trials may not be truly independent, but events that have occurred previously are set aside and a 'new trial' is begun. Each time period has the potential for confirming observations made in other time periods.

A striking observation from Figure 4.2 is that all three hazards decrease over time (after year 2). This is a reflection of the heterogeneity of breast cancer. The most aggressive tumours recur early, yielding the high hazards evident in the first few years. Once their tumours have recurred, patients are removed from the at-risk population. The remaining tumours tend to be less aggressive and so they recur at a lower rate.

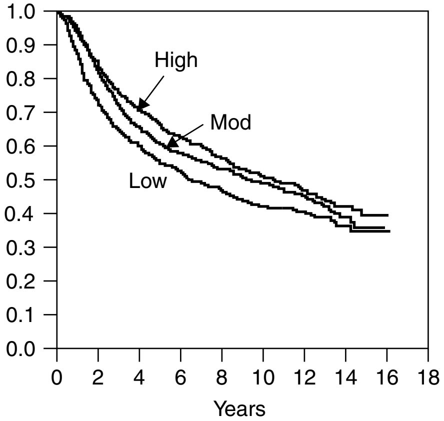


Figure 4.1. Disease-free survival proportion for the three CAF dose groups of CALGB 8541.

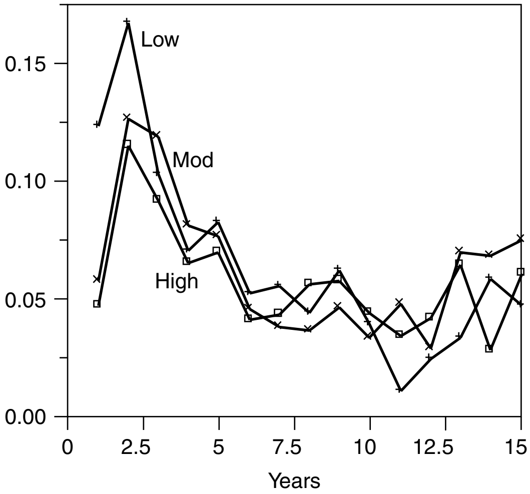


Figure 4.2. Hazards for the three CAF dose groups of CALGB 8541, derived from Figure 4.1.

As regards treatment-arm effect, the apparent benefit of a regimen of high-dose chemotherapy is restricted to the first five years or so. Actually, the hazard for a high dose is lower than those of the other two arms in each of the first six years (although it is not much lower in years 4, 5 and 6, and it is not much lower than that for a moderate-dose regimen in any of the six years). In view of the 'near independence' of the six time periods, this observation is impressive. Another important observation from Figure 4.2 is that after five years the risks of all three groups come together, with the annual risk of recurrence being approximately 5% in all three groups.

The reduction in hazard of recurrence for high versus low doses is 14% over the 18 years of follow-up (95% confidence interval: 6–22%). This is an average over these years (weighted over time because of differences in at-risk sample sizes over time), but since there is no reduction at all in the later years, the overall reduction is being carried by the early years. Restricting to the first three years, the reduction is 24% (13–33%). A benefit of chemotherapy that is restricted to the first few years is typical in breast cancer trials. An implication is that a hazard reduction seen early in a trial, say one with a median of three years of follow-up, will deteriorate over time. This is because the comparison will eventually involve averaging over periods where there is no longer a treatment benefit.

In the later years, the hazards of about 5% are very similar to the annual hazard for node negative breast cancer patients. Interestingly, convergence to about 5% applies irrespective of the number of positive lymph nodes. Figure 4.3 shows this effect. It gives hazard plots for three categories of positive nodes: 1–3, 4–9 and 10 or more (for the three dose groups combined). Early in the trial, patients with 10+ positive nodes have a very high annual recurrence rate of 20–30%. However, after five years or so, the annual hazard is about 5% in all three groups. A patient with a large number of positive nodes who has not experienced recurrence in the first five years or so has the same updated prognosis as a patient with a small number of positive nodes, including no positive nodes. The effects of both the number of positive nodes and dose of CAF have elapsed after five years.

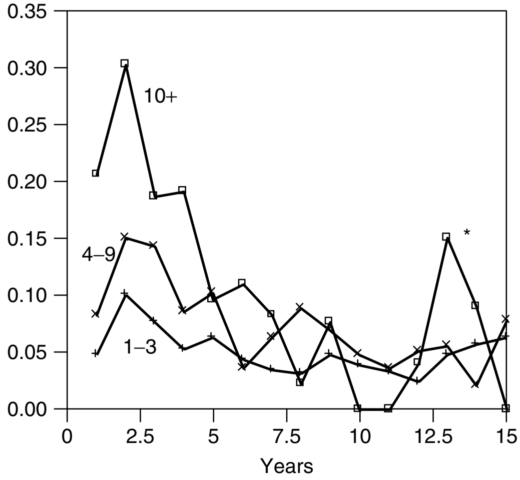


Figure 4.3. Hazards for the three categories of positive lymph nodes (1–3, 4–9 and 10 or more) for CALGB 8541. There are few patients at risk in the later years, especially in the 10+ group, and for two reasons. One is that this was the smallest group to start with (174 of the 1550 patients in the trial), and the other is that most recurred early. For example, the asterisk at 13 years indicates a time point at which there were only 24 patients at risk, and three of these recurred in the 13th year.

An important aspect of CALGB 8541 is the role of tumour HER-2/neu expression and in particular its interaction with dose of CAF.[10] HER-2/neu assessment was carried out for a subset of 992 patients from the original study. Its interaction with dose was shown to be significant in a multivariate proportional hazards model. But the manner of interaction is easiest to understand using hazards. Figure 4.4 shows the effect of dose of CAF separately for patients with HER-2/neu negative tumours ($n = 720$) and HER-2/neu positive tumours ($n = 272$). HER-2/neu negatives show no dose effect. The entire benefit of high dose over moderate dose and high dose over low dose that is observed in these patients is concentrated in patients whose tumours are HER-2/neu positive. Moreover, this benefit occurs through a reduction in hazard in

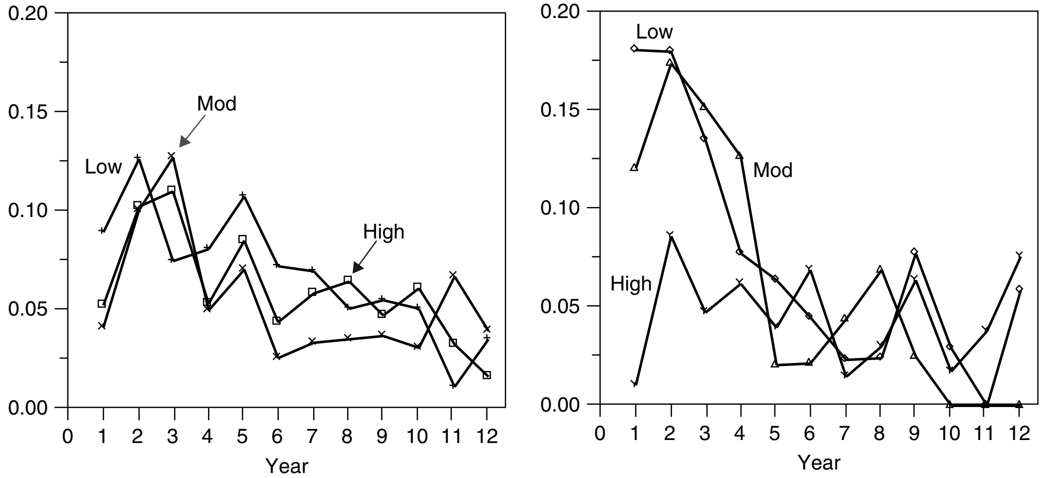


Figure 4.4. Annual disease-free survival hazards for a subset of patients ($n = 992$) in CALGB for whom expression of HER-2/neu in the patient's tumour was assessed. Patients in the left-hand panel had tumours that were HER-2/neu negative and the tumours of those in the right-hand panel were HER-2/neu positive.

each of the first three to four years. Again, each year is a separate study and so each of these years provides a separate confirmation of the overall conclusion. The hazard reduction in the first three years for high dose as compared with the other two groups combined was 65% among patients whose tumours were HER-2/neu positive. HER-2/neu overexpression apparently conveys a poor prognosis for lower doses but not for a high dose – it might even provide a favourable prognosis for a high dose.

Many of the above conclusions would have been difficult or impossible without considering hazards over time. A final comment regarding hazards relates to the common problem of predicting survival results into the future for patients already accrued to a trial. Consider Figure 4.1. Some patients have as little as 10 years of follow-up information. As more follow-up information becomes available, there will be no change in these curves prior to the 10-year time point, but they may change subsequent to 10 years. Because the focus is on patients who have not yet recurred, the way the curves will change depends on the hazards beyond 10 years. The information available about these hazards is shown in Figure 4.2. For predicting when and whether a patient recurs, hazards should be considered one year at a time, and based on the current year of follow-up.

ASSESSING LONG-TERM IMPACTS OF THERAPY

Showing that a cancer therapy is beneficial using logrank tests or proportional hazards regression models, or whatever other analysis one uses, does not allow for concluding the nature of the benefit. It may be that some patients are cured of their disease; or the therapy may delay the disease's progress in some patients; or the effect may be a mixture of the two. Deciding among these possibilities may be possible when all or almost all events occur in a modest amount of follow-up time. In primary breast cancer, a goodly proportion of patients never recur. Therefore, such a decision is difficult or impossible to make.

Figure 4.5 illustrates the difficulty in discriminating between cure and prolonging survival in breast cancer trials. Consider a clinical trial that is designed to evaluate a new therapy, one that may improve survival. Suppose further that in the population of interest, the current annual rate of breast

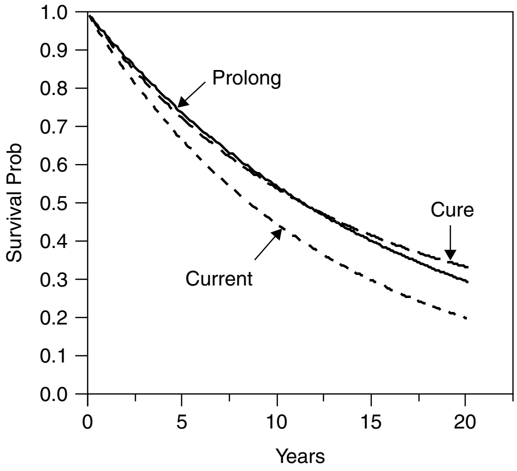


Figure 4.5. Hypothetical survival curves comparing cure and prolonging survival.

cancer mortality is 8%. The corresponding survival distribution is shown by the dashed survival curve labelled 'current' in Figure 4.5. It assumes exponential survival (although the same effect holds for any parametric form) and so the current median survival is 8.66 years. The goal of the new therapy is to improve this by 1/3 to 11.55 years. If this happens then it may be through prolonging every patient's survival by reducing the annual mortality rate to 6% – the curve labelled 'prolong' – or by leaving the annual rate unchanged (at 8%) for most of the patients but curing a fraction of them – the curve labelled 'cure'. To have the same median (11.55 years) as 'prolong' implies a cure rate of 17.1%.

In view of the sampling variability present in empirical survival information, it is impossible to discriminate between the 'prolong' and 'cure' curves shown in Figure 4.5 on the basis of results of even impossibly large clinical trials. Indeed, the critical part of the follow-up period for this discrimination is 20 years and beyond, and few trials have followed patients for this long. Moreover, information beyond 20 years is relatively sparse because earlier events and competing risks (such as cardiovascular disease) will have removed patients from the at-risk population. To make inferential matters worse, there is an enormous array of possible curves that are similar to the two shown in the figure, with some having cure rates and others not. Finally, the 'current' survival distribution assumes that all breast cancer is fatal (although survival times vary). More realistically, some breast cancer (including some invasive as well as *in situ* breast cancer) will never kill the patient. Deciding whether a new therapy cures some patients is even more difficult if a proportion of patients is assumed to have non-fatal disease.

This inability to distinguish between curing patients and prolonging survival has further implications in the evaluation of screening and diagnostic methods. It is possible that breast cancers become lethal or not in their very early development, as suggested by studies of tumour markers purported to identify especially aggressive tumours. If this is so, then early detection may not help, and the observed benefits of therapy, however substantial, may be the result of slowing the progress of the disease rather than curing it. Such slowing may be beneficial whether it comes early or late in the disease.

ADAPTIVE DESIGNS OF CLINICAL TRIALS

Adaptive designs have the dual goals of efficient learning from all relevant results and effective treatment of patients. They are more flexible than conventional designs, and have application in all phases of drug development. Such designs can be implemented using Bayesian methodology as a means to incorporate new information into the trial design.

Designs of clinical trials for breast cancer are usually static in the sense that the sample size and any prescription for assigning treatment, including the randomisation of patients, are fixed in advance. While designs may include stopping rules, such as the two-stage Phase II trial design of Simon,[44] or the interim comparisons in Phase III designs,[45,46] the criteria for early stopping are very conservative and therefore few trials actually stop early. The simplicity of trials with static design makes them solid inferential tools. The sample sizes tend to be large, with a straightforward treatment comparison as the objective. Despite their virtues, static trials result in slow and unnecessarily costly development of new therapeutic agents.

The tradition of drug development is to evaluate a single drug at a time. Given the fast pace of current new drug discovery (there are hundreds of known experimental drugs with potential benefits in breast cancer), these inefficient evaluation methods are no longer adequate. In addition to the traditional focus on false-positive and false-negative errors in standard drug testing, another kind of error applies to drugs not under investigation. Every such drug is a false neutral. Given the limited resources available to the medical establishment to develop new therapies, resource allocation must be approached in a more rational way. This is as true in breast cancer, for which a relatively large number of women are willing to participate in clinical trials, as it is for other forms of cancer. Pharmaceutical companies and medical researchers generally must be able to consider hundreds of drugs for development at the same time. Static trials inhibit the simultaneous processing of many drugs. They cannot efficiently address dose–response questions or prioritisation of similar agents when many drugs are under consideration. Dynamic designs that are integrated with the drug development process are necessary for reasonable progress in medical research.

Using an adaptive design means examining the accumulating data periodically – or even continually – with the goal of modifying the trial's design. These modifications depend on what the data show about the unknown hypotheses. Among the modifications possible are stopping early, restricting eligibility criteria, expanding accrual to additional sites, extending accrual beyond the trial's original sample size if its conclusion is still not clear, dropping arms or doses, and adding arms or doses. All of these possibilities are considered in light of the accumulating information.

Adaptive designs also include unbalanced randomisation, in which the degree of imbalance depends on the accumulating data. For example, arms that give more information about the hypothesis in question or that are performing better than other arms can be weighted more heavily.[47] Current (Bayesian) probabilities that each of several doses or agents surpass standard or placebo therapy are calculated. These calculations use all information from patients treated to date. A new patient is then assigned to treatment randomly, with weights proportional to these probabilities. The assignments involve some degree of randomisation, but all patients are more likely to receive treatments that are performing better. Those that are doing sufficiently poorly become inadmissible in the sense that their assignment weight becomes 0. When and if we learn that a new agent is effective (or ineffective), we stop the trial. Patients in the trial benefit from data collected *in the trial*. The explicit goal is to treat patients more effectively, but in addition we learn about the new agents more efficiently. Initially we evaluate each design's frequentist operating characteristics using Monte Carlo simulation, possibly modifying the parameters of the assignment algorithm to achieve the desired characteristics.

Adaptive designs are being used increasingly in cancer trials. This is true for trials sponsored by pharmaceutical companies, and more generally. A variety of trials at The University of Texas M.D. Anderson Cancer Center (MDACC) are prospectively adaptive. For example, we are building the foundation for a Phase II trial for evaluating drugs for breast cancer that is more a process than a trial. The idea is an extension of more general adaptive assignment strategies. We start with a number of treatment arms plus a control – possibly a standard therapy. We randomise to the arms and learn about their relative efficacy as the trial proceeds. Arms that perform better get used more often. An arm that performs sufficiently poorly gets dropped. An arm that does well enough graduates to Phase III, and if it does sufficiently well it might even replace the control. As more treatments become available, they are added to the mix and the process can continue indefinitely.

A trial of a new agent for treatment of metastatic breast cancer is being compared to the current standard therapy in a dynamic manner that allows the incorporation of newly available treatments in the randomisation process, as well as the elimination of treatments when a lack of improved efficacy can be established. Patients are randomised to treatments with weights proportional to the probability that a treatment is better than the standard therapy. The result is that superior therapies move through quickly and poorer therapies get dropped. Patients in the trial are provided with better treatment (when the arms are not equally good). Patients outside the trial get access to better treatments more rapidly.

Dose-finding trials of new agents are also conducted adaptively at MDACC, with dose assignment based on Bayesian updating of a model which relates dose and toxicity, using results from preceding patients. The model is the continual reassessment method or CRM.[48,49] Each patient is assigned to the dose having a probability of toxicity closest to some predetermined target value. This is the Bayesian posterior probability calculated from the data available up to that point (and so it is based on sufficient statistics).

The CRM more effectively finds the maximum tolerated dose (MTD) than does the conventional $3 + 3$ design.[50] A way in which both the $3 + 3$ and CRM designs are crude is the need to pause accrual while waiting for toxicity information.[51] Such pauses are inefficient and they cause logistical problems. Trials should be paused or stopped if there are safety concerns, not because the design cannot get out of its own way. In getting information about toxicity (or efficacy), there is seldom a magical dose that the next patient must get. All doses are potentially informative. Rather than stopping, one should use a design that models dose–response (toxicity and efficacy) and is able to assign a next dose even though patients previously treated are not yet fully evaluable. Other improvements to dose-finding methods are underway. These include the simultaneous incorporation of efficacy results into the design, and the use of toxicity severity rather than the usual assumption that toxicity is dichotomous.

CONCLUSION

Breast cancer clinical trials are not fundamentally different from those of other cancers. However, breast cancer stands out for several reasons. First, it is common, and it is becoming even more common with the improvements in and greater use of detection methods. That implies a greater ability to investigate the potential for therapeutic agents and combinations. As a consequence, there have been hundreds of randomised clinical trials conducted in breast cancer, more by far than in any other cancer. Second, it is a disease that is fatal in only a minority of cases. Third, patient advocates in the breast cancer community have been very influential, as both a research force and a political force, in lobbying for research funding. Fourth, breast cancer has been shown to be sensitive to a number of chemotherapies and hormonal therapies. The advances that have been made in breast cancer therapy are more impressive than for any other type of cancer, except for testicular cancer and some forms of leukaemia that commonly affect children. These advances have been built on a foundation of clinical trials.

REFERENCES

1. Surveillance, Epidemiology, and End Results Program (SEER) Public-Use Data (1973–1998), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, Bethesda, MD, August 2000 submission; released April 2001 online database, available at http://seer.cancer.gov.
2. Berry DA, Cronin KA, Plevritis SK, Fryback DG, ClarkeL, Zelen M, Mandelblatt JS, Yakovlev AY, Habbema JD, Feuer EJ; Cancer Intervention and Surveillance Modeling Network (CISNET) Collaborators. Effect of screening and adjuvant therapy on mortality from breast cancer. *N Engl J Med* (2005) **353**: 1784–92.
3. National Breast Cancer Coalition Fund (NBCCF). A patient guide to quality breast cancer care online information site. Available at http://www.natlbcc.org/nbccf.
4. American Joint Committee on Cancer (AJCC) online information site. Available at http://www.cancerstaging.org.
5. Singletary SE, Allred C, Ashley P, Bassett LW, Berry D, Bland KI, Borgen PI, Clark G, Edge SB, Hayes DF, Hughes LL, Hutter RBP, Morrow M, Page DL, Recht A, Theriault RL, Thor A, Weaver DL, Wieand HS, Greene FL. Revision of the American Joint Committee on Cancer Staging

System for Breast Cancer. *J Clin Oncol* (2002) **20**: 3628–36.

6. Greenberg PA, Hortobágyi GN, Smith TL, Ziegler LD, Frye DK, Buzdar AU. Long-term follow-up of patients with complete remission following combination chemotherapy for metastatic breast cancer. *J Clin Oncol* (1996) **14**: 2197–205.
7. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* (1998) **351**: 1451–67.
8. Buzdar A, Howell A. Advances in aromatase inhibition: clinical efficacy and tolerability in the treatment of breast cancer. *Clin Cancer Res* (2001) **7**: 2620–35.
9. Yamauchi H, Stearns V, Hayes DF. When is a tumor marker ready for prime time? A case study of c-erbB-2 as a predictive factor in breast cancer. *J Clin Oncol* (2001) **19**: 2334–56.
10. Thor AD, Berry DA, Budman DR, Muss HB, Kute T, Henderson IC, Barcos M, Cirrincione C, Edgerton S, Allred C, Norton L, Liu ET. erbB-2, p53, and efficacy of adjuvant therapy in lymph node-positive breast cancer. *J Natl Cancer Inst* (1998) **90**: 1346–60.
11. Paik S, Bryant J, Tan-Chiu E, Yothers G, Park C, Wickerham DL, Wolmark N. HER2 and choice of adjuvant chemotherapy for invasive breast cancer: National Surgical Adjuvant Breast and Bowel Project Protocol B-15. *J Natl Cancer Inst* (2000) **92**: 1991–8.
12. Ravdin PM, Green S, Albain KS, Boucher V, Ingle J, Pritchard K, Shepard L, Davidson N, Hayes DF, Clark GM, Martino S, Osborne CK, Allred DC. Initial report of the SWOG biological correlative study of c-erbB-2 expression as a predictor of outcome in a trial comparing adjuvant CAF T with tamoxifen (T) alone. *Proc Am Soc Clin Oncol* (1998) **17**: 97a.
13. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J, Norton L. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New Engl J Med* (2001) **344**: 783–92.
14. Halsted WS. The results of operations for the cure of cancer of the breast performed at the Johns Hopkins Hospital from June 1889 to January 1894. *Ann Surg* (1894) **20**: 497–555.
15. Halsted WS. The results of radical operations for the cure of carcinoma of the breast. *Ann Surg* (1907) **SLVI**: 1–19.
16. Buchanan EB. A century of breast cancer surgery. *Cancer Invest* (1996) **14**: 371–7.
17. Haagensen CD. A great leap backward in the treatment of carcinoma of the breast. *JAMA* (1973) **224**: 1181–3.
18. Fisher B, Montague E, Redmond C, Barton B, Borland D, Fisher ER, Deutsch M, Schwarz G, Margolese R, Donegan W, Volk H, Konvolinka C, Gardner B, Cohn I, Lesnick G, Cruz AB, Lawrence W, Nealon T, Butcher H, Lawton R. Comparison of radical mastectomy with alternative treatments for primary breast cancer: a first report of results from a prospective randomized clinical trial. *Cancer* (1977) **39**: 2827–39.
19. Fisher B, Jeong J-H, Anderson S, Bryant J, Fisher ER, Wolmark N. Twenty-five-year follow-up of a randomized trial comparing radical mastectomy, total mastectomy, and total mastectomy followed by irradiation. *New Engl J Med* (2002) **347**: 567–75.
20. Cooper RG. Combination chemotherapy in hormone-resistant breast cancer. *Proc Am Assoc Cancer Res* (1969) **10**: abstr 57.
21. Bonadonna G, Brusamolino E, Valagussa P, Rossi A, Brugnatelli L, Brambilla C, DeLena M, Tancini G, Bajetta E, Musumeci R, Veronesi U. Combination chemotherapy as an adjuvant treatment in operable breast cancer. *New Engl J Med* (1976) **294**: 405–10.
22. Hortobágyi GN. Anthracyclines in the treatment of cancer. *Drugs* (1997) **54**(Suppl): 1–7.
23. Weiss RB. The anthracyclines: will we ever find a better doxorubicin? *Semin Oncol* (1992) **19**: 670–86.
24. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Polychemotherapy for breast cancer: an overview of the randomised trials. *Lancet* (1998) **352**: 930–42.
25. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Tamoxifen for early breast cancer (Cochrane Review) online database, available at http://cochrane.de/cochrane. *Cochrane Database Syst Rev* (2001) **1**: CD000486.
26. Fisher B, Dignam J, Bryant J, Wolmark N. Five versus more than five years of tamoxifen for lymph node-negative breast cancer: updated findings from the National Surgical Adjuvant Breast and Bowel Project B-14 randomized trial. *J Natl Cancer Inst* (2001) **93**: 684–90.
27. Farquhar C, Basser R, Hetrick S, Lethaby A, Marjoribanks J. High dose chemotherapy and autologous bone marrow or stem cell transplantation versus conventional chemotherapy for women with metastatic breast cancer (Cochrane Review) online database, available at http://cochrane.de/cochrane. *Cochrane Database Syst Rev* (2003) **1**: CD003142.
28. Farquhar C, Basser R, Marjoribanks J, Lethaby A. High dose chemotherapy and autologous bone

marrow or stem cell transplantation versus conventional chemotherapy for women with early poor prognosis breast cancer (Cochrane Review) online database, available at http://cochrane.de/cochrane. *Cochrane Database Syst Rev* (2003) **1**: CD003139.

29. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1 – Breast cancer detection and death rates among women aged 40 to 49 years. *Can Med Assoc J* (1992) **147**: 1459–76.
30. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 2 – Breast cancer detection and death rates among women aged 50 to 59 years. *Can Med Assoc J* (1992) **147**: 1477–88.
31. Nystrom L, Andersson I, Bjurstam N, Frisell J, Nordenskjold B, Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomised trials erratum in *Lancet* (2002) **360**: 724. *Lancet* (2002) **359**: 909–19.
32. Gøtzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* (2000) **355**: 129–34.
33. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, Daly M, Wieand S, Tan-Chiu E, Ford L, Wolmark N. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst* (1998) **90**: 1371–88.
34. IBIS investigators. First results from the International Breast Cancer Intervention Study (IBIS-I): a randomised prevention trial. *Lancet* (2002) **360**: 817–24.
35. Vogel VG, Costantino JP, Wickerham DL, Cronin WM, Cecchini RS, Atkins JN, Bevers TB, Fehrenbacher L, Pajon ER Jr, Wade JL 3rd, Robidoux A, Margolese RG, James J, Lippman SM, Runowicz CD, Ganz PA, Reis SE, McCaskill-Stevens W, Ford LG, Jordan VC, Wolmark N; National Surgical Adjuvant Breast and Bowel Project (NSABP). Effects of tamoxifen vs raloxifene on the risk of developing invasive breast cancer and other disease outcomes: the NSABP Study of Tamoxifen and Raloxifene (STAR) P-2 trial. *JAMA* (2006) **295**: 2727–41.
36. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, Jeong J-H, Wolmark N. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *New Engl J Med* (2002) **347**: 1233–41.
37. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer: an overview of 61 randomized trials among 28,896 women. *New Engl J Med* (1988) **319**: 1681–91.
38. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). *Treatment of Early Breast Cancer*, Vol. I, *Worldwide Evidence 1985–1990*. Oxford: Oxford University Press (1990).
39. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *Lancet* (1992) **339**: 1–15; 71–85.
40. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of radiotherapy and surgery in early breast cancer: an overview of the randomized trials. *New Engl J Med* (1995) **333**: 1444–55.
41. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Ovarian ablation in early breast cancer: overview of the randomised trials. *Lancet* (1996) **348**: 1189–96.
42. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials. *Lancet* (2000) **355**: 1757–70.
43. Budman DR, Berry DA, Cirrincione CT, Henderson IC, Wood WC, Weiss RB, Ferree CR, Muss HB, Green MR, Norton L, Frei III E. Dose and dose intensity as determinants of outcome in the adjuvant treatment of breast cancer. *J Natl Cancer Inst* (1998) **90**: 1205–11.
44. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* (1989) **10**: 1–10.
45. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* (1979) **35**: 549–56.
46. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* (1983) **70**: 659–63.
47. Berry DA, Fristedt B. *Bandit Problems: Sequential Allocation of Experiments*. London: Chapman and Hall (1985).
48. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* (1990) **52**: 673–84.
49. Goodman S, Zahurak M, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. *Stat Med* (1995) **14**: 1149–61.
50. Ahn C. An evaluation of phase I cancer clinical trial designs. *Stat Med* (1998) **17**: 1537–49.
51. Thall PF, Lee JJ, Tseng C-H, Estey E. Accrual strategies for phase I trials with delayed patient outcome. *Stat Med* (1999) **18**: 1155–69.

5

Childhood Cancer

SHARON B. MURPHY[1] AND JONATHAN J. SHUSTER[2]
[1]Children's Cancer Research Institute, University of Texas Health Science Centre, San Antonio,
TX 78229-3900, USA
[2]College of Medicine, University of Florida, Gainesville, FL 32610-0212, USA

INTRODUCTION

There are substantial differences in the conduct of clinical cancer research in children compared with adults. First, childhood cancer is comparatively rare. According to statistics released by the National Cancer Institute SEER Program in 1999,[1] it is estimated that approximately 12 400 children and adolescents, younger than 20 years of age, are diagnosed annually with cancer in the United States. Stated another way, the average annual age-adjusted incidence rate for all childhood cancers is 150 per million persons, aged <20. This is under 2% of the total cancers diagnosed in the United States. Despite the rarity and notwithstanding the spectacular success in treatment of paediatric cancer, compared with incidence and mortality rates of cancer occurring among adults, cancer is the leading cause of death from disease among children and adolescents. Only accidents and firearms kill more children than cancer. Further, the distribution of cancer diagnoses in children is very different from that in adults. There are a number of major tumours, such as Wilm's tumour, retinoblastoma, rhabdomyosarcoma, neuroblastoma, Ewing's sarcoma and osteogenic sarcoma, for example, that are either exclusively or predominantly paediatric in nature. In contrast, carcinomas of differentiated epithelial tissues, like the aerodigestive tract or breast or prostate, do not occur in children. Thanks to the usual lack of co-morbid conditions and concomitant illnesses, children usually have a greater tolerance to cancer therapy than adults. Taken together with the differing spectrum of cancer seen, the host differences related to age necessitate that paediatric studies of anticancer drug dosage, efficacy and safety are needed. Recognising that children are not just small adults and that special considerations apply, the FDA has issued regulations mandating the testing of new drugs in paediatric patients.

Given the fact that modern treatments result in cure of 75–80% of all children and adolescents with cancer who are managed appropriately, the long-term consequences of therapy for children are also potentially much greater than in adults, as the therapy can interfere with normal growth and development, leaving them exposed for decades at risk for serious sequelae, major

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

organ disturbance, such as cardiac damage and cognitive dysfunction, or second malignancies. Children diagnosed with cancer generally also have less of a problem with competing mortality risks, when compared with adult cancer.

Given these factors, to adequately size childhood cancer research studies, a substantial proportion of the incident cases must be enrolled. In fact, in some situations, such as randomised Phase III trials of new regimens compared with already effective front-line treatments, nationwide multi-institutional trials are a necessity. A given trial will need to enrol nearly every child in the target population with the disease being studied for three to five years, with many years of follow-up needed to assess long-term outcomes. Accrual duration in childhood trials may be considerably longer than a corresponding adult trial. However, since clinical practice closely approximates that of the ongoing study, it is rare that progress from an external source ever renders the study question obsolete. Finally, given the rarity of childhood cancer, and the desirability of study designs of maximal efficiency, it is often desirable to conduct '2 × 2 factorial studies', where two interventions are used in the same trial (Standard vs. Standard + A vs. Standard + B vs. Standard + A + B). Such designs carry some risk where there is a qualitative interaction between the two interventions. For example, this would occur if the impact of A is highly dependent upon whether B is given or not. Hence the choice of randomised interventions needs to take this pitfall into account when such factorial designs are considered.

Based on previous trials and internal registry data of the major national paediatric cooperative oncology groups, Table 5.1 provides estimates of the potential accrual to the Children's Oncology Group (COG), a consortium of about 230 American, Canadian, European and Australasian medical centres. COG was formed in 2000 by the merger of the Pediatric Oncology Group (POG), the Children's Cancer Group (CCG), the Intergroup Rhabdomyosarcoma Study Group (IRSG) and the National Wilm's Tumor Study Group (NWTSG). Virtually every US or Canadian hospital with a childhood paediatric haematology/oncology division belongs to COG. Note that the anticipated annual accrual to COG trials does not mirror incidence figures, as there is a substantial gap between incident rates and rates of referral of newly diagnosed paediatric

Table 5.1. Major categories of paediatric cancer and projected annual accrual of the Children's Oncology Group

Leukaemia
Infant ALL[a]: 60
'Standard Risk' B-Precursor ALL[a]: 1100
'High Risk' B-Precursor ALL[a]: 600
T-Cell ALL[a]: 240
Philadelphia Chromosome Positive (Ph+) ALL[a]: 60
B-Cell (Sig+)ALL[a]: 45
ANLL[a]: 350

Lymphoma
Hodgkin Disease: 250
'Early Stage' NHL[a]: 90
'Advanced Stage' Lymphoblastic NHL[a]: 90
'Advanced Stage' Large Cell NHL[a]: 80
'Advanced Stage' Small Non-Cleaved Cell NHL[a]: 125

Brain Tumours
High Grade Glioma: 50
Low Grade Glioma: 140
Brainstem Glioma: 40
Medulloblastoma: 200
Ependymoma: 60

Sarcomas
Ewing's Sarcoma: 170
Osteosarcoma: 170
'Low Risk' Rhabdomyosarcoma: 60
'Intermediate Risk' Rhabdomyosarcoma: 100
'High Risk' Rhabdomyosarcoma: 30
Non-rhabdomyosarcoma soft tissue sarcoma: 150

Retinoblastoma
Group B Intraocular: 25
Group C/D Intraocular: 23
Unilateral Enucleated Intraocular: 120
Extra-ocular:15

Kidney
Wilm's Tumour: 480

Embryonal
'Low Risk' Neuroblastoma: 200
'Intermediate Risk' Neuroblastoma: 100
'High Risk' Neuroblastoma: 225
Hepatoblastoma: 65
Germ Cell Tumours: 25

[a] ALL = Acute Lymphoblastic Leukaemia
ANLL = Acute Non-Lymphoblastic Leukaemia
NHL = Non-Hodgkin Lymphoma.

cancer patients to member institutions of COG and participation in clinical trials. Ross *et al.*[2] analysed 21 026 incident paediatric cancer cases, diagnosed from 1989 to 1991, and compared observed to expected numbers of cases seen at member institutions of POG and CCG and found vastly different ratios (observed/expected) depending on age and site, and, to a much less extent, geographic region. According to this survey, 92% of children aged less than 15 years in the United States received their care at a CCG or POG institution, thereby providing at least a mechanism approximating population-based studies. However, the ratio (O/E) for 15–19 year olds was only 0.21, pointing to an adolescent gap in access to national cancer clinical trials at qualified institutions.

As seen in Table 5.1, the COG runs Phase III clinical trials in a wide variety of tumours, with accrual ranging from as few as 20 patients per year to over 1000 per year. The COG is also heavily involved in correlative science, pilot studies of potential Phase III interventions, as well as standard Phase I (dose escalation) studies and Phase II (early efficacy) studies. The COG places special emphasis on translational research (biologic correlation studies) and cancer control (supportive care studies to limit long-term side effects and epidemiologic studies to learn about the aetiology of childhood cancer). Due to the presumably genetic origin of most forms of childhood cancer and the short lag time between symptoms and diagnosis, prevention trials and screening trials are difficult to do in paediatric cancer, with neuroblastoma,[3] which is based on urinary screening for elevated levels of catecholamines, a notable exception.

This chapter is organised into several sections. In the first section, major accomplishments in the area of childhood cancer treatment are discussed. In the following section, examples are cited where translational research has affected the design of paediatric cancer trials. In the succeeding section, the typical methods of designing trials for the COG are presented. A special section dealing with the ethical aspects and unique considerations affecting the conduct of clinical trials in children is also included. The final section is devoted to a look into the future.

HISTORY AND PERSPECTIVES ON IMPORTANT PAEDIATRIC CANCER CLINICAL TRIALS

Statistically and clinically significant improvements have been achieved in all major forms of childhood cancers through conduct of well-organised single institution and cooperative group clinical trials which have resulted in sequential and steady improvement in survival rates since the 1960s when curative treatments were first devised. Documentation of the overall progress achieved by POG investigators has been reported, demonstrating significant improvements in overall survival (OS) and event-free survival (EFS) for 8 of 10 disease areas, in a sample of over 7000 children and adolescents treated between 1976 and 1989.[4] Similar results have been achieved by CCG and by European national paediatric cooperative clinical trials organisations. Figure 5.1 shows marked improvements in five-year survival rates for some diagnoses and more limited improvement for others. Improvements in survival rates have been mirrored by steadily declining rates for cancer-related mortality in children (Figure 5.2).[5] Of note, declines in mortality have been greater for lymphoid cancers (acute lymphoblastic leukaemia (ALL), non-Hodgkin lymphoma and Hodgkin disease) than for non-lymphoid cancers. Mortality for the former declined at a rate of 4.7% per year between 1979 and 1999, compared with only a 1.9% rate of annual decline for the latter. There is also evidence that children and adolescents with ALL, non-Hodgkin lymphoma (NHL), Wilm's tumour, medulloblastoma and rhabdomyosarcoma enjoy a significant survival advantage when treated according to well-defined protocols, compared with paediatric patients not enrolled on protocols and treated outside of paediatric cancer centres.[6] Most probably the inclusion benefit related to participation in clinical trials is a result of a number of factors, including the rigorous process of protocol development, incorporation of rapid pathology review and reference laboratories, defined

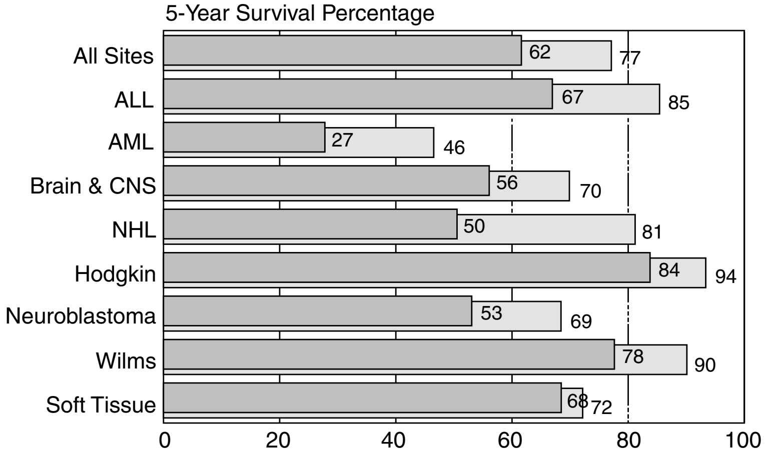


Figure 5.1. Five-year relative survival rates by site for 0–14 years, both sexes, all races, 1975–9 versus 1992–8. The top bar for each diagnosis is for 1975–9, and the lower bar is for 1992–8. Data are from the Surveillance, Epidemiology and End Results (SEER) Program of the NCI.

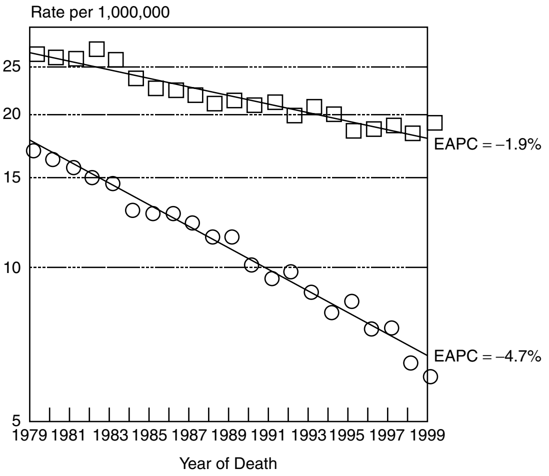


Figure 5.2. Declines in lymphoid versus non-lymphoid cancer mortality (age 0–14 years) for both sexes, all races, 1979–99. The top line (squares) is for non-lymphoid cancers, whereas the lower line (articles) is for lymphoid cancers. Lymphoid cancers include acute lymphoblastic leukaemia, non-Hodgkin lymphoma and Hodgkin disease. Rates are per 1 000 000 and are age-adjusted to the 2000 US standard. Data are from the Surveillance, Epidemiology and End Results (SEER) Program of the NCI. EAPC is estimated annual percentage change.

staging practices and procedures, on-study review of radiotherapy port films, and close monitoring for toxicity and efficacy. Some of the important advances achieved in treatment of paediatric cancers are listed in Table 5.2.

The NCI's SEER Program shows that the overall incidence of the most common form of paediatric malignancy, ALL, has been increasing. It increased from 19/106 person-years in 1973–77 to 28/106 person-years in 1993–98 ($P < 0.0001$);[7] however, success in treatment of ALL has been most gratifying. Indeed, a major reason for improvements in overall survival for childhood cancer in general is due to improvement in survival rates for ALL, which accounts for roughly a third of paediatric cancer.[1] With modern chemotherapy, 97–99% of children can be expected to attain complete remission, and it is not inconceivable to predict that modifications of the currently most successful protocols will boost long-term leukaemia-free survival rates to as high as 85–90%. Treatment success has been achieved through post-induction intensification/consolidation and re-induction treatments, effective treatments ('prophylaxis') for subclinical central nervous system leukaemia, and prolonged

Table 5.2. Examples of important advances resulting from paediatric cancer clinical trials

- Adjuvant chemotherapy improves survival from 20% to 70% in non-metastatic osteosarcoma of the extremity[8]
- Doxorubicin improves outcome when added to other chemotherapy for Ewing's sarcoma[9] and the addition of ifosfamide and etoposide to vincristine, adriamycin, cyclophosphamide and actinomycin results in greater benefit[10]
- Radiation therapy does not improve survival for patients receiving chemotherapy with Stage I and II, Wilm's tumour,[11,12] Stage I rhabdomyosarcoma[13] or localised non-Hodgkin lymphoma[14]
- Demonstration of improved event-free survival in high-risk neuroblastoma receiving myeloablative therapy in conjunction with autologous bone marrow transplantation and subsequent treatment with 13-*cis*-retinoic acid compared with chemotherapy alone[15]
- Attainment of 80% four-year event-free survival rates for standard risk B-precursor ALL[16]
- Achievement of 78% EFS for patients with loco-regional embryonal rhabdomyosarcoma through intensification of chemotherapy in Intergroup Rhabdomyosarcoma Study (IRS)-IV[17]
- Demonstration that dexamethasone is a more potent steroid than prednisone resulting in improved EFS in Standard Risk ALL[18]
- Augmented BFM-type post-induction intensified therapy results in improved EFS and survival in High Risk ALL patients who are slow responders[19]
- The use of adjuvant chemotherapy permits CSRT dose reduction to 2340 cGy with >75% survival for M0 medulloblastoma[20,21]
- Extent of resection is associated with outcome for children with high-grade glioma[22]
- The identification of the combination of anthracylines and cytarabine as the foundation of AML chemotherapy and the benefits of dose/time intensive therapy and matched related donor allogeneic BMT[23,24]

antimetabolite-based continuation treatments of 24–36 months' duration. Advances have been achieved by many single institutions and cooperative groups treating childhood leukaemias, including investigators at St Jude Children's Research Hospital in Memphis who pioneered a 'Total Therapy' curative approach beginning in the 1960s. It is beyond the scope of this chapter to review the treatment advances achieved through clinical trials for ALL by the BFM (Berlin–Frankfurt–Münster) Group, POG, CCG, the Dana Farber Consortium, the Medical Research Council/UKALL, the Dutch Childhood Leukaemia Study Group, the French Acute Lymphoblastic Leukaemia Cooperative Group (FRALLE) and the Italian Association of Paediatric Haematology–Oncology (AIEOP), but the interested reader may consult reviews summarising the spectacular progress achieved in treatment of ALL.[25]

The lymphomas, Hodgkin disease (HD) and non-Hodgkin lymphomas (NHLs) are the third most common form of paediatric malignancy, next in frequency behind leukaemias and tumours of the central nervous system. Currently 80–90% of all children and adolescents with malignant lymphomas are curable with optimal multidisciplinary management, based on immunopathologic classification, staging for determination of disease extent, and design and selection of risk-adapted therapies. Paediatric investigators at Stanford, beginning in 1970, first pioneered combined modality treatment for children with HD and demonstrated that low-dose involved field radiotherapy combined with multiple cycles of chemotherapy (MOPP or MOPP/ABVD) resulted in cure of 90% of paediatric patients.[26] Similarly outstanding rates of disease control with combined modality management of paediatric HD have since been reported by others, establishing the curability of HD in nearly all cases, such that the thrust of current trials in paediatric HD is towards reduction of serious late effects of HD treatments, such as secondary malignancies, particularly leukaemia, infertility, pulmonary fibrosis and restrictive lung disease, serious cardiac problems and premature death.

The NHLs occurring among children and adolescents are virtually all high-grade, diffuse malignancies, differing markedly from the distribution of histologic types typically seen among older adults. Staging systems in use for childhood and adult NHL also differ:[27] 90%

of localised NHLs, regardless of histology, are readily cured by nine weeks of chemotherapy without radiation.[28]

Progress in the treatment of paediatric solid tumours has been equally striking in the last 30 years as progress in treating childhood leukaemias and lymphomas, and may be attributable to development of accurate diagnostic methods and systems of disease staging and effective multimodal treatments combining surgery, chemotherapy and radiation. Cure rates for rhabdomyosarcoma have increased from approximately 25% in 1970 to greater than 75% currently, to 60–70% for non-metastatic bone sarcomas, to over 80% for Wilm's tumour, over 90% for retinoblastoma, over 90% for infants and children with localised neuroblastoma, and to over half of all children with brain tumours.

Aims of current trials are to increase or preserve high cure rates, decrease acute toxicity and long-term adverse sequelae of treatment, decrease costs and improve the quality of life for children with readily curable cancers. Patients with high risk or metastatic disease at diagnosis or those who recur after front-line therapies continue to pose challenges and should properly benefit from pilot trials and Phase I or II studies of new treatments.

PROGNOSTIC FACTORS, TRANSLATIONAL RESEARCH AND THERAPEUTICALLY RELEVANT RISK GROUPS

Successful childhood cancer research is in large part dependent upon its translational research programme. Over the past three decades, initial diagnosis and classification of childhood cancer has become far more sophisticated, as laboratory scientists have collaborated closely with clinical investigators. In addition, special biological and pharmacological studies, conducted during and after treatment, offer tools to clinical investigators that were never previously available. As a result, paediatric oncologists, surgeons, pathologists and collaborating statisticians have the opportunity and the obligation to design and stratify trials specifically for biologically defined, risk-adapted subsets of patients. For example, the National Wilm's Tumor Study-5, a therapeutic trial and biology study, was designed to reduce treatment intensity for the subgroup(s) of patients with the most favourable prognosis and intensify treatment for the patients with the least favourable prognosis, based on stage, histology (favourable or unfavourable, anaplastic, rhabdoid and clear cell types), tumour size and bilaterality, and to investigate the impact of loss of heterozygosity (LOH) of chromosome 16q and 1p on two-year relapse-free survival through collection of tumour and normal kidney tissue for DNA analysis and banking.

Perhaps the best example of important translational research that has led directly to therapeutic implications is in the collection of bone marrow specimens for cytogenetic studies in childhood ALL.[29] While classical karyotype analysis is typically informative in 60–70% of the patients, important genetic markers can now be identified by probes, using fluorescence *in situ* hybridisation (FISH) or polymerase chain reaction (PCR) in virtually all patients. Translocations, such as the t(4;11),[30–32] t(9;22)[33–35] and t(1;19),[36,37] confer an adverse prognosis and lead to targeting the patients for more aggressive therapy. On the other hand, patients with the cryptic t(12;21) genetic lesion encoding the TEL-AML1 transcript,[38–40] with hyperdiploid leukaemia identified by flow cytometric measurement of DNA index (typically 53+ chromosomes in their primary clone),[41,42] or with specific trisomies detected by FISH, such as 4, 10 and 17,[42,43] have a more favourable outcome and can be targeted for less intensive treatment. As an example of the latter, POG investigators designed a trial (#9201) with less intense chemotherapy for ALL patients with lesser risk of relapse, defined by initial white blood cell counts <50 000, age between 1 and 10 years, absence of CNS disease, and presence of one or both of the following: DNA index >1.16, and/or trisomies of chromosomes 4 and 10 by FISH.

In addition to the well-recognised prognostic importance of initial white blood cell count, age at diagnosis, extramedullary disease and blast cell genetic features in B-precursor ALL and their significance for stratification and trial design, the early response to therapy, presence of minimal residual disease (MRD), and pharmacologic and pharmacokinetic variables are also predictive of outcome. Slow early response to induction treatment is predictive of an adverse outcome and can be defined in several ways: slow clearance of circulating blast cells to one week of prednisone or multiagent induction, or greater than 25% marrow blasts on day 8 (or day 15) of treatment. Quantitation of MRD by immunologic methods or PCR assay of rearranged T-cell-receptor or immunoglobulin heavy-chain genes of leukaemic blasts as clonal markers of leukaemia in patients in clinical remission has been shown to identify patients at elevated risk for relapse, a factor which should be taken into account in assigning alternative treatment.[44,45] Wide variability in absorption of orally administrated chemotherapy, such as 6-mercaptopurine, and inter-patient variability in systemic exposure to both methotrexate and 6-mercaptopurine are important determinants of outcome in ALL.[46,47] Graham *et al.*[48] uncovered a pharmacologic interaction between methotrexate and cytosine arabinoside when given simultaneously, and demonstrated a correlation between host drug levels and adverse outcome. Individual variability in response to cancer treatment is surely related to genetic polymorphisms in drug-metabolising enzymes, transporters, receptors and other drug targets, and suggests that these genetic differences may form a solid scientific basis for optimising therapies within the context of clinical trials.[49] Current COG ALL studies are linked to a common classification/biology study, with the goal of integrating patient data into an organised framework to risk-stratify and assign patients to appropriate post-induction therapy, and to provide the necessary infrastructure for specimen banking for future research. The current ALL classification system was built on the experience of the two legacy groups CCG and POG. The outcome of over 8 000 patients was assessed in regard to clinical and biological variables tracked by both groups (NCI risk group, favourable trisomies, translocations, hypodiploidy and response to therapy). The goal was to define groups with differential risks of relapse in order to identify groups in which different therapeutic questions were appropriate. These efforts resulted in the identification of five distinct groups with different outcomes, including standard risk-low, standard risk-average, standard risk-high, high-risk and very high-risk (VHR) groups. The results of this effort have led to the current recommendations for the risk-based classification of B-precursor and T-cell ALL patients over 1 year of age. Risk assessment is based on clinical features (age, initial white blood cell count, immunophenotype), biological characteristics of the leukaemic blasts such as the presence or absence of specific genetic features (hypodiploidy, Trisomies 4/10/17, TEL-AML1, BCR-ABL, MLL translocations and/or fusion transcripts), early marrow response (day 8/15) as measured by morphology, and by assessment of MRD burden at the end of induction.

Given the plethora of prognostic factors now known for most paediatric malignancies, a pragmatic and rational approach to clinical trials design and stratification consists of risk assignment by a combination of clinical and biological factors identified through multivariate analysis to be of prognostic significance. Treatment is then tailored to risk status, commonly considering variables such as patient age, extent of disease and tumour biology. For example, childhood rhabdomyosarcoma is the most common malignant soft tissue tumour seen in children and accounts for about 3.5% of all cancer seen in children less than 15 years of age. The risk assignments used to direct treatment for the fifth generation of studies (shown in Table 5.3), conducted by the Intergroup Rhabdomyosarcoma Study Group (IRSG, since merged to be part of the COG), were based on prognostic factors identified in IRSG Studies I–IV, conducted from 1972 through 1991. These include (1) undetectable distant metastases at diagnosis; (2) primary sites in the orbit and non-parameningeal head/neck and genitourinary non-bladder/prostate regions; (3) grossly complete surgical removal of localised tumour

Table 5.3. Risk group assignments for Intergroup Rhabdomyosarcoma Study Group V

Risk (protocol)	Stage	Group	Site	Size	Age	Histology	Metastasis	Nodes	Treatment
Low, subgroup A	1	I	Favourable	a or b	<21	EMB	M0	N0	VA
(D9602)	1	II	Favourable	a or b	<21	EMB	M0	N0	VA + XRT
	1	III	Orbit only	a or b	<21	EMB	M0	N0	VA + XRT
	2	I	Unfavourable	a	<21	EMB	M0	N0 or NX	VA
Low, subgroup B	1	II	Favourable	a or b	<21	EMB	M0	N1	VAC + XRT
(D9602)	1	III	Orbit only	a or b	<21	EMB	M0	N1	VAC + XRT
	1	III	Favourable (excluding orbit)	a or b	<21	EMB	M0	N0 or N1 or NX	VAC + XRT
	2	II	Unfavourable	a	<21	EMB	M0	N0 or NX	VAC + XRT
	3	I or II	Unfavourable	a	<21	EMB	M0	N1	VAC (+XRT, Gp II)
	3	I or II	Unfavourable	b	<21	EMB	M0	N0 or N1 or NX	VAC (+XRT, Gp II)
Intermediate	2	III	Unfavourable	a	<21	EMB	M0	N0 or NX	VAC ± Topo + XRT
(D9803)	3	III	Unfavourable	a	<21	EMB	M0	N1	VAC ± Topo + XRT
	3	III	Unfavourable	b	<21	EMB	M0	N0 or N1 or NX	VAC ± Topo + XRT
	1 or 2 or 3	I or II or III	Favourable or unfavourable	a or b	<21	ALV/UDS	M0	N0 or N1 or NX	VAC ± Topo + XRT
	4	I or II or III or IV	Favourable or unfavourable	a or b	<10	EMB	M1	N0 or N1 or NX	VAC ± Topo + XRT
High (D9802)	4	IV	Favourable or unfavourable	a or b	≥10	EMB	M1	N0 or N1 or NX	CPT-11/vincristine, VAC + XRT
	4	IV	Favourable or unfavourable	a or b	<21	ALV/UDS	M1	N0 or N1 or NX	CPT-11/vincristine, VAC + XRT

Favourable = orbit/eyelid, head and neck (excluding parameningeal), genitourinary (not bladder or prostate) and biliary tract.
Unfavourable = bladder, prostate, extremity, parameningeal, trunk, retroperitoneal, pelvis, other.
a = tumour size ≤ 5 cm in diameter; b = tumour size >5 cm in diameter.
EMB = embryonal, botryoid, or spindle-cell rhabdomyosarcomas or ectomesenchymomas with embryonal RMS.
ALV = alveolar rhabdomyosarcomas or ectomesenchymomas with alveolar RMS, UDS = undifferentiated sarcomas.
N0 = regional nodes clinically not involved; N1 = regional nodes clinically involved; NX = node status unknown.
VAC = vincristine, actinomycin D, cyclophosphamide; XRT = radiotherapy; Topo = topotecan; Gp = Group; CPT-11 = irinotecan.
Source: Reproduced from Raney *et al.*[50] (p. 218), with permission.

at diagnosis; (4) embryonal/botryoid histology; (5) tumour size ≤ 5 cm; and (6) age younger than 10 years at diagnosis. Patients classified into low-, intermediate- and high-risk categories are predicted to have an estimated three-year EFS rate of 88%, 55–76% and <30%, respectively.[50] Subsequent COG studies of the treatment of rhabdomyosarcoma have sought to reduce the intensity of therapy and risk of late effects for patients with low-risk disease, while intensifying therapy for patients with high-risk disease.

Similarly, significant advances in translational research is neuroblastoma, which accounts for 8 to 10% of all childhood cancers, have resulted in a refined risk-related approach to therapy based on the age of the patient, the stage of the tumour according to the International Neuroblastoma Staging System (INSS), histopathologic features, the *MYCN* copy number status and the ploidy of tumour cells (Table 5.4). This risk group stratification is under revision based on the recent results of Attiyeh *et al.*[52] regarding the prognostic

Table 5.4. International Neuroblastoma Staging System (INSS)

Stage 1: Localised tumour continued to the area of origin; complete gross resection, with or without microscopic residual disease; identifiable ipsilateral and contralateral lymph node negative for tumour

Stage 2A: Unilateral with incomplete gross resection; identifiable ipsilateral and contralateral lymph node negative for tumour

Stage 2B: Unilateral with complete or incomplete gross resection; with ipsilateral lymph node positive for tumour; identifiable contralateral lymph node negative for tumour

Stage 3: Tumour infiltrating across midline with or without regional lymph node involvement; or unilateral tumour with contralateral lymph node involvement; or midline tumour with bilateral lymph node involvement

Stage 4: Dissemination of tumour to distant lymph nodes, bone marrow, liver or other organs except as defined in stage 4S

Stage 4S: Localised primary tumour as defined in Stage 1 or 2, with dissemination limited to liver, skin or bone marrow

COG neuroblastoma risk group and protocol assignment schema

INSS stage	Age (years)	*MYCN* status	Shimada histology	DNA ploidy	Risk group
1	0–21	Any	Any	Any	Low
2A and 2B	<1	Any	Any	Any	Low
	≥1–21	Non-amplified[a]	Any	NA	Low
	≥1–21	Amplified[b]	Favourable	NA	Low
	≥1–21	Amplified	Unfavourable	NA	High
3	<1	Non-amplified	Any	Any	Intermediate
	<1	Amplified	Any	Any	High
	≥1–21	Non-amplified	Favourable	NA	Intermediate
	≥1–21	Non-amplified	Unfavourable	NA	High
	≥1–21	Amplified	Any	NA	High
4	<1	Non-amplified	Any	Any	Intermediate
	<1	Amplified	Any	Any	High
	≥1–21	Any	Any	NA	High
4S	<1	Non-amplified	Favourable	>1	Low
	<1	Non-amplified	Any	1	Intermediate
	<1	Non-amplified	Unfavourable	Any	Intermediate
	<1	Amplified	Any	Any	High

[a] *MYCN* copy number ≤10.

[b] *MYCN* copy number >10.

INSS, International Neuroblastoma Staging System; NA, not applicable.

Source: Reproduced from Castleberry,[51] (pp. 926, 930), with permission from Elsevier.

value of unbalanced 11q LOH and 1p36 LOH as well as London *et al.*[53] regarding an age cut-off greater than 365 days.

For neuroblastoma patients in COG, the determination of risk group factors is performed in real time to facilitate rapid and accurate treatment assignment. Tumor specimens are submitted to central labs, the labs submit the results via a web-based data collection system viewable by the treatment institution, and the institution receives electronic notification of the treatment group assignment. With patient consent, leftover tumour specimens are banked for future research, and the International Neuroblastoma Virtual Tumor Bank (VTB) database permits linkage of laboratory results to patient clinical and outcomes data for statistical analysis.

Because childhood cancer is rare, national reference laboratories have been established to analyse and store samples from the membership of the COG as well as other large institutions and other international paediatric clinical trials organisations. Such laboratories help the research programme in terms of scientific expertise, quality control and correlative science. Few institutions can afford to maintain such laboratories solely for their own paediatric cancer patients, and web-based informatics applications afford access to the most sophisticated online resources and information even in smaller remote centres.

STUDY DESIGN FOR CHILDHOOD CANCER TRIALS

PHASE I STUDY DESIGN

Because childhood cancer is rare and the response to conventional treatment good, most children never experience recurrent disease and are thus not eligible for trials of new agents. Phase I trials are designed to estimate the maximal tolerated dose of a drug, to determine the nature and frequency of toxicities, and to define the drug pharmacokinetics. While eligibility varies, patients have typically failed front-line therapy and usually they will also have failed second-line therapy. Because of the small number of paediatric patients eligible for Phase I trials, most are accomplished as multi-institutional collaborations. Paediatric drug development requires separate Phase I studies (i.e. separate and distinct from studies done in adults) because paediatric patients may tolerate either higher or lower levels of drugs and may exhibit toxicities unique to children. Separate trials warranting emphasis may also reflect unique agents active in paediatric tumours, differing from agents that are of the highest priority for cancers common among adults.

The basic design is to begin at about 80% of the adult maximal tolerated dose. Patients are entered in cohorts and treated at increasing doses. At each level, three patients are typically accrued. If there is no dose-limiting toxicity amongst the three patients, the dose is raised to the next level (usually a 20–30% escalation), in successive cohorts of patients with no intrapatient dose escalation. If two or all three of these initially accrued patients experience dose-limiting toxicity (DLT), the maximum tolerated dose (MTD) will have been deemed exceeded. Finally, if one patient amongst the initial three patients experiences dose-limiting toxicity, an additional three patients are accrued. If six patients are needed, a dose escalation will occur if a total of one in six (i.e. zero of the next three) has dose-limiting toxicity. If two or more (i.e. one or more of the next three) experience dose-limiting toxicity, the maximal tolerated dose will be deemed to have been exceeded. The MTD is defined as the dose level immediately below the level at which two patients in three to six experience DLT. The definition of DLT can vary from study to study, but it generally falls into two categories: (a) Grade 3, 4 or 5 non-haematologic toxicity other than (1) Grade 3 nausea/vomiting, (2) Grade 3 transaminase elevation, and (3) Grade 3 fever/infection; and (b) Grade 4 myelosupression that lasts more than 7 days, which requires transfusions twice in 7 days, or causes a delay in therapy exceeding 14 days. While the study is temporarily closed after accrual of each set of three patients in order to assess patient-specific responses and toxicities, a patient reservation system is used to reserve places when and if the study reopens.

Phase I trials often require the evaluation of many dose levels. At a given dose level, the probabilities of declaring that the MTD has been exceeded are 9.3%, (50%) and [83%], when the true probabilities of dose-limiting toxicities are respectively 0.1, (0.3) and [0.5].

Consensus guidelines established by American and European investigators for the conduct of paediatric Phase I trials have been established.[54] A problem recently identified is the determination of MTDs in paediatric trials that are lower than those defined in adult patients, which may relate to differences in the intensity of prior therapy between adult and paediatric patients entered onto Phase I trials. There is a well-established association between prior therapy and reduced tolerance to myelotoxic drugs. If current paediatric Phase I trials in heavily pretreated patients define MTDs that tend to be lower than those determined in adult patients with minimal prior therapy, then application of the paediatric MTD to less heavily pretreated paediatric patients, e.g. in Phase II trials, may be problematic.

PHASE II STUDY DESIGN

The specific purpose of a Phase II trial is to determine activity, i.e. to develop estimates of the response rate of patients with specific tumour types to a particular drug or novel combination. Eligible patients typically will have relapsed on a front-line therapy, and the prospect of a cure is unlikely. Typically, the dependent variable is an objective all or none response variable such as achievement of a complete or partial (>50%) response. Interim results are masked from the participants until the study closes to accrual and response information for all patients has been established. There are three types of Phase II trial designs that depend upon the study objectives.

Design to Test Activity

The most common design is one to 'prove activity'. For these studies, a fixed objective response rate is specified for activity (null hypothesis), and the goal is to reject the hypothesis in favour of the alternate hypothesis that the response rate is greater than this fixed figure. Generally, since the number of Phase II agents that can be tested is large in comparison with patient availability, sequential designs are preferred. However, as Simon[55] pointed out, it is rarely advantageous to go beyond two stages. Two excellent references with regard to Phase II design are Simon[55] and Shuster.[56] The designs of Simon[55] stop at the first stage only if lack of activity is demonstrated. His argument is that patients should benefit from active drugs. However, in paediatrics, due to the relative scarcity of patients with recurrent disease, designs that stop early for either lack of activity or proven activity are preferred.

Design for Historical Comparison

Another strategy for defining efficacy would be to prove a response rate is superior to that seen in an historical control study. The response rate of the new study is statistically compared with that of the control therapy. Makuch and Simon[57] have provided methods to determine the sample size requirements for these studies. Chang *et al*.[58] have extended this to two-stage designs (i.e. a sequential approach that could save patient resources).

Design for Randomised Phase II Comparison

Due to a limited availability of patients, it is exceedingly rare that a randomised comparison of a new agent to a control is feasible in a paediatric Phase II study. However, such studies have been done. See McWilliams *et al*.[59] for an example from childhood neuroblastoma. As above, two-stage or group sequential designs are the preferred method. The programme EAST[60] can be used for designs that allow for both early acceptance and early rejection of the null hypothesis that the new treatment is equivalent to the control treatment.

In paediatric oncology, with limited patient numbers, only one or two cooperative Phase II trials are conducted with each new agent, and all malignancies refractory to standard therapy are typically combined into a single paediatric Phase II trial, usually stratified by histology. Not surprisingly, Phase II trials of novel multiagent regimens provide greater evidence of activity than single-agent Phase II trials and offer considerable possibility of therapeutic benefit.[61]

PHASE III STUDY DESIGN

These studies typically ask a randomised question about either survival or event-free survival (the time from study entry to the earliest of induction failure, relapse, second cancer, or death of any cause). Intent-to-treat[62] is the analysis of choice for efficacy, with other analysis done as secondary supportive inference. For treatment questions where the randomised divergence occurs considerably after study entry or where a significant number of failures are expected to occur before divergence, a delayed randomisation is typically done as close to the divergence point as possible. For these randomisations, the dependent variable would be event-free survival from the randomisation date.

Phase III studies are typically designed assuming either proportional hazards or the cure model of Sposto and Sather.[63] In either case, the designs are group sequential in nature with planned interim analyses. In the case of proportional hazards, the O'Brien–Fleming method[64] is used. The reader is referred to Shuster[65] for specific details. Nearly all Phase III childhood cancer trials are run either as two-armed studies or as 2 × 2 factorial studies. It is rare that sufficient numbers of paediatric cancer patients are available to conduct three-armed studies, except perhaps in ALL, the most commonly occurring malignancy. The type of questions utilised in 2 × 2 factorial studies must be such that the expectation is for no 'qualitative interaction' between the two interventions. A qualitative interaction between treatments A and B would occur if a standard regimen plus A is superior to the standard regimen alone, but the standard plus A plus B is inferior to the standard plus B. For example, if a study is to randomise leukaemia patients to receive or not receive regimen A, designed to have an impact on the CNS, while at the same time to receive or not receive regimen B, designed to have an impact on marrow remission, a factorial design would seem appropriate. Essentially, we can run two studies for the price of one. If the two interventions have much in common, this would be a contraindication for a factorial design. In contrast, if we wished to ask if the same drug had an impact in induction therapy (first intervention) and in maintenance therapy (second intervention), there is, at least intuitively, the plausibility that the advantage of both interventions over just one may be zero or even harmful.

Phase III studies done in cooperative groups are required by the NCI to have a Data Safety and Monitoring Board which reviews the study at a minimum of every six months for toxicity and at planned intervals for efficacy, until it releases the study to the study committee. The release can occur no sooner than the earlier of (1) all subjects have completed the planned intervention or (2) the study was closed early and a new intervention is needed for patients on one or both arms. Any release prior to the planned date of final analysis requires approval of the board. Double-blind Phase III studies are rarely feasible due to the toxic nature of cancer treatment. However, they are encouraged for studies of supportive care, as long as the intervention is given in a pill form, and has no major known side effects requiring special medical monitoring.

Equivalence questions are often posed for paediatric cancer. For such studies, a very high cure rate of at least 85% has been shown possible on a conventional regimen. The question posed is: can we do 'almost as well' with reduced therapy? To answer such questions with confidence requires large numbers, and it is rare that even the entire patient resources of COG are sufficient to address this in a randomised manner. For example, if a disease has a historical four-year remission rate of 90%, and an accrual rate of 200 patients per year, a randomised study would take six years of accrual (10-year duration) to have 95% power to detect a degradation to 85% under reduced therapy at $p = 0.20$, one-sided. (Note that the typical values of Type I error and power are reversed.) A single-arm study would require 315 patients to ask the same question of a fixed standard of 90% vs. a reduction to 85% (nearly a 75% reduction in sample size). While the benefits of reduced therapy may be obvious, such studies carry considerable risk and must be carefully monitored for early evidence that the reduction in therapy is unsafe and is associated with an inferior outcome.

ANCILLARY STUDIES

In paediatric cancer, there is considerable activity in translational research (see above). This can take the form of biologic studies, late effects, or in controlling acute side effects. These studies are designed on a case-by-case basis. Examples include the conduct of case–control 'tissue bank' studies to establish a promising prognostic marker. Cases are defined as patients failing a protocol (typically a relapse) and controls are long-term successes. These studies can be done using sequential designs, typically two-stage designs. Other typical studies might look at cognitive impairment (multivariate analysis of variance of neuropsychological variables), acute toxicity of a specified type (typical chi-square test), the prognostic significance of serial pharmacologically measured drug levels (time-dependent covariate in survival analysis), or exploratory analysis (e.g. microarrays).

ETHICAL AND OTHER SPECIAL CONSIDERATIONS AFFECTING CONDUCT OF TRIALS IN CHILDREN WITH CANCER

Children and adolescents constitute a special vulnerable population of research subjects, often grouped with other special classes, like the mentally retarded, mentally ill and prisoners. There are special federal protections which apply to all research involving children as subjects which are covered by Subpart D of Part 46 of Title 45 of the Code of Federal Regulations (45 CFR 46), requiring that institutional review boards (IRBs) give consideration to the degree of risk, the benefit to child subjects, the nature of the knowledge to be gained, permission of the parent or guardian, and the concurrence of the child subjects, known as assent. A child's capacity to give assent is conditioned by his or her developmental level.[66,67]

Subsequent to the promulgation of the original rules, adopted in 1983 and modified in 1991, there has been nearly continuous debate and controversy surrounding safeguards for all human subjects of research and for children especially. The tragic death of an 18-year-old research subject in 1999 in a gene-transfer trial at a major research university in which human subjects were not protected, adverse events had not been reported, and financial conflicts of interest were involved, served to trigger several new federal initiatives to further strengthen protection of human research subjects in clinical trials,[68] including the imposition of sanctions on investigators who fail to adhere to regulations. The federal Office for Protection from Research Risks (OPRR) has been reorganised, expanded and renamed the Office for Human Research Protections (OHRP) and transferred to the Office of the Secretary, Health and Human Services (HHS). The National Biothetics Advisory Commission, at the request of the President, has undertaken a sweeping examination of the ethical and policy issues in the oversight of human research in the United States (see www.bioethics.gov). As a result, the ethical and regulatory framework within which paediatric cancer clinical trials are conducted, now and in the future, will continue to evolve, and investigators must remain abreast.

Specific ethical issues impacting statisticians involved in collaborative research include ensuring confidentiality, data and safety monitoring, challenges in interpretation of interim analyses, and planning equivalence studies.[69] An equivalence question (e.g. what is the minimum therapy needed to produce cure?) has particular relevance for paediatric cancer trials which are (often) aimed at reduction of the acute or delayed effects of cancer treatment on the growing child.

Notwithstanding the strict ethical guidelines and regulations surrounding research in children, there is substantial and even increasing pressure to enrol children in clinical trials as a result of other federal policies and recent legislation, including the FDA's 1998 paediatric rule, the paediatric provisions of the FDA Modernization Act (FDAMA) of 1997, and the sweeping Children's Health Act of 2000 (PL 106–310), the sum of which is certain to increase paediatric clinical trials, particularly drug trials. Federal NIH policies promulgated in 1998 were aimed at increasing the participation of children in research so that adequate data would be developed to support the treatment for disorders affecting adults which also affect children, and rules mandated that children (i.e. individuals under age 21) must be included in all human subjects research unless there are scientific and ethical reasons not to include them. The FDA rules and regulations[70] require pharmaceutical manufacturers to assess the safety and effectiveness of new drugs and biologics in paediatric patients and establish powerful economic incentives for manufacturers (six months' extension of market exclusivity) on any drug for which the FDA requested paediatric studies (see www.fda.gov/cder/cancer for further information on regulatory aspects of paediatric oncology drug development).

In addition to ethical and regulatory issues which impact the conduct of paediatric trials, there are also practical problems associated with clinical cancer research in children. Due to an understandably greater concern for long-term adverse consequences of treatment in a population of patients, the majority of whom are likely to be cured and alive for decades at risk for late effects, it is absolutely essential that long-term follow-up and serial surveillance of survivors is built into the studies. While follow-up is essential, it is also exceedingly difficult and expensive to maintain, as children and adolescents grow up, go away to school, leave home, marry, change name, etc. The frequency and severity of late effects also tends to progress with time off treatment, making follow-up beyond 15 or 20 or 30 years critical and identification of risk factors for the development of these late consequences of treatment essential. For example, Lipshultz *et al.*[71] studied 120 survivors of childhood ALL or osteogenic sarcoma who had been treated with doxorubicin a mean of 8.1 years earlier (range 2–14 years) and compared their cardiac function with a control population, and evaluated the impact of gender, age at diagnosis, length of time since completion of therapy, and dosage and cumulative dose of doxorubicin on cardiac status. Calculating sex-specific standardised scores, or z scores (expressed as the number of standard deviations above or below the value for the normal controls), for cardiac contractility, wall thickness and afterload, the results of univariate and multivariate analysis showed that female sex and higher cumulative dose of doxorubicin were associated with depressed contractility, that there was an association between younger age at diagnosis and reduced left ventricular wall thickness and increased afterload, and that the prevalence and severity of abnormalities increased with longer follow-up.[71] Such studies typify the challenge of methodologic and statistical issues in the study of late effects of childhood cancer, the greatest challenge being data collection.

A LOOK INTO THE FUTURE OF CHILDHOOD CANCER RESEARCH

As cancer treatments become more successful there will be an increasing number of survivors. For these children who attain long-term survival, research is needed to reduce the sequelae of treatment and to enhance long-term quality of life. The same chemotherapy, radiation and surgery that are needed to achieve cure can cause long-term adverse effects. The establishment of the Late Effects Committee of COG was the formal attempt to critically analyse the potential effects of therapy in an organised manner; however, there are a multitude of challenges for researchers studying late effects. Foremost among these is the need to follow large numbers of survivors with relatively uniform treatment for many years from their initial cancer treatment to make meaningful inferences about the relationship between specific treatments and long-term adverse effects. Another challenge to researchers is the ability to maintain contact with survivors as they transition from childhood to adult medical care in our very mobile society and as they no longer return to the treatment facility at which they received their initial cancer treatment. Both of these factors result in high percentages of patients 'lost-to-follow-up' by the initial treating children's hospital over time. These challenges support the need for a research infrastructure that combines the patient base of many paediatric cancer centres and that has the capability to perform active follow-up of survivors without regard to their current residence or current site of medical care[72]. COG investigators are in the process of proposing to establish a childhood cancer Long-term Follow-up Center (LTFC) which will perform long-term follow-up for all patients seen at over 200 COG centres in the United States. The LFTC will perform annual follow-up with COG patients, utilising a variety of techniques to maintain currency of contact information and will employ other available resources to re-establish contact with patients (or their parents) for whom contact has been lost.

One would hope that future therapies for childhood cancer will be developed which would be more rational, less empirical and less toxic, relying more on strategies for growth control (e.g. anti-angiogenesis) and regulation of gene expression and cell proliferation, and/or induction of apoptotic pathways or blocking of anti-apoptotic signals, than on cytotoxic or ablative treatments. Assuming that deregulated and/or mutated cellular proto-oncogenes or loss of tumour suppressor genes are the proximate cause(s) of most forms of childhood cancer, then the genes and/or their protein products will very likely be the targets for the next generation of paediatric anticancer agents, many of which will likely be orphan drugs for orphan diseases.

With advances in translational research, the pie (universe of childhood cancer patients) will be divided into smaller but more homogeneous slices than ever before. International collaboration will probably be required in a substantial segment of cancer types in order to obtain sufficient patient numbers to conduct randomised trials. Enlightened partnerships between industry and academia, with the assistance of the FDA and NCI, will be needed for efficient development of new agents.

Finally, the skill sets necessary to conduct paediatric cancer research are expanding. Traditionally the field involved paediatric haematologist/oncologists, surgeons, radiation oncologists, pathologists, nurses, clinical research associates, pharmacologists, epidemiologists and biostatisticians. Today, diagnostic imagers, bench scientists, geneticists, pharmacists, clinical psychologists, health economists and others also play significant roles in the research. In the future, other fields of expertise will surely need to be added to the team. The cooperation of a multidisciplinary team and prompt referral of patients to paediatric cancer centres participating in clinical trials will be critical to achieving future goals of refining and improving therapy.

ACKNOWLEDGEMENTS

The authors wish to thank Tianni Zhou, Meenakshi Devidas, James Anderson, and Wendy London for assistance in updating this chapter.

REFERENCES

1. National Cancer Institute, SEER Program. *Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975–1995*. NIH Publ. No. 99–4649. Bethesda, MD: National Cancer Institute (1999).
2. Ross JA, Severson RK, Pollock BH, Robison LL. Childhood cancer in the United States. A geographic analysis of cases from the Pediatric Cooperative Clinical Trials Groups. *Cancer* (1996) **77**: 201–7.
3. Woods WG, Tuchman M, Robison LL *et al.* A population-based study of the usefulness of screening for neuroblastoma. *Lancet* (1996) **348**: 1682–7.

4. Pediatric Oncology Group. Progress against childhood cancer: the Pediatric Oncology Group experience. *Pediatrics* (1992) **89**(4): 597–600.
5. Smith MA, Anderson BD, Ries LA. Continuing improvements in outcome for children with cancer in the United States, 1979–1998. *Proc Am Soc Clin Oncol* (2002) **21**: 388a.
6. Murphy SB. The national impact of clinical cooperative group trials for pediatric cancer. *Med Pediatr Oncol* (1995) **24**: 279–80.
7. McNeil De, Cote TR, Clegg L, Mauer A. SEER update of incidence and trends in pediatric malignancies: acute lymphoblastic leukemia. *Med Pediatr Oncol* (2002) **39**: 554–7.
8. Smith M, Hare ML. An overview of progress in childhood cancer survival. *J Pediatr Oncol Nurs* (2004) **21**(3): 160–4.
9. Smith M. The impact of doxorubicin dose intensity on survival of patients with Ewing's sarcoma. *J Clin Oncol* (1991) **9**: 889–91.
10. Grier HE, Krailo MD, Tarbell NJ *et al*. The addition of ifosfamide and etoposide to standard chemotherapy in Ewing's sarcoma/primitive neuroectodermal tumor of bone: a Children's Cancer Group/Pediatric Oncology Group study. *New Engl J Med* (2003) **348**: 694–701.
11. D'Angio G, Breslow N, Beckwith B *et al*. Treatment of Wilms' tumor: results of the Third National Wilms' Tumor Study. *Cancer* (1989) **64**: 349–60.
12. D'Angio G, Evans A, Breslow N *et al*. The treatment of Wilms' tumor: results of the National Wilms' Tumor Study. *Cancer* (1976) **38**: 633–46.
13. Maurer H, Beltangady M, Gehan E *et al*. The intergroup rhabdomyosarcoma study – 1. *Cancer* (1988) **61**: 1215.
14. Link M, Donaldson S, Berard C, Shuster J, Murphy SB Results of treatment of childhood localized non-Hodgkin lymphoma with combination chemotherapy with or without radiotherapy. *New Engl J Med* (1990) **322**: 1169–74.
15. Matthay KK, Villablanca JG, Seeger RC *et al*. Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-*cis* retinoic acid. *New Engl J Med* (1999) **341**: 1165–73.
16. Smith M, Arthur D, Camitta B *et al*. Uniform approach to risk classification and treatment assignment for children with acute lymphoblastic leukemia. *J Clin Oncol* (1996) **14**(1): 18–24.
17. Baker KS, Anderson JR, Link MP *et al*. Benefit of intensified therapy for patients with local or regional embryonal rhabdomyosarcoma: results from the Intergroup Rhabdomyosarcoma study IV. *J Clin Oncol* (2000) **18**: 2427–34.
18. Bostrom B, Sensel M, Sather H *et al*. Dexamethasone versus prednisone and daily oral versus weekly intravenous mercaptopurine for patients with standard-risk acute lymphoblastic leukemia: a report from the Children's Cancer Group. *Blood* (2003) **101**: 3809–17.
19. Nachman J, Sather H, Sensel M *et al*. Augmented post-induction therapy for children with high risk acute lymphoblastic leukemia and a slow response to initial therapy. *New Engl J Med* (1998) **338**: 1663–71.
20. Packer RJ, Goldwein J, Nicholson HS. Treatment of children with medulloblastomas with reduced-dose craniospinal radiation therapy and adjuvant chemotherapy: a Children's Cancer Group study. *J Clin Oncol* (1999) **17**(7): 2127.
21. Thomas PRM, Deutsch M, Kepner JL *et al*. Low-stage medulloblastoma: final analysis of trial comparing standard-dose with reduced-dose neuraxis irradiation. *J Clin Oncol* (2000) **18**(16): 3004–11.
22. Wisoff JH, Boyett JM, Berger MS *et al*. Current neurosurgical management and the impact of the extent of resection in the treatment of malignant gliomas of childhood: a report of the Children's Cancer Group Trial No. CCG-945. *J Neurosurg* (1998) **89**: 52.
23. Woods WG, Kobrinsky N, Buckley JD, *et al*. Timed-sequential induction therapy improves postremission outcome in acute myeloid leukemia: a report from the Children's Cancer Group. *Blood* (1996) **87**: 4979–89.
24. Woods WG, Neudorf S, Gold S, Sanders J, Buckley JD, Barnard DR, Dusenbery K, DeSwarte J, Arthur D, Lange BJ, Kobrinsky NL A comparison of allogeneic bone marrow transplantation, autologous bone marrow transplantation, and aggressive chemotherapy in children with acute myeloid leukemia in remission: a report from the Children's Cancer Group. *Blood* 2001 **97**(1): 56–62.
25. Pui CH. *Childhood Leukemias*. Cambridge: Cambridge University Press (1999).
26. Donaldson SS, Link MP. Combined modality treatment with low-dose radiation and MOPP chemotherapy for children with Hodgkin disease. *J Clin Oncol* (1987) **5**(5): 742–9.
27. Murphy SB. Classification, staging, and end results of treatment of childhood non-Hodgkin lymphomas: dissimilarities from adults. *Sem Oncol* (1980) **7**: 332–9.
28. Link M, Shuster J, Donaldson S, Berard CW, Murphy SB. Treatment of children and young adults with early-stage non-Hodgkin lymphoma. *New Engl J Med* (1997) **337**(18): 1259–66.
29. Shuster JJ, Carroll AJ, Look TA *et al*. Management of cytogenetic data in multi-center leukemia trials. *Comput Methods Programs Biomed* (1993) **40**: 269–77.

30. Pui CH, Frankel LS, Carroll AJ *et al*. Clinical characteristics and treatment outcome of childhood acute lymphoblastic leukemia with the t(4;11)(q21;q23): a collaborative study of 40 cases. *Blood* (1991) **77**: 440–7.
31. Pui CH, Carroll AJ, Raimondi SC *et al*. Childhood acute lymphoblastic leukemia with the t(4;11)(q21;q23): an update. *Blood* (1994) **83**: 2384–5.
32. Heerema NA, Sather HN, Ge J *et al*. Cytogenetic studies of infant acute lymphoblastic leukemia: poor prognosis of infants with t(4;11) – a report of the Children's Cancer Group. *Leukemia* (1999) **13**: 679–86.
33. Ribeiro RC, Broniscer A, Rivera GK *et al*. Philadelphia chromosome-positive acute lymphoblastic leukemia in children: durable responses to chemotherapy associated with low initial white blood cell counts. *Leukemia* (1997) **11**: 1493–6.
34. Uckun FM, Nachman JB, Sather HN *et al*. Poor treatment outcome of Philadelphia chromosome-positive pediatric acute lymphoblastic leukemia despite intensive chemotherapy. *Leukemia Lymphoma* (1999) **33**: 101–6.
35. Arico M, Valsecchi MG, Camitta B *et al*. Outcome of treatment in children with Philadelphia chromosome-positive acute lymphoblastic leukemia. *New Engl J Med* (2000) **342**: 998–1006.
36. Crist WM, Carroll AJ, Shuster JJ *et al*. Poor prognosis of children with pre-B acute lymphoblastic leukemia is associated with the t(1;19)(q23;p13): a Pediatric Oncology Group study. *Blood* (1990) **76**: 117–22.
37. Uckun FM, Sensel MG, Sather HN *et al*. Clinical significance of translocation t(1;19) in childhood acute lymphoblastic leukemia in the context of contemporary therapies: a report from the Children's Cancer Group. *J Clin Oncol* (1998) **16**: 527–35.
38. Shurtleff SA, Buijs A, Behm FG *et al*. TEL/AML1 fusion resulting from a cryptic t(12;21) is the most common genetic lesion in pediatric ALL and defines a subgroup of patients with an excellent prognosis. *Leukemia* (1995) **9**: 1985–9.
39. Rubnitz JE, Shuster JJ, Land VJ *et al*. Case–control study suggests a favorable impact of TEL rearrangement in patients with B-lineage acute lymphoblastic leukemia treated with antimetabolite-based therapy: a Pediatric Oncology Group study. *Blood* (1997) **89**: 1143–6.
40. Raimondi SC, Shurtleff SA, Downing JR *et al*. 12p abnormalities and the TEL gene (ETV6) in childhood acute lymphoblastic leukemia. *Blood* (1997) **90**: 4559–66.
41. Trueworthy R, Shuster J, Look T *et al*. Ploidy of lymphoblasts is the strongest predictor of treatment outcome in B-progenitor cell acute lymphoblastic leukemia of childhood: a Pediatric Oncology Group study. *J Clin Oncol* (1992) **10**: 606–13.
42. Heerema NA, Sather HN, Sensel MG *et al*. Prognostic impact of trisomies of chromosomes 10, 17, and 5 among children with acute lymphoblastic leukemia and high hyperdiploidy (>50 chromosomes). *J Clin Oncol* (2000) **18**: 1876–87.
43. Harris MB, Shuster JJ, Carroll A *et al*. Trisomy of leukemic cell chromosomes 4 and 10 identifies children with B-progenitor cell acute lymphoblastic leukemia with a very low risk of treatment failure: a Pediatric Oncology Group study. *Blood* (1992) **79**: 3316–24.
44. Cave H, van der Werff ten Bosch J, Suciu S *et al*. Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia. European Organization for Research and Treatment of Cancer – Childhood Leukemia Cooperative Group. *New Engl J Med* (1998) **339**: 591–8.
45. Campana D, Pui CH. Detection of minimal residual disease in acute leukemia: methodologic advances and clinical significance. *Blood* (1995) **85**: 1416–34.
46. Evans WE, Crom WR, Abromowitch M *et al*. Clinical pharmacodynamics of high-dose methotrexate in acute lymphocytic leukemia. Identification of a relation between concentration and effect. *New Engl J Med* (1986) **314**: 471–7.
47. Koren G, Farrazini G, Sulh H *et al*. Systemic exposure to mercaptopurine as a prognostic factor in acute lymphocytic leukemia in children. *New Engl J Med* (1990) **323**: 17–21.
48. Graham ML, Shuster JJ, Kamen BA *et al*. Changes in red blood cell methotrexate pharmacology and their impact on outcome when cytarabine is infused with methotrexate in the treatment of acute lymphocytic leukemia in children: a Pediatric Oncology Group study. *Clin Cancer Res* (1996) **2**: 331–7.
49. Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* (1999) **286**: 487–91.
50. Raney RB, Anderson JR, Barr FG *et al*. Rhabdomyosarcoma and undifferentiated sarcoma in the first two decades of life: a selective review of intergroup Rhabdomyosarcoma Study Group experience and rationale for intergroup rhabdomyosarcoma study V. *J Pediatr Hematol/Oncol* (2001) **23**: 215–20.
51. Castleberry R. Biology and treatment of neuroblastoma. *Pediatr Clin North Am* (1997) **44**: 919–37.
52. Attiyeh EF, London WB, Mosse YP, Wang Q, Winter C, Khazi D, McGrady P, Seeger RC, Look AT, Shimada H, Brodeur GM, Cohn SL, Matthay KK, Maris JM Chromosome 11q deletion is

an independent marker for decreased survival probability in patients with neuroblastoma: a Children's Oncology Group study. *New Engl J Med* (2005) **353**: 2243–53.

53. London WB, Castleberry RP, Matthay KK, Look AT, Seeger B, Shimada H, Thorner P, Brodeur G, Maris JM, Reynolds CP, Cohn SL Evidence for an age cut-off greater than 365 days for neuroblastoma risk group stratification in the Children's Oncology Group. *J Clin Oncol* (2005) **23**(27): 6459–65.
54. Smith MD, Ho P, Ungerleider R *et al.* The conduct of Phase I trials in children with cancer. *J Clin Oncol* (1998) **16**: 966–78.
55. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clin Trials* (1989) **10**: 1–10.
56. Shuster JJ. Optimal two-stage designs for single arm Phase II cancer trials. *J Pharm Stat* (2002) **12**: 39–51.
57. Makuch R, Simon R. Sample size considerations for non-randomized comparative studies. *J Chronic Dis* (1980) **33**: 175–81.
58. Chang M, Shuster J, Kepner J. Group sequential designs for phase II trials. *Controlled Clin Trials* (1999) **20**: 353–64.
59. McWilliams N, Hayesm F, Green A *et al.* Cyclophosphamide/doxorubicin vs. cisplatin/teniposide in the treatment of children older than 12 months of age with disseminated neuroblastoma: a Pediatric Oncology Group randomized Phase II study. *Med Pediatr Oncol* (1995) **24**: 176–80.
60. Mehta C. *EaST (Early Stopping in Clinical Trials).* Cambridge, MA: Cytel Software Corp. (1992).
61. Weitman S, Ochoa S, Sullivan J *et al.* Pediatric Phase II cancer chemotherapy trials: a Pediatric Oncology Group study. *J Pediatr Hematol/Oncol* (1997) **19**: 187–91.
62. Armitage P. Controversies and achievements in clinical trials. *Controlled Clin Trials* (1984) **5**: 67–72.
63. Sposto R, Sather H. Determining the duration of comparative clinical trials while allowing for cure. *J Chronic Dis* (1985) **38**: 683–90.
64. O'Brien PC, Fleming TR A multiple testing procedure for clinical trials. *Biometrics* (1979) **35**: 549–56.
65. Shuster J. Power and sample size for phase III clinical trials of survival. In: Crowley J, ed, *Handbook of Statistics in Clinical Oncology*. New York: Marcel Dekker (2001). Chapter 7.
66. Jonsen AR Research involving children: recommendations of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *Pediatrics* (1978) **62**(2): 131–6.
67. Leikin SL. Minors' assent or dissent to medical treatment. *J Pediatr* (1983) **102**(2): 169–76.
68. Shalala D. Protecting research subjects – what must be done. *New Engl J Med* (2000) **343**(11): 808–10.
69. Shuster J, Urinches JP, Boyett JM Ethical issues in cooperative cancer therapy trials from a statistical viewpoint, II. Specific issues. *Am J Pediatr Hematol/Oncol* (1985) **7**(1): 64–70.
70. Hirschfeld S, Shapiro A, Dagher R *et al.* Pediatric oncology: regulatory initiatives. *The Oncologist* (2000) **5**: 441–4.
71. Lipshultz SE, Lipsitz SR, Mone SM *et al.* Female sex and higher drug dose as risk factors for late cardiotoxic effects of doxorubicin therapy for childhood cancer. *New Engl J Med* (1995) **332**(26): 1738–43.
72. Link MP, Goorin AM, Miser AW *et al.* The effect of adjuvant chemotherapy on relapse-free survival in patients with osteosarcoma of the extremity. *New Engl J Med* (1986) **314**: 1600–6.

6

Gastrointestinal Cancers

DANIEL SARGENT,[1] RICHARD GOLDBERG[2] AND PAUL LIMBURG[1]
[1]Mayo Clinic, Rochester, MN 55905, USA
[2]University of North Carolina, Chapel Hill, NC 27599, USA

INTRODUCTION

Cancers of the gastrointestinal tract account for approximately 20% of all new cancer cases in the United States, and the same proportion of cancer-related deaths. In this discussion we will use a broad definition of GI cancer, including any cancer of a digestive organ. In this definition we include cancers of the oesophagus, gastro-oesophageal junction, stomach, pancreas, gallbladder, bile duct, liver, small and large intestine, rectum and anus. Incident cancers of the oesophagus, stomach, pancreas, liver, large intestine and rectum all exceed 10 000 a year in the United States. In addition to the high prevalence and the large number of cancer sites within the GI tract, the prognosis of patients with GI cancers varies greatly. For example, patients with cancers of the large intestine, when discovered early in the course of disease, have five-year survival rates exceeding 90%. In contrast, the prognosis for patients with pancreatic cancer is very poor, with a five-year survival rate of less than 5% across all stages.

Incidence rates for GI cancers show a similar diversity. In the past 50 years, the incidence rates for liver and gastric cancers in the United States have fallen substantially. For example, in 1930, gastric cancer was the most common cancer diagnosis. By 1994, gastric cancer had fallen to 12th in incidence among cancers. In contrast, the rates of colon and rectal cancer have remained very stable. Incidence rates for GI cancers also vary greatly worldwide: gastric cancer is tenfold more prevalent in Asia than in the United States.

One common feature in all GI cancers is the prognostic importance of staging. The TNM system has been widely adopted to describe the patient's disease status at the time of detection, and has great relevance to the choice of therapy and eventual outcome in all GI cancers. The importance of early detection is clear, and some GI cancers are sufficiently frequent and amenable to detection to allow cost-effective screening.

In this chapter we will review, for the major sites of the GI tract, the important clinical trials that have been conducted. Whenever possible, we will highlight the methodological and design issues of these trials, in an effort to provide insight into their results. We will describe, through this review, how the current standard

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

treatments in each disease site have evolved, as well as presenting some of the most pressing issues for future research.

OESOPHAGEAL CANCER

Oesophageal cancer is an area where controversy as to the appropriate and optimal therapy is pervasive. In patients with localised disease (Stage I–III), the roles of surgery, radiotherapy and chemotherapy, alone or in combination, have been both advocated and questioned. In advanced disease, it seems clear that chemotherapy regimes have provided some degree of progress, albeit limited. The natural history of oesophageal cancer in developed countries has shifted from squamous histology which is declining in incidence to a preponderance of adenocarcinomas, a disease with a rapidly increasing incidence rate. Reasons for this shift remain obscure but may have a relationship to dietary shifts, increasing use of refrigeration, and increasing average body mass indices in the areas where these shifts are observed.

LOCALISED DISEASE

In the past two decades, a large number of randomised clinical trials, involving thousands of patients, have investigated the contributions of radiotherapy or chemotherapy, alone and in combination, in the pre-operative and post-operative settings versus definitive therapy without surgery. Pre-operative radiotherapy, as a single modality, has been shown in two relatively small randomised trials to provide no additional benefit compared with surgery. These two trials, reported by Launois *et al.*[1] and Gignoux *et al.*,[2] randomised 124 and 208 patients, respectively. Chemotherapy as a single modality added to surgery was investigated in 440 patients by Kelsen *et al.*[3] and shown to have no advantage over surgery alone. The larger sample size of this study lends credence to this result. The two modalities have also been compared with each other as single agents,[4] and no differences in patient outcomes were observed. Based on these results it seems clear that single modality neo-adjuvant or adjuvant therapy has limited if any impact on patient outcome.

Recently, interest has focused on combined radiochemotherapy regimens in the pre-operative setting. The results in this regard have been conflicting. Four studies have been conducted, three with negative results and one with a positive conclusion. Bosset *et al.*[5] randomised 297 patients to pre-operative chemoradiation followed by surgery versus surgery alone, and found no evidence of a difference in overall survival (a relative risk for survival between the two arms of 1.0), though they did observe an advantage in disease-free survival in the treated group. In smaller trials, Le Prise *et al.*[6] and Urba *et al.*[7] reached the same conclusion based on 86 and 100 randomised patients, respectively. In contrast, Walsh *et al.*,[8] in a trial of 113 patients, found a striking survival advantage for the combined modality pre-operative approach, with a median survival of 16 months in the multimodality arm compared with 11 months in the surgery alone arm ($p = 0.01$). However, the Walsh study has been criticised for several factors, including the small sample size, poorer than expected survival for the surgery-alone control group, and the fact that the study was stopped early at an unplanned interim analysis. In an effort to resolve this controversy, a large multicentre randomised trial was mounted in the United States, with an accrual goal of 500 patients. Unfortunately, accrual to the trial was very slow, and the trial was closed early, far short of its accrual goal. Currently, the combined modality pre-operative approach has been widely adopted, despite the conflicting evidence of benefit.

Additional controversy exists in this setting as to whether surgery itself is beneficial. The Radiation Therapy Oncology Group (RTOG) has conducted two randomised trials that have not included surgery as part of the treatment. Herskovic *et al.*[9] randomised 129 patients to radiation alone versus combined chemoradiotherapy. The study was stopped early (planned sample size of 150 patients) when the first planned interim analysis showed a significant survival advantage to the combined modality group. The RTOG then followed that study with a study comparing two doses of radiotherapy, both combined with chemotherapy.[10] This study was also stopped early, in this case due to a lack of any additional benefit in the high-dose radiation arm. No trials to date have compared a surgical approach with a non-surgical approach; such a trial would scientifically be highly desirable but its practical feasibility is questionable.

Based on these results, it is clear that there is no consensus as to a 'standard of care' for patients with localised oesophageal cancer, and that there is a great need for additional clinical trials. Historically, trials in this setting have tended to be small and underpowered for detecting moderate effects on outcome. Larger, more definitive trials should be conducted.

ADVANCED DISEASE

Trials in advanced oesophageal cancer have been plentiful, though attention in this setting has focused more on Phase II trials than randomised Phase III trials. A multitude of agents have been investigated, alone and in combination. It is clear that progress has been made principally when patients are treated with two or three agents in combination; over the last 20 years median survival for advanced oesophageal cancer has increased from 3 months to 6–9 months or greater. The emphasis on Phase II trials, in an attempt to find a promising new approach, is certainly appropriate given the modest results available from current chemotherapies.

GASTRIC CANCER

While the incidence of gastric cancer has declined in the United States over several decades, 21 600 new diagnoses and 12 400 deaths were still expected in 2002.[11] The nearly 40% cure rate that these numbers imply likely results from a better natural history than oesophageal or pancreatic cancer, early detection via endoscopy, improvements in surgery, and the post-operative use of chemotherapy with radiation for patients with resected disease. While gastric cancer is unusual among GI primary sites because of the large number of antineoplastic agents that show some activity (as measured by tumour response rate), in the advanced disease setting even the most active combination chemotherapy regimens result in remissions that generally last for only a few months and median survivals of less than one year.

LOCALISED DISEASE

The ideal operation for gastric cancer, including the issues of limited versus total gastrectomy and extended versus more limited lymph node dissection, has been a matter of controversy. Trials done in the 1980s and 1990s led to the conclusion that the most important surgical principle is achievement, when possible, of a pathologically negative resection (an R_0 resection). However, patients have improved post-operative quality of life if some of the stomach is retained, and most surgeons resect only as much of the stomach as is needed to achieve pathologically free margins. The rich lymphatic networks of the stomach can sometimes result in apparently clear margins, yet residual intralymphatic disease may be present in 'skip areas'. This has implications regarding post-operative treatment, and suggests a potential role for adjuvant radiation to the tumour bed and regional structures.

Many surgeons, particularly those in Japan, advocate extended lymph node dissections as a means to improve outcome due to the central location of the stomach with many lymph-node-bearing areas at risk for metastatic spread. In a landmark study the Dutch Gastric Cancer Group employed a single Japanese surgeon to train participating Dutch surgeons to perform the classical Japanese extended lymphadenectomy.[12] These investigators randomised 711 eligible patients to resection of the primary tumour with clear gastric margins and either standard (D1) or extended lymphadenectomy (D2). Three-year survival rates were 56% and 58% respectively for the two cohorts, suggesting no advantage to more aggressive surgery. The British Medical Research Council conducted a similar, albeit smaller (400 patients), trial that confirmed this finding.[13]

The adjuvant therapy of gastric cancer, mainly using 5-FU-based regimens, has been a matter of investigation for many years. Many randomised trials of chemotherapy versus surgery alone have been reported and these individual trials have generally been negative. A meta-analysis of 21 randomised controlled trials conducted worldwide which included 3962 patients with 1840 allocated to surgery alone and 2122 allocated to adjuvant chemotherapy, did show a modest potential benefit for treatment.[14] The odds ratio (OR) in favour of chemotherapy was 0.84 overall, but the principal benefit was confined to patients enrolled in trials done in Asia ($n = 888$ patients, OR 0.58) as opposed to Western patients ($n = 3074$, OR 0.96). This finding lends some support to the possibility of a geographically or ethnicity-based difference in the natural history of this disease, a finding supported by some epidemiologic evidence. It is also plausible that early stages of gastric cancer that are detected through screening programmes commonly employed in Japan may call for different management strategies than more advanced localised disease. Studies of post-operative radiation versus surgery alone have not shown any advantages, although interpretation of the limited data addressing this issue is problematic.

Adjuvant radiation and chemotherapy used in combination has recently been shown to be advantageous in North American patients. In a 603-patient study, patients were randomised to either surgery alone or to surgery followed by combined modality therapy.[15] In the treatment arm patients were given 5-FU plus leucovorin before and after 4500 cGy to the gastric bed (with radiosensitising 5-FU + leucovorin administered for four consecutive days at the beginning and three days near the end of the radiation). Adjuvant chemoradiotherapy led to a significant median survival advantage of 36 compared with 27 months ($p = 0.005$) and a reduction in local regional relapse rate to 67% compared with 82%. In addition to these outcome improvements, two important patterns of care findings were noted. The trial recommended but did not demand at least a D1 resection and noted that a D2 resection was preferred. However, when operative reports were analysed, only 10% of patients had D2 resections, 36% D1 resections, with the balance having less aggressive surgery. Second, pretreatment radiation field review by a single radiation oncologist indicated that 35% of submitted treatment plans contained major or minor deviations from the protocol, indicating a need for education of surgeons and radiation oncologists as to the preferred procedures in these settings. Some readers have raised the possibility that the chemotherapy and radiation were beneficial mainly because of suboptimal surgery in this cohort of patients.

In 2003 the preliminary results of the British MAGIC trial in which 503 patients were randomly assigned to three cycles of epirubicin, cisplatin and 5-FU (ECF) before and after surgery versus surgery alone were reported at the annual meeting of the American Society of Clinical Oncology. There was a statistically significant benefit to preoperative ECF chemotherapy with respect to disease-free survival (medians 12 versus 18 months, $p = 0.002$) and a borderline advantage with respect to median survival (18 versus 22 months, $p = 0.063$). No significant differences in perioperative morbidity or mortality were noted. Final results from this trial have not yet been reported. Pending the final report, the consideration of delivery of neo-adjuvant therapy in marginally resectable cases or patients with localised but unresectable disease seems warranted.[16]

ADVANCED DISEASE

Palliative therapy does make a meaningful difference for many patients with advanced gastric cancer, whose median survival with supportive care alone is around three months. One or more agents from virtually all classes of chemotherapy drugs have demonstrable activity, and median survivals approaching one year have been reported with several combination chemotherapy regimens. One example representative of modern

Phase III trials randomised patients to epirubicin, cisplatin and 5-FU (ECF) versus 5-FU, doxorubicin and methotrexate (FAMtx).[17] The overall response rate was 45% compared with 21% ($p = 0.0002$) and the overall survival was 8.9 months compared with 5.7 months ($p = 0.0009$) for ECF over FAMtx. Despite the intensive nature of these two regimens, and other combinations tested to date, the beneficial effects in terms of improved patient longevity have been modest. Earlier detection, improvements in the management of local regional disease, and the testing of new agents seem to provide the best avenues towards better outcomes.

PANCREATIC CANCER

Pancreatic cancer has a very poor prognosis. It affects approximately 27 000 new patients each year in the United States, and is fatal in approximately 95% of cases. As in all GI cancers, options for therapy include surgery, radiotherapy and chemotherapy, depending on the disease stage.

LOCALISED DISEASE

In the setting of resectable or locally advanced disease, both radiotherapy and chemotherapy, and the combination, have been tested extensively. Studies conducted prior to the mid-1990s tended to be small and underpowered, which has led to a variety of conflicting results.

In locally advanced disease, the Gastrointestinal Tumor Study Group (GITSG) randomised 227 patients to three arms: radiotherapy alone, or radiotherapy at two different dose levels given with chemotherapy (5-FU).[18] Accrual to the radiotherapy-alone arm was stopped early due to poor results. Two studies have investigated the need for chemoradiotherapy versus chemotherapy alone, with conflicting results. Klaassen *et al.*,[19] in a two-arm randomised study of 191 patients, found no advantage for combined therapy versus chemotherapy alone, while GITSG[20] reported that overall survival was improved with the addition of radiation to chemotherapy in a two-arm study of 43 patients. The small sample sizes of all these trials make definitive conclusions difficult, but there is little evidence to support a role for radiation alone in this setting.

In the setting of a complete surgical resection, several small randomised studies have suggested a benefit to post-operative chemotherapy or chemoradiotherapy. None of these trials enrolled greater than 114 patients, limiting the ability to draw conclusions. The recent report by Neoptolemos *et al.*[21] provided much more conclusive evidence in this regard. In this study, 289 eligible patients were randomised to receive post-operative chemotherapy (six months of post-surgical treatment), chemoradiotherapy (a 10-day course of radiotherapy accompanied by chemotherapy), both or neither (the design was not a true 2×2 factorial because clinicians were allowed to choose to participate in either one or both randomisations). In this study, there was no benefit to the chemoradiotherapy, while a clear benefit was observed for the chemotherapy group compared with the no-treatment group (five-year survival rates of 21% versus 8%, $p = 0.009$). This trial has been criticised for including patients with involved margins after surgery and also primaries arising in the ampulla and bile ducts. Interestingly, the authors of that study concluded that standard therapy should consist of curative surgery followed by adjuvant systemic chemotherapy, but designed their next trial to contain a no-treatment control arm with the randomisation schema of 5-FU with folinic acid or gemcitabine versus surgery alone.

ADVANCED DISEASE

Chemotherapy has been considered the standard of care in the United States for advanced pancreatic cancer, despite the lack of any randomised trial demonstrating a survival benefit for chemotherapy versus no treatment. The use of chemotherapy was justified by the occasional tumour response that was observed. Single agent therapy with 5-FU has been used as the control arm for multiple randomised trials, with the assumption that 5-FU was at worst a toxic placebo; thus if a new experimental regimen were shown superior to 5-FU, it would indeed have improved efficacy when compared with no treatment. Burris *et al.*[22] reported a Phase III randomised trial with 126 patients that showed an improved overall survival for gemcitabine compared with 5-FU alone (median survivals of 5.7 versus 4.4 months respectively, $p = 0.003$). The Burris trial established gemcitabine as a new standard of care in this setting. Ongoing and future trials will likely use gemcitabine as a base, comparing gemcitabine alone with a multi-drug chemotherapy regimen including gemcitabine.

A recently completed trial in pancreatic cancer can be used to illustrate the need for careful consideration of an agent prior to Phase III testing. Due in large part to the dismal prognosis and limited treatment options available for patients with pancreatic cancer, pressure has been applied to rapidly introducing novel agents into Phase III trials. The goal is to seek to speed the process of testing a new agent by avoiding the Phase II stage of testing. Such was the case in a randomised Phase III trial reported by Moore *et al.*,[23] where a novel agent (a matrix metalloproteinase inhibitor (MMPI)) was tested against gemcitabine in 277 patients. In this trial the MMPI had significantly inferior outcome compared with gemcitabine. The trial was carefully and appropriately designed to allow early stopping if the results were extreme, which in this case they were. A Phase II trial may have identified the lack of efficacy of this agent prior to its large-scale testing.

Multiple randomised Phase III trials of doublets of gemcitabine coupled with either another chemotherapy agent such as 5-FU or irinotecan, oxaliplatin, exatecan and pemetrexed, or a biologic such as tipifarnib (Zarnestra), compared with single agent gemcitabine have been reported. None has shown an advantage over single agent gemcitabine alone. Recently, Moore *et al.* reported in abstract form a modest but statistically significant improvement in overall survival from the addition of erlotinib, an oral reversible inhibitor of EGFR tyrosine kinase, to gemcitabine.[24] The overall survival was improved, with a hazard ratio of 0.81, $p = 0.025$, but the improvement in median overall survival was only approximately two weeks. Given this modest benefit, it is unclear if this combination will become standard of care; however, it did provide a signal that perhaps biologic agents are a more promising avenue to pursue in this disease. Ongoing trials are testing the addition of bevacuzimab or cetuximab, both monoclonal antibodies.

Clearly the results of Phase II trials must be promising to justify the commitment of patients and resources to Phase III trials in advanced pancreatic cancer. Despite the modest benefit seen with single agent gemcitabine, or the addition of erlotinib, in this setting, improving outcomes in pancreatic cancer is not easy. Advances in management of advanced disease have proven to be elusive as a consequence of its virulence and drug resistance, indicating that in pancreatic cancer research the integration of innovative therapeutic approaches is appropriate.

COLORECTAL CANCER

Colorectal cancer is the most common malignancy in the GI tract. Not surprisingly, it is also the GI cancer that has been the most extensively investigated in clinical trials.

Likely as the direct result of these intensive research efforts, considerable progress has been made in many facets of colorectal cancer, including chemoprevention, early detection and treatment.

CHEMOPREVENTION

Cancer chemoprevention can be defined as the use of nutritional or pharmaceutical agents to prevent, inhibit or reverse carcinogenesis at a pre-invasive stage of disease. Candidate agents are often identified through a combination of epidemiological and laboratory-based research. Since most subjects enrolled onto chemoprevention trials are generally healthy (except for their increased cancer risk), minimal toxicity represents an important criterion for selecting candidate agents. Colorectal adenomas are commonly employed as intermediate endpoint biomarkers to facilitate more rapid completion of colorectal cancer chemoprevention trials. To date, several colorectal cancer chemoprevention agents have been investigated, including fibre, antioxidant vitamins, calcium, nonsteroidal anti-inflammatory drugs (NSAIDs) and selective cyclooxygenase-2 enzyme inhibitors.

Dietary fibre represents a heterogeneous mixture of complex materials derived primarily from plant cell walls. Extensive observational data collected over more than three decades suggest that fibre might help to prevent colorectal neoplasia by diluting or adsorbing faecal carcinogens, reducing colonic transit time, altering bile acid metabolism, or increasing short-chain fatty acid production. However, high-fibre interventions have not been associated with a reduced risk for recurrent colorectal adenomas in five clinical trials.[25–29] In fact, one small randomised study observed a higher adenoma recurrence rate among subjects in the active fibre intervention group.[27] It remains possible that administration of dietary fibre at an earlier point in tumourigenesis (e.g. prior to first adenoma formation) might have a more appreciable anticarcinogenic effect. Nonetheless, the existing data do not support a major role for this agent in colorectal cancer chemoprevention.

Antioxidant vitamins such as the retinoids, carotenoids, ascorbic acid and alpha-tocopherol may prevent carcinogen formation by neutralising free radicals within the intestinal lumen. Although somewhat inconsistent, the preponderance of data from case–control and cohort studies supports an inverse association between antioxidant vitamin intake and colorectal cancer risk. Four colorectal cancer chemoprevention trials have investigated antioxidant vitamins at different doses and in various combinations. One relatively small study found that recurrent adenomas were less common among subjects treated with vitamin A (30 000 IU per day), vitamin C (1 g per day) and vitamin E (70 mg per day) compared with placebo over a mean intervention period of 17.8 months (6% versus 36%, respectively).[30] However, based on preliminary data, these investigators observed less striking effects from lower vitamin doses in a follow-up trial. Another three-year chemoprevention trial reported a 69% reduction in the number of recurrent colorectal polyps among subjects randomised to receive multiple antioxidants (beta-carotene, selenium, vitamin C, vitamin E) plus calcium versus placebo compounds.[31] In contrast, the Toronto Polyp Prevention Group ($n = 200$) and the Antioxidant Polyp Prevention Study ($n = 864$) found no significant chemopreventive benefits from vitamin C and vitamin E.[32,33] Of potential interest, stratified analyses from the latter trial did show that beta-carotene was associated with a 44% reduction in recurrent adenoma risk (RR = 0.56; 95% CI = 0.35–0.89) among subjects who were both non-smokers and non-drinkers.[34] However, secondary analyses from another large intervention trial of Finnish male smokers (the Alpha Tocopherol, Beta Carotene Cancer Prevention Study) reported no significant effect on CRC incidence from 5–8 years of beta-carotene (RR = 1.05; 95% CI = 0.75–1.47) or vitamin E (RR = 0.78; 95% CI = 0.55 − 1.09).[35] Thus, definitive evidence for a protective benefit from antioxidant vitamins on colorectal cancer risk remains to be demonstrated.

Calcium may serve as a colorectal cancer chemoprevention agent through at least two mechanisms: functionally removing toxic bile acids from the faecal stream and decreasing cellular proliferation in the large bowel mucosa. Data compiled from 24 observational studies yielded a summary risk estimate of 0.86 (95% CI = 0.74 − 0.98) for colorectal cancer among subjects with high versus low calcium intakes.[36] With respect to clinical trials, calcium 1600 mg per day (along with vitamins C and E, selenium and beta-carotene) was not associated with an appreciable difference in adenoma growth rate compared with placebo after 36 months of treatment among polyp-bearing subjects.[37] In contrast, the large Calcium Polyp Prevention Study reported that

48 months of calcium carbonate 3000 mg per day yielded a 15% reduction in recurrent adenoma risk (RR = 0.85; 95% CI = 0.74 − 0.98).[38] Also, the European Cancer Prevention Organization Study Group observed a decline in adenoma recurrence rate after three years of elemental calcium 2000 mg per day among post-polypectomy subjects, but this result was not statistically significant (RR = 0.66; 95% CI = 0.38 − 1.17).[39] Further data regarding the chemopreventive potential of calcium (and vitamin D) are being collected from the large Women's Health Initiative Clinical Trial,[40] which should help to clarify whether or not application of this agent to average-risk subjects has measurable value.

NSAIDs are a structurally diverse class of pharmaceutical agents that appear to reduce proliferation, delay cell cycle progression, and induce apoptosis in epithelially lined tissues. Extensive data from rodent models suggest that NSAID administration can reduce GI tumour incidence and/or multiplicity by up to 80%. In human populations, regular NSAID use has been associated with decreased colorectal cancer risk in numerous observational studies. NSAID chemoprevention of sporadic colorectal neoplasia has also been investigated in four randomised, controlled clinical trials. Among subjects with a history of curatively resected CRC, aspirin 325 mg per day was associated with a significantly lower adenoma recurrence rate compared with placebo (17% versus 27%; $p = 0.004$) after a median intervention period of 31 months.[41] However, in another study of subjects with prior colorectal adenomas, the same aspirin dose had no statistically significant effect on recurrent adenomas after a similar follow-up period.[42] Interestingly, a lower aspirin dose of 81 mg per day was associated with a 17% decrease in recurrent adenoma risk in this study (RR = 0.8; 95% CI = 0.7 − 1.0). The reasons for this counter-intuitive effect remain incompletely defined. Interim data from a third aspirin intervention trial, which used yet another dose of 300 mg per day, showed a borderline significant effect on recurrent adenoma risk (RR = 0.61; 95% CI = 0.37 − 0.99) after one year of follow-up among subjects with prior adenomas.[43] Lastly, the Physicians' Health Study ($n = 22\,071$ subjects) reported a null association between aspirin 325 mg per day and incident colorectal cancer based on secondary data analyses after both 5 years (RR = 1.15; 95% CI = 0.80 − 1.65)[44] and 12 years (RR = 1.03; 95% CI = 0.83 − 1.28)[45] of follow-up. Of note, certain limitations of the Physicians' Health Study trial design, such as the relatively low aspirin dose and lack of uniform colorectal cancer surveillance guidelines, may have hindered its ability to detect a protective association.

The chemopreventive effects of traditional NSAIDs are thought to result primarily from inhibition of cyclooxygenase-2 (COX-2). Selective COX-2 inhibitors like celecoxib and rofecoxib have been evaluated as potential colorectal cancer prevention agents. In the first trial to be reported, celecoxib 400 mg twice per day was associated with statistically significant reductions in both the mean number and total burden of colorectal polyps among subjects with familial adenomatous polyposis.[46] Final results from other ongoing and recently completed COX-2 inhibitor trials will help to further clarify the effect of these agents on sporadic colorectal neoplasia.

Although existing clinical trial data are limited, oestrogen also appears to modulate colorectal carcinogenesis. In the Women's Health Initiative clinical trial group, subjects who received equine estrogens 0.625 mg per day and medroxyprogesterone acetate 2.5 mg per day developed fewer incident colorectal cancers than subjects randomly assigned to placebo (43 versus 72 cases, respectively; $p = 0.003$).[47] However, more advanced stage colorectal cancers were diagnosed among subjects in the active intervention arm. A number of other candidate agents, including ursodeoxycholic acid, difluoromethylornithine and Bowman–Birk inhibitor, have shown promising results in cell culture experiments, animal model systems and/or observational studies. Further data regarding these (and other) potential colorectal cancer chemopreventive agents are anticipated in the near future as new Phase I, II and III clinical trials are organised and completed.

EARLY DETECTION

Due to a variety of factors, colorectal cancers are very amenable to early detection. First, the biology of colorectal carcinogenesis is becoming increasingly well understood, as evidenced by continued expansion of knowledge regarding the molecular events associated with different stages in the adenoma–carcinoma sequence. This relatively slow process typically requires several years to progress from normal mucosa to advanced neoplasia, which affords a clear opportunity for detecting lesions at an asymptomatic stage. Second, there are a variety of possible screening methods that range from non-invasive stool tests or imaging studies to invasive endoscopic evaluations. Third, due to the high incidence of colon cancer, such screening may be cost-effective in terms of screening costs versus years of life saved. Fourth, the high incidence of colon cancer provides a motivation for many individuals to seek out screening. Based on these and other considerations, several randomised trials of various screening methods have been conducted.

With respect to faecal occult blood testing, three large clinical trials have shown that regular screening may reduce colorectal cancer mortality by 13–33%.[48–50] In two trials from Europe, subjects ($n = 61\,933$ and $n = 150\,251$) were randomised to undergo screening every other year versus usual care. In the Minnesota Colon Cancer Study, subjects ($n = 46\,551$) were randomised to annual screening, biennial screening or usual care. Follow-up in these studies ranged from 11 to 18 years. Interestingly, only one trial found that programmatic screening was associated with a statistically significant reduction in colorectal cancer incidence.[50] These data suggest that pre-invasive adenomas (arguably the most relevant screening target) are poorly detected by faecal occult blood testing. Thus, despite the inclusion of faecal occult blood testing in widely endorsed colorectal cancer screening guidelines,[51,52] further pursuit of more sensitive and specific stool biomarkers is needed.

Direct examination of the distal colorectum by flexible sigmoidoscopy represents another option for colorectal cancer screening. However, this procedure is at least moderately invasive and may be associated with transient discomfort. As such, adherence to recommendations for initial and repeat flexible sigmoidoscopies was recently evaluated in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Among subjects randomised to the screening intervention arm ($n = 17\,713$), 83% completed the baseline flexible sigmoidoscopy. Additionally, 87% of subjects who were eligible for repeat testing after three years complied with the follow-up evaluation.[53] At present, the effects of flexible sigmoidoscopy screening on colorectal cancer incidence in the PLCO trial cohort remain unknown. An even larger flexible sigmoidoscopy screening trial is underway at 14 centres in the United Kingdom ($n = 170\,432$ randomised subjects).[54] When available, data from these two trials should be highly informative regarding the utility of flexible sigmoidoscopy screening to reduce colorectal cancer incidence rates in the general population.

Colonoscopy is currently the gold standard for structural evaluation of the large intestine. Cost-effectiveness models suggest that one-time screening colonoscopy between ages 50 and 54 years may be a rationale colorectal cancer prevention approach.[55] Existing early detection guidelines support a slightly more conservative strategy (i.e. colonoscopy every 10 years, in the absence of symptoms or other known risk factors). However, screening colonoscopy has not yet been investigated in a randomised clinical trial, with the exception of one ongoing feasibility study.[56] Two novel methods of colorectal cancer screening, CT colonography and DNA-based stool assays, are currently being tested in population-based clinical trials as well. Results from these studies are anticipated in the near future and may necessitate further modification of current early detection algorithms.

TREATMENT: COLON CANCER

Localised Disease

Surgery is the primary modality for the treatment of localised colon cancer. Depending on disease stage, surgery alone produces five-year survival rates of 50% to greater than 90%. As opposed to gastric and rectal cancer, however, the surgical technique for colon cancer resection has been the subject of limited investigation in randomised clinical trials. One large surgical trial randomised patients to either the standard 'open' colectomy or a 'minimally invasive' laparoscopically assisted colectomy.[57] The trial's primary endpoint was cancer recurrence, and it demonstrated equivalence of the two approaches. The trial also included extensive quality of life and cost-effectiveness assessments that were favourable for the laparoscopic technique.

The value of adjuvant treatment for patients with Stage III colon cancer (cancer able to be completely resected, but with positive lymph nodes in the resection specimen) is well accepted. The first trial conducted with a positive result was conducted by the North Central Treatment group, initiated in 1978.[58] This was a three-arm trial, with a sample size of approximately 135 patients per arm, comparing no post-surgical treatment with adjuvant treatment with either levamisole alone or 5-FU plus levamisole. The initial results of this trial indicated a moderate but statistically significant benefit for the 5-FU plus levamisole arm compared with control. Given the novelty of this result, in a decision that likely would never be made in the current day, the investigative team decided to embark on a larger, confirmatory trial prior to the release of the results to the oncology community. This confirmatory trial, known as Intergroup trial 0035, enrolled over 1200 patients to the same three arms as the initial trial. Intergroup 0035 clearly demonstrated improved overall survival in patients treated with adjuvant 5-FU and levamisole.[59] These findings led in part to the 1990 consensus statement from the National Cancer Institute that patients with Stage III colon cancer who are unable to enter a clinical trial should be offered adjuvant treatment with 5-FU plus levamisole unless there are contraindications.[60]

A number of clinical trials were in progress at the time of the publication of the beneficial results from the use of adjuvant 5-FU plus levamisole. Several of these trials included a no post-surgical treatment control arm, and thus these trials were closed prior to reaching their accrual goals due to ethical reasons, including five Phase III randomised trials testing 5-FU plus leucovorin versus no post-surgical treatment control. The results from three of these trials were pooled for analysis;[61] the other two were reported separately.[62,63] In each of these analyses, adjuvant 5-FU plus leucovorin showed a survival advantage compared with control. In subsequent studies, throughout the 1990s, various investigative groups conducted trials comparing various different schedules and combinations of 5-FU combined with either leucovorin or levamisole. None of these trials demonstrated a statistically significant improvement in survival between study arms, although through such trials it did become clear that 6 months of 5-FU plus leucovorin was at least as effective as 12 months of 5-FU plus levamisole.[64,65]

In the discussion in the preceding paragraph, all of the regimens discussed were based on the delivery of 5-FU as a short-term bolus infusion. Based on promising results in the advanced disease setting (as discussed below), multiple clinical trials have been conducted using regimens based on a long-term infusion with 5-FU. Intergroup trial 0153 directly compared a bolus with an infusional 5-FU-based regimen in a randomised Phase III trial of 1078 patients (terminated early at an interim analysis–original planned sample size of 1800 patients).[66] In this trial no difference in efficacy was observed between the arms, although the toxicity profile did differ substantially. Based on these results, two recent Phase III randomised trials in the United States have used control arms of six months of bolus 5-FU plus leucovorin. However, in Europe, regimens using short-term 24–48 hour 5-FU infusions are more popular.

Current efforts in the adjuvant treatment of Stage III colon cancer are directed towards improving the treatment options available in terms of both convenience (using oral therapy) and efficacy. New studies have randomised patients to treatments based on adding a new treatment to a 5-FU and leucovorin regimen. Four large Phase III trials have been completed comparing 5-FU and leucovorin with either a 5-FU, leucovorin and irinotecan regimen (trials C89803 and PETACC-3) or 5-FU, leucovorin and oxaliplatin (trials C-07 and MOSAIC). Results are now available for each of these trials. In the C89803 and PETACC-3 trials, no advantage in disease-free or overall survival was noted for the addition of irinotecan to bolus 5-FU and leucovorin.[67,68] In contrast, the MOSAIC and C-07 trials reported a statistically significant 5% increase in three-year disease-free survival (those trials' primary endpoint) for the addition of oxaliplatin.[69,70] These differing results suggest a greater activity of oxaliplatin compared with irinotecan in this setting, as oxaliplatin improved outcome when added to both an infusional (MOSAIC) and a bolus (C-07) 5-FU regimen, while irinotecan did not improve outcome when added to either method of 5-FU delivery (bolus 5-FU in the C89803 trial, infusional in the PETACC-3 trial). On the basis of these results, Oxaliplatin has been approved around the world for the adjuvant treatment of Stage III colon cancer, and the oxaliplatin and 5-FU/leucovorin combination is being used as the control arm for the current generation of adjuvant colon cancer trials, which test whether the addition of a biologic agent to chemotherapy can further improve outcomes (cetuximab in one trial, N0147, and bevicumab in the other, C-08).

Two large studies of oral therapy were recently reported. The first, trial C-06 from the NSABP, randomised 1608 patients to an oral regimen of UFT (uracil, oral folinic acid and tegafur) or bolus 5-FU/LV.[71] The X-ACT trial randomised 1987 patients to oral capecitabine versus bolus 5-FU/LV.[72] Both of these trials concluded an equivalent efficacy of the oral to the bolus regimens. Based on these data, coupled with the efficacy data from the MOSAIC trial discussed above, one may conclude that bolus 5-FU-based regimens will be of limited use in future clinical practice. Patients seeking maximally aggressive therapy will receive a combination chemotherapy regimen with 5-FU/LV and oxaliplatin, those seeking less aggressive therapy will opt for an oral regimen.

One additional insight into the conduct of clinical trials in GI cancers may be gained by examining the steady increase in the sample sizes that has occurred in Stage III colon clinical trials over the past two decades. In trials conduced in the early 1980s, sample sizes of 100–200 per arm were typical,[58,62] with some exceptions (such as the NSABP C-01 trial, with approximately 380 patients per arm).[73] With such a sample size, the study provided adequate power to detect only a relatively large effect. Fortunately, 5-FU, when combined with either levamisole or leucovorin, did provide a rather large effect, with a reduction in the hazard of death by approximately 25%.[74] However, the likelihood of a subsequent treatment advance of such a magnitude is unlikely, and smaller advances may indeed be clinically relevant. Therefore, more modern trials in Stage III disease have included sample sizes of 1600 (trial C89803),[67] 2400 (trial C-07) and 4900 patients for a four-arm trial (the QUASAR trial).[64] As therapy continues to improve, the sample size necessary to detect further incremental advances will continue to grow.

As opposed to the adjuvant treatment of Stage III disease, the benefit of adjuvant treatment for Stage II (node negative) disease is less clear. In many previous trials, patients with Stage II disease have been pooled together with Stage III patients. The sample size for such trials has typically been based on an analysis pooling the data from both patient groups. For a variety of reasons, patients with Stage III disease have typically constituted a majority of the enrolment to such trials; thus each individual trial has been underpowered to detect a moderate benefit of treatment in Stage II patients. Due to the limited sample size in each trial, two attempts have been made to pool data from several trials in order to gain a sufficient sample size to draw a definitive conclusion regarding the value of adjuvant therapy in Stage II disease. However, the two analyses have reported differing conclusions. One analysis, reported by Mamounas *et al.*,[75] pooled data from four randomised trials conducted by the NSABP. In none of these four trials was there a direct randomised comparison between treatment and control. In their analysis, the authors estimated the magnitude of the difference in outcome between the two study arms in each of the four studies. They then compared whether this difference in outcome differed by patient stage. The authors concluded that the treatment effect within each study was similar between Stage II and Stage III patients, and since it had been previously demonstrated that treatment is beneficial in Stage III patients, they concluded that treatment is also beneficial in Stage II patients.

The second investigation[76] used a more direct approach. In this analysis, the study team pooled the data from Stage II patients who had participated in five randomised trials of 5-FU plus leucovorin versus control. They found no statistically significant benefit of treatment, based on a pooled sample size of just over 1000 patients. Due to the excellent outcome of Stage II patients, with an approximately 80% five-year survival in untreated patients, even this pooled sample size had poor power to detect a small but possibly important improvement in patient outcome (only 60% power to detect an 85% five-year survival in treated patients). A large randomised trial of a monoclonal antibody in the setting of Stage II disease, with accrual of over 1700 patients (trial C9581), demonstrated no benefit to the antibody.[77]

Three recently reported studies have helped to greatly clarify the issue of optimal treatment for Stage II patients. The first, the QUASAR trial, reported at ASCO 2004, found a statistically significant 3% improvement in overall survival for 5-FU-based treatment compared with no post-surgical chemotherapy in a study of 3238 patients, the vast majority of which were Stage II.[78] The second was the MOSAIC trial, in which 40% of patients were Stage II. In this trial, no significant stage by treatment interaction was observed, and the benefit of oxaliplatin added to 5-FU/LV compared with 5-FU/LV alone in Stage II patients in terms of three year disease-free survival was 3%.[68] Finally, a pooled analysis of seven randomised trials with over 3300 total patients[79] demonstrated a statistically significant disease-free survival advantage for 5-FU-based therapy compared with control separately in both Stage II and Stage III patients. Taken together, these findings are actually very consistent with the previous analyses, and indicate that there is a small but real benefit to treating Stage II patients. The magnitude of the benefit (2–3% for 5-FU/LV, likely 4–5% for oxaliplatin + 5-FU/LV compared with no post-surgical chemotherapy) must be balanced against the risks and the side effects in making a decision to treat or not to treat each patient with Stage II disease.

Advanced Disease

It is likely that more clinical trials have been conducted in advanced colon cancer than in any other GI disease site. This is due to the high incidence of the cancer, and the fact that it is at least to some degree sensitive to chemotherapeutic agents. Trials in advanced colon cancer typically include patients with advanced rectal cancer, as the response to chemotherapy has not been shown to depend on the precise site of the patient's disease within the colorectum.

The drug 5-FU has been the mainstay of treatment for colorectal cancer for over 40 years. From 1950 to 1990, a multitude of trials were conducted in an effort to improve the efficacy of 5-FU-based regimens, by changing methods of administration, combining it with various supplemental agents (such as leucovorin or levamisole), or changing the dose and schedule. Regarding the timing of administration, the clear result of multiple studies is that, among regimens where 5-FU is delivered as a bolus injection, the particulars of the administration have a definite impact on toxicity, some impact on tumour response, but little impact on patient survival. The addition of leucovorin to 5-FU has been demonstrated in a meta-analysis to provide increased efficacy in terms of response rate compared with 5-FU alone.[80] In another meta-analysis, a schedule where the 5-FU is delivered by a continuous infusion has been shown to provide an advantage in both toxicity and overall survival compared with bolus schedules.[81,82] However, the improvement in median survival was modest at 0.8 months; thus many practitioners (at least in the United States) have continued to administer the bolus 5-FU-based regimens based on perceived benefits of patient and physician convenience.

After 40 years of testing variations on a 5-FU theme, multiple recent developments have added excitement to the advanced colorectal cancer clinical trials arena. The first is the introduction of oral 5-FU-based regimens. The oral method of delivery offers clear benefits in terms of patient preference. However, an oral approach would not likely be accepted if it did not provide at least equivalent efficacy to an IV approach. Therefore, two large equivalence trials have been conducted comparing an oral with an IV regimen. These two trials, one reported by Hoff *et al.*[83] and the other by van Cutsem *et al.*,[84] enrolled 605 and 602 patients respectively, and were formally designed to test the equivalence of the oral regimen to the IV approach. In both cases, formal equivalence was declared.

At almost the same time as the introduction of oral 5-FU-based agents for advanced colorectal cancer, new chemotherapeutic agents have been added to 5-FU with promising results. Based on results with the agent irinotecan in patients who had failed a 5-FU-based regimen,[85,86] trials with irinotecan were performed in the setting of patients with previously untreated advanced disease. As reported by Saltz *et al.*[87] and Douillard *et al.*,[88] irinotecan, when added to 5-FU and leucovorin, resulted in improved time to progression and overall survival when compared with 5-FU and leucovorin alone in first-line treatment of advanced disease. These two relatively large trials (683 and 387 patients, respectively) established a new standard of care in this setting. In the Saltz trial irinotecan was added to a bolus 5-FU schedule, while the Douillard trial added irinotecan to an infusional 5-FU regimen; thus the optimal method in which to give 5-FU with the new agent remained unclear.

Recently, the drug oxaliplatin has shown promising activity when combined with 5-FU and leucovorin in several studies,[89,90] with reported median survivals equalling or exceeding 18 months. In particular, trial N9741[90] tested the combination of oxaliplatin and infusional 5-FU/leucovorin, in a regimen known as FOLFOX4, against a control arm of irinotecan and bolus 5-FU/leucovorin, which had become standard of care based on the Saltz trial.[87] In this trial of 795 patients, FOLFOX4 was associated with a 4.5 month improvement in median survival, from 15.0 to 19.5 months, which was highly statistically significant. In addition the FOLFOX4 regimen had significantly increased time to tumour progression and response rate, and a lower 60-day all-cause mortality rate. This trial led to FDA approval of FOLFOX in the first-line setting.

In the N9741 trial,[90] approximately 60% of patients received treatment with irinotecan after disease progression or toxicity with the FOLFOX4 regimen, suggesting that sequential therapy with multiple agents may play an important role in extending survival. This supposition is supported by the recently reported trial by Tournigand *et al.*[91] where patients received either irinotecan or oxaliplatin in addition to infusional 5-FU/leucovorin, with planned crossover to the other agent after the failure of the first treatment. No difference was observed in outcome between the two treatment strategies, but a median survival of 21 months was achieved in both arms of the trial. The finding that access to all three agents (5-FU, irinotecan and oxaliplatin) improves survival was also supported by a recent pooled analysis.[92]

In addition to new chemotherapeutic approaches, targeted therapies, monoclonal antibodies, have also very recently shown great promise in the treatment of advanced colon cancer. Hurwitz *et al.* reported a Phase III trial comparing irinotecan, 5-FU and leucovorin with and without bevacizumab, a monoclonal antibody against vascular endothelial growth factor.[93] A total of 813 patients were randomised. The group treated with bevacizumab had a median survival of 20.3 months, compared with 15.6 months with standard irinotecan, 5-FU and leucovorin, which was highly statistically significant. In the second-line setting, the monoclonal antibody cetuximab, which specifically blocks the epidermal growth factor receptor, was shown to have 'resensitised' tumours that had become resistant to irinotecan.[94] Both of these monoclonal antibodies are currently being tested in a variety of first-line trials, in combination with both oral and infusional chemotherapy regimens.

The proven efficacy of second-line therapies in patients who fail initial therapy has complicated the design of first-line advanced disease trials. Traditionally, overall survival has been used as the primary endpoint for such studies, and extending the patient's longevity remains the ultimate goal. However, given that there are second- and even third-line therapies with proven benefit, the relative merits of overall survival as the primary outcome for a trial warrant reconsideration. Consider the design used in the Saltz trial,[87] where irinotecan plus 5-FU and leucovorin was compared with 5-FU and leucovorin. In this trial, patients who progressed on the 5-FU and leucovorin arm were able to receive irinotecan off study, as it was approved for the second-line indication. The availability of this effective second-line agent provided at least the theoretical possibility that the two primary study arms could show no difference in terms of overall survival, even though irinotecan was beneficial to patients on both arms of the study. For this reason, time to tumour progression was specified as the primary endpoint for the trial. In retrospect, the addition of irinotecan as a component of the initial treatment resulted in both improved time to progression and overall survival, making the result clear. However, these factors must be taken into consideration for future trials, where at minimum data on the use of second- and third-line therapy should be collected.

TREATMENT: RECTAL CANCER

Rectal cancer is second to colon cancer among GI malignancies in the number of new cases per year. When the initial diagnosis for rectal cancer is as advanced disease, i.e. not surgically completely resectable, its primary treatment is in the same manner as for advanced colon cancer. However, the optimal adjuvant treatment for rectal cancer is the issue of considerable study. Questions abound as to the importance of surgical technique, the value of radiation therapy, the optimal chemotherapy regimen and the timing of therapy, either pre- or post-resection.

Surgery/Adjuvant Therapy

Prior to 1990, external beam radiotherapy in the post-operative setting was considered by many as the standard of care in the United States, based primarily on an observed benefit in lowering the risk of local recurrence. In particular, radiation as a single agent added to surgery had never been shown to improve overall survival compared with surgery alone. In a randomised study of 204 patients, Krook *et al.*[95] demonstrated a benefit in overall survival of post-operative combined therapy with radiation and 5-FU and semustine compared with radiation alone. The 1990 NIH consensus statement concluded that 'Combined post-operative chemotherapy and radiation therapy improves local control and survival in Stage II and III patients and is recommended.'[60] In a subsequent study conducted by the US GI Intergroup, two questions were asked in a 2 × 2 factorial design: is semustine necessary, and can therapy be optimised by using continuous infusion 5-FU-based therapy as opposed to bolus? All patients in this study received radiation. This study of 680 patients concluded that (1) semustine is not necessary, and (2) infusional 5-FU-based therapy during the radiotherapy provides a survival advantage compared with bolus therapy.[96]

Two studies conducted by the National Surgical Breast and Bowel Project (NSABP) have questioned the value of radiation in the post-operative setting. In the Krook and O'Connell studies mentioned above, all patients received radiation, and the studies focused on the relative benefit of different chemotherapy regimens. In contrast, NSABP study R-01 tested three arms in a randomised manner in 574 patients: no post-surgical treatment, post-operative radiation and post-operative chemotherapy. A survival benefit was observed for the chemotherapy arm compared with the no-treatment arm, but this advantage was not observed in the radiation alone arm.[97] The NSABP followed this study with a two-arm randomised trial of 741 patients comparing chemotherapy alone with chemotherapy plus radiation.[98] The results of this trial showed no improvement in overall survival for the combined modality arm, although there was a statistically significant improvement in the rate of local recurrence associated with radiation. Despite these two consistent results, radiation continues to be commonly used in the post-operative treatment of Stage II and III rectal cancer.

Increasingly, practitioners are turning to delivering radiotherapy for rectal cancer in the pre-operative setting. There are several theoretical advantages to the pre-operative approach. Perhaps most importantly, from the patient's perspective, pre-operative therapy may shrink the tumour sufficiently to allow a sphincter-sparing resection. Pre-operative radiotherapy has been shown to improve outcome compared with no treatment in a large randomised trial of the Swedish Rectal Cancer Trial group. This trial randomised 1168 patients to a two-arm trial of a short course (25 Gy in one week) of pre-operative radiation compared with no pre-operative treatment, and showed a statistically significant improvement in both local recurrence rate and overall survival.[99] In the United States, the standard pre-operative regimen is to deliver the five-week post-operative course of 50.4 Gy pre-operatively. The efficacy of this approach has never been established in a randomised trial. A comparison of these two pre-operative approaches is clearly warranted.

Regardless of the specifics of the pre-operative approach, a burning question concerns whether the pre- or post-operative approach provides the best outcome. Two randomised trials have been attempted in the United States, and both were closed early far short of their accrual goals due to poor accrual. However, the Sauer *et al.* trial in Europe recently reported mature data on 421 patients randomised to one of the two approaches.[100] The 50.4 Gy long-course radiation was used in both arms, and both arms received the same chemotherapy regimen in combination with the radiation. The five-year overall survival rates did not differ between the two arms; however, the five-year local relapse rate was reduced from 13% in the post-operative arm to 6% in the pre-operative arm. Pre-operative therapy was also associated with significantly fewer acute and long-term side effects. Based on this trial, we expect the shift towards pre-operative therapy will be accelerated.

In addition to the controversies present in chemotherapy and radiation therapy, there is considerable interest in the optimal surgical management of this disease. In particular, the surgical approach of total mesorectal excision (TME) has been promoted as an important surgical advance. Based on case-series and other historical data, proponents of TME have claimed significant reductions in local recurrence rates and improved overall survival compared with standard surgery.[101] However, TME has never been tested against non-TME surgery in a randomised trial, and such a trial is unlikely to ever be conducted. In a large randomised trial of 1861 patients conducted by the Dutch Colorectal Cancer Group, pre-operative radiation was shown to reduce the rate of local recurrence compared with no radiation when all patients received TME surgery.[102] In this early report, with a median follow-up of two years in living patients, there was no improvement in overall survival for patients receiving radiation.

In summary, it is clear that rectal cancer is an area where randomised clinical trials have made several important contributions to improving patient outcomes. Post-operative chemotherapy and chemoradiotherapy, and pre-operative radiation therapy, have been shown to reduce the local recurrence rate and improve overall survival based on large randomised trials. It is also clear that considerable work remains to define the optimal timings and combinations of the different treatment modalities.

CASE STUDY: 5-FU PLUS LEUCOVORIN IN COLON CANCER

As is clear, the history of clinical trials in GI cancer is long and has been very successful. As an example illustrating several facets of both the past history of GI clinical trials and issues that will likely be faced again in future studies, here we present a case study of the development, establishment and replacement of what was once the US standard of care for both advanced and adjuvant colorectal cancer: the ‘Mayo Clinic’ bolus regimen of 5-FU and leucovorin delivered for five consecutive days every four or five weeks.

The activity of fluorinated pyrimidines in the treatment of GI cancers has been reviewed extensively; 5-Fluorouracil (5-FU) is the most ubiquitous of the fluorinated pyrimidines, which at least in part exert their antineoplastic effect by inhibiting the activity of the enzyme thymidylate synthase (TS), which in turn interferes with DNA synthesis in dividing cells. Often agents designed to improve the efficacy of fluorinated pyrimidines are combined with these agents in an effort to preferentially sensitise tumour cells relative to host cells to the agent(s). Leucovorin is an agent commonly used in such a setting. The Mayo regimen of 5-FU and leucovorin is thus a combination of an active chemotherapy agent, 5-FU, with a ‘biochemical modulator’ leucovorin.

Prior to the early 1980s, 5-FU was primarily administered as a single agent. Administered in this fashion, it was associated with limited activity and moderate toxicity. Response rates for metastatic colorectal cancer were low, in the neighbourhood of 10%, and these response were short-lived, lasting on an average a few months.

Based on pre-clinical laboratory studies,[103–105] the addition of leucovorin to cell culture with one of the metabolites of 5-FU, fluorodeoxuridylate monophosphate (FdUMP), resulted in enhanced binding to and inhibition of TS as compared with the binding when FdUMP was used alone. This improved inhibition of thymidylate synthase resulted in inhibited DNA synthesis and in enhanced tumour shrinkage. Depending on the model systems, optimal concentration of leucovorin ranged from 1 to 20 mmol/L.[106–110] These studies supported the use of leucovorin doses ranging from 10 to 600 mg/m^2 in clinical trials where leucovorin was added to 5-FU in an effort to improve on 5-FU’s single agent activity. While such laboratory studies provided basic information on the modulation of 5-FU using leucovorin, the applicability of these results to humans with colorectal cancer was unclear. Based on clinical experience, individuals with colorectal cancer clearly exhibit significant heterogeneity in their response to treatment. The sequence of administration of 5-FU and leucovorin, the optimal concentration of leucovorin and the appropriate interval of 5-FU and leucovorin administration were all variables to be studied to explore the efficacy of 5-FU and leucovorin in inhibiting tumour growth.

Early investigators studying the biochemical modulation of 5-FU with leucovorin in the treatment of colorectal and gastric cancers included Machover and colleagues.[111,112] The Machover regimen consisted of administering high-dose leucovorin at 200 mg/m^2/d prior to 5-FU at a dose of 370 mg/m^2/d, with both drugs given consecutively for five days. With this dose of leucovorin, the blood level is approximately 10–20 μmol/L.[113] In large part to lower the cost of the regimen (leucovorin was very expensive at the time), the ‘Mayo’ regimen was devised to use the identical 5-FU schedule to the Machover regimen, but to use low-dose leucovorin at a dose of 20 mg/m^2/d, which resulted in blood levels of 1–2 μmol/L.

This regimen was first tested as part of a randomised Phase II study in advanced unresectable colorectal cancer.[114,115] Three of the treatment arms are relevant for this discussion: (1) 5-FU as a single agent administered at a dose of 500 mg/m^2/d by IV bolus for five consecutive days every five weeks; (2) the Machover regimen repeated at four weeks, eight weeks and every five weeks thereafter; and (3) the Mayo regimen repeated at the same frequency as the Machover regimen. In this trial, provision was made in the protocol to escalate the 5-FU dose on any treatment arm if there was no observed myelosuppression or significant non-haematologic toxicity during the previous treatment course. When the toxicity was analysed after treatment of the first 100 patients, the starting dose of 5-FU for the Mayo regimen was increased to 425 mg/m^2/d in order to produce definite but tolerable toxicity that was of similar magnitude between the six treatment arms.[114] The original combination of low-dose leucovorin with 370 mg/m^2/d of 5-FU for five consecutive days was empiric; no formal Phase I trial of this regimen had ever been performed. In the 208 eligible patients entered on the three study arms of interest, the overall response rates were 10% for 5-FU alone, 26% for the Machover regimen and 43% for the Mayo regimen. Both leucovorin regimens demonstrated significant improvement in response rate and overall survival compared with 5-FU alone.

Concurrent to the previously mentioned study, investigators at the Roswell Park Memorial Cancer Institute (RPMI) began testing a regimen of leucovorin 500 mg/m^2/d with 5-FU 600 mg/m^2/d given for six consecutive weeks followed by a two-week rest period.[116] In a small study, the RPMI regimen was shown to significantly improve the tumour response rate compared with single agent 5-FU. Shortly thereafter, the RPMI and Mayo regimens were compared in a randomised trial of 366 patients.[117] In this trial, the objective response rates and overall survival were similar between the two arms. The toxicity profile of the two regimens did differ, but no clear winner was identified. Based largely on cost considerations, investigators from the Mayo Clinic and the North Central Cancer Treatment group chose to pursue the Mayo regimen for future testing.

The activity seen with the combination of leucovorin and 5-FU in the advanced disease setting naturally led to the evaluation of several of these regimens in the adjuvant treatment of patients with Stage II and III colon cancer. In a study that was suspended after accrual of 317 patients (based on the results of a large trial that demonstrated 5-FU plus levamisole was an effective treatment in this setting)[59], patients with resected Stage II or III colon cancer were randomised to the Mayo 5-FU plus leucovorin regimen for six months or to a no-treatment control arm.[63] The five-year survival for treated patients was 74%, compared with 63% in the control group ($p = 0.02$). This result established the efficacy of the Mayo 5-FU plus leucovorin regimen in the adjuvant setting.

Following this small study, a large trial was conducted to test four different combinations of 5-FU with leucovorin and/or levamisole in patients with Stage II and III colon cancer. The regimens included the Mayo 5-FU plus leucovorin regimen for 6 months, 5-FU plus levamisole for 12 months, 5-FU with high-dose leucovorin (the RPMI regimen) for 8 months, or 5-FU plus leucovorin plus levamisole for 12 months. In this study of 3759 patients, results were similar between the Mayo and RPMI 5-FU plus leucovorin programmes, and the 5-FU plus both leucovorin and levamisole regimen.[65] Based on the essentially identical activity profiles of these regimens, the choice between the two 5-FU and leucovorin regimens (Mayo and RPMI) has been based on issues related to schedule (some patients preferred weekly therapy over five consecutive days of treatment), cost (at the time of these studies leucovorin was expensive), toxicity profile and clinician's preference.

From the late 1980s until the year 2000, the Mayo regimen of 5-FU and leucovorin was regarded as the standard of care for both advanced and adjuvant colon cancer. As discussed previously, in the late 1990s and early 2000s, several randomised trials were conducted in both the United States and Europe in which infusion-based 5-FU regimens or regimens that combine 5-FU with CPT-11 or oxaliplatin have demonstrated improved patient outcomes compared with those seen with the Mayo regimen. In addition, the oral agent capecitabine has been approved as an alternative to IV 5-FU in advanced disease. Thus in the advanced disease setting, the Mayo 5-FU + leucovorin regimen has been replaced as the standard of care, indeed a welcome advance. In addition, in the adjuvant setting, the MOSAIC trial has demonstrated improved disease-free survival for a multiple drug combination,[69] and the X-ACT trial has shown the equivalence of an oral regimen to the Mayo regimen.[72] Future trials will use multiple drug regimens as the basis for comparison. As outcomes continue to improve, larger and larger trials will be required to establish the superiority of the next generation of treatments.

REFERENCES

1. Launois B, Delarue D, Campion JP, Kerbaol M. Preoperative radiotherapy for carcinoma of the esophagus. *Surg, Gynecol Obstet* (1981) **153**: 690–2.
2. Gignoux M, Roussel A, Paillot B, Gillet M, Schlag P, Favre J-P, Dalesio O, Buyse M, Duez N. The value of preoperative radiotherapy in esophageal cancer: results of a study of the E.O.R.T.C. *World J Surg* (1987) **11**: 426–32.
3. Kelsen DP, Ginsberg R, Pajak TF, Sheahan DG, Gunderson L, Mortimer J, Estes N, Haller DG, Ajani J, Kocha W, Minsky BD, Roth JA. Chemotherapy followed by surgery compared with surgery alone for localized esophageal cancer. *New Engl J Med* (1998) **339**: 1979–84.
4. Kelsen DP, Minsky BD, Smith M, Beitler J, Niedzwiecki D, Chapman D, Bains M, Burt M, Heelan R, Hilaris B. Preoperative therapy for esophageal cancer: a randomized comparison of chemotherapy versus radiation therapy. *J Clin Oncol* (1990) **8**: 1352–61.
5. Bosset J-F, Gignoux M, Triboulet J-P, Tiret E, Mantion G, Elias D, Lozach P, Ollier J-C, Pavy J-J, Mercier M, Sahmoud T. Chemoradiotherapy followed by surgery compared with surgery alone in squamous-cell cancer of the esophagus. *New Engl J Med* (1997) **337**: 161–7.
6. Le Prise E, Etienne PL, Meunier B, Maddern G, Hassel MB, Gedouin D, Boutin D, Campion JP, Launois B. A randomized study of chemotherapy, radiation therapy, and surgery versus surgery for localized squamous cell carcinoma of the esophagus. *Cancer* (1994) **73**: 1779–84.
7. Urba SG, Orringer MB, Turrisi A, Iannettoni M, Forastiere A, Strawderman M. Randomized trial of preoperative chemoradiation versus surgery alone in patients with locoregional esophageal carcinoma. *J Clin Oncol* (2001) **19**: 305–13.
8. Walsh TN, Noonan N, Hollywood D, Kelly A, Keeling N, Hennessy TPJ. A comparison of multimodal therapy and surgery for esophageal adenocarcinoma. *New Engl J Med* (1996) **335**: 462–7.
9. Herskovic A, Martz K, Al-Sarraf M, Leichman L, Brindle J, Vaitkevicius V, Cooper J, Byhardt R, Davis L, Emami B. Combined chemotherapy and radiotherapy compared with radiotherapy alone in patients with cancer of the esophagus. *New Engl J Med* (1992) **326**: 1593–8.
10. Minsky BD, Pajak TF, Ginsberg RJ, Pisansky TM, Martenson J, Komaki R, Okawara G, Rosenthal SA, Kelsen DP. INT 0123 (Radiation Therapy Oncology Group 94-05) phase III trial of combined-modality therapy for esophageal cancer: high-dose versus standard-dose radiation therapy. *J Clin Oncol*. (2002) **20**(5): 1167–74.
11. Jemal A, Thomas A, Murray T, Thun M. Cancer statistics, 2000. *CA Cancer J Clin* (2002) **52**: 23–47.
12. Bonenkamp JJ, Hermans J, Sasako M, van de Velde JH. Extended lymph node dissection for gastric cancer. *New Engl J Med* (1999) **340**: 908–14.
13. Cuschieri A, Weeden S, Fielding J, Bancewicz J, Craven J, Joypaul V, Sydes M, Fayers P. Patient survival after D1 and D2 resections for gastric cancer: long-term results of the MRC randomized surgical trial. Surgical Cooperative Group. *Br J Cancer* (1999) **79**: 1522–30.
14. Janunger KG, Hafstrom L, Nygren P, Glimelius B. A systematic overview of chemotherapy effects in gastric cancer. *Acta Oncol* (2001) **40**: 309–26.
15. Macdonald JS, Smalley SR, Benedetti J, Hundahl SA, Estes NC, Haller DG, Ajani JA, Gunderson LL, Jessup JM, Martenson JA. Chemoradiotherapy after surgery compared with surgery alone for adenocarcinoma of the stomach or gastroesophageal junction. *New Engl J Med* (2001) **345**: 725–30.
16. Allum W, Cunningham D, Weeden S. Perioperative chemotherapy in operable gastric and lower oesophageal cancer: a randomised controlled trial (the MAGIC trial, ISRCTN 93793971). *Proc ASCO* (2003) **22**: 249a.
17. Webb A, Cunningham D, Scarffe JH, Harper P, Norman A, Joffe JK, Hughes M, Mansi J, Findlay M, Hill A, Oates J, Nicolson M, Hickish T, O'Brien M, Iveson T, Watson M, Underhill C, Wardley A, Meehan M. Randomized trial comparing epirubicin, cisplatin, and fluorouracil versus fluorouracil, doxorubicin, and methotrexate in advanced esophagogastric cancer. *J Clin Oncol* (1997) **15**: 261–7.

18. Moertel CG, Frytak S, Hahn RG, O'Connell MJ, Reitemeier RJ, Rubin J, Schutt AJ, Weiland LH, Childs DS, Holbrook MA, Lavin PT, Livstone E, Spiro H, Knowlton A, Kalser M, Barkin J, Lessner H, Mann-Kaplan R, Ramming K, Douglas HO, Thomas P, Nave H, Bateman J, Lokich J, Chaffey J, Corson JM, Zamcheck N, Noval JW. Therapy of locally unresectable pancreatic carcinoma: a randomized comparison of high dose (6000 Rads) radiation alone, moderate dose radiation (4000 Rads +5-fluorouracil), and high dose radiation +5-fluorouracil. *Cancer* (1981) **48**: 1705–10.
19. Klaassen DJ, MacIntyre JM, Catton GE, Engstrom PF, Moertel CG. Treatment of locally unresectable cancer of the stomach and pancreas: a randomized comparison of 5-fluorouracil alone with radiation plus concurrent and maintenance 5-fluorouracil – an Eastern Cooperative Oncology Group Study. *J Clin Oncol* (1985) **3**: 373–8.
20. Gastrointestinal Tumor Study Group. Treatment of locally unresectable carcinoma of the pancreas: comparison of combined-modality therapy (chemotherapy plus radiotherapy) to chemotherapy alone. *J Natl Cancer Inst* (1988) **80**: 751–5.
21. Neoptolemos JP, Stocken DD, Friess H, Bassi C, Dunn J, Hickey H, Beger H, Fernandez-Cruz L, Dervenis C, Lacaine F, Falconi M, Pederzoli P, Pap A, Spooner D, Kerr DJ, Buchler MW, Members of the European Study Group for Pancreatic Cancer. A randomized trial of chemoradiotherapy and chemotherapy after resection of pancreatic cancer. *New Engl J Med* (2004) **350**: 1200–10.
22. Burris HA, Moore MJ, Andersen J, Green MR, Rothenberg ML, Modiano MR, Cripps MC, Portenoy RK, Storniolo AM, Tarassoff P, Nelson R, Dorr FA, Stephens CD, Von Hoff DD. Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. *J Clin Oncol* (1997) **15**: 2403–13.
23. Moore MJ, Hamm J, Dancey J, Eisenberg PD, Dagenais M, Fields A, Hagan K, Greenberg B, Colwell B, Zee B, Tu D, Ottaway J, Humphrey R, Seymour L, National Cancer Institute of Canada Clinical Trials Group. Comparison of gemcitabine versus the matrix metalloproteinase inhibitor BAY 12–9566 in patients with advanced or metastatic adenocarcinoma of the pancreas: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *J Clin Oncol* (2003) **21**(17): 3296–302.
24. Moore MJ, Goldstein D, Hamm J, Figer A, Hecht J, Gallinger S, Au H, Ding K, Christy-Bittel J, Parulekar W. Erlotinib plus gemcitabine compared to gemcitabine alone in patients with advanced pancreatic cancer. A Phase III trial of the National Cancer Institute of Canada Clinical Trials Group (NCIC-CTG). *ASCO Proc* (2005) 1s (abstr 1).
25. McKeown-Eyssen GE, Bright-See E, Bruce WR, Jazmaji V, Cohen LB, Pappas SC, Saibil FG. A randomized trial of a low fat high fibre diet in the recurrence of colorectal polyps. Toronto Polyp Prevention Group. *J Clin Epidemiol* (1994) **47**: 525–36.
26. MacLennan R, Macrae F, Bain C, Battistutta D, Chapuis P, Gratten H, Lambert J, Newland RC, Ngu M, Russell A. Randomized trial of intake of fat, fiber, and beta carotene to prevent colorectal adenomas. The Australian Polyp Prevention Project. *J Natl Cancer Inst* (1995) **87**: 1760–6.
27. Bonithon-Kopp C, Kronborg O, Giacosa A, Rath U, Faivre J. Calcium and fibre supplementation in prevention of colorectal adenoma recurrence: a randomised intervention trial. European Cancer Prevention Organisation Study Group. *Lancet* (2000) **356**: 1300–6.
28. Alberts DS, Martinez ME, Roe DJ, Guillen-Rodriguez JM, Marshall JR, van Leeuwen JB, Reid ME, Ritenbaugh C, Vargas PA, Bhattacharyya AB, Earnest DL, Sampliner RE. Lack of effect of a high-fiber cereal supplement on the recurrence of colorectal adenomas. Phoenix Colon Cancer Prevention Physicians' Network. *New Engl J Med* (2000) **342**: 1156–62.
29. Schatzkin A, Lanza E, Corle D, Lance P, Iber F, Caan B, Shike M, Weissfeld J, Burt R, Cooper MR, Kikendall JW, Cahill J. Lack of effect of a low-fat, high-fiber diet on the recurrence of colorectal adenomas. Polyp Prevention Trial Study Group. *New Engl J Med* (2000) **342**: 1149–55.
30. Ponz de Leon M, Roncucci L. Chemoprevention of colorectal tumors: role of lactulose and of other agents. *Scand J Gastroenterol Suppl* (1997) **222**: 72–5.
31. Hofstad B, Almendingen K, Vatn M, Andersen SN, Owen RW, Larsen S, Osnes M. Growth and recurrence of colorectal polyps: a double-blind 3-year intervention with calcium and antioxidants. *Digestion* (1998) **59**: 148–56.
32. McKeown-Eyssen G, Holloway C, Jazmaji V, Bright-See E, Dion P, Bruce WR. A randomized trial of vitamins C and E in the prevention of recurrence of colorectal polyps. *Cancer Res* (1988) **48**: 4701–5.
33. Greenberg ER, Baron JA, Tosteson TD, Freeman DH Jr, Beck GJ, Bond JH, Colacchio TA, Coller JA, Frankl HD, Haile RW. A clinical trial of antioxidant vitamins to prevent colorectal adenoma. Polyp Prevention Study Group. *New Engl J Med* (1994) **331**: 141–7.

34. Baron JA, Cole BF, Mott L *et al*. Neoplastic and antineoplastic effects of beta-carotene on colorectal adenoma recurrence: results of a randomized trial. *J Natl Cancer Inst* (2003) **95**(10): 717–22.
35. Albanes D, Malila N, Taylor PR *et al*. Effects of supplemental alpha-tocopherol and beta-carotene on colorectal cancer: results from a controlled trial (Finland). *Cancer Causes Control* (2000) **11**(3): 197–205.
36. Bergsma-Kadijk JA, van't Veer P, Kampman E, Burema J. Calcium does not protect against colorectal neoplasia. *Epidemiology* (1996) **7**: 590–7.
37. Hofstad B, Almendingen K, Vatn M *et al*. Growth and recurrence of colorectal polyps: a double-blind 3-year intervention with calcium and antioxidants. *Digestion* (1998) **59**(2): 148–56.
38. Baron JA, Beach M, Mandel JS, van Stolk RU, Haile RW, Sandler RS, Rothstein R, Summers RW, Snover DC, Beck GJ, Bond JH, Greenberg ER. Calcium supplements for the prevention of colorectal adenomas. Calcium Polyp Prevention Study Group. *New Engl J Med* (1999) **340**: 101–7.
39. Bonithon-Kopp C, Kronborg O, Giacosa A *et al*. Calcium and fibre supplementation in prevention of colorectal adenoma recurrence: a randomised intervention trial. European Cancer Prevention Organisation Study Group. *Lancet* (2000). **356**(9238): 1300–6.
40. The Women's Health Initiative Study Group. Design of the women's health initiative clinical trial and observational study. *Controlled Clin Trials* (1998) **19**: 61–109.
41. Sandler RS, Halabi S, Baron JA *et al*. A randomized trial of aspirin to prevent colorectal adenomas in patients with previous colorectal cancer. *N Engl J Med* (2003) **348**(10): 883–90.
42. Baron JA, Cole BF, Sandler RS *et al*. A randomized trial of aspirin to prevent colorectal adenomas. *N Engl J Med* (2003) **348**(10): 891–9.
43. Benamouzig R, Deyra J, Martin A *et al*. Daily soluble aspirin and prevention of colorectal adenoma recurrence: one-year results of the APACC trial. *Gastroenterology* (2003). **125**(2): 328–36.
44. Gann PH, Manson JE, Glynn RJ, Buring JE, Hennekens CH. Low-dose aspirin and incidence of colorectal tumors in a randomized trial. *J Natl Cancer Inst* (1993) **85**: 1220–4.
45. Sturmer T, Glynn RJ, Lee IM, Manson JE, Buring JE, Hennekens CH. Aspirin use and colorectal cancer: post-trial follow-up data from the Physicians' Health Study. *Ann Int Med* (1998) **128**: 713–20.
46. Steinbach G, Lynch PM, Phillips RK, Wallace MH, Hawk E, Gordon GB, Wakabayashi N, Saunders B, Shen Y, Fujimura T, Su LK, Levin B. The effect of celecoxib, a cyclooxygenase-2 inhibitor, in familial adenomatous polyposis. *New Engl J Med* (2000) **342**: 1946–52.
47. Chlebowski RT, Wactawski-Wende J, Ritenbaugh C *et al*. Estrogen plus progestin and colorectal cancer in postmenopausal women. *N Engl J Med* (2004) **350**(10): 991–1004.
48. Jorgensen OD, Kronborg O, Fenger C. A randomised study of screening for colorectal cancer using faecal occult blood testing: results after 13 years and seven biennial screening rounds. *Gut* (2002) **50**: 29–32.
49. Scholefield JH, Moss S, Sufi F, Mangham CM, Hardcastle JD. Effect of faecal occult blood screening on mortality from colorectal cancer: results from a randomised controlled trial. *Gut* (2002) **50**: 840–4.
50. Mandel JS, Church TR, Bond JH, Ederer F, Geisser MS, Mongin SJ, Snover DC, Schuman LM. The effect of fecal occult-blood screening on the incidence of colorectal cancer. *New Engl J Med* (2000) **343**: 1603–7.
51. Winawer SJ, Fletcher RH, Miller L, Godlee F, Stolar MH, Mulrow CD, Woolf SH, Glick SN, Ganiats TG, Bond JH, Rosen L, Zapka JG, Olsen SJ, Giardiello FM, Sisk JE, Van Antwerp R, Brown-Davis C, Marciniak DA, Mayer RJ. Colorectal cancer screening: clinical guidelines and rationale. *Gastroenterology* (1997) **112**: 594–642.
52. Smith RA, Cokkinides V, von Eschenbach AC, Levin B, Cohen C, Runowicz CD, Sener S, Saslow D, Eyre HJ. American Cancer Society guidelines for the early detection of cancer. *CA Cancer J Clin* (2002) **52**: 8–22.
53. Weissfeld JL, Ling BS, Schoen RE, Bresalier RS, Riley T, Prorok PC. Adherence to repeat screening flexible sigmoidoscopy in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. *Cancer* (2002) **94**: 2569–76.
54. UK Flexible Sigmoidoscopy Screening Trial Investigators. Single flexible sigmoidoscopy screening to prevent colorectal cancer: baseline findings of a UK multicentre randomised trial. *Lancet* (2002) **359**: 1291–300.
55. Ness RM, Holmes AM, Klein R, Dittus R. Cost-utility of one-time colonoscopic screening for colorectal cancer at various ages. *Am J Gastroenterol* (2000) **95**: 1800–11.
56. Winawer SJ, Zauber AG, Church TR, Mandelson M, Feld A, Bond JH. National Colonoscopy Study (NCS) preliminary findings: a randomized clinical trial of general population screening colonoscopy. *Gastroenterology* (2002) T1560.

57. The Writing Committee of the Clinical Outcomes of Surgical Therapy Study Group responsible for the reporting of the Laparoscopic Colectomy Trial. A comparative evaluation of laparoscopic-assisted versus open colectomy for colon cancer. *New Engl J Med* (2004) **350**: 2050–9.
58. Laurie JA, Moertel CG, Fleming TR, Wieand HS, Leigh JE, Rubin J, McCormack GW, Gerstner JB, Krook JE, Mailliard J, Twito DI, Morton RF, Tschetter LK, Barlow JF. Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. *J Clin Oncol* (1989) **7**: 1447–56.
59. Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Goodman PJ, Ungerleider JS, Emerson WA, Tormey DC, Glick JH, Veeder MH, Mailliard JA. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New Engl J Med* (1990) **322**: 352–8.
60. NIH Consensus Conference. Adjuvant therapy for patients with colon and rectal cancer. *J Am Med Assoc* (1990) **264**: 1444–50.
61. International Multicentre Pooled Analysis of Colon Cancer Trials (IMPACT) Investigators. Efficacy of adjuvant fluorouracil and folinic acid in colon cancer. *Lancet* (1995) **345**: 939–44.
62. Francini G, Petrioli R, Lorenzini L, Mancini S, Armenio S, Tanzini G, Marsili S, Aquino A, Marzocca G, Civitelli S, Mariani L, De Sando D, Bovenga S, Lorenz M. Folinic acid and 5-fluorouracil as adjuvant chemotherapy in colon cancer. *Gastroenterology* (1994) **106**: 899–906.
63. O'Connell MJ, Mailliard JA, Kahn MJ, Macdonald JS, Haller DG, Mayer RJ, Wieand S. Controlled trial of fluorouracil and low-dose leucovorin given for 6 months as postoperative adjuvant therapy for colon cancer. *J Clin Oncol* (1997) **15**: 246–50.
64. QUASAR Collaborative Group. Comparison of fluorouracil with additional levamisole, higher-dose folinic acid, or both, as adjuvant chemotherapy for colorectal cancer: a randomised trial. *Lancet* (2000) **355**: 1588–96.
65. Haller DG, Catalano PJ, Macdonald JS, O'Rourke MA, Frontiera MS, Jackson DV, Mayer RJ. Phase III Study of Flourouracil, Leucovorin, and Levamisole in High-Risk Stage II and III Colon Cancer: Final Report of Intergroup 0089 *J Clin Oncol* (2005) **23**: 8671–8678.
66. Poplin E, Benedetti N, Estes D, Haller DG, Mayer R, Goldberg R, Macdonald J. Phase III randomized trial of bolus 5-FU/leucovorin/levamisole versus 5-FU continuous/infusion/levamisole as adjuvant therapy for high risk colon cancer (SWOG 9415/INT-0153). *Proc Am Soc Clin Oncol* (2000) **19**: 240a.
67. Saltz LB, Niedzwiecki D, Hollis D, Goldberg RM, Hantel A, Thomas JP, Fields ALA, Carver G, Mayer RJ. Irinotecan plus fluorouracil/leucovorin (IFL) versus fluorouracil/leucovorin alone (FL) in stage III colon cancer (intergroup trial CALGB C89803). *ASCO Proc* (2004) 246 (abstr 3500).
68. van Cutsem E, Labianca R, Hossfeld D, Bodoky G, Roth A, Aranda E, Nordlinger B, Assadourian S, Wang K, Cunningham D. Randomized phase III trial comparing infused irinotecan/5-fluorouracil (5-FU)/folinic acid (IF) versus 5-FU/FA (F) in stage III colon cancer patients (pts) (PETACC 3). *ASCO Proc* (2005) 3s (abstr LBA8).
69. Andre T, Boni C, Mounedji-Boudiaf L, Navarro M, Tabernero J, Hickish T, Topham C, Zaninelli M, Clingan P, Bridgewater J, Tabah-Fisch I, de Gramont A. Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. *New Engl J Med* (2004) **350**: 2343–51.
70. Wolmark N, Wieand S, Kuebler JP, Colangelo L, Smith RE. A phase III trial comparing FULV to FULV + oxaliplatin in stage II or III carcinoma of the colon: Results of NSABP Protocol C-07. *ASCO Proc* (2005) 246s (abstr LBA3500).
71. Wolmark N, Wieand S, Lembersky B, Colangelo L, Smith R, Pazdur R. A phase III trial comparing oral UFT to FULV in stage II and III carcinoma of the colon: Results of NSABP protocol C-06. *J Clin Oncol* (2004) **22**: 3508.
72. Twelves C, Wong A, Nowacki MP, Abt M, Burris H 3rd, Carrato A, Cassidy J, Cervantes A, Fagerberg J, Georgoulias V, Husseini F, Jodrell D, Koralewski P, Kroning H, Maroun J, Marschner N, McKendrick J, Pawlicki M, Rosso R, Schuller J, Seitz JF, Stabuc B, Tujakowski J, Van Hazel G, Zaluski J, Scheithauer W. Capecitabine as adjuvant treatment for stage III colon cancer. *New Engl J Med* (2005) **352**(26): 2696–704.
73. Wolmark N, Fisher B, Rockette H. Postoperative adjuvant chemotherapy or BCG for colon cancer: results from NSABP protocol C-01. *J Natl Cancer Inst* (1988) **80**: 30–6.
74. Sargent DJ, Goldberg RM, Jacobson SD, Macdonald JS, Labianca R, Haller DG, Shepherd LE, Seitz JF, Francini G. A pooled analysis of adjuvant chemotherapy for resected colon cancer in elderly patients. *New Engl J Med* (2001) **345**: 1091–7.
75. Mamounas E, Wieand S, Wolmark N, Bear HD, Atkins JN, Song K, Jones J, Rockette H. Comparative efficacy of adjuvant chemotherapy in patients with Dukes' B versus Dukes' C colon

cancer: results from four National Surgical Adjuvant Breast and Bowel Project Adjuvant Studies (C-01, C-02, C-03, and C-04). *J Clin Oncol* (1999) **17**: 1349–55.

76. International Multicentre Pooled Analysis of B2 Colon Cancer Trials (IMPACT B2) Investigators. Efficacy of adjuvant fluorouracil and folinic acid in B2 colon cancer. *J Clin Oncol* (1999) **17**: 1356–63.
77. Colacchio TA, Niedzwiecki D, Compton C, Warren R, Benson AIB, Goldberg R, Kerr D, Fields A, Hollis D, Mayer R. Phase III trial of adjuvant immunotherapy with MOAb 17-1A following resection for stage II adenocarcinoma of the colon (CALGB 9581). *ASCO Proc* (2004) 14s (abstr 3522).
78. Gray RG, Barnwell J, Hills R, McConkey C, Williams N, Kerr D. QUASAR: a randomized study of adjuvant chemotherapy (CT) vs observation including 3238 colorectal cancer patients. *ASCO Proc* (2004) 246 (abstr 3501).
79. Gill S, Loprinzi CL, Sargent DJ, Thome SD, Alberts SR, Haller DG, Benedetti J, Francini G, Shepherd LE, Seitz JF, Labianca R, Chen W, Cha SS, Heldebrandt MP, Goldberg RM. A pooled analysis of 5-FU based adjuvant therapy for stage II and III colon cancer: who benefits and by how much? *J Clin Oncol* (2004) **22**(10): 1797–1806.
80. Advanced Colorectal Cancer Meta-analysis Project. Modulation of fluorouracil by leucovorin in patients with advanced colorectal cancer: evidence in terms of response rate. *J Clin Oncol* (1992) **10**: 896–903.
81. Meta-Analysis Group in Cancer. Efficacy of intravenous continuous infusion of fluorouracil compared with bolus administration in advanced colorectal cancer. *J Clin Oncol* (1998) **16**: 301–8.
82. Meta-Analysis Group in Cancer. Toxicity of fluorouracil in patients with advanced colorectal cancer: effect of administration schedule and prognostic factors. *J Clin Oncol* (1998) **16**: 3537–41.
83. Hoff PM, Ansari R, Batist G. Comparison of oral capecitabine versus intravenous fluorouracil plus leucovorin as first-line treatment in 605 patients with metastatic colorectal cancer: results of a randomized phase III study. *J Clin Oncol* (2001) **19**: 2282–92.
84. Van Cutsem E, Twelves C, Cassidy J. Oral capecitabine compared with intravenous fluorouracil plus leucovorin in patients with metastatic colorectal cancer: results of a large phase III study. *J Clin Oncol* (2001) **19**: 4097–106.
85. Rougier P, Van Cutsem E, Bajetta E. Randomised trial of irinotecan versus fluorouracil by continuous infusion after fluorouracil failure in patients with metastatic colorectal cancer. *Lancet* (1998) **352**: 1407–12.
86. Cunningham D, Pyrhonen S, James RD. Randomised trial of irinotecan plus supportive care versus supportive care alone after fluorouracil failure for patients with metastatic colorectal cancer. *Lancet* (1998) **352**: 1413–18.
87. Saltz LB, Cox JV, Blanke C, Rosen LS, Fehrenbacher L, Moore MJ, Maroun JA, Ackland SP, Locker PK, Pirotta N, Elfring GL, Miller LL. Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. *New Engl J Med* (2000) **343**: 905–14.
88. Douillard JY, Cunningham D, Roth AD, Navarro M, James RD, Karasek P, Jandik P, Iveson R, Carmichael J, Alakl M, Gruia G, Awad L, Rougier P. Irinotecan combined with fluorouracil compared with fluorouracil alone as first-line treatment for metastatic colorectal cancer: a multicentre randomised trial. *Lancet* (2000) **355**: 1041–7.
89. De Gramont A, Figer A, Seymour M, Homerin M, Hmissi A, Cassidy J, Boni C, Cortes-Funes H, Cervantes A, Freyer G, Papamichael D, Le Bail N, Louvet C, Hendler D, de Braud F, Wilson C, Morvan F, Bonetti A. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. *J Clin Oncol* (2000) **18**: 2938–47.
90. Goldberg RM, Sargent DJ, Morton RF, Fuchs CS, Ramanathan RK, Williamson SK, Findlay BP, Pitot HC, Alberts SR. A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *J Clin Oncol* (2004) **22**: 1–8.
91. Tournigand C, Andre T, Achille E, Lledo G, Flesh M, Mery-Mignard D, Quinaux E, Couteau C, Buyse M, Ganem G, Landi B, Colin P, Louvet C, de Gramont A. FOLFIRI followed by FOLFOX6 or the reverse sequence in advanced colorectal cancer: a randomized GERCOR study. *J Clin Oncol* (2004) **22**(2): 229–37.
92. Grothey A, Sargent DJ, Goldberg RM, Schmoll H-J. Survival of patients with advanced colorectal cancer improves with the availability of 5-Fluorouracil/Leucovorin, Irinotecan, and Oxaliplatin in the course of their treatment. *J Clin Oncol* (2004) **22**: 1209–14.
93. Hurwitz H, Fehrenbacher L, Novotny W, Cartwright T, Hainsworth J, Heim W, Berlin J, Baron A, Griffing S, Holmgren E, Ferrara N, Fyfe G, Rogers B, Ross R, Kabbinavar F. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *New Engl J Med* (2004) **350**: 2335–42.

94. Cunningham D, Humblet Y, Siena S, Khayat D, Bleiberg H, Santoro A, Bets D, Mueser M, Harstrick A, Verslype C, Chau I, Van Cutsem E. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *New Engl J Med* (2004) **351**: 337–45.
95. Krook JE, Moertel CG, Gunderson LL, Wieand S, Collins RT, Beart RW, Kubista TP, Poon MA, Meyers WC, Mailliard JA, Twito DI, Morton RF, Veeder MH, Witzig TE, Cha S, Vidyarthi SC. Effective surgical adjuvant therapy for high-risk rectal carcinoma. *New Engl J Med* (1991) **324**: 709–15.
96. O'Connell MJ, Martenson JA, Wieand S, Krook JE, Macdonald JS, Haller DG, Mayer RJ, Gunderson LL, Rich TA. Improving adjuvant therapy for rectal cancer by combining protracted-infusion fluorouracil with radiation therapy after curative surgery. *New Engl J Med* (1994) **331**: 502–7.
97. Fisher B, Wolmark N, Rockette H, Redmond C, Deutsch M, Wickerham DL. Postoperative adjuvant chemotherapy or radiation therapy for rectal cancer: results from NSABP Protocol R-01. *J Natl Cancer Inst* (1988) **80**: 21–9.
98. Wolmark N, Wieand S, Hyams DM, Colangelo L, Dimitrov NV, Romond EH, Wexler M, Prager D, Cruz AB Jr, Gordon PH, Petrelli NJ, Deutsch M, Mamounas E, Wickerham DL, Fisher ER, Rockette H, Fisher B. Randomized trial of postoperative adjuvant chemotherapy with or without radiotherapy for carcinoma of the rectum: National Surgical Adjuvant Breast and Bowel Project Protocol R-02. *J Natl Cancer Inst* (2000) **92**: 388–96.
99. Swedish Rectal Cancer Trial. Improved survival with preoperative radiotherapy in resectable rectal cancer. *New Engl J Med* (1997) **336**: 980–7.
100. Sauer R, Becker H, Hohenberger W, Rodel C, Wittekind C, Fietkau R, Martus P, Tschmelitsch J, Hager E, Hess CF, Karstens J-H, Liersch T, Schmidberger H, Raab R. Preoperative versus postoperative chemoradiotherapy for rectal cancer. *New Engl J M* (2004) **351**: 1731–40.
101. Enker WE, Thaler HT, Cranor ML, Polyak T. Total mesorectal excision in the operative treatment of carcinoma of the rectum. *J Am Coll Surg* (1995) **181**: 335–46.
102. Kapiteijn E, Marijnen CAM, Nagtegaal ID, Putter H, Steup WH, Wiggers T, Rutten HJT, Pahlman L, Glimelius B, Han J, van Krieken HJM, Leer JWH, van de Velde JH. Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer. *New Engl J Med* (2001) **345**: 638–46.
103. Ullman B, Lee M, Martin DW Jr, Santi DV. Cytotoxicity of 5-fluoro-2-deoxyuridine: requirement for reduced folate co-factors and antagonism by methotrexate. *Proc Natl Acad Sci* (1978) **75**: 980–3.
104. Evans RM, Laskin JD, Hakala MT. Effect of excess folates and deoxyinosine on the activity and site of action of 5-fluorouracil. *Cancer Res* (1981) **41**: 3288–95.
105. Waxman S, Bruckner H. The enhancement of 5-fluorouracil antimetabolic activity by leucovorin, menadione, and alpha-tocopherol. *Eur J Cancer Clin Oncol* (1982) **18**: 685–92.
106. Keyomarsi K, Moran RG. Folinic augmentation of the effects of fluoropyrimidines on murine and human leukemic cells. *Cancer Res* (1986) **46**: 5229–35.
107. Houghton JA, Williams LG, Loftin SK. Relationship between the dose rate of [6RS] leucovorin administration, plasma concentrations of reduced folates, and pools of 5,10 methylenetetrahydrafolates and tetrahydrafolates in human colon adenocarcinoma xenografts. *Cancer Res* (1990) **50**: 3493–502.
108. Houghton JA, Williams LG, Loftin SK. Factors that influence the therapeutic activity of 5-fluorouracil [6RS] leucovorin combinations in colon adenocarcinoma xenografts. *Cancer Chemother Pharmacol* (1992) **30**: 423–32.
109. Houghton JA, Williams LG, Cheshire PJ. Influence of the dose of [6RS] leucovorin on reduced folate pools and 5-fluorouracil-mediated thymidylate synthase inhibition in human colon adenocarcinoma xenografts. *Cancer Res* (1990) **50**: 3940–6.
110. Cao S, Frank C, Rustum YM. Role of fluoropyrimidine scheduling and (6R,S) leucovorin dose in a preclinical animal model of colorectal carcinoma. *J Natl Cancer Inst* (1996) **88**: 430–6.
111. Machover D, Goldschmidt E, Chollet P. Treatment of advanced colorectal and gastric adenocarcinomas with 5-fluorouracil and high-dose folinic acid. *J Clin Oncol* (1986) **4**: 685–96.
112. Machover D, Schwartenberg L, Goldschmidt E. Treatment of advanced colorectal and gastric adenocarcinomas with 5-FU combined with high-dose folinic acid: a pilot study. *Cancer Treat Rep* (1982) **66**: 1803–7.
113. Rustum YM, Trave F, Zakrzewski SF. Biochemical and pharmacologic basis for potentiation of 5-fluorouracil action by leucovorin. *NCI Monogr* (1987) **5**: 165–70.
114. Poon MA, O'Connell MJ, Moertel CG. Biochemical modulation of fluorouracil: evidence of significant improvement in survival and quality of life in patients with advanced colorectal carcinoma. *J Clin Oncol* (1989) **7**: 1407–18.

115. O'Connell MJ. A phase III trial of 5-fluorouracil and leucovorin in the treatment of advanced colorectal cancer: a Mayo Clinic/North Central Cancer Treatment Group study. *Cancer* (1989) **63**: 1026–30.

116. Petrelli N, Herrara L, Rustum Y. A prospective randomized trial of 5-fluorouracil vs. 5-fluorouracil and high dose leucovorin +5-fluorouracil and methotrexate in previously untreated patients with advanced colorectal carcinoma. *J Clin Oncol* (1987) **5**: 1559–65.

117. Buroker TR, O'Connell MJ, Wieand S. Randomized comparison of two schedules of fluorouracil and leucovorin in the treatment of advanced colorectal cancer. *J Clin Oncol* (1994) **12**: 14–20.

7

Melanoma

PING-YU LIU[1] AND VERNON K. SONDAK[2]
[1]Fred Hutchinson Cancer Research Centre, Seattle, WA 98109-1024, USA
[2]H. Lee Moffitt Cancer Centre, Tampa, FL 33612, USA

INTRODUCTION

Randomised Phase III clinical trials are the gold standard for medical decision making, particularly where a modest incremental benefit is sought (such as with surgical treatment of regional nodes or with systemic adjuvant therapies). The results of such randomised studies in melanoma have repeatedly demonstrated the limitations of conclusions drawn based on retrospective data. However, there is sometimes marked disagreement among clinicians in their interpretation of Phase III trial results. Nowhere is this more evident than in the arena of adjuvant therapy of resected 'high-risk' melanoma. In this chapter, we will review the results of several key randomised trials: vaccine trials that demonstrate the limitations of non-randomised data, and interferon trials that illustrate the potential for conflicting clinical interpretations of the same trial data.

BACKGROUND

A basic familiarity with malignant melanoma is required in order to understand the statistical and clinical issues presented herein.

The prognosis of localised cutaneous melanoma is based on several well-defined factors. Pathologic analysis of the primary tumour can predict the likelihood of regional and distant metastasis and death from melanoma. Clinically localised melanomas are grouped into three prognostic categories based on the thickness of the primary tumour as measured by the pathologist using a micrometer built into the microscope eyepiece (Breslow's thickness). Melanomas less than 1.0 mm in thickness have an overall excellent prognosis with relatively minimal intervention and are considered 'low-risk' lesions. Melanomas between 1.0 and 3.9 mm are considered to be intermediate risk, while melanomas 4.0 mm or greater are considered 'high-risk' tumours. The presence of ulceration of the primary tumour increases the risk of metastasis and death within any given thickness category.[1]

The thickness is highly predictive of the risk of regional lymph node metastasis, with nodal involvement in $<5\%$ of melanomas that are <1.0 mm versus $>30\%$ in melanomas ≥ 4.0 mm. Intermediate-thickness melanomas have an intermediate risk of nodal spread, of the order of 20%.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

The prognostic significance of the presence of nodal metastasis far outweighs the significance of tumour thickness: a thin or intermediate-thickness melanoma with nodal metastases generally has a worse prognosis than a thick melanoma with negative nodes. Once nodal metastasis has been documented, the number of involved nodes is the strongest predictor of subsequent outcome, along with the manner of detection of the metastasis. Melanoma in clinically enlarged nodes portends a worse prognosis than melanoma in clinically normal nodes.[1]

The mainstay of treatment for localised or regionally metastatic melanoma is surgery. Adequate wide excision of the primary tumour site (generally taking a margin of 1 to 2 cm of normal skin around the visible edge of the melanoma or biopsy scar) is highly efficacious in controlling disease at the primary site.[2,3]

Three main options are available for staging regional nodes in patients with cutaneous melanoma: clinical staging, surgical staging by complete (elective) lymph node dissection, and surgical staging by sentinel lymph node biopsy.

CLINICAL STAGING

Physical examination is the mainstay of clinical staging of the regional nodes. Any palpable lymph nodes that are ≥1 cm in maximum diameter or very hard or fixed to adjacent structures must be considered highly suspicious for metastatic involvement. Unfortunately, both the specificity and sensitivity of physical examination for detecting melanoma nodal metastases are low. In muscular or obese patients, even relatively large lymph node metastases can be missed on physical examination. Lymph nodes may be enlarged after a biopsy procedure due to reactive hyperplasia without containing metastasis. Most importantly, metastatic involvement of normal-sized lymph nodes cannot be reliably identified by physical examination.

Radiologic studies – computed tomography (CT), positron emission tomography (PET) and ultrasonography – are also available to clinically stage the regional nodes. CT shares many of the deficiencies of physical examination: enlarged nodes may not be malignant, and normal-sized nodes harbouring metastases will be deemed normal. PET is more sensitive than CT for differentiating melanoma-containing nodes from reactive nodes, but is still not able to identify microscopic foci of melanoma in normal nodes.[4] Currently, neither PET nor CT are routinely recommended for clinical staging. Ultrasonography, which involves no ionising radiation, has recently emerged as an alternative for the evaluation of clinically normal nodes, but its sensitivity and specificity remain to be defined in large-scale trials.[5]

For patients with low-risk melanomas, i.e. those that are <1 mm in Breslow's depth and have no evidence of ulceration or significant regression, clinical staging by physical examination is standard practice. Currently, surgical staging is used in the majority of patients with higher-risk lesions. For any patient with clinically evident nodal involvement, a complete therapeutic lymph node dissection is associated with cure in about 20% to 40% of patients.

SURGICAL STAGING BY COMPLETE (ELECTIVE) LYMPH NODE DISSECTION

Elective removal of clinically normal regional nodes identifies evidence of metastasis about 20% of the time, and is clearly a more accurate determinant of nodal status than clinical staging. Retrospective reviews suggested a survival advantage for elective node dissection compared with clinical staging with subsequent therapeutic node dissection at the time of nodal recurrence.[6] To date, however, no prospective study has demonstrated an overall survival advantage for elective node dissection.[3,7] Although the lack of a demonstrated benefit is not the same as the demonstration of no benefit, elective dissection of clinically normal nodes is not considered standard practice for cutaneous melanoma at the present time. It is clear, however, that elective node dissection results in durable regional disease control in the vast majority of patients, and failures within the dissected nodal basin are quite uncommon.

SURGICAL STAGING BY SENTINEL LYMPH NODE BIOPSY

Sentinel lymph node biopsy is based on the concept that lymphatic fluid from an area of skin drains specifically to an initial node or nodes ('sentinel nodes') prior to disseminating to other nodes in the same or nearby basins. Morton *et al.* described a reliable method for identification and removal of the sentinel node draining the site of a cutaneous melanoma.[8] They showed conclusively that the pathologic status of the sentinel node accurately determines whether melanoma cells have metastasised to that specific lymph node basin.[9] An important aspect of sentinel node biopsy is a detailed histologic examination of the sentinel lymph nodes. Generally, this examination is more thorough than is practical to perform on the larger number of nodes obtained during elective node dissection. This more detailed pathologic analysis, combined with the ability to identify sentinel nodes that are outside the defined boundaries of a regional basin, makes sentinel node biopsy the most sensitive and specific test for nodal metastasis currently available. The prognostic value of sentinel node status has been demonstrated in multiple studies. In published multivariate analyses, histologic status of the sentinel nodes is the most powerful predictor of disease-specific survival.[10] Overall, five-year disease-specific survival is >80% for patients with negative sentinel nodes, compared with about 50% for patients with one or more positive sentinel nodes. Importantly, patients with positive sentinel nodes go on to elective complete lymph node dissection. Among patients with negative sentinel nodes, only 4% or fewer ultimately experience a clinically evident relapse within the nodal basin. Thus, sentinel node biopsy matches the excellent regional control achieved by elective node dissection while subjecting fewer patients to the morbidity of the complete node dissection procedure.

A large-scale randomised trial, involving over 2000 patients, has recently been completed comparing wide excision alone with wide excision plus sentinel node biopsy.[11] Complete node dissection was performed if the sentinel node was found to be involved with tumour, or if clinically involved nodes developed after wide excision alone or wide excision and negative sentinel node biopsy. Final results have not been yet been published, but initial presentations and abstracts describing the data illustrate the difficulties inherent in interpreting trials of this type. Only a minority (roughly 20%) of study patients actually have involved lymph nodes, hence for all the patients on the wide excision-alone (control) arm and 80% of the patients on the wide excision plus sentinel node biopsy arm (investigational arm) the delivered treatment is essentially the same. Hence, even if sentinel node biopsy has an impact on melanoma recurrence and survival for node-positive patients, the ability to detect a difference in outcome specifically related to the intervention under study is inherently limited. The outcome difference of most interest – between a patient with nodal involvement randomised to wide excision only and a patient with the same extent of nodal involvement randomised to wide excision plus sentinel node biopsy – can only be assessed indirectly (with potential biases), since we do not know at baseline which observation arm patients are node-positive and which node-positive patients on the sentinel node biopsy arm are falsely deemed node-negative.

ADJUVANT THERAPY FOR MELANOMA

The development of effective adjuvant therapy has been a long-standing goal of melanoma researchers, and the subject of over 100 randomised clinical trials involving a host of different agents.[12] Adjuvant therapy is the systemic or regional administration of drugs or radiation to patients after apparently successful surgery, in an effort to minimise the risk of subsequent recurrence. Although many patients are cured by surgery, some benefit from adjuvant treatment while others will relapse regardless of adjunctive measures. Currently there are no predictive methods to distinguish one group of patients from another, therefore it is necessary to treat all patients in hopes of gaining an incremental benefit for a select few. Hence, in addition to the overall level of efficacy, clinicians evaluate toxicity, convenience, cost-effectiveness and the prospects of post-relapse salvage therapy when deciding whether to employ adjuvant therapy. Virtually all of these factors can be determined accurately only in randomised trials.

In 1995, high-dose interferon-α2b (IFN-α2b) was approved by the United States Food and Drug Administration, based on the positive results of a single randomised Phase III clinical trial, E1684. The FDA's decision was considered controversial at the time. Subsequent randomised trials involving the same basic interferon regimen have not only failed to put this controversy to rest, but in fact enhanced it.

ADJUVANT INTERFERON CLINICAL TRIALS

E1684

Eastern Cooperative Oncology Group (ECOG) trial E1684, with 280 eligible patients with thick primary ($\geq$4.00 mm) or node-positive melanoma who were randomly assigned after surgery to observation or post-operative adjuvant treatment with IFN-α2b for one year, demonstrated statistically significant improvements in relapse-free and overall survival for patients randomised to the interferon arm. IFN-α2b therapy increased the median relapse-free survival by nine months (1.72 years for IFN-α2b patients versus 0.98 years for observation patients) and produced a relative 42% improvement in the five-year relapse-free survival rate (37% for IFN-α2b patients versus 26% for observation patients). In addition, IFN-α2b therapy significantly increased median overall survival by one year (3.82 years for IFN-α2b patients versus 2.78 years for observation patients) and produced a 24% relative improvement in the five-year overall survival rate (46% for IFN-α2b patients versus 37% for observation patients).[13] Side effects were common and frequently severe, but even when adjusted for time with toxicity, the results favoured adjuvant IFN-α2b therapy.[14]

A late analysis of outcome data for patients on this trial has been performed.[15] After a median follow-up of 12.6 years, patients randomised to the interferon arm have a sustained improvement in relapse-free survival that remains statistically significant. The overall survival difference between the two arms, however, is no longer statistically significant. Cause of death outcome data are not available to assist in interpreting whether this phenomenon can be attributed to an excess of late deaths from tumour on the interferon arm (unlikely by virtue of the absence of a corresponding excess in relapses), to late toxicity of interferon prompting an increase in non-tumour deaths, or to chance imbalances in the factors that contribute to non-melanoma mortality (e.g. more smokers, more diabetics or more patients with an elevated baseline cholesterol in the interferon arm). No such factors were stratified for or even recorded in the study data to shed light on this particular issue.

E1690

A subsequent Intergroup adjuvant therapy trial, E1690, also compared high-dose IFN-α2b with observation after complete resection of all known disease.[16] This was a three-arm trial involving 608 eligible patients. The eligibility criteria were the same as for E1684, except for the fact that elective node dissection was not required for patients entered onto E1690 with thick primary melanomas and clinically negative nodes. Results of this trial confirmed the relapse-free survival advantage seen in E1684 but with no survival advantage observed.

E1694

In light of the discordant survival results in E1684 and E1690 as initially reported, the initial results of another Intergroup trial, E1694, have received intense scrutiny. This trial compared one year of high-dose interferon not with an observation control as in the two earlier studies, but rather with two years of a ganglioside vaccine called GMK. This was the largest of the three trials, with 774 eligible patients between two study arms. For the first time, staging of the lymph nodes by sentinel node biopsy was performed in a significant fraction of patients. Gangliosides are carbohydrate antigens found on the surface of melanoma cells, as well as normal cells of neural crest origin and tumour cells of other types. A pilot randomised trial suggested a relapse-free survival benefit in patients who were treated with purified ganglioside GM2 (the specific ganglioside in the GMK vaccine) plus BCG compared with those treated with BCG alone.[17] In May 2000, the E1694 trial's independent Data Safety Monitoring Committee concluded that the high-dose interferon arm was associated with highly significantly improved relapse-free and overall survival, and mandated that the study results be disclosed early.[18]

CLINICAL CONSIDERATIONS

RELAPSE-FREE SURVIVAL VERSUS OVERALL SURVIVAL

It has been the authors' experience that clinicians tend to view clinical trial results as dichotomous: that is, 'positive' or 'negative'. Moreover, particularly for adjuvant therapy trials, the acceptance of a clinical trial as 'positive' is often restricted to trials demonstrating a statistically significant benefit in overall survival. From this perspective, there seems to be an obvious discrepancy among the two observation-controlled trials: E1684 demonstrated seemingly striking benefits from the high-dose interferon regimen in both relapse-free and overall survival, whereas E1690 validated only the relapse-free survival benefit with no survival difference. However, the importance of relapse-free survival may be worth closer examination in the current setting.

From the statistician's standpoint, it is widely accepted that, compared with overall survival, disease relapse is a less objective endpoint because it depends on the definition of relapse as well as the frequency and method of detection. Defining relapse is less of an issue in the adjuvant setting since patients enter the study with no detectable disease and thereafter any new disease found is considered a relapse. In a well-conducted clinical trial the interval and method of disease assessment are specified in the protocol and generally complied with by trialists, thereby rendering relapse-free survival a somewhat more reliable endpoint than would be the case in other situations. From the purely clinical viewpoint, patients have made clear that they are willing to accept even toxic adjuvant therapies that provide improvements in relapse-free survival, even if they do not result in any prolongation of overall survival. This observation has been directly validated in melanoma patients,[19] and represents the perception that time spent without signs or symptoms of recurrent cancer is inherently of value even in the absence of prolongation of total lifespan. In addition, relapse-free survival often represents a truer reflection of the biologic activity of an adjuvant therapy since randomised trials rarely include rigorous controls on post-relapse salvage therapy. The confounding effect of such treatment on overall survival is unknown and not assessable.

RECONCILING THE STUDY RESULTS BASED ON CLINICAL CONSIDERATIONS

Two of the three randomised Phase III trials of high-dose interferon, E1684 and E1690, demonstrate a relapse-free survival advantage in their original publications. The third trial, E1694, also shows a relapse-free survival benefit but compared to GMK vaccine and not observation as the control treatment. The implication of this design difference is discussed in detail below. Nevertheless, many consider there is uniformity of evidence that high-dose interferon has biologic activity in at least delaying relapse after surgical therapy. This fact alone, combined with the lack of proven alternatives, is enough for many patients to choose interferon therapy in the absence of consensus regarding the overall survival benefit.

Crossover to interferon therapy upon relapse might have partially affected the outcome of at least one study. The original trial, E1684, was unlikely to have been affected by crossover for two reasons. Surgical staging of the regional nodes by complete (elective or therapeutic) node dissection was required. Hence, few patients were likely to experience regional relapse or other resectable recurrence, where secondary resection and delayed adjuvant interferon could be employed. Most relapses occurred in nonresectable distant sites. In recent medical practice, interferon is rarely employed for the treatment of measurable metastatic disease.

In contrast, the E1690 trial required only clinical staging of the regional nodes, and surgery was not required for patients with thick primary tumours and clinically negative nodes. Among all relapsed patients ($n = 114$ in the high-dose interferon arm and $n = 121$ in the observation control arm), 54% on high-dose interferon and 45% on observation experienced regional recurrence only. Retrospective data collection indicated more patients relapsing on the observation arm received subsequent interferon-α-containing regimens (31% vs. 15%) and/or biochemotherapy (17% vs. 6%).

While there is some evidence of differential post-relapse treatment received, concluding that the lack of interferon survival benefit observed in E1690 is due to these differences is not justified. Making this conclusion presupposes survival efficacy from these salvage therapies, which cannot be substantiated with currently available data. In addition, comparing outcomes by post-relapse treatment groups provides little useful information because patients were not randomised to salvage treatment strategies upon relapse. As is inherent in observational data, unknown patient selection factors cannot be accounted for by analysis techniques and their impact can easily remain even after adjusting for known prognostic factors. Therefore, although available data appear compatible with the notion that initial observation after surgery followed by high-dose interferon in case of resectable relapse presents an alternative strategy to routine use of adjuvant high-dose interferon, this study offers no proof for the conjecture. The conservative conclusion is that salvage treatment difference is a possible confounding factor that limits the confidence regarding the lack of overall survival benefit of high-dose interferon from study E1690.

STATISTICAL CONSIDERATIONS

Although clinical factors clearly impact on the interpretation of the three trials, our main goal is to examine the statistical aspects of these trials to determine the extent to which they actually present ‘conflicting’ information. We focus first on E1684 and E1690.

STATISTICAL TESTS EMPLOYED AND PRESENTATION OF RESULTS

One source of confusion could be due to the fact that one-sided p-values were presented for E1684 but two-sided p-values were presented for E1690. Unless otherwise specified, we use all two-sided p-values (p_2) in this discussion. In addition, all hazard ratios are expressed as observation arm versus treatment arm ratios. Thus, a hazard ratio >1 indicates an excess of hazard in the observation arm, or treatment advantage.

Another possible source of confusion could be the fact that, in E1684, statistically significant p-values for relapse-free and overall survival differences by the stratified log rank test (stratified for disease burden and presentation at initial diagnosis versus recurrent nodal disease status) were reported (Table 2 of Ref. 13). But when Cox regression analysis was performed, further adjusting for age, time from diagnosis to randomisation and ulceration status of the primary tumour, a significant interferon over observation benefit was presented only for those with nodal disease (Table 4 of Ref. 13). The hazard ratio for this patient subset was 1.64 for relapse-free survival and 1.49 for overall survival with $p_2 = 0.003$ and 0.02 respectively. However, these hazard ratios (presented in their reciprocals as interferon over observation ratios in actuality) were labelled ‘Treatment with IFN’ without reference to the positive nodal disease subset. An interaction term between the interferon treatment and the thick primary with no nodal disease patient group was actually included in the Cox models and the results were presented in the same table with the label ‘CS1/PS1 + IFN’. The hazard ratios were 0.36 and 0.34 respectively for relapse-free survival and overall survival. These interaction hazard ratios translated into observation over interferon hazard ratios of 0.59 and 0.50 for relapse-free and overall survival in patients with thick primary tumours and pathologically negative nodes, reflecting the occurrence that interferon-treated patients fared worse than the observation patients in this subset. For the readers who did not appreciate these details of the Cox modelling, the hazard ratios for the nodal disease subset could have been over-interpreted as the Cox model treatment effects for the study as a whole, which were not presented in the original publication. Such misinterpretation might have contributed to an exaggerated impression of the overall survival benefit from E1684.

TRIAL SIZE, OVERALL RESULTS AND OTHER ASPECTS

To interpret the combined results of E1684 and E1690, it is useful to compare the study parameters and overall results. Tables 7.1–7.4 are based mainly on Refs 15 and 16. Since there was no low-dose interferon arm in E1684, only the high-dose interferon and observation arms of E1690 are included in the tables. Due to the limitations of data availability, all randomised patients regardless of eligibility determination are presented for consistency.

The tables indicate that when E1690 results became available, the study had 50% more patients than E1684, reflecting the wider participation of the US Melanoma Intergroup. The patient enrolment periods were non-overlapping. Although the updated data for E1684 had longer follow-up at the time of E1690 publication, more events were analysed for E1690 from the larger sample size and the fact that few events occurred after five years. The main known patient characteristic difference was in the distribution of disease stage. There were more node-negative patients (26% vs. 11%) and fewer recurrent disease patients (63% vs. 50%) in E1690, representing a somewhat more favourable prognosis. It may be worth pointing

Table 7.1. E1684 and E1690 study characteristics

Study	E1684	E1690*
Participating groups	ECOG**	ECOG, SWOG***, CALGB****, MDACC*****
Patient accrual period	1984–90	1991–5
N (all randomised)	286	427

*High-dose interferon and observation arms only.
**Eastern Cooperative Oncology Group.
***Southwest Oncology Group.
****Cancer and Acute Leukemia Group B.
*****MD Anderson Cancer Center.

Table 7.2. E1684 and E1690 patient disease stage distribution

Disease stage	T4 N0	T1-4 N+ (occult)	T1-4 N+ (overt)	N+ Recurrent
E1684	11%	12%	14%	63%
E1690	26%	11%	12%	50%

Table 7.3. E1684 and E1690 results when first published

Study	Median follow-up (years)	Events	Hazard ratio	95% Confidence interval	*p*-value*
Relapse-free survival					
E1684	6.9	197	1.43	(1.08, 1.89)	0.005
E1690	4.3	241	1.28	(1.0, 1.65)	0.05
Overall survival					
E1684	6.9	175	1.32	(0.98, 1.77)	0.07
E1690	4.3	190	1.0	(0.75, 1.33)	1.00

*Two-sided *p*-value by stratified log rank test.

Table 7.4. E1684 and E1690 results with data updated to April 2001

Study	Median follow-up (years)	Events	Hazard ratio	95% Confidence interval*	*p*-value**
Relapse-free survival					
E1684	12.6	201	1.38	(1.05, 1.81)	0.02
E1690	6.6	248	1.24	(0.97, 1.59)	0.09
Overall survival					
E1684	12.6	188	1.22	(0.91, 1.63)	0.18
E1690	6.6	211	1.00	(0.71, 1.42)	0.98

*Estimated from hazard ratios and *p*-values presented in Kirkwood *et al.*[15]
**Two-sided *p*-value by stratified log rank test.

out that, among those with nodal disease, there did not appear to be survival differences between newly diagnosed and recurrent disease patients. The more favourable relapse and survival experiences of the E1690 patients compared with those in E1684 (all randomised patients: median relapse-free survival of 2.3 years vs. 1.4 years and median overall survival of 7.0 years vs. 3.2 years)[15] remain largely unexplained by known factors.

Regarding the treatment outcomes, the magnitude of the interferon effect was smaller in E1690 than in E1684. For data at original publications, the hazard ratio for relapse-free survival was 1.43 for E1684 vs. 1.28 for E1690, and it was 1.32 vs. 1.00 for overall survival. The overall survival benefit associated with high-dose interferon was of borderline statistical significance for E1684 and there was no survival difference in E1690. The outcomes of both of these trials were updated to April 2001, which resulted in small increases in event counts for both trials.[15] With the passage of time, the statistical significance of differences observed initially lessened: the relapse-free survival benefit originally seen in E1684 was the only remaining statistically significant treatment effect in the updated analysis of the two trials. There was no longer a significant relapse-free survival treatment effect in E1690 and, for E1684, there was not a significant overall survival effect despite the continued relapse-free survival prolongation. The authors commented that the diminished survival effect in E1684 could have been related to deaths from competing causes in an aging study group. Of the 111 E1684 survivors at the original analysis, 48 were from the observation arm and 63 from the interferon arm. Of these, 3/48 and 10/63 died before the April 2001 analysis. While certainly plausible, it is difficult to assess the validity of the death due to aging conjecture without the cause of death information. The latest combined results from these two trials seem to indicate that, for node-positive and thick primary, node-negative melanoma patients, there is still evidence that treatment with high-dose interferon prolongs relapse-free survival. Survival benefit, if it exists, would be more limited.

It is worth pointing out that E1690 was designed with not one but two primary comparisons, comparing high-dose interferon and low-dose interferon with observation (but not to each other) with a two-sided *p*-value of 0.025 for each comparison to maintain an overall Type I error rate of 0.05 for the study. When the results were presented, however, *p*-values less than 0.05 were treated as statistically significant for each comparison, representing a study-wide, two-sided error rate of 0.10. Also, per design the study was sized so that the power for each individual comparison was 0.83. In other words, should the true magnitude of benefit from both low- and high-dose interferon be the same as assumed by design, the power to detect both effects in the same study was approximately 0.83 × 0.83 or 0.69 for a study-wide Type II error rate of 0.31. With the inflated Type I error rate in the end, the overall power would increase somewhat but would likely remain less than adequate for detecting reasonable effects from both treatment arms. Hence, the question about the low-dose interferon regimen's treatment effect was essentially unanswered in this study, yet clinicians seem to have uniformly concluded, in part because of data from European studies, that low-dose interferon is inactive in E1690.

WHAT DOES E1694 TELL US?

E1694 was designed to detect a GMK vaccine benefit over interferon as the contemporary treatment standard. As is often practised with superiority designs, the trial would be stopped at planned interim analyses if the hypothesised vaccine benefit could be definitively ruled out. This provision was incorporated in the study design in the following manner. Instead of the typical, highly stringent interim *p*-value requirements, the GMK vaccine needed only to be inferior to interferon at a fixed, one-sided *p*-value of 0.05 for relapse-free survival in order to consider study termination at interim analyses. Such evidence might not establish the vaccine inferiority but would certainly rule out its superiority.

Considering the substantially more favourable vaccine toxicity profile, a more appropriate trial design might have sought to demonstrate the equivalence of the two agents in their efficacy rather than the superiority of the vaccine. In fact, instead of using the above, protocol-specified stopping rule, the Data and Safety Monitoring Committee in this case seemed to have followed the equivalence principle and disclosed the study results only when there was decisive evidence that the GMK vaccine was inferior to high-dose interferon in both relapse-free survival (one-sided *p*-value 0.0015) and overall survival (one-sided *p*-value 0.009).[18] Because no observation control arm was incorporated in the study design, the clinical interpretation of E1694 in this respect is subject to debate. Obviously, if it were known that the GMK vaccine had some level of clinical efficacy, the finding that high-dose interferon was significantly better in both disease-free and overall survival would be of great clinical significance and would substantiate the benefits identified in the initial E1684 trial. Without this knowledge, one has to maintain the possibility of a deleterious vaccine effect and abide by the principle that the study cannot be used to draw definitive conclusions regarding the non-design comparison of interferon versus observation.

To date, no credible evidence exists that the GMK vaccine is either beneficial or deleterious. It is likely that the GMK vaccine acted essentially as placebo and the study provided further validation that high-dose interferon was efficacious over no treatment in both relapse-free and overall survival. But we do not know this for certain. As the dramatic survival difference between E1684 and E1690 patients (median approximately 3.3 years vs. 5.9 years for observation patients in respective study) amply illustrates,[16] comparison of patient outcomes in the GMK vaccine arm with historical controls from the other two trials offers few clues to the efficacy of the vaccine.

Data were presented that, among the vaccine-treated patients, those displaying antibody responses had a trend towards favourable outcomes.[20] Assuming that the analyses corrected for the inherent responder versus non-responder comparison bias,[21] the results still cannot be used to establish a causal relationship between vaccine response and favourable outcome. As pointed out in numerous publications, response to treatment could simply serve as a selection mechanism wherein responders represented a better prognosis group. One may contend that it is difficult to reconcile a trend in favour of antibody responders with speculations of a deleterious effect of the vaccine resulting from production of 'blocking' antibodies. However, it is known that effects of prognostic factors such as disease stage can easily overwhelm any treatment effects. There will ultimately be more information on which to base an evaluation of the potential effects – good or bad – of the GMK vaccine. A randomised Phase III clinical trial in Europe compares treatment with this vaccine to observation among patients with Stage II melanoma.

DID ANY SUBSET OF PATIENTS BENEFIT MORE FROM INTERFERON?

The predominant subcategories of high-risk melanoma patients are those having thick primary tumours with clinically or pathologically negative nodes and those having any thickness melanoma with documented involvement of the nodes. Among the node-positive patients, subsets include those with 1, 2 to 3 and ≥4 nodes; patients with clinically evident versus microscopic nodal involvement; and patients found to have nodal involvement at the time of initial presentation versus those developing recurrent disease in the nodes.

The initial findings of E1684 indicated that the subset of patients with thick primary tumours and pathologically negative nodes had no benefit, and perhaps even a detrimental effect on relapse-free survival (observation vs. interferon hazard ratio of 0.59, as previously mentioned), from adjuvant interferon.[13] The veracity of this finding was called into question from the outset, because of the small number of node-negative patients (a total of 31 out of 280 eligible patients, or 11%) and an imbalance in a major prognostic factor (ulceration of the primary tumour) biasing the results in favour of the observation arm. In contrast, subset analysis of the results of trial E1690 found that the relapse-free survival benefit for patients with thick primary tumours and clinically negative nodes (making up 25% of the eligible patient population) was identical to that for the study population as a whole.[16] Subset analysis of E1694 showed the greatest interferon over vaccine benefit for the subset of thick, node-negative patients.[18]

Indeed, in each of the three clinical trials, subset analysis indicated a different group as obtaining the most benefit from high-dose interferon: the subset with one single positive node in E1684; the subset with two to three positive nodes in E1690; and the node-negative subset in E1694. The authors properly suggested that, taken together, there was no indication of preferential treatment effect in any one subset.[18] These results exemplify the lack of reliability of subset results, a phenomenon previously discussed in regard to other melanoma clinical trials.[22] Without appropriate study size for adequate power within subsets, and control for inflated Type I errors stemming from multiple testing, *post hoc* subset analyses suffer both high false-positive and high false-negative rates.

WHAT CAN META-ANALYSES TELL US?

When there are multiple trials investigating the 'same' treatment with results that are not always consistent, meta-analyses are often conducted in attempt to reach a conclusion. The term 'meta-analysis' here refers to formal statistical analysis of combined data from several trials. Since different trials are almost never the same with respect to patient inclusion criteria, treatment dose and schedule, primary outcome definition, and other trial conduct and data management practices, analysing the combined data as if they arose from a single trial is rarely justified. Therefore results from such attempts are often better viewed as hypothesis-generating rather than conclusive. For a general discussion of the principles and common pitfalls of meta-analysis, see Green *et al.*[23]

On the subject of interferon treatment for melanoma, we will comment on one meta-analysis paper and one systematic review article. Both included data from E1684 and E1690's original publications.[13,16] Wheatley *et al.*[24] performed a meta-analysis in the adjuvant setting on 11 trials comparing interferon with observation and one additional trial which had the same GMK vaccine treatment for the control arm and two interferon arms with different starting times. The authors concluded an interferon-α benefit for relapse-free survival with no clear evidence of an overall survival benefit. The patient populations of the 12 trials encompassed patients with primary tumour thickness as low as 1.5 mm and no clinical nodal involvement, to those with any degree of nodal involvement (all primary tumour thickness allowed) but no evidence of systemic metastases, to patients with distant but resectable disease. Recurrent nodal disease was disallowed in three trials. The first and perhaps the most positive trial for interferon (E1684) was the only trial that mandated complete regional lymphadenectomy for node-negative patients. Three of the trials included two interferon arms with different total doses and/or schedules. The planned total interferon doses ranged from 182 MU to 3500 MU among the trials, with treatment durations from three months to three years. As discussed below, with this much patient and treatment diversity, we do not think a meta-analysis is likely to be more meaningful than the results of any individual trial.

Assuming the analyses were performed accurately, since patients did not come from a single trial, confidence intervals calculated from the pooled meta-analysis would generally be overly narrow and precise, leading to an appearance of statistical significance which would not have the same interpretation as that derived from a single randomised trial. This would be true despite stratifying by trial in the pooled analyses. Obviously the treatment effect with a narrow confidence interval from the meta-analysis would not apply to a patient population as broad as one that included all trial populations. A single trial on such a heterogeneous population would produce larger variations in its results. This limitation is typically present in all meta-analysis attempts. Chi-square p-values of >0.05 from tests of heterogeneity would not represent outcome homogeneity and do not justify a pooled analysis from the statistical perspective because of the large number of trials included (e.g. 14, with two three-arm trials each included twice) and the tests' consequent lack of power to detect differences. Statistics aside, the interpretations of a pooled analysis resulting from trials with such diverse patient groups, many of which of less than adequate sizes, are necessarily unclear.

Similarly, with the tremendous variation in treatment doses and schedules, the meta-analysis provided no guidance on treatment recommendations. The authors tried to reduce the treatment differences by dividing the trials into dose groups and conducting separate meta-analyses within each group. Still, substantial patient population and treatment heterogeneity remained, even in the high-dose trials, as exemplified by the relapse-free and overall survival differences between E1684 and E1690, and treatment differences between these two trials (total dose of 3500 MU over 52 weeks) and NCCTG 83-7052 (total dose of 1350 MU over three months). The other trial in the high-dose group was the small E2696 trial ($n = 72$) with vaccine as part of the treatment for all three arms.

Regarding the analyses, without access to individual patient data, published or 'estimated' counts of disease relapses were analysed in place of the duration of relapse-free survival. By the authors' description under 'estimation of treatment effects', the Cochran–Mantel–Haenszel test[25] for combining independent 2×2 tables was used to pool the results from different trials. However, we were not able to reproduce Wheatley *et al.*'s p-values. For example, using the standard chi-square test for 2×2 tables, we obtained the two-sided p-values for disease relapse proportions from E1684, E1690, NCCTG 83-7052 and E2696 (i.e. the 'high-dose group') to be 0.03, 0.15, 0.31 and 0.13, respectively, as opposed to 0.005, 0.05, 0.2 and 0.03 presented in Figure 1 of the article. These smaller p-values might have been those from these studies' original publications where the time-to-event data for each patient were analysed.[13,16] Combining these four studies' binary relapse data presented in Wheatley *et al.* by the Cochran–Mantel–Haenszel procedure yielded a p-value of 0.003 instead of the 0.00009 presented by Wheatley *et al.* Therefore it is unclear to us which pooled analysis techniques were used. Lastly, from a data maturity perspective, the median follow-up ranged from 1.6 years to 6.9 years among the 12 trials and was less than 3.5 years for 4 trials, indicating less than mature data. As shown earlier in this chapter for E1684 and E1690, treatment effects stabilise only when the data reach sufficient maturity. Trial conduct and presentation quality were also not assessed. The latter would be important when the analysis was based on published results and not raw data. Generally speaking, meta-analysis efforts would benefit from sensitivity analyses where certain trials are down-weighted or left out. The authors concluded that the meta-analysis results provided clear evidence of interferon-α's relapse-free survival benefit due to reduced random errors and greater statistical reliability achieved by larger numbers of events. We, however, do not feel this meta-analysis resulted in a greater clarity regarding interferon-α's role as an adjuvant treatment for melanoma than is provided by analysis of individual trials.

In contrast, citing design heterogeneity as the reason, Lens and Daws[26] did not perform a meta-analysis but conducted a 'systematic review' of randomised trials comparing regimens with and without interferon as adjuvant treatments for melanoma. Patients must not have had distant metastatic disease and the trials had to have been designed as investigations of interferon-α as monotherapy using interferon-α only. Eight trials, all included in Wheatley *et al.* were selected. It is interesting to note that three of the trials presented as meeting these criteria by Wheatley *et al.* (EORTC 18 952, EORTC 18 871, DKG-80) were excluded. Trial characteristics were discussed in detail with conduct and presentation quality assessed in nine areas. To apply uniform treatment measuring scales, also without individual patient data, the authors derived summary statistics from published binary rates such as the five-year disease-free survival and overall survival rates and disease relapse proportions for each included trial. The statistics reported included relative risk reduction, absolute risk reduction, 'number needed to treat' and odds ratio. As comparisons of binary rates, it is well known that these statistics are less powerful than procedures based on time-to-event data such as the log rank test. Therefore, negative findings by the authors for an individual study would not invalidate positive findings from the original analysis if the latter used more powerful tests. In fact, it might have been helpful had the original analyses results been included in the review.

In summary, while informed literature review can be very helpful, we quote Green *et al.* that 'in settling therapeutic issues, a meta-analysis is a poor substitute for one large, well-conducted trial . . . the expectation that a meta-analysis will be done does not justify designing studies that are too small to detect realistic differences with adequate power'.[23]

ADJUVANT VACCINE CLINICAL TRIALS

CANVAXIN

Allogeneic vaccines are composed of intact or modified melanoma cells selected for the presence of shared antigens found on a large percentage of melanomas. They may be more inherently recognisable to the patient's immune system than an autologous cell preparation. Allogeneic vaccines are much more amenable to standardisation and large-scale manufacture than autologous vaccines, and hence are more readily evaluated in large-scale randomised trials. One allogeneic vaccine that has now been evaluated in two randomised Phase III trials is Canvaxin, a polyvalent irradiated melanoma vaccine originally developed by Dr Donald Morton.[27,28] This vaccine has been studied in randomised Phase III trials in two groups of patients at high risk of recurrence after surgery, those with resected Stage III and resected Stage IV melanoma. But prior to the initiation of those studies, a wealth of non-randomised data was acquired that strongly suggested the value of Canvaxin as a post-surgical adjuvant therapy.

Morton and colleagues treated 935 resected Stage III melanoma patients with Canvaxin and compared their outcome with a control group of 1667 patients who received similar surgical therapy but no vaccine. Median and five-year survival rates were significantly higher for the vaccine-treated patients than the controls (56.4 months versus 31.9 months median and 49% versus 37% alive at five years, respectively; $p = 0.0001$).[29] This apparent benefit persisted even after matching vaccine patients to controls for up to seven known prognostic variables.

Similar benefit was seen in non-randomised studies with this vaccine in patients with resected Stage IV melanoma.[30]

Recently, the Data Safety Monitoring Board overseeing these two studies determined that the trials were sufficiently unlikely to result in a determination of vaccine efficacy so each trial was ended and all protocol treatments were discontinued. Any single randomised trial, if conducted exactly according to the study protocol under conditions that closely mirror the original statistical assumptions of the study designers, has a 5% (1 in 20) chance of being falsely positive, but usually between a 10% and 20% chance of yielding a false-negative result when in fact a difference of the anticipated magnitude does exist. That is because most studies are designed with a statistical power of between 80% and 90%, and the value of 1 minus the statistical power is the likelihood of a false-negative trial result. If in fact the underlying truth or the conditions of accrual do not match the prespecified expectations, the likelihood of a false-negative study can rise dramatically higher. If, as in the case of Canvaxin, two clinical trials are conducted testing the same agent in similar (albeit not identical) study populations, the likelihood that *both* would be falsely negative due to chance alone decreases to only 4% (1 in 25 pairs of clinical trials), assuming both studies are conducted at the 80% power level. Were both studies to be conducted at the 90% power level, the likelihood of both being falsely negative by chance alone would only be 1% (1 per 100 trial pairs), even though the chance of a single trial being falsely negative was 10%. Thus, it seems unlikely but by no means impossible that the two vaccine trials were both falsely negative on the basis of chance alone. If subsequent data reveal that the underlying assumptions used to formulate the power calculations for these two studies were flawed, however, this possibility may need to be revisited. In the absence of a fuller understanding of the specific study results, it must be concluded that the non-randomised data that indicated such a strong vaccine effect actually more likely reflected a strong selection bias for the most favourable patients to be treated with the vaccine.

CONCLUSIONS

While randomised clinical trials remain the gold standard for assessing the benefit of new therapies in melanoma and most other diseases, our discussion illustrates the inescapable fact that the results of randomised trials do not end all controversies regarding therapeutic approaches. They clearly teach us that non-randomised data, especially retrospective data, are best relied upon to select therapies for evaluation in randomised trials – not as a substitute for such trials. Even the most sophisticated manipulations of retrospective data cannot compensate for unseen but systematic biases between the groups chosen for treatment with a new therapy and the historical or concurrent 'controls' not selected to receive that therapy. On the other hand, randomised trials, even of moderately large size, have major limitations in their ability to resolve subset differences – especially subsets that cannot be identified prior to randomisation.

Three randomised trials evaluating high-dose interferon, involving over 1600 patients, have been conducted in the United States, yet its treatment benefit remains controversial. The combined evidence to date suggests that, for high-risk melanoma patients, treatment with high-dose interferon is likely to prolong relapse-free survival. Evidence for a survival benefit is far less certain. There is no credible evidence to suggest that interferon exerts a differential effect in different subsets of 'high-risk' patients.

There are many reasons why high-dose interferon has not been uniformly embraced by physicians and patients around the world, even though it is the only adjuvant therapy yet shown to have any sustained impact on relapse-free survival. When the three trials are looked at in the light of statistical principles, what seem to be glaring differences are more plausibly regarded as understandable variations reflecting trial design and analysis, combined with the fluctuations inherent in human clinical trials conducted over time in similar yet subtly different patient populations.

While it is easy to conclude that further research is necessary to determine if high-dose interferon-α2b improves overall survival, there is in fact little chance that definitive further research will take place. Only one current clinical trial, the Sunbelt Melanoma Trial, is comparing one year of high-dose interferon with a control group. This study includes only patients with a single positive sentinel node identified at the time of initial presentation.[31] As such, it is composed of a far more homogeneous patient population than any prior clinical trial, potentially enhancing the scientific validity. Of note, this group now constitutes by far the largest fraction of 'high-risk' melanoma patients being seen and treated in the United States today, yet less than 10% of participants in the three prior trials combined were from this category. Unfortunately, this trial is likely to be small compared with the most recent Intergroup trials and, regardless of the results, it will not directly address the role of interferon in all of the other high-risk categories.

It is now over 20 years since the design of clinical trial E1684, and more than a decade since the FDA's approval of high-dose interferon-α for the adjuvant therapy of high-risk melanoma, and we may never fully know to what extent this toxic and inconvenient regimen improves overall survival. The implications of that statement are profound, and the burden they place on clinical trialists is clear: design and analyse your trials carefully to have the greatest probability of a clear and unambiguous result.

REFERENCES

1. Balch CM, Soong SJ, Gershenwald JE *et al*. Prognostic factors analysis of 17,600 melanoma patients: validation of the American Joint Committee on Cancer melanoma staging system. *J Clin Oncol* (2001) **19**: 3622–34.
2. Veronesi U, Cascinelli N. Narrow excision (1-cm margin): a safe procedure for thin cutaneous melanoma. *Arch Surg* (1991) **126**: 438–41.
3. Balch CM, Soong S, Ross MI *et al*. Long-term results of a multi-institutional randomized trial comparing prognostic factors and surgical results for intermediate thickness melanomas (1.0 to 4.0 mm). Intergroup Melanoma Surgical Trial. *Ann Surg Oncol* (2000) **7**: 87–97.
4. Wagner JD, Schauwecker D, Davidson D *et al*. Prospective study of fluorodeoxyglucose-positron emission tomography imaging of lymph node basins in melanoma patients undergoing sentinel node biopsy. *J Clin Oncol* (1999) **17**: 1508–15.
5. Starritt EC, Uren RF, Scolyre RA *et al*. Ultrasound examination of sentinel nodes in the initial assessment of patients with primary cutaneous melanoma. *Ann Surg Oncol* (2005) **12**: 18–23.
6. Balch CM. The role of elective lymph node dissection in melanoma: rationale, results, and controversies. *J Clin Oncol* (1988) **6**: 163–72.
7. Cascinelli N, Morabito A, Santinami M, MacKie RM, Belli F. Immediate or delayed dissection of regional nodes in patients with melanoma of the trunk: a randomised trial. WHO Melanoma Programme. *Lancet* (1998) **351**(9105): 793–6.
8. Morton D, Wen D, Wong J *et al*. Technical details of intraoperative lymphatic mapping for early stage melanoma. *Arch Surg* (1992) **127**: 392–9.
9. Morton DL, Thompson JF, Essner R *et al*. Validation of the accuracy of intraoperative lymphatic mapping and sentinel lymphadenectomy for early-stage melanoma: a multicenter trial. Multicenter Selective Lymphadenectomy Trial Group. *Ann Surg* (1999) **230**: 453–65.
10. Gershenwald JE, Thompson W, Mansfield PF *et al*. Multi-institutional melanoma lymphatic mapping experience: the prognostic value of sentinel lymph node status in 612 stage I or II melanoma patients. *J Clin Oncol* (1999) **17**: 976–83.
11. Morton DL, Thompson JF, Cochran AJ *et al*. Interim results of the Multicenter Selective Lymphadenectomy Trial (MSLT-I) in clinical stage I melanoma. *J Clin Oncol* (2005) **23**: 710s.
12. Sondak VK, Wolfe JA. Adjuvant therapy for melanoma. *Curr Opin Oncol* (1997) **9**: 189–204.
13. Kirkwood JM, Strawderman MH, Ernstoff MS *et al*. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group trial EST 1684. *J Clin Oncol* (1996) **14**: 7–17.
14. Cole BF, Gelber RD, Kirkwood JM *et al*. Quality-of-life–adjusted survival analysis of interferon alfa-2b adjuvant treatment of high-risk resected cutaneous melanoma: an Eastern Cooperative Oncology Group study. *J Clin Oncol* (1996) **14**: 2666–73.
15. Kirkwood JM, Manola J, Ibrahim J *et al*. A pooled analysis of ECOG and Intergroup trials of adjuvant high-dose interferon for melanoma. *Clin Cancer Res* (2004) **10**: 1670–7.
16. Kirkwood JM, Ibrahim JG, Sondak VK *et al*. High- and low-dose interferon alfa-2b in high-risk

melanoma: first analysis of Intergroup trial E1690/S9111/C9190. *J Clin Oncol* (2000) **18**: 2444–54.

17. Livingston PO, Wong GYC, Adluri S *et al*. Improved survival in stage III melanoma patients with GM2 antibodies: a randomized trial of adjuvant vaccination with GM2 ganglioside. *J Clin Oncol* (1994) **12**: 1036–44.
18. Kirkwood JM, Ibrahim JG, Sosman JA *et al*. High-dose interferon alfa-2b significantly prolongs relapse-free and overall survival compared with the GM2-KLH/QS-21 vaccine in patients with resected stage IIB–III melanoma: results of Intergroup trial E1694/S9512/C509801. *J Clin Oncol* (2001) **19**: 2370–80.
19. Kilbridge KL, Weeks JC, Sober AJ *et al*. Patient preferences for adjuvant interferon alfa-2b treatment. *J Clin Oncol* (2001) **19**: 812–23.
20. Kirkwood JM, Ibrahim J, Lawson DH *et al*. High-dose interferon alfa-2b does not diminish antibody response to GM2 vaccination in patients with resected melanoma: results of the multicenter Eastern Cooperative Oncology Group phase II trial E2696. *J Clin Oncol* (2001) **19**: 1430–6.
21. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol* (1983) **1**: 710–19.
22. Sondak VK. Multi-institutional melanoma vaccine trial. *Ann Surg Oncol* (1996) **3**: 588–9.
23. Green S, Benedetti J, Crowley J. *Clinical Trials in Oncology*, 2nd edn. Boca Raton, FL: Chapman and Hall/CRC (2003).
24. Wheatley K, Ives N, Hancock B *et al*. Does adjuvant interferon-α for high-risk melanoma provide a worthwhile benefit? A meta-analysis of the randomised trials. *Cancer Treat Rev* (2003) **29**: 241–52.
25. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* (1959) **22**: 719–48.
26. Lens MB, Dawes M. Interferon alfa therapy for malignant melanomas: a systematic review of randomized controlled trials. *J Clin Oncol* (2002) **20**: 1818–25.
27. Hsueh EC, Morton DL. Antigen-based immunotherapy of melanoma: Canvaxin therapeutic polyvalent cancer vaccine. *Semin Cancer Biol* (2003) **13**: 401–7.
28. Motl SE. Technology evaluation: Canvaxin, John Wayne Cancer Institute/Cancer Vax. *Curr Opin Mol Ther* (2004) **6**: 104–11.
29. Morton DL, Hsueh EC, Essner R *et al*. Prolonged survival of patients receiving active immunotherapy with Canvaxin therapeutic polyvalent vaccine after complete resection of melanoma metastatic to regional lymph nodes. *Ann Surg* (2002) **236**: 438–49.
30. Hsueh EC, Essner R, Foshag LJ *et al*. Prolonged survival after complete resection of disseminated melanoma of patients and active immunotherapy with a therapeutic cancer vaccine. *J Clin Oncol* (2002) **20**: 4549–54.
31. McMasters KM, Noyes RD, Reintgen DS *et al*. Lessons learned from the Sunbelt Melanoma Trial. *J Surg Oncol* (2004) **86**: 212–23.

8

Respiratory Cancers

JOAN H. SCHILLER[1] AND KYUNGMANN KIM[2]
[1]Division of Hematology/Oncology, University of Texas Southwestern Medical Center, Dallas, Texas, USA
[2]Department of Biostatistics & Medical Informatics, University of Wisconsin–Madison, Wisconsin, USA

INTRODUCTION

Carcinoma of the lung is responsible for more than 165 000 deaths each year in the United States. This represents one-third of all deaths due to cancer, and more than the number of deaths due to breast, colon and prostate cancers combined. The incidence of the disease continues to rise, particularly in women and blacks, and thus is likely to present a significant public health problem for years to come.

Lung cancer consists of four major histological types: adenocarcinoma, squamous cell carcinoma, large-cell carcinoma and small-cell carcinoma. Because of the unique biological features of small cell lung cancer (SCLC), its staging and treatment differ radically from the other three types of lung cancer, which collectively are called non-small-cell lung cancer (NSCLC).

Clinical trials have resulted in significant seminal trials which have resulted in changes in the management of these patients as described below.

SCREENING

Three US randomised screening studies failed to detect an impact of screening high-risk patients with chest radiographs or sputum cytology on mortality, although earlier stage cancers were detected in the screened groups.[1–3] These studies have been criticised for a number of potential methodological and statistical problems, such as over-diagnosis, and analysing data by survival rather than mortality.[4]

Recently, several clinical studies have demonstrated that early stage lung cancers can be detected with the use of spiral CT that would not have been detected by routine chest X-ray.[5] Spiral CT is a CT scan which does not evaluate the mediastinum, employs low doses of radiation, and can be completed within one patient 'breath'. Because it can be done rapidly, and does not require a radiologist to be present, it is being used in some centres to screen for lung cancers in high-risk populations. Preliminary results suggest that those patients whose disease is discovered while still asymptomatic and still small have an improved survival compared with those patients who are not screened and who present with symptoms.[6,7] However, it has not been determined if the early detection of small tumours results in improved survival that is not a result

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

of lead time or length time bias.[4,8] In addition, many patients (over 50%) have abnormalities on their screening CT of unclear significance. The work-up and follow-up of these non-diagnostic abnormalities may be prohibitive. The National Cancer Institute has recently completed a large randomised study comparing spiral CT vs. chest radiograph which should help answer the question of the role of spiral CTs in screening high-risk patients for lung cancer.[9]

NON-SMALL-CELL LUNG CANCER

The prognosis and treatment of NSCLC are dependent primarily on stage of disease at the time of diagnosis. Stage, in turn, is dependent upon the size of the tumour (T), location of nodal metastases (N), if any, and presence or absence of distant metastases (M). The current TNM staging classification is shown in Table 8.1, and the stage grouping in Table 8.2.[10]

STAGE I DISEASE

A lobectomy is the treatment of choice for Stage I NSCLC, with cure rates of 60%–80% reported. Within Stage I, patients with T2, N0 disease do not fare as well as those with T1, N0 cancers. In approximately 20% of patients with

Table 8.1. TNM staging

Primary tumour (T)	
TX	Tumour proven by the presence of malignant cells in bronchopulmonary secretions but not visualised roentgenographically or bronchoscopically, or any tumour that cannot be assessed as in a retreatment staging
T0	No evidence of primary tumor
Tis	Carcinoma *in situ*
T1	A tumour that is 3.0 cm or less in greatest dimension, surrounded by lung or visceral pleura, and without evidence of invasion proximal to a lobar bronchus at bronchoscopy
T2	A tumour more than 3.0 cm in greatest dimension, or a tumour of any size that either invades the visceral pleura or has associated atelectasis or obstructive pneumonitis extending to the hilar region. At bronchoscopy, the proximal extent of demonstrable tumour must be within a lobar bronchus or at least 2.0 cm distal to the carina. Any associated atelectasis or obstructive pneumonitis must involve less than an entire lung
T3	A tumour of any size with direct extension into the chest wall (including superior sulcus tumours), diaphragm, or the mediastinal pleura or pericardium without involving the heart, great vessels, trachea, esophagus or vertebral body, or a tumour in the main bronchus within 2 cm of the carina without involving the carina
T4	A tumour of any size with invasion of the mediastinum, heart, great vessels, trachea, oesophagus, vertebral body or carina; presence of malignant pleural effusion; or a satellite nodule within the same lobe
Nodal involvement (N)	
NX	Regional lymph nodes cannot be assessed
N0	No demonstrable metastasis to regional lymph nodes
N1	Metastasis to lymph nodes in the peribronchial or the ipsilateral hilar region, or both, including direct extension
N2	Metastasis to ipsilateral mediastinal lymph nodes and subcarinal lymph nodes
N3	Metastasis to contralateral mediastinal lymph nodes, contralateral hilar lymph nodes, ipsilateral or contralateral scalene or supraclavicular lymph nodes
Distant metastasis (M)	
M0	No distant metastasis
M1	Distant metastasis, including pulmonary nodule not in the same lobe as the primary tumour

Table 8.2. The 1997 revisions to the International Staging Classification for lung cancer

		Five-year survival (%)	
Stage	TNM subset	Clinical stage	Pathological stage
IA	T1, N0, M0	61	67
IB	T2, N0, M0	38	57
IIA	T1, N1, M0	34	55
IIB	T2, N1, M0 T3, N0, M0	24	39
IIIA	T3, N1, M0 T1-3, N2, M0	9	25
IIIB	T4, any N, M0 any T, N3, M0	13	23
IV	Any T, any N, M1	1	–

medical contraindications to surgery but with adequate pulmonary function, high-dose radiotherapy will result in cure. No role of adjuvant chemotherapy for Stage I NSCLC has been identified. Patients with a resected Stage I NSCLC are at high risk for the development of second lung cancers (approximately 1%/year), prompting a number of ongoing clinical trials looking at the role of chemoprevention. Surprisingly, several randomised studies have demonstrated that the use of Vitamin A or one of its derivatives at best does not prevent lung cancer in smokers, and, at worst, may increase the risk of developing it.[11–13] Preliminary studies have suggested that selenium may reduce the incidence of lung cancer, and total cancer mortality. In a multi-centre, double-blind, randomised, placebo-controlled trial, 1312 patients were randomised to receive either selenium or placebo. The selenium group had fewer total carcinomas, including lung cancer ($\mathrm{RR} = 0.54$; 95% confidence interval $= 0.30 - 0.98$; $p = 0.04$).[14] This has formed the basis for an intergroup chemoprevention trial which is now ongoing.

STAGE II AND 'NON-BULKY' IIIA DISEASE

Treatment of locally advanced NSCLC is one of the most controversial issues in the management of lung cancer. Treatment options include surgery for less advanced disease, or radiotherapy, either of which has been given with or without chemotherapy for control of micrometastases. Interpretation of the results of clinical trials involving patients with locally advanced disease has been clouded by a number of issues, including changing diagnostic techniques, different staging systems, and heterogeneous patient populations that may have disease that ranges from 'non-bulky' Stage IIIA (clinical N1 nodes, with N2 nodes discovered only at the time of surgery or mediastinoscopy), to 'bulky N2' nodes (enlarged adenopathy clearly visible on chest X-ray films or multiple nodal-level involvement), to clearly inoperable Stage IIIB disease.

Post-operative Thoracic Radiotherapy

The treatment for Stage II and selected IIIA NSCLC patients is surgical resection. However, many of these patients will relapse, prompting numerous trials evaluating the role of post-operative radiotherapy or chemotherapy. A meta-analysis examining the role of post-operative radiotherapy (PORT) found that patients randomised to receive PORT actually had an inferior survival to those randomised to observation alone.[15] In a meta-analysis of 2128 patients in nine clinical trials of PORT, a 7% survival decrement from radiation was identified. However, this particular analysis included a number of trials from the 1960s and 1970s when staging was highly inaccurate and relatively outmoded radiation therapy technologies were utilised. In addition, several of the trials included in this report aggressively treated patients with no evidence of nodal involvement or those with early nodal involvement only, a group that by today's standards would not be subjected to PORT. More recent studies looking at the role of PORT have concluded that it does not prolong survival, but does enhance local control. The most comprehensive randomised trial in this regard was performed by the Lung Cancer Study Group, and it demonstrated major improvement in intrathoracic disease control.[16] For those patients receiving thoracic radiotherapy, intrathoracic failure rate was only 3%, compared with 43% for patients not receiving PORT, although no significant survival advantage was identified.

Adjuvant Chemotherapy

Given the propensity of these resected patients to relapse with distant disease, adjuvant post-operative chemotherapy has been of significant interest. A meta-analysis published in 1995 found a small improvement in survival with post-operative adjuvant chemotherapy that borderlines on statistical significance ($p = 0.08$).[17] Although several other randomised studies have also been negative,[18–20] three other positive randomised studies were recently reported for patients with Stage IB or Stage II disease.[21–23] It is not clear as to why three trials were negative and three positive, although possible reasons include more homogeneity of the patient population in the positive studies, and differences in post-operative radiation and chemotherapy treatment and regimens.

Preoperative Chemotherapy Plus Surgery

There have been two small randomised studies involving surgery with or without pre-operative chemotherapy which popularised this approach. Both involved 60 patients, and both report response rates of 35%–62% following induction chemotherapy. Both have also reported prolonged survival, prompting early closure of both trials. In the European trial, the median survival time was 26 months for patients receiving pre-operative chemotherapy plus surgery, compared with 8 months for patients treated with surgery alone.[24] In the MD Anderson trial, the median survival of the 32 patients randomised to the surgery-alone group was 11 months compared with 64 months in the 28 patients randomised to the combined-modality arm.[25] Of note, however, updated results of the MD Anderson trial, while still statistically significant, showed a narrowing of the survival curves, with a median survival of 14 months and 21 months for the surgery alone and combined modality arms, respectively.[26]

A larger trial has recently been reported, in which 355 patients with Stage I, II or IIIA disease were randomised to three cycles of chemotherapy followed by surgery, or to surgery alone.[27] Median survival (37 months and 26 months, respectively) and two-year survival (52% and 59%, respectively) were not statistically different between the two groups. However, a subset analysis in which patients who died with 150 days of peri-operative problems were excluded revealed a 0.77 reduction in risk which was statistically significant ($p = 0.03$). Other subset analysis looked at outcome by patient stage, and found that the patients with N0/N1 disease who received chemo/surgery had a hazard ratio of 0.68, compared with patients with N2 disease, where the hazard ratio was 1.04.

An intergroup study evaluating chemo/RT vs. chemo/RT followed by surgery in patients with pathologically confirmed N2 nodes has recently been completed; preliminary results suggest no difference in overall survival, although chemo/RT followed by surgery yielded superior progression-free survival (PFS) (log rank $p = 0.02$): median PFS, 14.0 vs. 11.7 months; three-year PFS, 29% vs. 19%.[28] There were more early non-cancer deaths in the surgery arm, but overall survival curves cross at the median (22.1 months for the chemo/RT/surgery arm vs. 21.7 months for the chemo/RT arm), so that by year 3 survival favoured the chemo/RT/surgery arm (38% vs. 33%). Longer follow-up is needed to determine if surgery significantly prolongs overall survival in IIIA (pN2) NSCLC.

LOCALLY ADVANCED ('BULKY') STAGE IIIA/IIIB DISEASE

The optimal treatment for bulky Stage IIIA and Stage IIIB disease is also controversial. Current investigational efforts are directed at identifying the optimal combined-modality approach, involving treatments directed at local control of the disease (surgery or radiotherapy) and micrometastatic disease (chemotherapy). Possibilities include radiotherapy only, preoperative chemotherapy or chemotherapy plus radiotherapy.

Chemotherapy Plus Radiation Therapy

Chemotherapy plus radiotherapy is the treatment of choice for patients with bulky or inoperable Stage III disease. Two randomised studies have demonstrated an improvement in median and long-term survival with chemotherapy followed by radiation therapy versus radiotherapy alone.[29,30] More recently, two randomised trials have shown that concurrent chemoradiotherapy results in prolonged survival (albeit at the expense of enhanced toxicity) compared with sequential treatment.[31,32] Other active areas of investigation include choice of chemotherapy, fractionation and treatment fields.

Recently, weekly, low-dose 'sensitising' chemotherapy plus radiation therapy has become popular, primarily due to lower toxicities when administered with radiotherapy than standard dose chemotherapy.[33] A Phase II study randomised (1) two cycles of induction paclitaxel/carboplatin followed by thoracic radiotherapy (TRT) with (2) two cycles of the same induction chemotherapy followed by weekly, low-dose paclitaxel/carboplatin and concurrent TRT with (3) concurrent chemo/TRT followed by two cycles of consolidation chemotherapy. Median survival was highest in the concurrent/consolidation arm. However, a Phase III CALGB study in which patients with Stage III NSCLC were randomized to concurrent chemo/RT or induction paclitaxel/carboplatin followed by concurrent chemo/RT reported median survival of 11.4 months for the concurrent arm versus 11.4 months for the induction/concurrent arm ($p = 0.154$).[34] One-year survival estimates are 48% (41%–57%) and 54% (47%–62%) respectively. The median survival achieved in each of the treatment groups was low compared with the randomised Phase II results, illustrating the importance of randomised Phase III studies.

STAGE IV DISEASE

Several meta-analyses have demonstrated that chemotherapy improves survival in patients with metastatic NSCLC (approximately 10% one-year survival untreated vs. 35%–40% one-year survival with treatment),[35,36] particularly if the chemotherapy is platin-based.[17] In the past 10 years, numerous different cytotoxic drugs have become available for the treatment of lung cancer patients. These include, among others, vinorelbine, the taxanes (docetaxel, paclitaxel), gemcitabine and the topoisomerase I inhibitors (irinotecan and topotecan). In general, these studies have shown that third-generation drugs improve survival over older regimens.[37] The combination of two drugs is superior to one drug, but three drugs are not superior to two in prolonging survival.[38–43] However, there is probably little difference in outcome between agents when combined with cisplatin, although there are clear differences in toxicity and cost.[44,45] Debate continues as to whether cisplatin is superior to carboplatin in the treatment of advanced disease,[46] and whether platin-based chemotherapy is superior to non-platin-based.[47]

SECOND-LINE CHEMOTHERAPY

Docetaxel was approved for the second-line treatment of NSCLC based upon two clinical trials. One trial compared two doses of docetaxel with best supportive care, and found an improvement in median and long-term survival, despite a low response rate of 7%.[48] The other trial compared docetaxel with either vinorelbine or ifosfamide (the treatment physician was allowed to choose), and found an improvement in long-term, although not median, survival.[49] Pemetrexed was recently approved for the treatment of second-line NSCLC based on a randomised non-inferiority study which showed that it was not inferior to docetaxel, but was less toxic.[50]

'TARGETED' THERAPIES

Given the overall poor results with standard cytotoxic therapies, and the number of advances that have been made recently in our understanding of the biology of cancer, a strong interest has emerged in targeting pathways unique to neoplastic cells. One such example is the epidermal growth factor receptor (EGFR), which has been found to be expressed in the majority of patients with lung cancer. Two drugs which have been approved for inhibition of this pathway are gefitinib (Iressa) and erlotinib (Tarceva). These drugs are small molecules which inhibit activation of the EGFR by inhibiting phosphorylation of the tyrosine kinase residues on the internal domain of the receptor, and hence are known as EGFR TKIs. Recently, mutations at the site in which these drugs bind in the EGFR have been identified, and predict very strongly for response.[51,52] These mutations are most commonly found in non-smoking women with a subtype of adenocarcinoma called bronchioalveolar carcinoma, particularly of Asian background.

A Phase III trial of erlotinib in patients with second- or third-line disease demonstrated prolonged survival compared with best supportive care alone,[53] whereas somewhat surprisingly gefitinib did not. Both drugs reduce symptoms and improve quality of life, even in patients who do not have an objective response. Neither drug has shown a survival benefit when combined with chemotherapy, despite pre-clinical studies demonstrating otherwise.[54–56] Reasons for this are unclear, but include patient selection issues as well as potentially antagonist effects secondary to schedule-dependent interaction between the EGFR TKIs and G2/M blocking chemotherapeutic agents, such as the taxanes.[57] Clinical trials are ongoing to determine which patient populations are most likely to respond, and the role of these agents in the metastatic, locally advanced and adjuvant settings. Other ongoing clinical trials in lung cancer involve humanised antibodies to the EGFR or TKIs which are irreversible or inhibit more than one member of the Her2neu family of receptors.

Although patients with EGFR mutations are most likely to respond to these drugs, it is not clear as to whether a survival benefit is likely to be limited to these patients. Statistical modelling of the Phase III erlotinib trial suggests that the number of the responding patients were unlikely to account for all of the survival benefit, suggesting that even patients with stable disease may benefit. This, in turn, would suggest that some patients without mutations may also derive a survival benefit without experiencing a classic 50% or more shrinkage of tumour. The same kind of statistical modelling can be used to assess whether patients without mutations may also derive a survival benefit. Implication of molecular heterogeneity of lung cancer for design of clinical trials in terms of sample size requirements will be discussed later.

Bevacizumab, the monoclonal antibody to vascular endothelial growth factor (VEGF) ligand, also appears particularly promising as a targeted therapy for lung cancer. The addition of bevacizumab to fluorouracil-based combination chemotherapy resulted in a statistically significant and clinically meaningful improvement in survival among patients with metastatic colorectal cancer when compared with irinotecan, bolus fluorouracil and leucovorin (IFL) alone.[58] A randomised Phase II trial that compared chemotherapy with or without bevacizumab was recently reported in previously untreated NSCLC patients. Compared with the control arm, treatment with carboplatin and paclitaxel plus bevacizumab (15 mg/kg) resulted in a higher response rate (31.5% vs. 18.8%), longer median time to progression (7.4 vs. 4.2 months) and a modest increase in survival (17.7 vs. 14.9 months). Of the 19 control patients that crossed over to single-agent bevacizumab, five experienced stable disease, and one-year survival was 47%. Bleeding was the most prominent adverse event and was associated with squamous cell histology, tumour necrosis and cavitation, and disease location close to major blood vessels.[59] A Phase III study of carboplatin and paclitaxel with or without bevacizumab in non-squamous NSCLC was recently completed by the Eastern Cooperative Oncology Group (ECOG); results are pending.

A number of other anti-angiogenic drugs are under development. These include thalidomide, currently being studied in a Phase III study by ECOG, antibodies to the VEGF receptor (VEGF TRAP), and small-molecule tyrosine kinase inhibitors of one or more of the VEGF receptors. The majority of these trials are completing Phase I studies and are being moved into the Phase II clinical trial arena. Some of the TKIs inhibit not only the VEGF receptor, but also other signalling pathways, such as EGF, platelet-derived growth factor (PDGF) or c-kit receptors.

Clearly, these targeted agents have promising activity in the treatment of NSCLC. Continued efforts need to be made at identifying optimal ways of combining them with other targeted agents or other cytotoxic drugs, as well as identifying which patient populations are most likely to benefit.

SMALL-CELL LUNG CANCER

SCLC differs from NSCLC in a number of important ways. First, it has a more rapid clinical course and natural history, with the rapid development of metastases, symptoms and death. Untreated, the median survival time for patients with local disease is typically 12–15 weeks, and for those with advanced disease, 6–9 weeks. Second, it exhibits features of neuroendocrine differentiation in many patients (which may be distinguishable histopathologically), and is associated with paraneoplastic syndromes. Third, unlike NSCLC, SCLC is exquisitely sensitive to both chemotherapy and radiotherapy, although resistant disease often develops. Because of the rapid development of distant disease and its extreme sensitivity to the cytotoxic effects of chemotherapy, this mode of therapy forms the backbone of treatment for this disease.

FIRST-LINE THERAPY

A number of combination chemotherapeutic regimens are available for SCLC. With these chemotherapy regimens, overall response rates of 75%–90% and complete response rates of 50% for localised disease can be anticipated. For extensive-stage disease, overall response rates of about 75% with complete response rates of 25% are common. Despite these high response rates, however, the median survival time remains about 14 months for limited-stage disease, and 7–9 months for extensive-stage disease. Less than 5% of extensive-stage patients have long-term survival (>2 years). As with NSCLC, three drugs do not appear to be more effective than two.[60]

A Phase III randomised trial was recently reported, in which patients with SCLC were randomised to the control arm of etoposide and cisplatin, versus cisplatin and the topoisomerase I inhibitor, irinotecan (CPT11).[61] Median survival and one-year survival were 420 days and 60% in the cisplatin/irinotecan arm, and 300 days and 40% in cisplatin/etoposide arm. If ongoing Phase III studies confirm these results, cisplatin/irinotecan would become the first combination chemotherapy to improve survival in SCLC patients in decades.

SECOND-LINE THERAPY

No curative regimens for patients with recurrent disease have been identified. Topotecan has a 20%–40% response rate in patients with sensitive SCLC (those patients who relapsed two or more months after their first-line therapy) with a median survival of 22–27 weeks. For patients with refractory disease (progressed through or within two months of completion of first-line therapy), the response rate in Phase II studies is only 3%–11%. Median survival is about 20 weeks.[62] Results of a randomised trial comparing topotecan with CAV (cyclophosphamide, adriamycin and vincristine) in patients who relapsed or progressed two or more months from completion of first-line chemotherapy revealed no difference in response rates, duration of response, or survival between the two groups.[63]

CHEMOTHERAPY PLUS CHEST IRRADIATION

Numerous studies have been done with chemotherapy and thoracic radiotherapy for patients with limited-stage SCLC. Conflicting results have been attributed to differences in chemotherapy regimens and different schedules integrating chemotherapy and thoracic radiation (concurrent, sequential and sandwich approach). Two recent meta-analyses concluded that thoracic radiation does result in a small but significant improvement in survival and major control of the disease in the chest, although no conclusions could be made regarding the optimal sequencing of chemotherapy and thoracic radiation.[35,64]

PROPHYLACTIC CRANIAL IRRADIATION

Numerous trials have demonstrated that prophylactic cranial irradiation (PCI) does not enhance survival, but does decrease the risk of brain metastases without a decrease in mental function.[65] However, a recent meta-analysis demonstrated a small but statistically and clinically significant improvement in survival with PCI.[66]

CLINICAL TRIAL METHODS IN RESPIRATORY CANCERS

With traditional cytoreductive and cytotoxic chemotherapy, there are well-established and accepted experimental designs for Phase I, II and III clinical trials. These experimental designs are generally based on the paradigm that with increased myelosuppression, tumour cells are more likely to be killed, leading to shrinkage of tumours and that there is a monotonically increasing dose–response and dose–toxicity relationship. It is also assumed that tumour shrinkage will eventually lead to clinical benefit such as prolonged survival or improved quality of life. Tumour shrinkage has been used as a surrogate endpoint for clinical benefit despite lack of its validation universally.

Advances in molecular biology and cancer genetics coupled with biotechnology are bringing forth a number of new drugs which appear to target molecular pathways such as cancer initiation, angiogenesis, invasion or metastasis. Examples include anti-angiogenesis drugs, EGFR antibodies or tyrosine kinase inhibitors, VEGF receptor inhibitors, matrix metalloproteinase inhibitors, PDGF receptor inhibitors, and other so-called molecular targeted therapies. These new drugs are generally not expected to shrink tumours. Instead they are expected to inhibit tumour growth or prevent metastasis by targeting specific pathways involved in tumour progression. With the emergence of these new classes of drugs with entirely different mode of action and expected therapeutic effects, the traditional designs for Phase I, II and III clinical trials may not be adequate. These new classes of drugs will challenge the existing paradigm for experimental design, conduct and analysis of Phase I, II and III clinical trials in cancer.

In lung cancer as in other solid tumours, treatment decision is typically guided by the histology and the staging of disease. In clinical trials, the same considerations determine the inclusion and exclusion criteria. Despite this traditional practice, it has long been suspected that underlying heterogeneity at the molecular level is the reason for widely differing outcomes in response to the identical treatment in both safety and efficacy. Recent advances in molecular biology and genetics and in pharmacogenetics appear to hold keys to future drug development.[67]

New classes of therapies and drugs and advances in pharmacogenetics, along with heterogeneity of cancer based on molecular findings, represent a fundamental paradigm shift at the core of experimental designs for clinical trials in cancer. Obviously this paradigm shift poses serious challenges to clinical investigators and statisticians alike.[68] These challenges are certainly not unique to lung cancer. Heterogeneity of NSCLC patients in EGFR mutations and failure of gefitinib combined with standard chemotherapy in prolonging survival despite pre-clinical data highlight issues in clinical trial design with targeted therapies.[68–70] Lessons from gefitinib trials and trials of other targeted therapies may shed light on how to design clinical trials of targeted therapies and on the role of pharmacogenetics in future drug development.

In what follows, we will discuss how targeted therapies and pharmacogenetics may change the clinical trials in the future in terms of study endpoints, target and trial design.

PHASE I CLINICAL TRIALS

In typical Phase I clinical trials with acute dose-limiting toxicities as the primary endpoint, a standard dose-escalation scheme with cohorts of a fixed number of patients treated at each of sequential dose levels is used to estimate the so-called maximum tolerated dose (MTD) or safe dose to be used in subsequent Phase II efficacy studies. Despite its shortcomings, this standard dose-escalation design has been the mainstay for Phase I cancer clinical trials.

With targeted therapies, it is unclear whether there is a clear dose–toxicity and dose–response relationship to help guide us in determining the most appropriate dose for Phase II efficacy and Phase III effectiveness clinical trials. As a consequence it is questionable whether the paradigm for dose-escalation designs for cytotoxic drugs for Phase I clinical trials is suitable or even relevant for targeted therapies as acute toxicities and MTD may not be meaningful with such therapy. This obviously calls for alternative methods for estimating a safe and effective dose in Phase I clinical trials. With targeted therapies, it was suggested that a biological endpoint other than acute toxicity may be used in Phase I trials to define the so-called optimal biologic dose for subsequent Phase II trials.[71] As targeted therapies are believed to be considerably less toxic than conventional chemotherapy and are perceived as predominantly cytostatic as against cytotoxic, it may be appropriate to use a randomised design for Phase I trials.

PHASE II CLINICAL TRIALS

In Phase II clinical trials with cytotoxic chemotherapy, multi-stage designs with objective tumour response defined as shrinkage of tumour by more than 50%, i.e. complete or partial response, as the primary endpoint are widely used. These are essentially sequential designs in the sense that a decision to treat additional patients for establishment of clinical efficacy is conditional on the observed clinical efficacy or safety with the patients from the previous stages. This is primarily to avoid treating patients with seemingly inefficacious therapy. Also single-arm designs are sometimes used in which comparisons are made with historical control data in terms of time to disease progression.

Targeted therapies are often cytostatic, rather than cytoreductive. Therefore, it may be reasonable to use instead objective tumour response plus stable disease as the primary endpoint. With this expanded primary endpoint, traditional multistage designs may still be used provided that reliable historical data on the expanded primary endpoint are available. Sequentially measured times to disease progression before and after treatment with targeted therapies within each patient who has failed previous treatment may be used in a Phase II setting where statistical hypotheses regarding a hazard ratio of times to disease progression before and after treatment with targeted therapies can be tested.[72]

Considering the heterogeneity of cancer, one may wish to distinguish antiproliferative activity attributable to targeted therapies from less aggressive disease in Phase II screening trials. In that setting, one may use a randomised discontinuation design in which all patients are treated initially with targeted therapies and only those whose disease is stable are randomised in a double-blind fashion to the same targeted therapy vs. placebo.[73]

Before proceeding to Phase III clinical trials, it is critical to identify predictive markers of response to the therapies and to validate the molecular target. As in chemoprevention trials, pre-surgery models are very useful models for an exploratory clinical trial in which (1) a core biopsy is obtained to identify surgical candidates, (2) an optimal biologic dose of a targeted therapy is given for a short duration, typically 2 to 4 weeks, before surgery, (3) surgery is performed and tumour and normal tissues are obtained, and (4) finally, assays are performed on pre-treatment biopsy samples and post-treatment tissue samples.[69] Changes in biomarkers such as Ki-67, TUNEL, EGFR and phosphorylated EGFR for an EGFR inhibitor may be evaluated for identification of the molecular targets, i.e. predictive markers of response to the therapies, and for validation of the target by analysis of their association with clinical outcome. However, difficulties in obtaining adequate tumour biopsies in lung cancer make this approach difficult.

PHASE III CLINICAL TRIALS

Overall survival typically being the ultimate criterion for evaluation of the effectiveness of cancer treatment in Phase III clinical trials, a traditional randomised controlled design with time to death due to all causes as the primary endpoint has become recognised as a gold standard for establishment of new standard therapies in cancer. Depending on the disease and the therapies under investigation, however, other endpoints such as time to disease progression, symptom benefit or quality of life may be appropriate as a surrogate endpoint despite the problems associated with the surrogate endpoint. With targeted therapies, biomarkers appear to be an attractive alternative as an endpoint. As was pointed out above, however, they need to be validated before being used as a surrogate endpoint for clinical benefit.

Gefitinib was approved for marketing based on two randomised Phase II trials in patients with advanced NSCLC who had failed platinum- or docetaxel-containing regimen or both,[74,75] despite negative preliminary results from two randomised, controlled Phase III trials in chemo-naive patients with NSCLC.[54,55] Failure of gefitinib in prolonging survival of patients with NSCLC when combined with standard chemotherapy has raised a number of questions about the promise of targeted therapies. A number of explanations for gefitinib's failure have been proposed such as inadequate dosing resulting in suboptimal target modulation, antagonism between gefitinib and chemotherapy, sensitivity of tumour cells to both gefitinib and chemotherapy, and dilution of effect of gefitinib due to heterogeneity in sensitivity to EGFR inhibition.[70]

Two independent pharmacogenetic investigations into the molecular heterogeneity revealed EGFR mutations as the cause of different responses to gefinitib in NSCLC patients.[51,52] These findings appear to provide partial explanation for the failure of gefitinib, and have a serious implication for design of Phase III clinical trials in terms of sample size requirements and power.[70,76,77] If targeted therapies confer different benefits to patients depending on their pharmacogenetic profile and if patients are not screened for susceptibility to targeted therapies, power of the trial to detect a truly beneficial effect can be seriously undermined and sample size requirements can be many orders of magnitude larger than typical, thus rendering such clinical trials impossible to conduct realistically. This is primarily due to dilution of treatment effects resulting from mixture of patients with different responses to treatment. The extent of decrease in power and increase in sample size is dependent on the proportion of patients with the target and the effect size of therapeutic benefits in patients with the target and in those without the target.

An alternative approach is to pre-screen patients with the target who are most likely to benefit from the targeted therapies and to enrol only those patients into Phase III clinical trials. This will reduce the number of patients to be treated in the trials significantly; however, depending on the prevalence of patients with the target, a large number of patients have to be pre-screened. All in all this requires validated molecular targets in the tissue of interest in the patient population, ready accessibility of the tissue, and validated assays to screen patients with the target.

Exploratory clinical trials such as that described in the previous subsection are a key to prospectively identifying the target of therapy and validating it before the information can be used in the design of Phase III clinical trials. Successful testing of the targeted therapies in Phase III clinical trials requires careful design, conduct and analysis of Phase II clinical trials where patients' pharmacogenetics and molecular targets have to be rigorously investigated to provide necessary information for adequate design of Phase III clinical trials with targeted therapies.

REFERENCES

1. Frost JK, Ball WC, Levin M, Tockman MS, Baker RR, Carter D, Eggelston JC, Erozan YS, Gupta PK, Khouri NF, Marsh BR, Stitik FP. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study. *Am Rev Respir Dis* (1984) **130**: 549–54.
2. Flehinger BJ, Melamed MR, Zaman MB, Heelan RT, Perchick WB, Martini N. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Memorial Sloan-Kettering study. *Am Rev Respir Dis* (1984) **130**: 555–60.
3. Fontana RS, Sanderson DR, Taylor WF, Woolner LB, Miller WE, Muhm JR, Uhlenhopp MA. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Mayo Clinic Study. *Am Rev Respir Dis* (1984) **130**: 561–5.
4. Patz EF, Goodman PC, Bepler G. Screening for lung cancer. *New Engl J Med* (2000) **343**: 1627–33.
5. Henschke CI, Naidich DP, Yankelevitz DF, McGuinness G, McCauley DI, Smith JP, Libby DM, Pasmantier MW, Vazquez M, Koizumi J, Flieder D, Altorki NK, Miettinen OS. Early Lung Cancer Action Project: Initial findings on repeat screening. *Cancer* (2001) **92**: 153–9.
6. Henschke CI, Yankelevitz D, Libby D, Kimmel M. CT screening for lung cancer: the first ten years. *Cancer J* (2002) **8**: 47S–54S.
7. Henschke CI, Yankelevitz D, Smith J, Miettinen O. Screening for lung cancer: the early lung cancer action approach. *Lung Cancer* (2002) **35**: 143–8.
8. Bach FH, Kelley M, Tate RC, McCrory DC. Screening for lung cancer: a review of the literature. *Chest* (2003) **123**: 72S–82S.
9. Gohagan J, Marcus P, Fagerstrom R, Pinsky P, Kramer B, Prorok P. Baseline findings of a randomized feasibility trial of lung cancer screening with spiral CT scan vs chest radiograph: The Lung Screening Study of the National Cancer Institute. *Chest* (2004) **126**: 114–21.
10. Mountain CF. Revisions in the International System for Staging of Lung Cancer. *Chest* (1997) **111**: 1710–17.
11. van Zandwijk N, Dalesio O, Pastorino U, de Vries N, van Tinteren H. Euroscan, a randomized trial of vitamin A and N-acetylcysteine in patients with head & neck cancer or lung cancer. *J Natl Cancer Inst* (2000) **92**: 977–86.
12. Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, Keogh JP, Meyskens FL, Valanis B, Williams JH, Barnhart S, Hammar S. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *New Engl J Med* (1996) **334**: 1150–5.
13. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New Engl J Med* (1994) **330**: 1029–35.
14. Clark LC, Combs GF, Turnbull BW, Slate EH, Chalker DK, Chow J, Davis LS, Glover RA, Graham GF, Gross EG, Krongrad A, Lesher JL, Park HK, Sanders BB, Smith CL, Taylor JR. Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin: a randomized controlled trial. *J Am Med Assoc* (1996) **276**: 1957–63.
15. PORT Meta-analysis Trialists Group. Postoperative radiotherapy in non-small-cell lung cancer: systematic review and meta-analysis of individual patient data from nine randomised controlled trials. *Lancet* (1998) **352**: 257–63.
16. The Lung Cancer Study Group. Effects of postoperative mediastinal radiation on completely resected stage II and stage III epidermoid cancer of the lung. *New Engl J Med* (1986) **315**: 1377–81.
17. Non-small Cell Lung Cancer Collaborative Group. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials. *Br Med J* (1995) **311**: 899–909.
18. Waller D, Fairlamb DJ, Gower N, Milroy R, Peake MD, Rudd RM, Spiro SG, Stephens RJ. The Big Lung Trial (BLT). Determining the value of cisplatin-based chemotherapy for all patients with non-small cell lung cancer (NSCLC): preliminary results in the surgical setting. *Proc Am Soc Clin Oncol* (2003) **22**: 632a.
19. Scagliotti GV, Fossati R, Torri V, Crino L, Giaccone G, Silvano G, Martelli M, Clerici M, Cognetti F, Tonato M. Randomized study of adjuvant chemotherapy for completely resected stage I, II, or IIIA non-small-cell lung cancer. *JNCI Cancer Spectrum* (2003) **95**: 1453–61.
20. Keller SM, Adak S, Wagner H, Herskovic A, Komaki R, Brooks BJ, Perry MC, Livingston RB, Johnson DH for the Eastern Cooperative Oncology Group. A randomized trial of postoperative adjuvant therapy in patients with completely resected

stage II or IIIA non-small cell lung cancer. *New Engl J Med* (2000) **343**: 1217–22.

21. Strauss G, Herndon J, Maddaus M, Johnstone D, Johnson E, Watson D, Sugarbaker D, Schilsky R, Green M. Randomized clinical trial of adjuvant chemotherapy with paclitaxel and carboplatin following resection in Stage IB non-small cell lung cancer (NSCLC): Report of Cancer and Leukemia Group B (CALGB) Protocol 9633. *J Clin Oncol* (2004) **22**(14S): 621 (abstr 7019).
22. Winton T, Livingston R, Johnson D, Rigas J, Cormier Y, Butts C, Ding K, Seymour L, Magoski N, Shepherd FA. A prospective randomised trial of adjuvant vinorelbine and cisplatin in completely resected stage 1B and II non small cell lung cancer (NSCLC) Intergroup JBR.10. *New Engl J Med* (2005) **352**: 2589–97.
23. Kato H, Ichinose Y, Ohta M, Hata E, Tsubota N, Tada H, Watanabe Y, Wada H, Tsuboi M, Hamajima N. A randomized trial of adjuvant chemotherapy with uracil-tegafur for adenocarcinoma of the lung. *New Engl J Med* (2004) **350**: 1713–21.
24. Rosell R, Gomez-Codina J, Camps C, Maestre J, Padille J, Canto A, Mate JL, Li S, Roig J, Olazabal A, Canela M, Ariza A, Skacel Z, Morera-Prat J, Abad A. A randomized trial comparing preoperative chemotherapy plus surgery with surgery alone in patients with non-small-cell lung cancer. *New Engl J Med* (1994) **330**: 153–8.
25. Roth JA, Fossella F, Komaki R, Ryan MB, Putnam JB, Lee JS, Dhingra H, DeCaro L, Chasen M, McGavran M, Atkinson EN, Hong WK. A randomized trial comparing perioperative chemotherapy and surgery with surgery alone in resectable stage IIIA non-small-cell lung cancer. *J Natl Cancer Inst* (1994) **86**: 673–80.
26. Roth JA, Atkinson EN, Fossella F, Komaki R, Ryan MB, Putnam JB, Lee JS, Dhingra H, DeCaro L, Chasen M, Hong WK. Long-term follow-up of patients enrolled in a randomized trial comparing perioperative chemotherapy and surgery alone in resectable stage IIIA non-small-cell lung cancer. *Lung Cancer* (1998) **21**: 1–6.
27. Depierre A, Milleron B, Moro-Sibilot D, Chevret S, Quoix E, Lebeau B, Braun D, Breton J-L, Lemarie E, Gouva S, Paillot N, Brechot J-M, Janicot H, Lebas F-X, Terrioux P, Clavier J, Foucher P, Monchatre M, Coetmeur D, Level M-C, Leclerc P, Blanchon F, Rodier J-M, Thiberville L, Villeneuve A, Westeel V, Chastang C. Preoperative chemotherapy followed by surgery compared with primary surgery in resectable stage I (except T1N0), II, and IIIa non-small cell lung cancer. *J Clin Oncol* (2002) **20**: 247–53.
28. Albain K, Scott C, Rusch V, Turrisi A, Shepherd FA, Smith C, Gandara D, Johnson D, Green M, Miller R. Phase III comparison of concurrent chemotherapy plus radiotherapy (CT/RT) and CT/RT followed by surgical resection for stage IIIA (pN2) non-small cell lung cancer (NSCLC): initial results from intergroup trial 0139 (RTOG 93-09). *Proc Am Soc Clin Oncol* (2003) **22**: 621 (abstr 2497).
29. Dillman RO, Herndon J, Seagren SL, Eaton WL, Green MR. Improved survival in stage III non-small cell lung cancer: seven-year follow-up of cancer and leukemia group B (CALGB) 8433. *J Natl Cancer Inst* (1996) **88**: 1210–15.
30. Sause W, Kolesar P, Taylor S, Johnson D, Livingston R, Komaki R, Emami E, Curran W, Byhardt R, Fisher B. Final results of phase III trial in regionally advanced unresectable non-small cell lung cancer: radiation Therapy Oncology Group (RTOG) 88-08 and Eastern Cooperative Oncology Group (ECOG) 4588. *Chest* (2000) **117**: 358–64.
31. Furuse K, Fukuoka M, Kawahara M, Nishikawa H, Takada Y, Kudoh S, Katagami N, Ariyoshi Y. Phase III study of concurrent versus sequential thoracic radiotherapy in combination with mitomycin, vindesine, and cisplatin in unresectable stage III non-small-cell lung cancer. *J Clin Oncol* (1999) **17**: 2692–9.
32. Curran W, Scott C, Langer C, Komaki R, Lee J, Hauser S, Movsas B, Wasserman T, Russell A, Byhardt R, Sause W, Cox J. Phase III comparison of sequential vs. concurrent chemoradiation for pts with unresected stage III non-small cell lung cancer (NSCLC): initial report of the Radiation Therapy Oncology Group (RTOG) 9410. *Proc Am Soc Clin Oncol* (2000) **19**: 484 (abstr 1891).
33. Choy H, Akerley W, Safran H, Graziano S, Chung C, Williams T, Cole B, Kennedy T. Multiinstitutional phase ii trial of paclitaxel, carboplatin, and concurrent radiation therapy for locally advanced non-small-cell lung cancer. *J Clin Oncol* (1998) **16**: 3316–22.
34. Vokes E, Herndon J, Kelley M, Watson D, Cicchetti M, Green M. Induction chemotherapy followed by concomitant chemoradiotherapy (CT/XRT) versus CT/XRT alone for regionally advanced unresectable non-small cell lung cancer (NSCLC): initial analysis of a randomized phase III trial. *J Clin Oncol* (2004) **22**(14S): 616 (abstr 7005).
35. Soquet PJ, Chauvin F, Boissel JP, Cellerino R, Cormier Y, Ganz PA, Kaasa S, Pater JL, Quoix E, Rapp E, Tunarello D, Williams J, Woods BL, Bernard JP. Polychemotherapy in advanced non-small cell lung cancer: a meta-analysis. *Lancet* (1993) **342**: 19–21.
36. Marino P, Pampallona S, Preatoni A, Cantoni A, Inveinezzi F. Chemotherapy vs. supportive care in advanced non-small cell lung cancer: results of a

meta-analysis of the literature. *Chest* (1994) **106**: 861–5.

37. Baggstrom M, Socinski M, Hensing T, Poole C. Third generation chemotherapy regimens (3GR) improve survival over second generation regimens (2GR) in stage IIIB/IV non-small cell lung cancer (NSCLC): a meta-analysis of the published literature. *Proc Am Soc Clin Oncol* (2002) **21**: 306 (abstr 1222).
38. Delbaldo C, Syz N, Michiels S, Le Chevalier T, Pignon J-P. Adding a second or a third drug to a chemotherapy regimen in patients with advanced non-small-cell lung carcinoma (NSCLC): a meta-analysis of the literature. *Proc Am Soc Clin Oncol* (2003) **22**: 623 (abstr 2507).
39. Baggstrom MQ, Socinski MA, Hensing TA, Poole C. Addressing the optimal number of cytotoxic agents in stage IIIB/IV non-small cell lung cancer (NSCLC): A meta-analysis of the published literature. *Proc Am Soc Clin Oncol* (2003) **22**: 624 (abstr 2510).
40. Sandler A, Nemunaitis J, Denham C, von Pawel J, Cormier Y, Gatzemeier U, Mattson K, Manegold C, Palmer M, Gregor A, Nguyen B, Niyikiza C, Einhorn L. Phase III trial of gemcitabine plus cisplatin versus cisplatin alone in patients with locally advanced or metastatic non-small cell lung cancer. *J Clin Oncol* (2000) **18**: 122–30.
41. Wozniak AJ, Crowley JJ, Balcerzak SP, Weiss GR, Spiridonidis CH, Baker LH, Albain KS, Kelly K, Taylor SA, Gandara DR, Livingston RB. Randomized trial comparing cisplatin with cisplatin plus vinorelbine in the treatment of advanced non-small-cell lung cancer: a Southwest Oncology Group study. *J Clin Oncol* (1998) **16**: 2459–65.
42. Lilenbaum RC, Herndon J, List M, Desch C, Watson D, Holland J, Weeks JC, Green MR. Single agent (SA) versus combination chemotherapy (CC) in advanced non-small cell lung cancer: a CALGB randomized trial of efficacy, quality of life (QoL), and cost effectiveness. *Proc Am Soc Clin Oncol* (2002) **21**: 1 (abstr 2).
43. Sederholm C. Gemcitabine (G) compared with gemcitabine plus carboplatin (GC) in advanced non-small cell lung cancer (NSCLC): a phase III study by the Swedish Lung Cancer Study Group (SLUSG). *Proc Am Soc Clin Oncol* (2002) **21**: 291 (abstr 1162).
44. Kelly KJ, Crowley J, Bunn PA, Presant CA, Grevstad PK, Moinpour CM, Ramsey SD, Wozniak AJ, Weiss GR, Moore DF, Israel VK, Livingston RB, Gandara DR. Randomized phase III trial of paclitaxel plus carboplatin versus vinorelbine plus cisplatin in the treatment of patients with advanced non-small-cell lung cancer: a Southwest Oncology Group trial. *J Clin Oncol* (2001) **19**: 3210–18.
45. Schiller JH, Harrington D, Belani CP, Langer C, Sandler A, Krook J, Zhu Z, Johnson DH for the Eastern Cooperative Oncology Group. Comparison of four chemotherapy regimens for advanced non-small cell lung cancer. *New Engl J Med* (2002) **346**: 92–8.
46. Zojwalla N, Raftopoulos H, Gralla RJ. Are cisplatin and carboplatin equivalent in the treatment of non-small cell lung carcinoma (NSCLC)? Results of a comprehensive review of randomized studies in over 2300 patients. *J Clin Oncol* (2004) **22**(14S): 632 (abstr 7068).
47. D'Addario G, Pintilie M, Leighl NB, Feld R, Cerny T, Shepherd FA. Platinum-based versus non-platinum-based chemotherapy in advanced non-small cell lung cancer (NSCLC): a meta-analysis of the published literature. *J Clin Oncol* (2005) **23**: 2926–36.
48. Shepherd FA, Dancey J, Ramlau R, Mattson K, Gralla R, O'Rourke M, Levitan N, Gressot L, Vincent M, Burkes R, Coughlin S, Kim Y, Berille J. Prospective randomized trial of docetaxel versus best supportive care in patients with non-small cell lung cancer previously treated with platinum-based chemotherapy. *J Clin Oncol* (2000) **18**: 2095–2103.
49. Fossella FV, DeVore R, Kerr RN, Crawford J, Natale RR, Dunphy F, Kalman L, Miller V, Lee JS, Moore M, Gandara D, Karp D, Vokes E, Kris M, Kim Y, Gamza F, Hammershaimb L. Randomized phase III trial of docetaxel versus vinorelbine or ifosfamide in patients with advanced non-small cell lung cancer previously treated with platinum-containing chemotherapy regimens. *J Clin Oncol* (2000) **18**: 2354–62.
50. Hanna N, Shepherd FA, Fossella FV, Pereira JR, De Marinis F, von Pawel J, Gatzemeier U, Tsao TYC, Pless M, Muller T, Lim H-L, Desch C, Szondy K, Gervais R, Shaharyar, Manegold C, Paul S, Paoletti P, Einhorn L, Bunn PA. Randomized phase III trial of pemetrexed versus docetaxel in patients with non-small-cell lung cancer previously treated with chemotherapy. *J Clin Oncol* (2004) **22**: 1589–97.
51. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* (2004) **304**: 1497–1500.
52. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, Louis DN, Christiani DC, Settleman J, Haber DA. Activating

mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *New Engl J Med* (2004) **350**: 2129–39.

53. Shepherd FA, Pereira JR, Ciuleanu E, Tan EH, Hirsh V, Thongprasert S, Campos D, Maoleekoonpiroj S, Smylie M, Martins R, van Kooten M, Dediu M, Findlay B, Tu D, Johnson D, Bezjak A, Clark G, Santabarbara P, Seymour L, for the National Cancer Institute of Canada Clinical Trials Group. Erlotinib in previously treated non-small cell lung cancer. *New Engl J Med* (2005) **353**: 123–32.
54. Giaccone G, Herbst RS, Manegold C, Scagliotti G, Rosell R, Miller V, Natale RB, Schiller JH, Von Pawel J, Pluzanska A, Gatzemeier U, Grous J, Ochs JS, Averbuch SD, Wolf MK, Rennie P, Fandi A, Johnson DH. Gefitinib in combination with gemcitabine and cisplatin in advanced non-small-cell lung cancer: a phase III trial – INTACT 1. *J Clin Oncol* (2004) **22**: 777–84.
55. Herbst RS, Giaccone G, Schiller JH, Natale RB, Miller V, Manegold C, Scagliotti G, Rosell R, Oliff I, Reeves JA, Wolf MK, Krebs AD, Averbuch SD, Ochs JS, Grous J, Fandi A, Johnson DH. Gefitinib in combination with paclitaxel and carboplatin in advanced non-small-cell lung cancer: a phase III trial – INTACT 2. *J Clin Oncol* (2004) **22**: 785–94.
56. Sirotnak FM, Zakowski MF, Miller VA, Scher HI, Kris MG. Efficacy of cytotoxic agents against human tumor xenografts is markedly enhanced by coadministration of ZD1839 (Iressa), an inhibitor of EGFR tyrosine kinase. *Clin Cancer Res* (2000) **6**: 4885–92.
57. Piperdi B, Ling YH, Kroog G, Perez-Soler R. Schedule-dependent interaction between epidermal growth factor inhibitors (EGFRI) and G2/M blocking chemotherapeutic agents (G2/MB) on human NSCLC cell lines in vitro. *J Clin Oncol* (2004) **22**(14S): 623 (abstr 7028).
58. Hurwitz H, Fehrenbacher L, Novotny W, Cartwright T, Hainsworth J, Heim W, Berlin J, Baron A, Griffing S, Holmgren E, Ferrara N, Fyfe G, Rogers B, Ross R, Kabbinavar F. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *New Engl J Med* (2004) **350**: 2335–42.
59. Johnson DH, Fehrenbacher L, Novotny WF, Herbst RS, Nemunaitis JJ, Jablons DM, Langer CJ, DeVore RF, Gaudreault J, Damico LA, Holmgren E, Kabbinavar R. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. *J Clin Oncol* (2004) **22**: 2184–91.
60. Niell H, Herndon J, Miller A, Watson D, Sandler A, Kelly K, Marks R, Green M. Randomized phase III intergroup trial (CALGB 9732) of etoposide (VP-16) and cisplatin (DDP) with or without paclitaxel (TAX) and G-CSF in patients with extensive stage small cell lung cancer (ED-SCLC). *Proc Am Soc Clin Oncol* (2002) **21**: 293 (abstr 1169).
61. Noda K, Nishiwaki Y, Kawahara M, Negoro S, Sugiura T, Yokoyama A, Fukuoka M, Mori K, Watanabe K, Tamura T, Yamamoto S, Saijo N for the Japan Clinical Oncology Group. Irinotecan plus cisplatin compared with etoposide plus cisplatin for extensive small cell lung cancer. *New Engl J Med* (2002) **346**: 85–91.
62. Schiller JH, Kim K, Hutson P, DeVore R, Glick J, Stewart J, Johnson D. Phase II study of topotecan in patients with extensive-stage small cell carcinoma of the lung: an Eastern Cooperative Oncology Group trial (E1592). *J Clin Oncol* (1996) **14**: 2345–52.
63. von Pawel J, Schiller JH, Sheperd F, Kleisbauer J, Chrysson N, Steward D, Fields S, Clark P, Palmer M, Depierre A, Carmichael J, Krebs J, Ross G, Gralla R. Topotecan versus CAV for the treatment of recurrent small cell lung cancer. *J Clin Oncol* (1999) **17**: 658–67.
64. Pignon J, Arriagada R, Ihde D, Johnson D, Perry M, Souhami R, Brodin O, Joss R, Kies M, Lebeau B, Onoshi T, Osterlind K, Tattersall M, Wagner H. A meta-analysis of thoracic radiotherapy for small-cell lung cancer. *New Engl J Med* (1992) **327**: 1618–24.
65. Arriagada R, Le Chevalier T, Borie F, Riviere A, Chomy P, Monnet I, Tardivon A, Viader F, Tarayre M, Benhamou S. Prophylactic cranial irradiation for patients with small-cell lung cancer in complete remission. *J Natl Cancer Inst* (1995) **87**: 183–90.
66. Auperin A, Arriagada R, Pignon J, Pechoux C, Gregor A, Stephens R, Kristjansen P, Johnson B, Ueoka H, Wagner H, Aisner J. Prophylactic cranial irradiation for patients with small cell lung cancer in complete remission. *New Engl J Med* (1999) **341**: 476–84.
67. Roses AD. Pharmacogenetics and future drug development and delivery. *Lancet* (2000) **355**: 1358–61.
68. Schiller JH. Clinical trial design issues in the era of targeted therapies. *Clin Cancer Res* (2004) **10**: 4281s–2s.
69. Arteaga CL, Baselga J. Clinical trial design and end points for epidermal growth factor receptor-targeted therapies: implications for drug development and practice. *Clin Cancer Res* (2003) **9**: 1579–89.

70. Dancey JE, Freidlin B. Targeting epidermal growth factor receptor – are we missing the mark? *Lancet* (2003) **362**: 62–4.
71. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trials designs for cytostatic agents: are new approaches needed? *J Clin Oncol* (2001) **19**: 265–72.
72. Mick R, Crowley JJ, Carroll RJ. Phase II clinical trial design for noncytotoxic anticancer agents for which time to disease progression is the primary endpoint. *Controlled Clin Trials* (2000) **21**: 343–59.
73. Rosner GL, Stadler W, Ratain MJ. Randomized discontinuation design: application to cytostatic antineoplastic agents. *J Clin Oncol* (2002) **20**: 4478–84.
74. Fukuoka M, Yano S, Giaccone G, Tamura T, Nakagawa K, Douillard J-Y, Nishiwaki Y, Vansteenkiste J, Kudo S, Rischin D, Eek R, Horai T, Noda K, Takata I, Smit E, Averbuch S, Macleod A, Feyereislova A, Dong R-P, Baselga J. Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer. *J Clin Oncol* (2003) **21**: 2237–46.
75. Kris MG, Natale RB, Herbst RS, Lynch TJ, Prager D, Belani CP, Schiller JH, Kelly K, Spiridonidis H, Sandler A, Albain KS, Cella D, Wolf MK, Averbuch SD, Ochs JJ, Kay AC. Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non-small cell lung cancer. *J Am Med Assoc* (2003) **290**: 2149–58.
76. Fijal BA, Hall JM, Witte JS. Clinical trials in the genomic era: effects of protective genotypes on sample size and duration of trial. *Controlled Clin Trials* (2000) **21**: 7–20.
77. Betensky RA, Louis DN, Cairncross JG. Influence of unrecognized molecular heterogeneity on randomized clinical trials. *J Clin Oncol* (2002) **20**: 2495–9.

9

Adrenal Cortical Carcinoma

ISHA MALIK AND PETROS PERROS
Endocrine Unit, Freeman Hospital, Newcastle upon Tyne NE7 7DN, UK

INTRODUCTION

Adrenal cortical carcinomas (ACC) are rare, aggressive malignancies, generally carrying a poor prognosis, and represent 0.05 to 0.2% of all cancers. The annual incidence is approximately 2 per million.[1] ACC have a bimodal age distribution, with a peak in the first decade of life and in the fourth to fifth decade. The median age at diagnosis in adults is 44 years. Children with ACC are younger than 5 years in 75% of cases, at presentation.[2] The female to male ratio is approximately 3:1.

Radical surgical excision is the only means by which cure or long-term survival may be achieved when detected in early stages, but only 30% of these malignancies are confined to the adrenal gland at the time of diagnosis, therefore the majority require additional treatment modalities including palliative chemotherapy and radiotherapy.[3] Unfortunately, the literature provides no robust evidence of a clear survival benefit achieved by non-surgical treatments, largely because of lack of large series.

Clinical manifestations of the disease vary depending on the functionality of the tumour: 50 to 70% of patients with ACC have endocrine symptoms at presentation, with Cushing‘s syndrome being commonest in adults, followed by mixed virilisation/Cushing‘s syndrome, and pure virilisation syndrome[4,5]. Children with ACC characteristically present with virilisation. Rarer endocrine presentations include feminisation, hyperaldosteronism, hypoglycaemia, non-glucocorticoid-related insulin resistance and hypercalcaemia. Non-functioning tumours may present with fever, weight loss, abdominal pain, fullness and tenderness, or are incidentally detected as an adrenal mass on imaging studies. Adrenal cancers can also present as metastatic lesions involving lung, liver, lymph node, bone, kidneys, or other sites.

The majority of ACC are sporadic. The Li–Fraumeni syndrome, familial adenomatous polyposis and Carney’s complex are genetic syndromes associated with ACC[6,7]. Although an adrenal hyperplasia-to-adenoma-to-carcinoma sequence has not been proven, a few cases of congenital adrenal hyperplasia and Beck–Wiedemann syndrome have been reported to be associated with adrenal adenomas and carcinomas[8].

Endocrine studies are essential to determine the functionality of adrenal masses pre-operatively.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

Table 9.1. Staging of ACC

Tumour criteria	
T1	Tumour diameter ≤5 cm with no local invasion
T2	Tumour diameter >5 cm with no local invasion
T3	Tumour of any size with local extension but not involving adjacent organs
T4	Tumour of any size with local invasion of adjacent organs
Lymph node criteria	
N0	No regional lymph node involvement
N1	Positive regional lymph nodes
Metastasis criteria	
M0	No distant metastasis
M1	Distant metastasis
Stages	
Stage I	T1, N0, M0
Stage II	T2, N0, M0
Stage III	T1 or T2, N1, M0 or T3, N0, M0
Stage IV	T3 or T4, N1, M0 or any T, M1

Many patients, who do not present with endocrine symptoms, may display evidence of subclinical autonomous hormone secretion.[9] Detailed endocrine assessment pre-operatively is imperative so that hormone markers may subsequently be used for monitoring. Furthermore, subclinical Cushing's syndrome may be followed by hypoadrenalism and an adrenal crisis after tumour resection due to suppression of the contralateral adrenal gland, if this diagnosis is not made pre-operatively.[10] The evaluation of adrenal masses should also include screening tests to rule out phaeochromocytomas.

Staging is shown in Table 9.1. In a series of 113 patients with ACC four-year survival was 37%.[11] Patients who presented as Stage I and II had a five-year survival of 60%, whereas those with late stage disease had a five-year survival of 10%.

TREATMENT

To date, no randomised controlled studies have been performed, due to the rarity of the tumour.

SURGERY

Surgical resection is the only potentially curative treatment, particularly if the disease is detected early (Stage I or II). Retrospective studies have identified completeness of resection and stage at presentation, as the two most significant prognostic indicators,[11–13] thus in the case of localised and regional disease (Stages I–III and Stage IV without metastases), complete surgical excision, including regional lymph nodes and involved adjacent organs, offers the best chance of long-term disease-free survival. Laparoscopic resection is increasingly used for small tumours.[14] Despite complete resection, approximately two-thirds of patients develop recurrence within two years of treatment, and 85% eventually develop local recurrence or distant metastases.[15]

Surgery also has a role in patients with recurrent disease, particularly when symptomatic. In a retrospective comparison of patients with recurrent ACC, 18 patients were treated with chemotherapy (primarily mitotane) and 15 patients with surgical resection in addition to similar chemotherapy.[16] Although no patient was cured, resection was associated with a slight prolongation of survival and good palliation of Cushing's symptoms.[16] Similar observations were made by another group who reported that the five-year survival in patients with recurrent disease who underwent surgery was significantly better in those who underwent surgery (49.7% versus 8.3%).[17] Based on these data an aggressive surgical approach against recurrent ACC seems justifiable.

RADIOTHERAPY

The role of radiotherapy is confined to unresectable tumours causing localised symptoms.[18]

CHEMOTHERAPY

Mitotane

Mitotane has an adrenolytic effect possibly by inhibiting mitochondrial activity in adrenal cortical cells. It decreases production of cortisol by blocking 11-beta hydroxylation and increases the peripheral metabolism of steroids. Side effects are common and a barrier to tolerating high doses. They include dizziness, drowsiness, nausea, diarrhoea, loss of appetite, headache and muscle aches. Hypoadrenalism may also be induced by mitotane.

Mitotane has been shown to control endocrine hypersecretion in approximately 80% of patients with hormonally active tumours.[19] There is therefore a clear indication for mitotane therapy in patients with recurrent or metastatic disease who are symptomatic as a result of steroid hypersecretion.

The use of mitotane in patients with recurrent disease who have no steroid hypersecretion is controversial. A beneficial effect of this drug on survival has never been demonstrated and tumour regression is only seen in 19–34% of patients.[19]

Mitotane has been advocated for patients who have no evidence of residual disease after radical surgery in the hope that this therapy prolongs disease-free survival. This is, however, anecdotal[20] and such a decision has to be balanced against the frequent side effects and their impact on quality of life.

Further controversies in this area include the dose of mitotane (low dose 1.5–2 g/day versus high dose 6–10 g/day) and whether treatment should be guided by serum mitotane levels. Neurotoxicity has been associated with levels above 20 mg/l.[21] Drug plasma levels above 14 mg/l may result in enhanced efficacy.[22]

Mitotane is a useful treatment in patients with recurrent or metastatic disease who have steroid hypersecretion, though any effect on tumour regression is weak (Table 9.2). The evidence for low-dose mitotane in patients who are disease free is anecdotal and further studies are required

Table 9.2. Overview of clinical trials on adjuvant mitotane therapy with surgery

Study	Year	Institute	No. of Patients	Conclusion
Schteingart *et al.*[23]	1982	Division of Endocrinology & Metabolism, University of Michigan, Ann Arbor, USA	6	Low-dose long-term mitotane therapy effective Longer mean survival achieved
Pommier *et al.*[13]	1992	Department of Surgery, Memorial Sloan-Kettering Cancer Centre, New York, USA	73	Value of adjuvant mitotane therapy unproven
Vassilopoulou-Sellin *et al.*[31]	1993	M.D. Anderson Cancer Center, University of Texas, Houston, USA	19	Adjuvant mitotane not beneficial
Haak *et al.*[22]	1994	Department of Endocrinology, University Hospital Leiden, The Netherlands	96	No survival benefit Mitotane effective only when high serum levels achieved
Kasperlik–Zaluska *et al.*[32]	1995	Department of Endocrinology, Centre of Postgraduate Medical Education, Warsaw, Poland	52	Effective form of therapy Improved survival rate achieved
Dickstein *et al.*[20]	1998	Division of Endocrinology, Bnai Zion Medical Centre, Haifa, Israel	4	Low dose adjuvant mitotane therapy (1.5–2 g daily) effective Longer disease free survival achieved

in order to evaluate this strategy.[20,23] Should mitotane therapy be employed, monitoring of serum levels is desirable, aiming for levels of 14–20 mg/l.

Mitotane-Based Combination Chemotherapy

Several multi-centre trials have studied the use of mitotane in combination with other cytotoxic agents. These include mitotane and cisplatin,[24] mitotane and doxorubicin,[25] mitotane, etoposide, doxorubicin and cisplatin[26] and mitotane, doxorubicin, etoposide and vincristine[27] with complete responses of 5% at best and partial responses of 22-44%. Generally the results are poor and limited by high toxicity.

Combination Chemotherapy with Other Cytotoxic Agents

In cases where mitotane fails, chemotherapeutic regimens containing cisplatin alone or in combination with other cytotoxic agents have been used.

Schlumberger *et al.* treated 14 patients with progressive metastatic ACC with combination of 5-fluorouracil, doxorubicin and cisplatin.[28] The overall response rate was 23%. One patient achieved complete remission, which lasted for 42 months, and two patients achieved partial remission, which lasted for 6–11 months.

Another study included 11 patients treated with advanced progressive ACC who received cyclophosphamide, doxorubicin and cisplatin combination therapy.[29] Two partial responses and six disease stabilizations were observed. The total group had a median survival of 10 months.

Other Therapies in the Treatment of ACC

Novel approaches are being tried with anti-angiogenic and anti-chemotactic agents, tariquidar, suramin, antineoplastons and tamoxifen.[30]

CONCLUSION

Surgery is the mainstay of treatment for ACC. Mitotane has a role in managing patients with recurrent or metastatic disease with autonomous steroid hypersecretion. Several chemotherapeutic agents and regimens have been tried in metastatic disease, but are of limited value.

REFERENCES

1. *Third National Cancer Survey: Incidence Data National Cancer Institute*, Monograph 41, DHEW Publication No. (NIH) 75–787. US Department of Health, Education & Welfare: Public Health Service, National Institute of Health, National Cancer Institute, Bethesda, MD (1975).
2. Boscaro M, Fallo F, Barzon L. Adrenocortical carcinoma: epidemiology and natural history. *Minerva Endocrinol* (1995) **20**: 89–94.
3. Norton JA. Adrenal tumors. In: DeVita VT Jr, Hellman S, Rosenberg SA, eds, *Cancer: Principles and Practise of Oncology* 5th edn. Philadelphia, PA: Lippincott-Raven, (1997) 1659–77.
4. Icard P, Chapuis Y, Andreassia B *et al.* Adrenal carcinoma in surgically treated patients: a retrospective study on 156 cases by French Association of Endocrine surgery. *Surgery* (1992) **112**: 972–9.
5. Luton JP, Cerdas S, Billaud L *et al.* Clinical features of adrenalcortical carcinoma, prognostic factors, and the effect of mitotane therapy. *New Engl J Med* (1990) **322**: 1195–1201.
6. Gicquel C, Bertherat, Le Bouc Y. Pathogenesis of adrenocortical incidentalomas and genetic syndromes associated with adrenocortical neoplasms. *Endocrinol Metab Clin North Am* (2000) **29**: 1–13.
7. Wakatsuki S, Sasano H, Matsui T. Adrenocortical tumour in a patient with familial adenomatous polyposis: a case associated with a complete inactivating mutation of the APC gene and unusual histologic features. *Hum Pathol* (1998) **29**: 302–6.
8. Shimsi M, Ross F, Goodman A. Virlising adrenocortical tumor superimposed on congenital adrenocortical hyperplasia. *Am J Med* (1992) **93**: 338–42.
9. Bernini G, Moretti A, Iacconi P *et al.* Anthropometric, haemodynamic, humoral and hormonal evaluation in patients with incidental adrenocortical adenomas before and after surgery. *Eur J Endocrinol* (2003) **148**: 213–19.
10. Raut CP, Lee JE. Unusual functioning endocrine tumors. *Curr Treat Options Oncol* (2004) **5**: 327–34.
11. Schulick RD, Brennan MF. Long-term survival after complete resection and repeat resection in patients with adrenocortical carcinoma. *Ann Surg Oncol* (1999) **6**: 719–26.
12. Lee JE, Berger DG, El-Nagaar AK *et al.* Surgical management, DNA content, and the patient

survival in adrenal cortical carcinoma. *Surgery* (1995) **118**: 1090–8.

13. Pommier RF, Brennan MF. An eleven-year experience with adrenal carcinoma. *Surgery* (1992) **112**: 963–71.
14. Metro MJ. Cleveland Clinic reports experience with laparoscopic radical adrenalectomy. *J Urol* (2005) **173**: 519–25.
15. National Cancer Institute: Adrenocortical Carcinoma (PDQ): Treatment, 3 June 2003 (available at http.//www.nci.nih.gov/cancerinfo/pdq/treatment/adrenocortical/healthprofessional).
16. Jensen JC, Pass HI, Sindelar WF *et al*. Recurrent or metastatic disease in select patients with adrenocortical carcinoma: aggressive resection vs chemotherapy. *Arch Surg* (1991) **126**: 457–61.
17. Bellantone R, Ferrante A, Boscherini M *et al*. Role of reoperation in recurrence of adrenal cortical carcinoma: results from 188 cases collected in the Italian National Registry for Adrenal Cortical Carcinoma. *Surgery* (1997) **122**: 1212–18.
18. Kopf D, Goretzki PE, Lehnert H. Clinical management of malignant adrenal tumors. *J Cancer Res Clin Oncol* (2001) **127**: 143–55.
19. Lubitz JA, Freeman L, Okun R. Mitotane use in inoperable adrenal cortical carcinoma. *J Am Med Assoc* (1973) **223**: 1109–12.
20. Dickstein G, Hechner CS, Arad E *et al*. Is there a role for low doses of mitotane (o,p′-DDD) as adjuvant therapy in adrenocortical carcinoma? *J Clin Endocrinol Metab* (1998) **83**: 3100–3.
21. Van Slooten H, Moolenaar AJ, van Seters AP, Smeenk D. The treatment of adrenocortical carcinoma with o,p’-DDD: prognostic implications of serum level monitoring. *Eur J Cancer Clin Oncol* (1984) **20**: 47–53.
22. Haak HR, Cornelisse CJ, Hermans J *et al*. Nuclear DNA content and morphological characteristics in the prognosis of adrenalcortical carcinoma. *Br J Cancer* (1993) **68**: 151–5.
23. Schteingart DE, Motazedi A, Noonan RA, Thompson NW. Treatment of adrenal carcinomas. *Arch Surg* (1982) **117**: 1142–6.

24 Bukowski RM, Wolfe M, Levine HS *et al*. Phase II trial of mitotane and cisplatin in patients with adrenal carcinoma: a Southwest Oncology Group study. *J Clin Oncol* (1993) **11**: 161–5.

25. Decker RA, Elson P, Hogan TF *et al*. Eastern Cooperative Oncology Group Study 1879: mitotane and adriamycin in patients with advanced adrenocortical carcinoma. *Surgery* (1991) **110**: 1006–13.
26. Berruti A, Terzolo M, Sperone P *et al*. Etoposide, doxorubicin and cisplatin plus mitotance in the treatment of advanced adrenocortical carcinoma: a large prospectivephase II trial. Endocr Relat Caver (2005) **12**: 657–66.
27. Abraham J, Bakke S, Rutt A *et al*. A phase II trial of combination chemotherapy and surgical resection for the treatment of metastatic adrenocortical cancer: continuous infusion doxorubicin, vincristine, and etoposide with daily mitotane as a P-glycoprotein antagonist. *Cancer* (2002) **94**: 2333–43.
28. Schlumberger M, Ostronoff M, Bellaiche M *et al*. 5-fluorouracil, doxorubicin, and cisplatin regimen in adrenal cortical carcinoma. *Cancer* (1988) **61**: 1492–4.
29. Van Slooten H, Van Oosterom AT. CAP (Cyclophosphamide, doxorubicin and cisplatin) regimen in adrenal cortical carcinoma. *Cancer Treat Rep* (1983) **67**: 377–9.
30. See www.clinicaltrials.gov.
31. Vassilopoulou-Sellin R, Guinee VF, Klein MJ *et al*. Impact of adjuvant mitotane on the clinical course of patients with adrenal cortical cancer. *Cancer* (1993) **71**: 3119–23.
32. Kasperlik-Zaluska AA, Migdalska BM, Zgliczynsalic S *et al*. Adrenocortical carcinoma: a clinical study and treatment results of 52 patients. *Cancer* (1995) **75**: 2587–91.

10

Thyroid Cancer

VIJAYARAMAN ARUTCHELVAM AND PETROS PERROS
Endocrine Unit, Freeman Hospital, Newcastle upon Tyne NE7 7DN, UK

INTRODUCTION

Thyroid cancer is the commonest endocrine malignancy.[1] The incidence is 2.3/100 000 in women to 0.9/100 000 in men. Every year approximately 900 new cases and 250 deaths are recorded in England and Wales. For the US population the lifetime risk of thyroid carcinoma is 1%[2] and the most recent prevalence estimate is 292 555.[3,4] Mean age at diagnosis is 45–50 years. The commonest types are papillary and follicular carcinomas,[5] often referred to as differentiated thyroid carcinomas (DTC). The overall 10-year survival rate for middle-aged adults with DTC is 80–90%; however, 5–20% of patients develop recurrences, 10–15% distant metastases and 9% die of their disease.[1]

The evidence base for management of DTC consists of studies of large patient cohorts in which therapy has not been randomly assigned. There are no prospective randomised trials of treatment with survival endpoints. No such evidence is likely to be forthcoming in the foreseeable future because of the relative rarity of the disease the long survival of most patients. The central tenet of recent guidelines is that patients with DTC should be managed by multidisciplinary teams,[6–8] provided in well-resourced centres. This chapter will focus on recent trial data addressing management aspects of DTC.

TUMOUR STAGING AND PROGNOSTIC SCORING

There are several staging and clinical prognostic scoring strategies and most use age as an important parameter to identify cancer-related mortality risk.[6,9–12] The TNM classification (size of primary *t*umour, regional lymph *n*ode involvement and distant *m*etastases) is widely used (Table 10.1).[13] The 6th edition of the International Union Against Cancer TNM classification published in 2002,[14] has introduced some modifications, particularly for tumour size (Table 10.1). The impact of this change was studied in a retrospective survey.[16] The new classification resulted in T1 tumours having a slightly worse prognosis, but no significant impact on disease management was noted. Other staging systems are also used by some centres.[17,18]

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

Table 10.1

The TNM system 5th edition[13]			The TNM system 6th edition[14]	
Tx	Primary tumour cannot be assessed		Tx	Same as 5th edition
T0	No evidence of primary tumour		T0	Same as 5th edition
T1	Tumour $\leq$ 1 cm diameter confined to thyroid		T1	Tumour $\leq$ 2 cm diameter confined to thyroid
T2	Tumour $>$ 1 cm and $\leq$ 4 cm confined to thyroid		T2	Tumour $>$ 2 cm and $\leq$ 4 cm confined to thyroid
T3	Tumour $>$ 4 cm confined to thyroid			
T4	Any size with extrathyroidal extension		T3	Tumour $>$ 4 cm confined to thyroid or any tumour size with minimal extrathyroidal extension
			T4a	Tumour invades subcutaneous tissues, larynx, trachea, oesophagus or recurrent laryngeal nerve
			T4b	Tumour invades prevertebral fascia, mediastinal vessels or encases carotid artery
Nx	Regional lymph nodes cannot be assessed		Nx	Same as 5th edition
N0	No regional lymph node metastases		N0	Same as 5th edition
N1	Regional lymph node metastases		N1	Same as 5th edition
N1a	Metastasis in unilateral cervical nodes		N1a	Metastasis in level VI
N1b	Metastasis in bilateral, midline, contralateral or mediastinal lymph nodes		N1b	Metastasis in other unilateral, bilateral, contralateral or mediastinal lymph nodes
Mx	Distant metastases cannot be assessed		Mx	Same as 5th edition
M0	No distant metastasis		M0	Same as 5th edition
M1	Distant metastasis		M1	Same as 5th edition
		10 year cancer-specific mortality[15]		
Stage I	M0 $<$ 45 years T1 $\geq$ 45 years	1.7%		Tx Nx M0 $<$ 45 years T1 N0 M0 $\geq$ 45 years
Stage II	M1 $<$ 45 years T2 or T3 $\geq$ 45 years	15.8%		Tx Nx M1 $<$ 45 years T2 N0 M0 $\geq$ 45 years
Stage III	T4 or N1 $\geq$ 45 years	30%		T3 N0 M00 $\geq$ 45 years T1-3 N1a M0 $\geq$ 45 years
Stage IV	M1 $\geq$ 45 years	60.9%		A T1-3 N1b M0 $\geq$ 45 years T4a N0-1 M0 $\geq$ 45 years B T4b Nx M0 $\geq$ 45 years C Tx Nx M1 $\geq$ 45 years

THYROID SURGERY

The surgical management of DTC remains controversial. Total thyroidectomy has been associated with higher rates of post-operative morbidity than more conservative surgery. However, the complication rates of total thyroidectomy compared with lesser forms of surgery in centres where experienced surgeons perform large numbers of operations are comparable[19] and increasingly total thyroidectomy is advocated as the surgery of choice.

In one series, low-risk patients with papillary thyroid carcinoma had no improvement in survival rates after undergoing more extensive surgery than lobectomy.[20] However, the same centre in a subsequent study found a greater incidence of local recurrences and nodal metastases after unilateral lobectomy compared with bilateral thyroid lobe resection.[21] Most centres therefore would advise total thyroidectomy for tumours of T2 or higher staging.

Minimally invasive video-assisted thyroidectomy (MIVAT) is a relatively new technique, which has recently been used in patients with thyroid cancer. A prospective study[22] randomly assigned 33 patients with a thyroid nodule proven to be papillary carcinoma to MIVAT or to conventional near total thyroidectomy. ^{131}I uptake and serum thyroglobulin were measured one month after surgery. There were no statistically different outcomes between the two groups. Another study[23] reported that total thyroid resection was achieved in 20/24 patients based on postoperative ultrasonography, ^{131}I uptake scan and serum thyroglobulin levels. Smaller scars and more rapid recovery are potential advantages of this surgical approach.

POST-OPERATIVE ^{131}I ABLATION

Patients who have undergone thyroidectomy are usually referred for ^{131}I ablation to destroy any remaining normal and malignant thyroid cells.[1] ^{131}I ablation also increases the sensitivity of subsequent ^{131}I total body scanning and the specificity of serum thyroglobulin measurements. Studies demonstrate decreased recurrence and disease-specific mortality when ^{131}I ablation is administered, but the data are mainly from high-risk populations.[19,24–26]

Total ablation can be achieved with either high (~3700 MBq) or low (~1100 MBq) doses of ^{131}I.[1] A meta-analysis[27] revealed a statistically significant advantage for high dose over low ablative ^{131}I and a pooled reduction in relative risk of failure of the high dose of about 27%. Ablation was more likely to be achieved after near total thyroidectomy rather than lobectomy. A randomised study of 509 patients[28] comparing eight different ^{131}I doses ranging from 185 to 1850 MBq showed relatively low ablation rates (68%) with doses less than 925 MBq compared with higher doses (81.6%) and little difference between 925 and 1850 MBq.

A recent retrospective analysis of 225 patients[29] treated with 3500 MBq of ^{131}I showed a higher ablation rate after total thyroidectomy compared with lobectomy (98% vs. 90%) and a markedly longer biological half-life of ^{131}I after lobectomy than total thyroidectomy, thus illustrating the advantages of total thyroidectomy in patients who require subsequent ^{131}I ablation.

SUPPRESSIVE THYROXINE THERAPY

Inhibition of TSH secretion with thyroxine therapy is thought to be beneficial as the growth of thyroid tumour cells is controlled by TSH.[30] The target serum TSH is usually $\leq$0.1 mU/L. A recent study on the effects of suppressive thyroxine therapy on bones showed no significant bone loss at any site[31] in either sex, which is highly reassuring.

DIAGNOSTIC USE OF RECOMBINANT HUMAN TSH (rhTSH)

Monitoring of patients with DTC requires iodine whole-body scans and thyroglobulin measurements under conditions of TSH stimulation, classically achieved by thyroxine withdrawal for 4–6 weeks. The main drawback of thyroxine withdrawal is that most patients experience significant hypothyroid symptoms and impaired quality of life for several weeks. An alternative to thyroxine withdrawal is administration of rhTSH intramuscularly, while the patient continues thyroid hormone suppressive therapy. rhTSH is now established as an alternative to thyroxine withdrawal for diagnostic purposes following a randomised study comparing the two alternatives.[32] These findings were confirmed by other studies.[33,34] The combined use of rhTSH-stimulated serum thyroglobulin measurement and high-definition neck ultrasound was shown to have a sensitivity of 92.7% and a negative predictive value of 99%.[35] As a result of such data,[33–35] radioiodine whole-body scanning is being used less frequently. Recent consensus statements based on cumulative experience suggest an extremely low risk of disease recurrence in selected patients who have been treated with total thyroidectomy and ^{131}I ablation and have an undetectable serum thyroglobulin 72 hours after rhTSH.[36,37] Because of the heavy reliance on serum thyroglobulin measurements and the frequent assay interference by antithyroglobulin antibodies, it is crucial that validated assays calibrated against the international standard are used and antithyroglobulin antibodies are measured using a quantitative assay.[38,39]

THERAPEUTIC USE OF rhTSH

Following case reports of the use of rhTSH to aid ^{131}I ablation or therapy, several additional studies have established the efficacy of rhTSH for therapy.[40] A recent prospective randomised study of 60 patients with DTC who had total thyroidectomy showed no differences in ablation rates between thyroid hormone withdrawal and rhTSH using 3700 MBq of ^{131}I.[41] As a result of this study rhTSH has recently been granted European Marketing approval for use in thyroid cancer ablation.

FLUORINE-18 FLUORODEOXYGLUCOSE (FDG) POSITRON EMISSION TOMOGRAPHY (PET) SCANNING

In a study of 19 patients with DTC comparing PET with ^{99}Tc MIBI single-photon emission tomography and post-therapy ^{131}I whole-body scanning, PET scanning in the hypothyroid state was superior to the others in detecting metastatic disease.[42] In a study of 33 patients suspected of having recurrent papillary thyroid cancer, PET/CT fusion imaging had a sensitivity in identifying recurrences of 66%, a specificity of 100%, a positive predictive value of 100%, and a negative predictive value of 27%.[43]

CONCLUSION

The management of DTC is challenging because of lack of prospective randomised trials of treatment due to its typically prolonged course and relative infrequency. When diagnosed early, and treated with total thyroidectomy, radioiodine ablation and suppressive thyroxine therapy, survival approximates that of the background population. rhTSH is likely to replace thyroxine withdrawal for diagnostic monitoring and postoperative ^{131}I ablation in most patients as its efficacy appears to be equivalent. Early detection of recurrent or metastatic disease can be achieved by measurement of serum thyroglobulin under TSH stimulation. PET scanning is likely to be used increasingly for localisation of metastases as it is superior to other imaging modalities.

REFERENCES

1. Mazzaferri EL, Kloos RT. Clinical review 128: current approaches to primary therapy for papillary and follicular thyroid cancer. *J Clin Endocrinol Metab* (2001) **86**: 1447–63.
2. Ries LAG, Eisner MP, Kosary CL *et al.* (eds). *SEER Cancer statistics review*, 1975–2001, National Cancer Institute. Bethesda, MD (2004) (available at http://seer.cancer.gov/csr/1975-2001/).
3. Jemal A, Tiwari RC, Murray T *et al.* Cancer statistics, 2004. *CA Cancer J Clin* (2004) **54**: 8–29.
4. Gloeckler Ries LA, Reichman ME, Lewis DR *et al.* Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. *Oncologist* (2003) **8**: 541–52.
5. Hedinger C, Williams ED, Sobin LH. Histological typing of thyroid tumours. In: *International Histological Classification of Tumours*, 2nd edn, World Health Organization. New York: Springer (1988) 1–20.
6. British Thyroid Association and Royal College of Physicians. Guidelines for the management of thyroid cancer in adults (2002) (available at http://www.british-thyroid-association.org/guidelines.htm).
7. National Comprehensive Cancer Network. Practice guidelines in oncology-thyroid carcinoma (2004) (available at http://www.nccn.org/professionals/physician_gls/PDF/thyroid.pdf).
8. National Institute of Clinical Excellence. Improving outcomes in head and neck cancers – the manual (2004) (available at http://www.nice.org.uk/page.aspx?o=233550).

9. Sherman SI, Brierley JD, Sperling M *et al*. Prospective multicenter study of treatment of thyroid carcinoma: initial analysis of staging and outcome. *Cancer* (1998) **83**: 1012–21.
10. Hay ID, Bergstralh EJ, Goellner JR *et al*. Predicting outcome in papillary thyroid carcinoma: development of a reliable prognostic scoring system in a cohort of 1779 patients surgically treated at one institution during 1940 through 1989. *Surgery* (1993) **114**: 1050–58.
11. Greene FL, Page DL, Fleming ID *et al*. (eds). *AJCC Cancer staging manual*, 6th edn. New York: Springer (2002) 77–87.
12. Cady B. Our AMES is true: how an old concept still hits the mark, or risk group assignment points the arrow to rational therapy selection in differentiated thyroid cancer. *Am J Surg* (1997) **174**: 462–8.
13. UICC 1997. Wittekind Ch, Wagner G (eds). *TNM Klasifikation maligner Tumoren*, 5th edn. Berlin: Springer (1997).
14. UICC 2002. Sobin LH, Wittekind Ch (eds). *TNM Classification of malignant tumors*, 6th edn. New York: Wiley-Liss, (2002) 52–6.
15. Loh KC, Greenspan FS, Gee L, Miller TR, Yeo PP. Pathological tumor-node-metastasis (pTNM) staging for papillary and follicular thyroid carcinomas: a retrospective analysis of 700 patients. *J Clin Endocrinol Metab* (1997) **82**: 3553–62.
16. Dobert N, Menzel C, Oeschger S, Grunwald F. Differentiated thyroid carcinoma: the new UICC 6th edition TNM classification system in a retrospective analysis of 169 patients. *Thyroid* (2004) **14**: 65–70.
17. Van De Velde CJH, Hamming JF, Goslings BM *et al*. Report of the consensus development conference on the management of differentiated thyroid cancer in the Netherlands. *Eur J Cancer Clin Oncol* (1988) **24**: 287–92.
18. Baldet L, Manderscheid JC, Glinoer D *et al*. The management of differentiated thyroid cancer in Europe in 1988: results of an international survey. *Acta Endocrinol (Copenh)* (1989) **120**: 547–58.
19. Tsang RW, Brierley JD, Simpson WJ *et al*. The effects of surgery, radioiodine and external radiotherapy on the clinical outcome of patients with differentiated thyroid carcinoma. *Cancer* (1998) **82**: 375–88.
20. Hay ID, Grant CS, Taylor WF, McConahey WM. Ipsilateral lobectomy vs bilateral lobar resection in papillary thyroid carcinoma: a retrospective analysis of surgical outcome using a novel prognostic system. *Surgery* (1987) **102**: 1088–95.
21. Hay ID, Grant CS, Bergstralh EJ *et al*. Unilateral total lobectomy: is it sufficient surgical treatment for patients with AMES low-risk papillary thyroid carcinoma? *Surgery* (1998) **124**: 958–66.
22. Miccoli P, Elisei R, Materazzi G *et al*. Minimally invasive video assisted thyroidectomy for papillary carcinoma: a prospective study of its completeness. *Surgery* (2002) **132**: 1070–74.
23. Bellantone R, Lombardi CP, Raffaelli M *et al*. Video assisted thyroidectomy for papillary carcinoma. *Surg Endosc* (2003) **17**: 1604–8.
24. DeGroot LJ, Kaplan EL, Straus FH *et al*. Does the method of management of papillary thyroid carcinoma make a difference in outcome? *World J Surg* (1994) **133**: 89–93.
25. Taylor T, Specker B, Robbins J *et al*. Outcome after treatment of high risk papillary and non Hurthle follicular thyroid carcinoma. *Ann Int Med* (1998) **129**: 622–7.
26. Mazzaferri EL. Thyroid remnant 131-I ablation for papillary and follicular thyroid carcinoma. *Thyroid* (1997) **7**: 265–71.
27. Doi SA, Woodhouse NJ. Ablation of the thyroid remnant and 131-I dose in differentiated thyroid cancer. *Clin Endocrinol* (2000) **52**: 765–73.
28. Bal CS, Kumar A, Pant GS. Radio iodine dose for remnant ablation in differentiated thyroid carcinoma: a randomized clinical trial in 509 patients. *J Clin Endocrinol Metab* (2004) **89**: 1666–73.
29. Hoyes KP, Owens SE, Millns MM *et al*. Differentiated thyroid cancer – a clinical feasibility study. *Nucl Med Commun* (2004) **25**: 245–51.
30. Pujol P, Daures JP, Nsakala N *et al*. Degree of thyrotropin suppression as a prognostic determinant in differentiated thyroid cancer. *J Clin Endocrinol Metab* (1996) **81**: 4318–23.
31. Rosen HN, Moses AC, Garber J *et al*. Randomized trial of pamidronate in patients with thyroid cancer: bone density is not reduced by suppressive doses of thyroxine, but is increased by cyclic intravenous pamidronate. *J Clin Endocrinol Metab* (1998) **83**: 2324–30.
32. Haugen BR, Pacini F, Reiners C *et al*. A comparison of recombinant human thyrotrophin and thyroid hormone withdrawal for the detection of thyroid remnant or cancer. *J Clin Endocrinol Metab* (1999) **84**: 3877–85.
33. Robbins RJ, Tuttle RM, Sharaf RN *et al*. Preparation of recombinant human thyrotrophin or thyroid hormone withdrawal are comparable for the detection of residual differentiated thyroid carcinoma. *J Clin Endocrinol Metab* (2001) **86**: 619–25.
34. Pacini F, Molinaro E, Lippi F *et al*. Prediction of disease status by recombinant human TSH-stimulated serum Tg in the post surgical follow up of differentiated thyroid carcinoma. *J Clin Endocrinol Metab* (2001) **86**: 5686–90.
35. Pacini F, Molinaro E, Castagna MG *et al*. Recombinant human thyrotropin-stimulated serum thyroglobulin combined with neck ultrasonography

has the highest sensitivity in monitoring differentiated thyroid carcinoma. *J Clin Endocrinol Metab* (2003) **88**: 3668–73.

36. Mazzaferri EL, Robbins RJ, Spencer CA *et al.* A consensus report of the role of serum thyroglobulin as a monitoring method for low-risk patients with papillary thyroid carcinoma. *J Clin Endocrinol Metab* (2003) **88**: 1433–41.
37. Schlumberger M, Pacini F, Wiersinga WM *et al.* Follow-up and management of differentiated thyroid carcinoma: a European perspective in clinical practice. *Eur J Endocrinol* (2004) **151**: 539–48.
38. Weightman DR, Mallick UK, Fenwich D *et al.* Discordant serum thyroglobulin results generated by two classes of assay in patients with thyroid carcinoma: correlation with clinical outcome after 3 years of follow-up. *Cancer* (2003) **98**: 41–7.
39. Mazzaferri EL, Robbins RJ, Spencer CA *et al.* A consensus report of the role of serum thyroglobulin as a monitoring method for low-risk patients with papillary thyroid carcinoma. *J Clin Endocrinol Metab* (2003) **88**: 1433–41.
40. Luster M, Lippi F, Jarzab B *et al.* rhTSH-aided radioiodine ablation and treatment of differentiated thyroid carcinoma: a comprehensive review. *Endocr Relat Cancer* (2005) **12**: 49–64.
41. Pacini F, Ladenson PW, Schlumberger M, *et al.* Radioiodine ablation of thyroid remnants after preparation with recombinant human thyrotropin in differentiated thyroid carcinoma: results of an international, randomized, controlled study. *J Clin Endocrinol Metab* (2006) **91**: 926–32.
42. Iwata M, Kasagi K, Misaki T *et al.* Comparison of whole-body 18F-FDG PET, 99mTc-MIBI SPET, and post-therapeutic 131I-Na scintigraphy in the detection of metastatic thyroid cancer. *Eur J Nucl Med Mol Imaging* (2004) **31**: 491–8.
43. Nahas Z, Goldenberg D, Fakhry C *et al.* The role of positron emission tomography/computed tomography in the management of recurrent papillary thyroid carcinoma. *Laryngoscope* (2005) **115**: 237–43.

11

Urologic Cancers

EDITH D. CANBY-HAGINO,[1] CATHERINE M. TANGEN,[2] AND IAN M. THOMPSON[1]
[1]The Department of Urology, University of Texas Health Science Center at San Antonio and the San Antonio Cancer Institute, San Antonio, Texas, USA
[2]The Southwest Oncology Group and the Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

Urologic oncology covers the spectrum of a number of organ sites including neoplasms of the kidney, adrenal, ureter, bladder, urethra, testis, penis and adnexal structures. This collection of tumour sites encompasses 44% of all cancers diagnosed in men and 5% of cancers diagnosed in women.[1] Prostate cancer, diagnosed in 230 110 men in 2004 and expected to be diagnosed in 17% of US men during their lifetimes, is the most common non-dermatologic malignancy in men.[1] Through the work of national and international organisations in the conduct of clinical trials, urologic cancers have been the focus of major advances ranging from cancer prevention, early detection, treatment of primary and recurrent disease, as well as development of investigational therapies. Because these tumours can affect important physical functions including urinary function as well as sexual function, the urologic research community has actively evaluated the disease and its treatment and impact on quality of life (QOL). In this chapter, we will review issues related to the spectrum of clinical trials (prevention to early detection to treatment) and highlight these efforts in different anatomic sites. Challenges in study design and execution as well as opportunities for advances with new clinical trials will be discussed.

PRIMARY PREVENTION OF GENITOURINARY MALIGNANCIES

To date, clinical trials for primary prevention of GU malignancies have focused on tumours of the prostate and bladder with the former being the only site with an intervention developed for general population disease prevention. While there are a number of etiologies for tumours of the kidney and testis, their relatively low frequency, the lack of targeted agents, and a poor understanding of the etiology of the majority of tumors have prevented this approach.

BLADDER CANCER PREVENTION

Bladder cancer provides a major opportunity for disease prevention in a high-risk population. Patients who have had a bladder tumour – pathologically, a urothelial neoplasm, arising from the

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

urothelium lining of the bladder – have about a 50% risk of disease recurrence.[2] Modifying factors for recurrence include tumour grade, number of tumours at the previous episode, and frequency of tumour recurrence in the past.[3,4] It is because of this risk of recurrence that patients with a history of bladder cancer undergo regular surveillance endoscopic examinations (cystoscopy) of the bladder to identify tumours early, while they are amenable to local treatment (generally, resection or scraping of the bladder wall to eliminate the tumour).

It is thus possible to identify a population of subjects who have had a previous bladder tumour, and then use additional risk factors to identify those at risk of recurrence. It is then also the standard-of-care for these individuals to be surveyed approximately every three months with cystoscopic examinations for disease recurrence.

This patient scenario has been the subject of multiple Phase II and Phase III studies that have substantially affected patterns of care and dramatically improved patient outcomes. Initially, treatment for high-risk, superficial bladder cancer involved repeated resections and ultimately removal of the bladder (cystectomy). With a series of Phase III clinical trials in the 1980s and 1990s, a sequence of intravesical therapies including Thiotepa, Mitomycin C, Adriamycin and BCG were examined with the ultimate recognition of the superiority of BCG to reduce both the frequency of tumour recurrences and the risk of tumour progression to muscle invasive disease.[5–16]. The rate of discovery of new compounds has slowed recently as additional therapies have had minimal to modest degrees of improvement in outcomes.[17–22] Additional randomised trials have examined single-dose intravesical instillation of chemotherapies to reduce the frequency of tumour recurrence.[23] Because of the relatively high hazard rate for disease recurrence and somewhat lower but significant risk of disease progression over time, bladder cancer prevention is an attractive opportunity for clinical trials. The endpoints chosen for these trials – disease recurrence and progression – are clinically meaningful as tumour recurrence is associated with morbidity (need for invasive surgery, anaesthesia), cost and mortality (associated with development of invasive disease).

PROSTATE CANCER PREVENTION

Prostate cancer prevention has been the subject of one completed Phase III clinical trial and there are currently three other ongoing Phase III trials testing four specific agents. The challenges to the design of prostate cancer prevention trials are enormous and fall into several major categories, as follows.

ENDPOINT

Prostate cancer is a disease of unknown biologic consequence in the majority of men. Studies of observation for localised prostate cancer have increasingly demonstrated extreme variability in the natural history with the majority of most patients in these studies never suffering from disease-related morbidity or mortality.[24,25] This heterogeneous disease progression pattern of prostate cancer is perhaps the greatest confound and challenge to clinical trial design. In the ideal world, an endpoint with the greatest degree of biologic significance – for example, survival or disease-specific survival – would be selected for prevention trials. Unfortunately, if a 25% reduction in risk of death is anticipated, study populations as great as 50 000 with follow-up of 20–30 years would be required. Studies of such size would be prohibitively expensive and the risk of patient drop-out over the course of time would further confound results. A second option would be an endpoint that serves as a surrogate for survival. There are a range of surrogate endpoints that could reduce the duration of follow-up required. One, which is probably a valid endpoint, is the development of metastatic disease. Unless there is differential ascertainment or detection of disease in study groups, metastatic disease is generally associated with a median survival of 29.9–33.5 months and thus could shorten the duration of a trial by perhaps 2.5 years.[26] Unfortunately, in the context of a prevention trial, this reduction would be minimal when compared with the overall duration of follow-up.[27] With the growing body of evidence that prostate specific antigen (PSA) kinetics (e.g. rate of increase, doubling time) are related to the risk of progression and death, this measure may ultimately prove to be a valid surrogate.[28,29] Despite its promise, surrogacy for any PSA measure for a prevention clinical trial has yet to be established and there is a real potential that some interventions could affect PSA production and invalidate this surrogate endpoint. (Examples of these interactions include suramin and five alpha-reductase inhibitors such as finasteride and dutasteride.)[30–37].

At the current time, despite the variable natural history of prostate cancer, disease incidence is probably the most valid study endpoint for prevention trials for prostate cancer. In the United States, for example, about 85% of men who are diagnosed with prostate cancer elect treatment.[38] As treatment is associated with substantial cost as well as morbidity, prevention of these meaningful outcomes, despite the unknown impact on population survival, has been acknowledged as a meaningful and important goal.[39]

STUDY POPULATION

If a study based on prostate cancer incidence is undertaken, the sample size required to seek a 25% reduction in disease incidence relative to placebo in the general population at risk is about 10 000 to 12 000. Durations of such studies are between 7 and 12 years, depending on the postulated efficacy of the agent, the incidence rate in the control group and the rate of subject accrual. Expected levels of compliance and loss to follow-up and death from other causes need to be taken into account. See Table 11.1 for an illustration of how assumptions about incidence rates and treatment effect sizes can affect sample size of a trial. With such huge numbers of subjects required, one approach to reduce the sample size is to select a population at higher risk for disease and hence a higher anticipated incidence rate. Three populations have been suggested including African American men, men with a family history of prostate cancer, and individuals with an elevated PSA and a previous negative prostate biopsy. The first two of these groups clearly have an increased risk of a diagnosis of prostate cancer, but it is not clear at this time that men with an elevated PSA and a negative biopsy actually have a substantially higher risk than men with a normal PSA but who have not had a biopsy. In our analysis of the results of the Prostate Cancer Prevention Trial, we found that men with a PSA < 4.0 ng/ml had an overall risk of prostate cancer on biopsy of about 15%.[40] This is not substantially less than the risk of cancer on repeat biopsy after a previous 10–12 core prostate biopsy.[41,42] Because the etiology of prostate cancer in these high-risk populations may be

Table 11.1. Sample size requirements for a two-arm prostate cancer prevention trial where clinical diagnosis is the primary endpoint

Incidence in placebo Group	Relative risk for active treatment relative to placebo	Total sample size needed*
4%	0.50	8 620
6%	0.50	5 725
8%	0.50	4 275
4%	0.75	16 250
6%	0.75	10 800
8%	0.75	8 070

* Assumes two-sided 0.05 test with 90% power and no competing risks or loss to follow-up or lack of compliance. Uniform accrual for four years and four additional years of follow-up is also assumed. Accounting for these additional factors would increase the sample size.

different from the general public and because the efficacy of interventions may also be different, there are problems with studying only high-risk populations: A preventive effect may be found that may not be seen in the general population *or* no effect may be seen in the smaller, high-risk population, but a true preventive effect may be seen in the general population. Recruitment may also be slower. It is for these reasons that the two largest prostate cancer prevention studies have been general-population-based.

The challenges in the design of a prostate cancer prevention trial were perhaps no greater than in the first Phase III study – the Prostate Cancer Prevention Trial (PCPT). This study was initiated based on a confluence of events: (1) a dramatic increase in prostate cancers detected by the use of PSA in the late 1980s and early 1990s, (2) the understanding that androgens play a major part in prostate carcinogenesis, (3) the side effects with traditional hormonal therapies, and (4) the discovery and FDA registration of finasteride, an inhibitor of five alpha reductase, the enzyme that converts testosterone to the more potent androgen, dihydrotestosterone, the major androgen in the prostate. Perhaps the greatest challenge to this study was the fact that the agent being tested – finasteride–affected PSA. At the time the study was initiated, it was generally known that finasteride reduced PSA by approximately 50%. As such, without controlling for this reduction and using a prostate cancer incidence endpoint, a reduction in prostate cancers detected would be noted but this would be attributed to the artefact of the fall in PSA. To control for this detection bias, two unique steps were taken. It was acknowledged that to maintain current standard-of-care existing in the United States, it would be necessary to provide subjects and their physicians with PSA results on an annual basis and that these results may lead to prostate cancer diagnosed based upon elevated PSA values. There were concerns about the previously reported 'multiply-by-two' conversion factor (based on the 50% PSA fall with finasteride) because the observation period was only 24 months, and PCPT was a seven-year study. An incorrect adjustment factor could introduce confounds and differential ascertainment. This problem was corrected by a 'PSA indexing' procedure: on an annual basis, all men in the placebo group of the study with a PSA above 4.0 ng/ml were recommended to consider a prostate biopsy. The fraction of the total placebo group above 4.0 ng/ml was then calculated. Then, the PSA cutpoint in the finasteride group was adjusted so that the same fraction of men in the treatment group received a recommendation for biopsy as in the placebo group. Practically, a doubling of PSA effectively corrected for this bias for the first three years. Ultimately, the PSA in the finasteride group was multiplied by 2.3 in subsequent years to ensure the same number of biopsies in the two study groups throughout the duration of the study.

Despite this correction to the PSA value, study coordinators could not be certain that other biases in disease ascertainment and detection did not exist. For example, we know that finasteride reduced the size of the prostate gland. This may alter the gland texture, reducing or increasing the proportion of abnormal digital rectal examinations. These and other potential confounds led to a second design characteristic of the study – an end-of-study prostate biopsy. This biopsy was planned in all individuals not previously diagnosed with prostate cancer who reached the seven-year mark on study. It was anticipated that 60% of men randomised in the study would have an endpoint ascertained (interim cancer or end-of-study biopsy). The study design is depicted in Figure 11.1.

The results of the PCPT were reported in 2003.[43] The study was closed early, based on recommendations of an independent Data and Safety Monitoring Committee because there was convincing evidence that the study objective had been met – there was a 24.8% reduction in period prevalence of the disease. Three additional observations of importance were made in the initial data presentation. First, as had been expected, there was an increase in sexual side effects seen with finasteride. Of interest, however, sexual side effects were extremely common

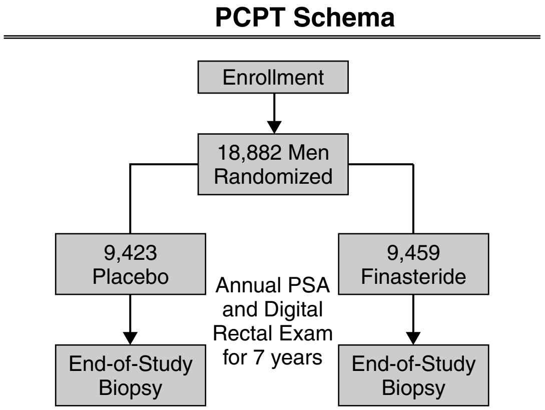


Figure 11.1. Design of the Prostate Cancer Prevention Trial.

in the placebo group. Erectile dysfunction, for example, was seen in 67.4% of men receiving finasteride and in 61.5% of those receiving placebo over the course of the study. Second, urinary side effects and complications were less common with finasteride with men receiving the drug having fewer episodes of prostatitis, BPH symptoms, suffering from urinary retention, or requiring transurethral resection of the prostate. The most important observation beyond the primary objective, however, was the tumour grade in the two groups, grade representing one of the most important surrogates for disease prognosis. The rate of high-grade (Gleason 7–10) disease among men with an endpoint was 5.1% in the placebo group compared with 6.4% in the finasteride group. For biopsies performed for cause (elevated PSA or abnormal digital rectal examination), the rate of high-grade disease was 7.6% in the placebo group compared with 11.5% in the finasteride group. Because of the lower number of tumours in the finasteride group, high-grade disease constituted a much greater proportion of tumours in this group compared with placebo: 37.0% versus 22.2% of the biopsies performed for cause.[43]

Despite the complexity of the study design to optimally compare finasteride and placebo, a variety of biases remained operational, several of which had been anticipated. First, there was a reduction in prostate gland volume from an average of 33.6 cm^3 to 25.5 cm^3, a reduction of 24%.[43] Such a gland reduction, with a similar number of prostate cores biopsies in the two study groups, would be expected to oversample the prostates in the finasteride group of subjects, biasing *against finasteride*.

There are very complex design issues associated with prevention trials, especially those in the area of prostate cancer. Designing and executing these trials is a formidable undertaking and it must be understood from the outset that there is no perfect study design, but that each of the components of the trial must be fully understood at the outset and that potential confounds be planned for in advance, building biases in such a fashion as can be understood as the trial proceeds. Ultimately, the study's primary objective and critical assumptions, the latter being continuously overseen by an independent and well-chosen Data Safety Monitoring Board, will affect how the study is ultimately understood by the medical community.

SECONDARY PREVENTION (SCREENING) FOR GU MALIGNANCIES

In almost no other area of oncology is early detection of cancer as widely debated and as aggressively investigated as in urologic cancer. Prostate cancer has the most widely performed cancer screening test: PSA, a test that is performed annually in over half of the US male population over age 65.[44] In bladder cancer, multiple urinary studies are currently approved for monitoring for bladder cancer recurrence. In kidney cancer, the majority of tumours treated today are found serendipitously, by 'accidental screening' during abdominal imaging with CT, ultrasound or MRI.[45] In each of these areas, important design considerations must be well understood.

BLADDER CANCER

The most common tumour of the bladder is urothelial carcinoma and represents the second most frequent cancer of the GU tract. Two factors related to this disease are important vis-á-vis early detection. First, after the diagnosis of bladder cancer and tumour resection, a person has a 50% risk of disease recurrence. Second, while approximately 75% of all tumours are superficial at presentation (a tumour with a generally low biologic potential and which can generally be controlled with tumour resection alone), 80-90% of invasive tumours, tumours that have a high risk of mortality, are invasive at the time of diagnosis.[46,47] These two observations provide two opportunities for clinically relevant early detection. There are perhaps 5 million individuals with a history of bladder cancer in the United States who undergo regular cystoscopic examinations (from every 3 to 12 months, depending on their pattern of time from previous tumour).[1,48]. An early detection marker (probably a urine test) with high test sensitivity could potentially eliminate these expensive and invasive tests. Currently available urinary tests for this purpose include BTA stat, BTA Trak, NMP-22, DD23, Immunocyt, Quanticyt, Accu-Dx and UroVysion.[49–51] Unfortunately, the concept of 'no free lunch' (there are no clinical tests that maintain both high sensitivity and specificity; there is always a trade-off between the two) has led to poor adoption of these non-invasive tests clinically and the persistence of cystoscopy for early detection and surveillance.

The other potential opportunity for early detection of urothelial carcinoma is population-based screening of high-risk individuals. The basis for this is the improved outcomes of high-risk tumours when detected prior to development of muscle-invasive disease. The attractive nature of these high-risk tumours is that they provide an attractive target for early detection due to their high degree of genetic losses or gains, allowing many signatures to be recognised in urine. The challenges to this approach, however, are the fairly low prevalence of these tumours as well as their potential for rapid development and progression. Thus, small decrements in test specificity (in the presence of low prevalence) would lead to a high risk of false-positive tests. Additionally, the rapid development of these tumours may preclude successful outcomes due to unreasonable frequency of testing that would be required. One method to improve these outcomes would be to focus on high-risk populations such as smokers, individuals with hematuria, individuals with other smoking-related malignancies (e.g. head and neck cancers), or workers in high-risk industries (e.g. exposure to aromatic compounds). To date, very few studies have taken this approach.[52–55].

PROSTATE CANCER

For prostate cancer, early detection is as or more complex than is prevention. First and foremost, it must be recognised that the benefit of screening for prostate cancer is unknown. Currently ongoing is the Prostate, Lung, Colorectal, and Ovarian Cancer (PLCO) prospective clinical trial in which 75 000 men have been randomised to screening with PSA and digital rectal examination or the community standard-of-care. This study is designed to examine the improvement in survival of prostate cancer screening.

Previous studies, examining PSA screening, allowed the test itself to prompt prostate biopsy and individuals with a 'normal' PSA did not have a prostate biopsy. It was for this reason that the test was perceived to have a very high level of specificity. Our analysis of end-of-study biopsies from the PCPT led to a better understanding that PSA cannot be understood as a dichotomous ('positive' or 'negative') test but that PSA demonstrates a range of risk at all levels of the test.[40] This paper has called into question the reliability of conclusions drawn from previous biomarker studies in which controls were not systematically biopsied. (In the PCPT cohort of men with PSA values <4.0 ng/ml, approximately 15% had prostate cancer on biopsy.) Thus, 15% or more of controls in these studies may have actually been cases. Fortunately, Platz and colleagues have shown that small proportions of misclassification will not obscure most associations.[56]

Two other important concerns in these biomarker discovery and validation studies are the concepts of *missing cancer on biopsy* as well as *predictive sensitivity*. With regards to the former, if men with a negative biopsy are included as 'controls' for biomarker studies, this is confounded by the fact that as many as 15-20% of these men would be found to have prostate cancer if the biopsy were repeated. Predictive sensitivity describes the concept that a biomarker may not be indicative of an existing condition, but may predict the likelihood of a future event. For example, in the case of prostate cancer biomarker 'controls' with negative biopsies, one worries about a biomarker that may be elevated in a man whose prostate is undergoing premalignant changes and may display those changes in the serum proteome, for example. The serum protein may thus be outside the 'normal' range, yet at biopsy no cancer is detected. However, if the individual being studied were to be followed for additional years with a subsequent biopsy, the test might then change from a false positive to a true positive (perhaps best labelled as an *anticipatory true positive*).

Perhaps the greatest challenge to prostate cancer biomarker research is the concept of tumour biologic activity or *prognosis*. In the PCPT, examining a very low-risk group of men (92% were white, 85% had no family history of prostate cancer, and all had a PSA at study entry that was less than 3.0 ng/ml), a total of 24% of men on placebo were ultimately found to have prostate cancer within a seven-year period. This observation, in the face of a 3% lifetime mortality risk from prostate cancer, suggests a tremendous potential for overdiagnosis. Thus, if it is determined that the 'ideal' prostate cancer biomarker has 100% sensitivity, the roll-out of this marker would have profound implications on US medicine, leading to dramatic overtreatment of the disease with its attendant cost and morbidity.[57] A superior biomarker then would merge *diagnosis* with *prognosis*. Thus, in the ideal system, the results would be thus:

> In a man without prostate cancer – the test is negative.
>
> In a man with prostate cancer at low risk of dying from the disease – the test is negative.
>
> In a man with prostate cancer that poses a high risk during his lifetime – the test is positive.

To date, no prostate cancer biomarker tests have attempted to merge diagnosis with prognosis. An ongoing study of the Early Detection Research Network examining protein profiling with Surface Enhanced Laser Desorption and Ionization – Time-of-flight, has taken this step. The study is examining three groups of men: men without prostate cancer (as demonstrated by a negative prostate biopsy or with very low levels of PSA), men with low-grade (Gleason 6 and less) prostate cancer, as well as men with high-grade (Gleason 7 and higher) prostate cancer. It is the ultimate goal of this study to determine if there is a unique serum protein pattern for high-grade tumours that separates them from low-grade disease and then again from men without prostate cancer.[58]

ISSUES REGARDING TREATMENT TRIALS FOR GU MALIGNANCIES

PROSTATE CANCER

Major challenges exist related to optimal treatment of prostate cancer. At diagnosis, prostate cancers can be categorised into three groups: clinically localized (no evidence of extraprostatic disease), locally advanced, or metastatic (in most cases, metastases to bone). Presently, three-quarters of all newly diagnosed patients present with clinically localised prostate cancer. For this most common group of tumours there is only one completed clinical trial that provides guidance for individual patients and, in the case of this study, the results do not provide concrete guidance.[59] In this randomised comparison of surveillance and prostatectomy with a median followup of 6.2 years (with enrolment beginning in 1989 and follow-up concluding in 2001), three somewhat divergent observations were made. First, development of metastatic disease was significantly higher in men on surveillance: 27.3% versus 13.4% for prostatectomy ($p = 0.03$). On the other hand, while prostate cancer specific mortality was slightly greater in men on surveillance (13.6% for surveillance versus 7.1% for radical prostatectomy, $p = 0.02$), there was no difference in overall survival (28.3% versus 22.0%, $p = 0.31$). Clearly, the duration of follow-up is inadequate for conclusions to be drawn regarding differences. A similar study from the United States – PIVOT–is currently completing its follow-up, also for a mortality endpoint.[60]

The design, execution and early results of these studies illustrate the multiple challenges of executing meaningful comparisons of treatment in localised prostate cancer. A major obstacle to subject recruitment is the general impression of physicians and patients that individual patients desire to select their treatment and are unwilling to be randomised. Despite considerable evidence that all treatment options have risks and impact on QOL (primarily related to sexual, urinary and bowel function), it is common for purveyors of 'information' related to treatment to characterise one treatment or another as being vastly different.[57,61–63] Adding to this problem is the substantial difference in character of the primary treatment options: surveillance, surgery and radiation. Non-randomised trials and patient selection bias (e.g. younger patients with fewer co-morbidities tend to receive surgery) make it dangerous to draw comparisons across trials. Exacerbating this design challenge is PSA-driven treatment contamination. An example of this is found in the PIVOT study design.[60] For men randomised to surveillance, it is currently standard-of-care for PSA to be monitored on a relatively frequent basis in the United States With evidence that PSA increases over time in men even without prostate cancer, it is challenging for a man with a diagnosis of prostate cancer to watch while his PSA increases. In a study from Toronto of 206 men on surveillance, Fleshner and colleagues found that 52% opted for treatment within four years.[64] PIVOT manages this challenge by blinding participants and physicians to PSA measurements, until PSA reaches a level of 50 ng/ml (or 25 ng/ml for participants using finasteride).

PSA, while a very sensitive measure that is related to risk of disease recurrence, has led to considerable problems with study design for almost all meaningful studies related to prostate cancer outcomes. PSA increases after radiation or surgery often lead to adjuvant treatments – none of which have been proven to alter the natural history of the disease. Additionally, the level of PSA at which a PSA recurrence correlates with actual disease progression is not clear and is truly a matter of sensitivity–specificity trade-offs. For example, while a detectable PSA after radical prostatectomy is generally synonymous with disease recurrence, low levels of PSA can occur without progression and there are occasional patients who have slowly rising PSA levels in whom detectable disease never occurs. After radiation, the picture is even more complicated. In general, the lower the PSA nadir after radiotherapy, the more durable the cancer control. However, there is no generally accepted value which indicates that the disease has persisted. The most commonly accepted evidence of disease recurrence after radiotherapy is a rising value of PSA which has been defined by consensus.[65] Nevertheless, there are multiple problems with this definition.[66] Finally, the lack of a correlation between a PSA nadir value or endpoint with one treatment and the value or endpoint with another makes this endpoint impossible to use to compare treatments for localised prostate cancer.[67]

Unfortunately, while PSA has not been validated as a surrogate marker in prostate cancer and therefore cannot be used to compare treatments for localised prostate cancer, it is clinically used as the most common endpoint to prompt secondary therapy. Thus, a patient who has had a radical prostatectomy and has a detectable or rising PSA will commonly receive secondary therapy, often with hormones or radiation. Similarly, a rising PSA after radiation often leads to secondary hormonal therapy fairly early in the disease process. Thus, studies that are designed to compare one treatment versus another in terms of survival or metastasis-free survival find themselves comparing an uncontrolled group of subsequent treatments versus another uncontrolled group, making comparison of the differences in primary treatments very difficult.

Illustrating the challenge with mounting clinical trials for localised prostate cancer are the results of two Phase III trials comparing radiation and surgery. The first, Southwest Oncology Group study 8890, was designed to compare external beam radiotherapy and radical prostatectomy with a goal of randomising 900–1000 patients. The study accrued a total of six patients in 21 months and was thereafter closed. Nine years later, the American College of Surgeons Oncology Group initiated a randomised trial (Z0070) comparing brachytherapy with radical prostatectomy.[68] Despite considerable efforts and resources to recruit patients, including attempts to open the study in the United Kingdom, the study recruited 56 of the total of 1980 needed within 5.5 years and ultimately closed to accrual within 17 months after it opened. These two Herculean efforts by dedicated investigators across the United States illustrate the considerable difficulties with the conduct of clinical trials in clinically localised disease.

Successes have been achieved in clinical trials in prostate cancer. A series of studies comparing different strategies of administrating radiotherapy have been rapidly accomplished by the Radiation Therapy Oncology Group (RTOG). This success was generally built upon the widespread acceptance of two designs: (1) radiotherapy type A versus radiotherapy type B, and (2) radiation alone versus radiation plus additional therapy. That hormonal therapy improves survival with radiation therapy, primarily in locally advanced prostate cancer, was demonstrated by several radiotherapy clinical trials.[69–71]. Interestingly, in a trial coordinated by RTOG (92-02), investigators showed that men who were randomised to receive more extensive hormonal treatment with radiation had a 50% greater risk of death after PSA progression compared with those randomised to less hormonal treatment. This result brings into question whether PSA progression can be used as a surrogate for survival when different hormonal regimens are being compared in a trial.[72]

A truly novel clinical trial has nearly completed accrual. This study, JPR3, which is being coordinated by the NCI of Canada's Clinical Trials Group, compares radiation plus hormonal therapy versus hormonal therapy alone for locally advanced prostate cancer.[73] While somewhat difficult to accrue patients to this trial due to the substantial difference in treatment administered, accrual has been steady and the conclusions that will result will have far-reaching conclusions for the management of prostate cancer. Even challenging comparisons such as adjuvant radiotherapy for high-risk patients after surgery have been accomplished (Southwest Oncology Group study 8794, comparing observation with adjuvant radiotherapy). Unfortunately, a general tenet of these trials poses a major challenge to their success: the earlier in the disease one studies, the larger the sample size, the longer the follow-up, and the greater potential of confounding from subsequent treatments. Attesting to this fact is the timeline to SWOG 8794: accrual was initiated in 1988 and completed in 1997, with planned publication of results in 2005. A total of 17 years of study accrual and follow-up before results are known, even for high-risk localised disease, is a sobering impediment to organisations that plan studies in locally advanced prostate cancer.

METASTATIC PROSTATE CANCER – HORMONE-NAÏVE

In the realm of metastatic prostate cancer, clinical trials have been considerably more successful. A review of OVID finds a total of 41 Phase III trials completed in this stage of disease. To conduct this search, we explored prostate neoplasms (exploded, all subheadings), limited to Phase III clinical trials. Duplicate reports and those addressing non-metastatic disease were then excluded. The bulk of these studies have examined differing combinations of hormonal therapy for advanced prostate cancer. The success of these studies was often based on the availability of new medications or combinations of medications as well as the poor outcomes of this disease stage, leading many patients to seek out clinical trials. Additionally, with an average survival of approximately 30 months, study outcomes are obtained quickly.

METASTATIC DISEASE – HORMONE-REFRACTORY

Unfortunately, until recently, the clinical impact of these Phase III studies could be reasonably noted to be 'minimal' with little evidence that overall survival of metastatic prostate cancer had been changed. This was generally due to the lack of active agents in disease that was no longer responsive to hormonal therapy. This changed in the Spring of 2004 with the presentation of the results of two Phase III clinical trials. While the studies had a number of design differences, the upshot was similar: the administration of docetaxel improved the survival of patients with hormone refractory prostate cancer.[74,75] With the publication of these two studies, the landscape of clinical trials for advanced prostate cancer has changed considerably and in a number of ways. First, a new 'standard-of-care' appears to have emerged – docetaxel for hormone refractory prostate cancer. Second, this then allows for comparisons of this single agent with other new agents that appear to have activity, often through other pathways, for this disease. These new agents could potentially include atrasentan (an endothelin receptor antagonist), bevacizumab (a VEGF inhibitor), and Genasense (a bcl-2 antisense oligonucleotide), to name just a few. The challenges for clinical trialists at this time will be to prioritise these combination therapies in a rational manner and to potentially select populations of patients who will most likely benefit from combination therapy. Translational studies that examine markers of effect linked with surrogates for disease will almost certainly assist in this prioritisation process.

IMPLICATIONS OF STUDIES IN METASTATIC DISEASE

There are a number of implications of these new observations in advanced prostate cancer for earlier stages of the disease. It is very clear now that, using a variety of methods to risk-stratify patients with early-stage prostate cancer, cohorts of patients who have a high risk of failure of initial therapy (radiotherapy or surgery) can be identified. These risk stratifications are generally based on Gleason score, PSA and tumour stage and may be refined further with results from future biologic studies.[76] With the understanding that 'monotherapy' in these patients is likely to fail, adjuvant therapy trials are currently ongoing or are being designed. One such study, SWOG 9921, is enrolling patients at high risk of recurrence after radical prostatectomy, based on high Gleason score or pathologic stage. All patients receive two years of hormonal therapy and are randomised to receive chemotherapy with mitoxantrone and prednisone or no chemotherapy with the primary endpoint being survival. Several other clinical trials using this design either are enrolling patients or will begin accrual shortly.

As can be seen, the challenges to successful clinical trials in prostate cancer – at all stages – are substantial. PSA, sample size, study duration (including both duration of therapy and duration of follow-up until the primary endpoint is reached), patient preferences (e.g. non-interest in study enrolment as well as crossover to the other study arm (drop-ins and drop-outs)), physician bias, cost, and multiple other considerations are major hurdles to their design, execution and completion. Despite these hurdles, progress has been made and patients are beginning to reap the rewards of these studies. With one man in six diagnosed with prostate cancer in his lifetime and with most clinical questions regarding their treatment lacking high-quality evidence (results of randomised clinical trials) regarding optimal treatment, a continued focus on completion of these trials must be a high priority.

BLADDER CANCER

Treatment trials for bladder cancer can be most easily considered to fall into three groups: superficial disease, muscle-invasive disease and metastatic disease. In superficial disease, the primary focus has been on the urothelial neoplasm that has a high risk of recurrence or progression. Tumour grade and depth of invasion of the superficial component of the bladder wall are the primary determinants of such risk. As endpoints, disease recurrence and progression are quite meaningful endpoints. In the event of disease recurrence, the patient then faces repeat hospitalisation for tumour resection while for progression to muscle-invasive disease, radical cystectomy is a common outcome. In the 1980s and 1990s, a series of phase III clinical trials was completed, enrolling such high-risk patients and using these endpoints. These studies tested a variety of therapeutic agents that were generally instilled into the bladder, ultimately demonstrating that the most effective of these agents was Bacillus Calmette Guerin (BCG).[77] This study design remains of value as new agents are developed that may have equal or superior efficacy than BCG and with less toxicity (due to the irritative urinary symptoms and potential of infection with BCG).

When urothelial carcinoma has invaded the muscular wall of the bladder, the outlook for the patient changes considerably. Without treatment, the risk of metastatic disease and death is considerable. Even with radical cystectomy, there is an 8% risk of isolated local recurrence and a 35% risk of subsequent metastatic disease (and almost universal death from the disease).[78] Evidence that chemotherapy has efficacy in this disease led to studies examining chemotherapy applied in a neoadjuvant role – prior to cystectomy. One such study, SWOG 8710, randomised patients with muscle-invasive disease to MVAC (methotrexate, vinblastine, adriamycin and cisplatin) chemotherapy prior to cystectomy or to cystectomy alone.[79] With a total of 317 patients enrolled, risk of death was reduced by 25% with MVAC. Clearly, however, not all patients with muscle-invasive disease are appropriate candidates for neoadjuvant chemotherapy given that some patients undergoing radical cystectomy will not progress and that chemotherapy will delay a curative procedure in others. As pathologic T0 status at cystectomy appears to be a reasonable surrogate for survival, Phase II studies are currently screening new compounds seeking to achieve this endpoint. With the understanding that molecular risk factors such as p53 mutations help to identify patients who are more likely to benefit from additional systematic cytotoxic therapy, 4B951 – a Phase III Intergroup Trial – was activated in 2001. Following cystectomy for muscle-invasive transitional cell carcinoma, tumours are evaluated for p53 status. Patients with mutant p53 are randomised to chemotherapy vs. observation, while patients with wild-type p53 are observed.

Rapid advances in molecular biology have led to exploitation of molecular targets which promote bladder cancer cell proliferation, invasion and metastasis in patients with metastatic and/or unresectable transitional cell carcinoma. These include proteasomal inhibitors such as bortezomib, epidermal growth factor receptor inhibitors such as trastuzamab and gefinitib, and farnesyl transferase inhibitors such as tipifarnib which inhibit activity of *Ras* gene products (see www.cancer.gov). Some show promising antitumour activity as single agents in laboratory and early clinical trials. However, their greatest utility may be in enhancing efficacy and overcoming tumour resistance to traditional chemotherapeutic agents. Numerous clinical trials are underway exploring their utility alone and in concert with a variety of agents with known activity against urothelial malignancies.

KIDNEY CANCER

The landscape of kidney cancer has changed dramatically in the past two decades, due to two major events – imaging and molecular biology. Prior to the 1980s, the majority of kidney tumours were diagnosed at the time patients presented with symptoms, often with an abdominal mass, hematuria and/or flank pain. Patients were imaged with crude technologies and were often found to have metastatic disease. Those with clinically localised disease were generally treated with radical nephrectomy but disease progression and death were common. With the advent of imaging modalities including ultrasound, CT and MRI as well as their widespread application for almost any patient complaint, patients with solid renal masses were increasingly detected. Such ‘serendipitously detected’ kidney tumours were found to have dramatically different outcomes with most treated patients never suffering disease recurrence.[45]

Due to continuing improvement in the quality of imaging studies and their proliferation in use, the size of tumours detected has decreased dramatically. This size reduction has led to an increasing realisation that partial nephrectomy is as effective as radical nephrectomy for the control of localised disease.[80] The conclusions of these clinical series of partial nephrectomy have been recently called into question through two observations. The first of these is the realisation that smaller renal lesions found by imaging studies are more likely to be benign in distinction to the preponderance of malignant renal cell carcinomas in large renal masses. More important, was a recent series of patients with solid renal masses who were followed in a surveillance protocol. The authors found that of 32 solid renal masses with a median follow-up of 27.9 months, 25% either did not change in size or decreased in size, and only 34% either doubled in size or reached a diameter of 4 cm.[81] These data would suggest that as imaging studies detect smaller and smaller tumours, there is the risk that 'tumours' with minimal biologic significance will be detected for which treatment may be unnecessary.

Perhaps explaining these findings is the growing body of evidence that all renal tumours are not the same. Specific genetic mutations (both germline and somatic) have been linked to distinct subtypes of renal cell carcinoma and are prognostic to some degree.[82] Study of clear cell renal tumours associated with Von-Hippel Lindau disease led to identification of the VHL tumour-suppressor gene on chromosome 3, in both inherited and sporadic tumours. Bi-allelic loss of VHL gene function leads to failure of proteasomal degradation of hypoxia-inducible factor-1α (HIF-1α), and subsequent overexpression of a variety of tumour-promoting factors, including VEGF, glucose transporter 1 (GLUT-1), platelet-derived growth factor (PDGF), transforming growth factor-α (TGF-α) and erythropoietin. Mono-allelic loss of VHL function in animal models results in either no tumours or development of small and slowly growing tumours.[83]

Similar studies of Type I papillary renal cell carcinoma, first in hereditary and then in sporadic cases, led to the identification of mutations in the MET gene on chromosome 7.[82] MET is a proto-oncogene, and the mutations in MET associated with Type I papillary renal cell carcinoma activate tyrosine kinase domains which increase signalling in response to hepatocyte growth factor (HGF). Studies of chromophobe and oncocytic renal tumours led to identification of the BHD gene on chromosome 17.[82] The role of BHD in the development of renal tumours has not yet been clearly identified.

Clinical trials in renal cell carcinoma have, up to the present, been hampered by a lack of effective therapeutic agents. Perhaps the most successful, the paradigm-changing study (S8949) examined the impact of treatment of the primary tumour on the survival of patients with metastatic renal cell carcinoma. The basis for this study was the repeated observation that patients with advanced renal tumours who received immunotherapy were most likely to respond to treatment if their primary (the involved kidney) had been removed. The study design thus called for patients with metastatic renal cancer either to be randomised to immunotherapy with interferon-alpha or to undergo radical nephrectomy prior to immunotherapy. The results were striking with a 27% improvement in median survival in patients treated with 'cytoreductive nephrectomy'.[84]

Background data related to tumour size and molecular genetics of the disease demonstrate that not only are all renal tumours not alike with differing prognoses, but also their biologic characteristics may determine outcomes more so than treatment. The identification of specific mutations associated with distinct subtypes of renal cell carcinoma has led to 'molecular staging' of kidney tumours in clinical trials. Trials are currently underway to investigate numerous molecules which target particular growth factors or signalling mechanisms specific to particular forms of renal tumours.

In the case of clear cell carcinoma, associated with mutations in the VHL gene, therapies are in development which target HIF-1α and its attendant growth factors.[85] These include agents which promote proteasomal degradation of HIF (bortezomib), those which inhibit HIF synthesis (rapamycin and YC-1), monoclonal antibodies to VEGF (bevacizumab) and EGF (ABX-EGF), and signal transduction inhibitors of VEGF, PDGF and c-kit tyrosine kinase.

A renaissance is also underway in immunotherapy for renal cell carcinoma. Treatment of advanced renal cell carcinoma with relatively non-specific immunotherapies with interleukins and interferons has tantalising potential but disappointing efficacy. Agents designed to enhance these non-specific therapies include MDX-010, a monoclonal antibody which inhibits down-regulation of cytotoxic T lymphocyte activity. More specific immune therapies in development include stem cell transplantation, autologous dendritic cell stimulation, patient-specific anti-tumour antibodies and stimulation of host immune response through vaccination (see www.cancer.gov). The concept of 'molecular staging' has also been applied to immunotherapy. An NCI-sponsored trial currently open to accrual is investigating individualised HLA-appropriate immunisations which target a subset of clear cell tumours expressing Fibroblast Growth Factor-5.

As in prostate cancer, the challenge with this proliferation of agents will be to use the agents in a rational manner. Questions that must be answered include:

1. Do different subtypes of renal tumours respond differently to different agents (i.e. is there a best agent for each different subtype)?
2. Is there a role for use of these agents in high-risk localised disease?
3. Should these agents be combined? If so, which combinations for which tumours?

QOL IN CLINICAL TRIALS OF GU MALIGNANCIES

One important aspect of clinical trials in GU malignancies that is often ignored is the aspect of QOL. In perhaps no other set of organ sites is QOL more affected by the tumour and its treatment with sexual, bowel and urinary function as well as body image affected on a regular basis. Oftentimes, clinical trials examine treatment options with a single focus on survival, ignoring the QOL trade-off that often comes with treatment.

There are many challenges to the assessment of QOL in clinical trials for GU malignancies – challenges that are often so significant that these endpoints are generally not measured. The hurdles to these QOL assessments include:

1. Cost – Such studies add tremendously to the overall cost of clinical trials.
2. Complexity – QOL studies add considerable patient and investigator time and may compromise either accrual or study adherence if they are too burdensome.
3. Assessment frequency and integration of values – This is a major challenge of studies comparing treatments that are substantially different. An example of this might be radiation therapy versus radical prostatectomy examining urinary incontinence. As stress urinary incontinence is present in many patients immediately after surgery, early assessments may indicate major impact on QOL whereas later assessments would be much less.
4. Missing data – In the advanced disease setting, missing QOL assessments are a very real problem for making correct conclusions about changes in QOL over time and making comparisons between treatments. Patients who are deteriorating quickly are more likely to have incomplete QOL data, which results in a biased representation of the study population for post-baseline timepoints. There are some statistical methods that can be used to evaluate patterns of missing data and potential bias, but great care must be taken in the interpretation of results.
5. Comparison of QOL endpoints – If we examine urinary function again in surgery and radiation for prostate cancer, the *type* of dysfunction is different in the two: patients after

surgery most often would have stress urinary incontinence while patients after radiation would be more likely to have urgency incontinence or an increase in urinary obstructive symptoms. How these two widely different symptom complexes are compared is a major challenge.

There is a general agreement that QOL studies should include global assessments of the patient's QOL but should also include measures that are most likely to be associated with the individual treatments. Thus, in many GU clinical trials, validated instruments that examine sexual and urinary function are often paired with global QOL instruments. The common observation of a lack of concordance between patient-reported and physician-reported QOL mandates that these trials include validated instruments that are completed by patients themselves.

Perhaps the greatest challenge in QOL clinical trials is the integration of the results of the therapeutic trial itself with the QOL outcomes. Certainly, if one treatment proves superior from a therapeutic standpoint and the QOL outcome parallels this outcome, this is not a problem. The problem arises when the results diverge and there is a therapy–QOL trade-off.

A very practical example of the importance of QOL can be found in the recent taxotere studies for hormone-refractory prostate cancer. These studies reported a two and three month improvement in survival in the taxotere arm over the least effective or other arm of the study.[74,75] In both studies, there was evidence that these arms had a greater degree of side effects and toxicity. QOL was evaluated in both studies. Eisenberger *et al*. showed a greater proportion of patients on the three week taxotere arm had pain relief relative to the mitoxantrone arm, whereas the Petrylak *et al*. study showed similar pain relief for both groups. The difference may be due to the fact that the first study included prednisone on all arms whereas the second study did not combine prednisone with taxotere. It is clear from these two recent reports that, upon completion of these Phase III trials, it will be common for both patients and physicians to continue to struggle with the question 'Is the improvement in survival worth the toxicity associated with the drug?'

ACKNOWLEDGEMENTS

Supported by funding from the National Cancer Institute, Early Detection Research Network (U01 CA86402).

REFERENCES

1. Jemal A *et al*. Cancer statistics, 2004. *Ca Cancer J Clin* (2004) **54**(1): 8–29.
2. Droller MJ. Bladder cancer: state-of-the-art care. *Ca Cancer J Clin* (1998) **48**(5): 269–84.
3. Donat SM. Evaluation and follow-up strategies for superficial bladder cancer. *Urol Clin North Am* (2003) **30**(4): 765–76.
4. Ali-El-Dein B *et al*. Superficial bladder tumours: analysis of prognostic factors and construction of a predictive index. *BJU Int* (2003) **92**(4): 393–9.
5. Malmstrom P. Improved patient outcomes with BCG immunotherapy vs. chemotherapy – Swedish and worldwide experience. *Eur Urol* (2000) **37** (Suppl 1): 16–20.
6. Huben RP. Intravesical chemotherapy versus immunotherapy for superficial bladder cancer. *Semin Urol Oncol* (1996) **14**(1 Suppl 1): 17–22.
7. Lamm DL. Optimal BCG treatment of superficial bladder cancer as defined by American trials. *Eur Urol* (1992) **21** (Suppl 2): 12–16.
8. Ohe H *et al*. Intravesical instillation of adriamycin for bladder tumors. *Cancer Chemother Pharmacol* (1992) **30** (Suppl): S45–9.
9. Brosman SA. Bacillus Calmette-Guerin immunotherapy. Techniques and results. *Urol Clin North Am* (1992) **19**(3): 557–64.
10. Richie JP. Intravesical chemotherapy. Treatment selection, techniques, and results. *Urol Clin North Am* (1992) **19**(3): 521–7.
11. Bouffioux C *et al*. Adjuvant chemotherapy of recurrent superficial transitional cell carcinoma: results of a European organization for research on treatment of cancer randomized trial comparing intravesical instillation of thiotepa, doxorubicin and cisplatin. The European Organization for Research on Treatment of Cancer Genitourinary Group. *J Urol* (1992) **148**(2 Pt 1): 297–301.
12. Lamm DL *et al*. A randomized trial of intravesical doxorubicin and immunotherapy with bacille Calmette-Guerin for transitional-cell carcinoma of

the bladder. *New Engl J Med* (1991) **325**(17): 1205–9.

13. Khanna OP *et al*. Multicenter study of superficial bladder cancer treated with intravesical bacillus Calmette-Guerin or adriamycin. Results of long-term follow-up. *Urology* (1991) **38**(3): 271–9.
14. Khanna OP *et al*. Multicenter study of superficial bladder cancer treated with intravesical bacillus Calmette-Guerin or adriamycin. *Urology* (1990) **35**(2): 101–8.
15. Martinez-Pineiro JA *et al*. Bacillus Calmette-Guerin versus doxorubicin versus thiotepa: a randomized prospective study in 202 patients with superficial bladder cancer. *J Uro* (1990) **143**(3): 502–6.
16. Martinez-Pineiro JA *et al*. Intravesical therapy comparing BCG, adriamycin, and thiotepa in 200 patients with superficial bladder cancer: a randomized prospective study. *Prog Clin Biol Res* (1989) **310**: 237–52.
17. Steinberg G *et al*. Efficacy and safety of valrubicin for the treatment of Bacillus Calmette-Guerin refractory carcinoma *in situ* of the bladder. The Valrubicin Study Group. *J Urol* (2000) **163**(3): 761–7.
18. Anonymous. Valrubicin for bladder cancer. *Med Lett Drugs Ther* (1999) **41**(1049): 32.
19. Mohanty NK *et al*. Combined low-dose intravesical immunotherapy (BCG + interferon alpha-2b) in the management of superficial transitional cell carcinoma of the urinary bladder: a five-year follow-up. *J Chemother* (2002) **14**(2): 194–7.
20. Brown DH, Wagner TT, Bahnson RR. Interferons and bladder cancer. *Urol Clin North Am* (2000) **27**(1): 171–8.
21. Belldegrun AS *et al*. Superficial bladder cancer: the role of interferon-alpha [erratum appears in J Urol (1998) **160**(4): 1444]. *J Urol* (1998) **159**(6): 1793–1801.
22. Jimenez-Cruz JF *et al*. Intravesical immunoprophylaxis in recurrent superficial bladder cancer (Stage T1): multicenter trial comparing bacille Calmette-Guerin and interferon-alpha. *Urology* (1997) **50**(4): 529–35.
23. Sylvester RJ, Oosterlinck W., van der Meijden AP. A single immediate postoperative instillation of chemotherapy decreases the risk of recurrence in patients with stage Ta T1 bladder cancer: a meta-analysis of published results of randomized clinical trials. *J Urol* (2004) **171**(6 Pt 1): 2186–90.
24. Johansson JE *et al*. Natural history of early, localized prostate cancer. *JAMA* (2004) **291**(22): 2713–19.
25. Albertsen PC *et al*. Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer. *JAMA* (1998) **280**(11): 975–80.
26. Crawford ED *et al*. A controlled trial of leuprolide with and without flutamide in prostatic carcinoma.[erratum appears in New Engl J Med (1989) **321**(20): 1420]. *New Engl J Med* (1989) **321**(7): 419–24.
27. Eisenberger MA *et al*. Bilateral orchiectomy with or without flutamide for metastatic prostate cancer. *New Engl J Med* (1998) **339**(15): 1036–42.
28. Loberg RD *et al*. Prostate-specific antigen doubling time and survival in patients with advanced metastatic prostate cancer. *Urology* (2003) **62**(Suppl 1) 128–33.
29. D'Amico AV *et al*. Determinants of prostate cancer specific survival following radiation therapy during the prostate specific antigen era. *J Urol* (2003) **170**(6 Pt 2): S42–6; discussion S46–7.
30. Small EJ *et al*. Randomized study of three different doses of suramin administered with a fixed dosing schedule in patients with advanced prostate cancer: results of intergroup 0159, cancer and leukemia group B 9480. *J Clin Oncol* (2002) **20**(16): 3369–75.
31. Sridhara R *et al*. Evaluation of biomarkers of survival response in hormone-refractory prostate cancer patients treated with suramin. *Cancer Epidemiol Biomarkers Prev* (1998) **7**(7): 631–4.
32. Thalmann GN *et al*. Suramin-induced decrease in prostate-specific antigen expression with no effect on tumor growth in the LNCaP model of human prostate cancer. *J Nat Cancer Inst* (1996) **88**(12): 794–801.
33. Roehrborn CG *et al*. Efficacy and safety of dutasteride in the four-year treatment of men with benign prostatic hyperplasia. *Urology* (2004) **63**(4): 709–15.
34. Zhu YS *et al*. Androgen-induced prostate-specific antigen gene expression is mediated via dihydrotestosterone in LNCaP cells. *J Androl* (2003) **24**(5): 681–7.
35. Pannek J *et al*. Influence of finasteride on free and total serum prostate specific antigen levels in men with benign prostatic hyperplasia. *J Urol* (1998) **159**(2): 449–53.
36. Oesterling JE *et al*. Biologic variability of prostate-specific antigen and its usefulness as a marker for prostate cancer: effects of finasteride. The Finasteride PSA Study Group. *Urology* (1997) **50**(1): 13–18.
37. Wang LG *et al*. Down-regulation of prostate-specific antigen expression by finasteride through inhibition of complex formation between androgen receptor and steroid receptor-binding consensus in the promoter of the PSA gene in LNCaP cells. *Cancer Res* (1997) **57**(4): 714–19.
38. Cooperberg MR *et al*. The contemporary management of prostate cancer in the United States: lessons from the cancer of the prostate strategic

urologic research endeavor (CapSURE), a national disease registry. *J Urol* (2004) **171**(4): 1393–401.

39. Ford LG *et al.* Prevention and early detection clinical trials: opportunities for primary care providers and their patients. *Ca Cancer J Clin* (2003) **53**(2): 82–101.
40. Thompson IM *et al.* Prevalence of prostate cancer among men with a prostate-specific antigen level < or = 4.0 ng per milliliter. *New Engl J Med* (2004) **350**(22): 2239–46.
41. Singh H *et al.* Predictors of prostate cancer after initial negative systematic 12 core biopsy. *J Urol* (2004) **171**(5): 1850–4.
42. Mian BM *et al.* Predictors of cancer in repeat extended multisite prostate biopsy in men with previous negative extended multisite biopsy. *Urology* (2002) **60**(5): 836–40.
43. Thompson IM *et al.* The influence of finasteride on the development of prostate cancer. *New Engl J Med* (2003) **349**(3): 215–24.
44. Ross LE *et al.* Prostate-specific antigen test use reported in the 2000 National Health Interview Survey. *Prev Med* (2004) **38**(6): 732–44.
45. Leslie JA, Prihoda T., Thompson IM. Serendipitous renal cell carcinoma in the post-CT era: continued evidence in improved outcomes. *Urol Oncol* (2003) **21**(1): 39–44.
46. Kaye KW, Lange PH. Mode of presentation of invasive bladder cancer: reassessment of the problem. *J Urol* (1982) **128**(1): 31–3.
47. Hopkins SC, Ford KS, Soloway MS. Invasive bladder cancer: support for screening. *J Urol* (1983) **130**(1): 61–4.
48. Anonymous. *U.S. and World Population Clocks – POPClocks.* US Census Bureau (2004).
49. Sawczuk IS *et al.* DD23 Biomarker: a prospective clinical assessment in routine urinary cytology specimens from patients being monitored for TCC. *Urol Oncol* (2002) **7**(5): 185–90.
50. Dey P. Urinary markers of bladder carcinoma. *Clin Chim Acta* (2004) **340**(1–2): 57–65.
51. Mian C *et al.* Liquid-based cytology as a tool for the performance of uCyt+ and Urovysion Multicolour-FISH in the detection of urothelial carcinoma. *Cytopathology* (2003) **14**(6): 338–42.
52. Felknor SA *et al.* Bladder cancer screening program for a petrochemical cohort with potential exposure to beta-napthylamine. *J Occup Environ Med* (2003) **45**(3): 289–94.
53. Marsh GM, Cassidy LD, The Drake Health Registry Study: findings from fifteen years of continuous bladder cancer screening. *Am J Ind Med* (2003) **43**(2): 142–8.
54. Shinka T *et al.* Clinical study on urothelial tumors of dye workers in Wakayama City. *J Urol* (1991) **146**(6): 1504–7.
55. Crosby JH *et al.* Cytologic detection of urothelial cancer and other abnormalities in a cohort of workers exposed to aromatic amines. *Acta Cyt* (1991) **35**(3): 263–8.
56. Platz E, DeMarzo A, Giovannucci E. Prostate cancer association studies: pitfalls and solutions to cancer misclassification in the PSA era. *J Cell Biochem* (2004) **91**: 553–71.
57. Bermejo C *et al.* Localized prostate cancer: quality of life meets Whitmore's legacy. *J Natl Cancer Inst* (2004) **96**(18): 1348–9.
58. Grizzle WE *et al.* Serum protein expression profiling for cancer detection: validation of a SELDI-based approach for prostate cancer. *Di Markers* (2003) **19**(4–5): 185–95.
59. Holmberg L *et al.* A randomized trial comparing radical prostatectomy with watchful waiting in early prostate cancer. *New Engl J Med* (2002) **347**(11): 781–9.
60. Wilt TJ, Brawer MK. The Prostate Cancer Intervention Versus Observation Trial: a randomized trial comparing radical prostatectomy versus expectant management for the treatment of clinically localized prostate cancer. *J Urol* (1994) **152**(5 Pt 2): 1910–14.
61. Hall JD *et al.* Why patients choose prostatectomy or brachytherapy for localized prostate cancer: results of a descriptive survey. *Urology* (2003) **61**(2): 402–7.
62. Patel HR, Mirsadraee S, Emberton M. The patient's dilemma: prostate cancer treatment choices. *J Urol* (2003) **169**(3): 828–33.
63. Potosky AL *et al.* Five-year outcomes after prostatectomy or radiotherapy for prostate cancer: the prostate cancer outcomes study. *J Natl Cancer Inst* (2004) **96**(18): 1358–67.
64. Choo R *et al.* Feasibility study: watchful waiting for localized low to intermediate grade prostate carcinoma with selective delayed intervention based on prostate specific antigen, histological and/or clinical progression. *J Urol* (2002) **167**(4): 1664–9.
65. Horwitz EM *et al.* The correlation between the ASTRO Consensus Panel definition of biochemical failure and clinical outcome for patients with prostate cancer treated with external beam irradiation. American Society of Therapeutic Radiology and Oncology. *Int J Radiat Oncol Biol Phys* (1998) **41**(2): 267–72.
66. Williams SG. Ambiguities within the ASTRO consensus definition of biochemical failure: never assume all is equal. *Int J Radiat Oncol Biol Phys* (2004) **58**(4): 1083–92.
67. Amling CL *et al.* Defining prostate specific antigen progression after radical prostatectomy: what is the most appropriate cut point?. *J Urol* (2001) **165**(4): 1146–51.

68. American College of Surgeons Oncology Group website. www.acosog.org.
69. Bolla M *et al*. Long-term results with immediate androgen suppression and external irradiation in patients with locally advanced prostate cancer (an EORTC study): a phase III randomised trial. *Lancet* (2002) **360**(9327): 103–6.
70. Pilepich MV *et al*. Phase III radiation therapy oncology group (RTOG) trial 86-10 of androgen deprivation adjuvant to definitive radiotherapy in locally advanced carcinoma of the prostate. *Int J Radiat Oncol Biol Phys* (2001) **50**(5): 1243–52.
71. D'Amico, AV *et al*. 6-month androgen suppression plus radiation therapy vs radiation therapy alone for patients with clinically localized prostate cancer: a randomized controlled trial. *JAMA* (2004) **292**(7): 821–7.
72. Sandler H *et al*. Can biochemical failure (ASTRO definition) be used as a surrogate endpoint for prostate cancer survival in phase III localized prostate cancer clinical trials? Analysis of RTOG protocol 92-02. *Proc Am Soc Clin Oncol* (2003) **22** (abstr 1529): 381.
73. Anonymous. *Intergroup (NCIC CTG, CUOG, ECOG, CALGB, SWOG) Phase III randomized trial comparing total androgen blockade versus total androgen blockade plus pelvic irradiation in clinical stage T3 – 4, N0, M0 adenocarcinoma of the prostate*. www.swog.org.
74. Eisenberger MA *et al*. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. *New Engl J Med* (2004) **351**(15): 1502–12.
75. Petrylak DP *et al*. Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *New Engl J Med* (2004) **351**(15): 1513–1520.
76. D'Amico AV *et al*. Determinants of prostate cancer-specific survival after radiation therapy for patients with clinically localized prostate cancer. *J Clin Oncol* (2002) **20**(23): 4567–73.
77. Lamm DL *et al*. Maintenance bacillus Calmette-Guerin immunotherapy for recurrent TA, T1 and carcinoma *in situ* transitional cell carcinoma of the bladder: a randomized Southwest Oncology Group Study. *J Urol* (2000) **163**(4): 1124–9.
78. Madersbacher S. *et al*. Radical cystectomy for bladder cancer today – a homogeneous series without neoadjuvant therapy. *J Clin Oncol* (2003) **21**(4): 690–6.
79. Grossman HB *et al*. Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer. [erratum appears in New Engl J Med (2003) **349**(19): 1880]. *New Engl J Med* (2003) **349**(9): 859–66.
80. Fergany AF, Hafez KS, Novick AC, Long-term results of nephron sparing surgery for localized renal cell carcinoma: 10-year followup. *J Urol* (2000) **163**(2): 442–5.
81. Volpe A *et al*. The natural history of incidentally detected small renal masses. *Cancer* (2004) **100**(4): 738–45.
82. Linehan WM, Walther MM, Zbar B, The genetic basis of cancer of the kidney. *J Urol* (2003) **170**(6 Pt 1): 2163–72.
83. Gnarra JR. von Hippel-Lindau gene mutations in human and rodent renal tumors–association with clear cell phenotype. *J N Cancer Inst* (1998) **90**(22): 1685–7.
84. Flanigan RC *et al*. Nephrectomy followed by interferon alfa-2b compared with interferon alfa-2b alone for metastatic renal-cell cancer. *New Engl J Med* (2001) **345**(23): 1655–9.
85. Pantuck AJ *et al*. Pathobiology, prognosis, and targeted therapy for renal cell carcinoma: exploiting the hypoxia-induced pathway. *Clin Cancer Res* (2003) **9**(13): 4641–52.

12

Haematologic Cancers: Challenges in Developing New Therapeutic Approaches

CHARLES A. SCHIFFER[1] AND STEPHEN L. GEORGE[2]
[1]Division of Hematology/Oncology, Karmanos Cancer Institute, Wayne State University School of Medicine, Detroit, Michigan, USA
[2]Department of Biostatistics and Bioinformatics, CALGB Statistical Center, Duke University Medical Center, Durham, North Carolina, USA

INTRODUCTION

The haematologic malignancies are a diverse group of disorders which include diseases involving many different biologically distinct cells of origin, with clinical courses that can range from days to weeks or years to decades, and which have varying incidence and outcome in patients of different ages. With a few exceptions, there is a considerable need for improvements in treatment and outcome in patients affected by these disorders. Study design issues vary according to the clinical characteristics of particular haematologic cancers. This chapter examines the challenges in developing new therapeutic approaches in haematologic malignancies by focusing on the particular example of acute myeloid leukaemia (AML) in older patients.

AML has been studied extensively both in the clinic and in the laboratory, in part because of the accessibility of cells for *in vitro* testing, and has served as a model for the elucidation of many of the principles of anti-neoplastic therapy and infectious disease and transfusion medicine supportive care. Complete remission after initial therapy is achieved in about two-thirds of patients, a significant fraction of whom can be cured with additional post-remission treatment. Unfortunately, however, the great majority of patients eventually relapse and succumb to complications of the disease and its treatment. AML occurs across the entire age spectrum but is most common in adults greater than 60 years of age and the prognosis is particularly poor in these individuals. Indeed, it is arguable that the prognosis of older patients with AML has not changed in the last 15 years. Therefore, it is important to evaluate new therapies in as efficient a fashion as possible. A number of issues serve as impediments to new drug development in haematologic malignancies, some of which are idiosyncratic to AML. This chapter will review some of these

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

problems and suggest and discuss some of the statistical issues in trial design.

BACKGROUND

AML occurs as a consequence of an acquired mutation in a haematopoietic stem cell which results in a failure of normal maturation and differentiation of myeloid cells and an accumulation of juvenile leukaemia cells or “blasts”. By mechanisms which are not well understood, the expansion of this malignant clone suppresses normal blood cell formation. Patients usually present with symptoms related to an absence of normal blood cells including weakness and fatigue due to anaemia, infection because of decreased number of normal white blood cells and bleeding because of marked decreases in the number of platelets. Because this is a systemic disease, initial treatment is with chemotherapy, generally including an anthracycline and cytarabine (ara-c), administered for three and seven days respectively. This induces a period of low blood counts for three to four weeks during which time the patient is at risk for bleeding and infection and invariably requires transfusions of red blood cells, platelets and the use of systemic antibiotics. Should therapy be successful, a “complete remission” is obtained which is defined as normal blood counts and bone marrow with no evidence of leukaemia.[1] It is known that small amounts of leukaemia remain, which cannot be detected morphologically with the microscope, because without further post-remission therapy, leukaemia invariably relapses, generally within six months. The complete remission rate is approximately 75% in younger patients but only approximately 50% in patients greater than 60 years of age.

Major improvements in treating infectious diseases and the use of transfusions and anti-emetic supportive care have left drug-resistant leukaemia (i.e. persistence of leukaemia after treatment) as the major cause of initial treatment failure. Large numbers of randomised trials have been performed in patients with AML including comparative evaluations of different doses and types of chemotherapeutic agents, the use of haematopoietic growth factors, stem cell transplantation in first remission and modulation of various mechanisms of intrinsic drug resistance.[2–9] Some of these trials have, in fact, changed the standard of care in younger patients. Intensive post-remission chemotherapy regimens have improved outcome in younger patients such that approximately 35% of patients less than 50 years of age remain disease free after three years with almost all of these patients being functionally “cured” of their leukaemia. In most studies, however, the newer approaches failed to improve remission rates and overall survival. For example, randomised trials evaluating higher dose therapy in first remission with either autologous or allogenic stem cell rescue have produced results similar to those achieved with chemotherapy alone, although the causes of treatment failure vary, with lower rates of relapse with transplant approaches offset by higher-treatment-associated mortality.[7]

Clinical trials by cooperative cancer treatment groups around the world have shown remarkably consistent and poor results in patients 60 years and older, with complete remission rates of around 50% with less than 10% of patients remaining in long-term remission and surviving for four years.[2,8,9,10] There is also an additional selection process favouring inclusion of “better risk” older patients on these trials, in that older patients with poorer performance status and organ dysfunction are often excluded from protocol treatment. Thus, the overall results in the general population of older patients with AML is probably even worse than that observed on clinical trials. Increasing the complexity further, AML is a biologically extremely heterogeneous disorder. Multiple subtypes can be identified by cytogenetic or molecular studies of the leukaemia cells. Some of these subtypes (predominately found in younger patients) have an excellent prognosis with cure rates of approximately 60%, whereas other subtypes, generally characterised by chromosome loss and duplication, have few, if any, patients cured with chemotherapy alone.[11] The latter are much more common in older patients, particularly those in whom AML developed as a progression of a prior bone marrow disorder either of a myeloproliferative nature or more commonly following a myelodysplastic syndrome.

Such cytogenetic and molecular characterisation of AML can be critical, as evidenced by the remarkable results achieved in recent years in acute progranulocytic leukaemia (APL), a subgroup representing about 10% of patients with AML, predominantly in younger adults and children.[12] All patients with APL have a balanced translocation between chromosomes 15 and 17 resulting in a mutation of the gene coding for the retinoic acid nuclear receptor. By complex mechanisms, this confers unique sensitivity to an oral retinoid, all-*trans* retinoic acid (ATRA), which has appreciably fewer side effects than traditional chemotherapy. A series of randomised trials have elucidated the optimal means of combining ATRA with chemotherapy such that more than 70% of patients with this previously devastating leukaemia can be cured.[13] Interestingly, APL is also uniquely sensitive to arsenical compounds. It is hoped that similar strategies with different compounds can be discovered for other AML variants with discrete activating mutations as has recently also been achieved in patients with chronic myeloid leukaemia (CML) with the tyrosine kinase inhibitor imatinib mesylate, which specifically targets the abnormal enzyme produced by the bcr/abl mutation characteristic of CML.[14]

Some studies have suggested differential responsiveness of AML subtypes to different types of chemotherapy.[15] In particular, patients with more favourable balanced translocations seem to benefit from high-dose ara-c-based consolidation therapy. In contrast, older patients with chromosomal changes associated with poor prognosis do not benefit from these more intensive chemotherapeutic approaches and experience greater, and sometimes fatal, side effects from intensification of therapy. Because of these issues, many treatment groups have devised different therapeutic approaches for older and younger patients.

IMPROVING THERAPY FOR OLDER PATIENTS WITH AML

Outcomes are inferior in patients greater than the age of 60 years as a consequence of more intrinsic leukaemia cell drug resistance and more baseline organ dysfunction than are encountered in younger individuals. New therapeutic approaches should focus both on increasing remission rates as well as on prolonging remission and enhancing the cure fraction of such patients. Many studies of new agents for AML have focused on older patients because of the large numbers of such patients as well as the feeling that the overall results of therapy are so poor that it would be possible to rapidly identify truly active agents because differences with historical or randomised controls would be obvious.

However, there are a number of both practical and biologic issues complicating the conduct of such trials in both the initial induction and post-remission settings:

- A focus on patients with highly resistant disease represents a particularly high hurdle for new therapies and treatments. Modest, but nonetheless important, benefits which could be of value to other AML patients could be missed by studying only patients in very poor prognostic groups.
- AML in older individuals is biologically and clinically heterogeneous. Some therapies might be appropriate only for certain AML subtypes and positive effects can be missed when tested in the overall AML population. This may be particularly true for newer molecularly "targeted" therapies.
- Evaluation of post-remission manipulations is made more difficult by the low complete response rate, so that less than 50% of older patients initially entered on trial are eligible for post-remission treatment.
- Many patients are not candidates for intensive therapy because of compromised organ function from toxicities encountered during induction, as well as the observation that many older patients do not recover normal blood counts

even after a significant antileukaemic response during induction.

- Many older individuals decline post-remission treatment, preferring to spend their remaining time outside of the hospital, as far from aggressive medical ministrations as possible. Thus, randomised studies of new therapies introduced post-remission need larger numbers of enrolled patients to account for this drop-off in patients as the study progresses. This represents a major issue in the United States since only a small fraction of such patients are captured for clinical trials.

These problems are particularly relevant today, because there is no shortage of new agents which merit evaluation in AML. In addition to a continued supply of cytotoxic drugs, there will be large numbers of anti-angiogenesis compounds, immune modulators, signal transduction inhibitors (either with specific or more generic enzymatic targets), as well as new and less toxic approaches to stem cell transplantation. It is likely that additional targeted drugs will be identified as molecular characterisation of the leukaemias becomes more sophisticated. Many of the non-cytotoxic therapies also have the allure of oral treatment with potentially fewer side effects, a feature of particular importance to older individuals.

STATISTICAL ISSUES IN DESIGN AND ANALYSIS

Because of the nature of AML and its treatment, several statistical issues in the design and analysis of clinical trials need special attention. Five of these are discussed in this section: outcome measures, factorial designs, competing risks, statistical modelling and randomised Phase II designs.

OUTCOME MEASURES

There are various choices for outcome measures in clinical trials involving AML patients. The primary ones are:

- Response rate – The proportion of patients who achieve an initial clinical response to the induction therapy is referred to as the response rate. In older AML patients, as in all leukaemia patients, the critical category is the complete response (CR) rate, although ‘partial’ responses are sometimes included in Phase II trials where attention is focused on identifying the antileukaemic effect of an agent, no matter how small. Achievement of a complete response is a *sine qua non* for long-term control of disease. However, the CR rate is a very imperfect surrogate for the more meaningful clinical outcome measures described below, has been defined differently by different leukaemia treatment groups, and should never be used as a substitute for them, especially in Phase III clinical trials. The primary role for the CR rate is as a measure of clinical activity in Phase II trials.
- Event-free survival (EFS) – This is the time from the start of study until a failure to respond, relapse (for those achieving a CR), or death, whichever occurs first. EFS is a good measure of the overall control of disease from the start of therapy and combines the effects of induction and post-remission therapies as well as deaths from toxicity of treatment. In a Phase III trial, all randomised patients contribute to the analysis of EFS under the usual ‘intent-to-treat’ approach. Standard techniques for time-to-event data (‘survival’ methods) are used in design and analysis of EFS.
- Disease-free (or relapse-free) survival (DFS) – This is a standard outcome measure in trials of adjuvant therapy for solid tumours, but in AML trials, DFS refers to the survival time spent free of disease. Thus, DFS is applicable only to patients who achieve a CR. It is defined as the time from achieving a CR to relapse or death, whichever occurs first. Since patients who fail to achieve a CR are excluded, this measure is unsuitable as an overall assessment of therapy. However, it is useful for comparing two or more post-remission therapies as long as it is recognised that the distribution of

DFS is not representative of the result to be expected for all patients.

- Length of remission (LR) – The length of remission is ordinarily defined as the time from achieving CR to the time of relapse, with deaths in remission counted as censored observations. This measure suffers from the same problems as DFS and, in addition, the usual Kaplan–Meier estimation is no longer valid (see discussion below on competing risks).
- Overall survival (OS) – The time from the start of study to death is an obviously critical outcome measure for any generally fatal disease like AML in older adults. It has the virtue of being unambiguously defined and captures a result of obvious significance. However, there are often difficulties in interpretation, particularly if multiple therapies are given, or if patients 'cross over' to the alternative therapy after relapse. Nevertheless, the importance of overall survival is so fundamental that it should always be reported, even if it is not used as the primary outcome measure.
- Other outcome measures – There are some other measures occasionally used in AML studies, particularly measures of quality of life (QOL). Some of these attempt to measure survival or related time-to-event measures adjusted for quality of life. For example, the QTWIST method discounts survival time spent with an unacceptable level of adverse symptoms due to treatment.[16–18] Such methods attempt to quantify the generally accepted notion that simply prolonging survival is not a sufficient objective. Improved quality of life is equally important. Physicians and patients virtually always incorporate qualitative assessments of this type in the decision about whether to pursue treatment in individual patients because of the intensity of standard treatment for AML and the poorer tolerance of such treatment in older patients.

FACTORIAL DESIGNS

The treatment phases for AML are conventionally divided into an initial phase of 'induction' therapy and, for those achieving a complete remission, a post-remission 'consolidation' or 'maintenance' therapy phase. If an agent can be safely added to the usual dose of conventional therapy, it might be most efficient to utilise the new therapy in both induction and consolidation, thereby perhaps maximising the chance to detect antileukaemic activity. Alternatively, one may elect to evaluate a different new agent post-remission, depending on the unique features of the agents being tested. Studies comparing therapies in each of these two phases utilise so-called factorial designs, in which patients are randomly assigned to one of two or more induction therapies (the first 'factor') and then to one of two or more maintenance therapies (the second factor). With two possible treatment assignments in each phase, this is a 2 × 2 factorial design, a common and well-known statistical approach. Much has been written about this design applied in the clinical trials setting. Because the second randomisation is applicable only for patients who respond to the induction therapy, only about 50% of older AML patients entered on study would be medically suitable and eligible for the second randomisation.

It is typical to separate the objectives of such studies into a comparison of induction regimens with respect to response rates, and, separately, a comparison of maintenance regimens with respect to the LR, DFS or OS. For example, CALGB 8923 was a randomised clinical trial of this type involving AML patients at least 60 years old.[3,10] The induction phase involved a randomisation between GM-CSF, a haematopoietic growth factor, and placebo following initial chemotherapy. The hypothesis was that the GM-CSF would reduce infectious complications and perhaps increase the response rate. Responding patients were to be randomised to receive one of two post-remission regimens, cytarabine alone or a combination of cytarabine and mitoxantrone. Overall, 388 participants were randomised to the induction therapies, 205 (53%) achieved a CR, but only 169 (44%) were randomised in the post-remission phase.

One of the problems with the usual approach to these designs is that there is no direct estimation or testing of the four possible treatment policies implied in the design, although one fairly recent paper deals directly with this issue, making efficient use of data from all patients.[19] The policies are defined by selecting one of two induction therapies followed by one of two post-remission therapies, if a response is obtained and the patient consents to continue. There are also issues related to when the randomisation to the post-remission therapy should be done. For example, if both randomisations are done at the time of study entry with a planned 'intent to treat' analysis, then the inevitable (and anticipated) large patient drop-out can substantially complicate evaluation of the second therapeutic manoeuvre. Further, in the design of such studies, it is usually assumed that there is little or no statistical interaction between the two factors. Even a small negative interaction can greatly complicate the analysis and influence initial calculations of sample size.[20]

The time course of study 8923 is also instructive when considering the difficulties of rapidly testing the effectiveness of new treatments in this disease. The study was conceived and written in 1988, initiated in 1989, and accrued patients fairly rapidly over a three-year period of time. The results of the initial induction phase were then published in 1995 with the final results of the post-remission treatments published in 2001. Although some endpoints such as initial response rates were known earlier, the need for more prolonged follow-up for assessment of OS and DFS can delay planning and implementation of subsequent approaches.

COMPETING RISKS

For some trials of therapy for older AML patients, it is informative to use the techniques of 'competing risks' analysis.[21–23] That is, rather than using a composite measure such as EFS, one can break this measure into its constituent parts by considering the time to each outcome separately. The term 'competing risks' refers to the various risks of failure, each competing with the others. This terminology originally arose in the context of analysing various causes of death, but it is applicable more generally. In contrast to most cancers, both the clinical manifestations of the AML itself and the antileukemic treatment can result in death, sometimes before response to induction therapy can be assessed, as well as in patients in CR. This is particularly true in older patients. The fundamental problem from a statistical perspective is that the risks cannot be assumed to be operating independently from each other. Thus, methods which assume such independence, such as the Kaplan–Meier life table analysis, which treats other risks as independent censoring mechanisms, are inherently flawed. One way to properly account for the dependence is through the use of the 'cumulative incidence' curve, a topic that has been extensively explored in recent years. In AML studies, cumulative incidence of relapse curves have been commonly used to assess the effect of different treatments, such as transplantation, in biologically distinct subgroups of patients.

STATISTICAL MODELS

Statistical models are frequently used in AML trials. The usual time-to-event measures (EFS, DFS, OS) are often handled non-parametrically in the primary analysis (e.g. Kaplan–Meier estimates, log rank tests, etc.), but the semiparametric proportional hazards regression model is most commonly used to adjust for covariates in the analysis. In addition, in younger patients with AML, there has been increasing use of so-called 'cure' models, in which it is hypothesised that an unknown fraction (c) of patients are 'cured' (or at least will have long-term control of disease), and the rest ($1 - c$) are not.[24–26] Interest then focuses on estimating c, in comparing the cure rates of various treatments, on identifying factors predictive of c, and on identifying prognostic factors for the time to failure in the patients not cured. In older patients with AML, this model has not been used as much due to the obviously low value of c, but it has been important in other leukaemias.

RANDOMISED PHASE II TRIALS

While most new therapies, unless their effects are dramatic, will eventually have to be tested in randomised trials, more rapid approaches are needed to screen amongst multiple new compounds so as to identify the treatment which most deserves the resources and costs inherent in an evaluation in a Phase III trial. Traditionally, these steps include sequential Phase I and Phase II trials with eventual selection of a dose and schedule of an agent to be evaluated further. It should be noted that doses and schedules of new agents are often derived from underpowered Phase I and II trials, usually done in patients with relapsed or refractory AML. These trials can be difficult to conduct and evaluate. Response rates are usually low and it can be problematic to accurately judge dose-limiting toxicity because many patients enter such trials with pre-existing morbidity due to the disease itself or residua from prior unsuccessful treatments. Thus, a selected dose or schedule may not be optimal, particularly when used in a patient population which may differ in age or stage of disease from the earlier trials.

Recent experience with putatively "targeted" drugs suggests that most will have to be combined with cytotoxic agents to produce maximal benefit. Let us assume that a number of new non-cytotoxic agents are ready for Phase II evaluation in combination with standard chemotherapy. Randomised Phase II trials are one approach that might be used in this setting.[27–29] Such trials have the administrative advantage that institutional review processes have to be conducted once, rather than on multiple occasions if the trials are done separately. It is also possible to add new study arms as new treatments come along. The randomisation allocates different risk factors amongst the treatment arms (with the caveat that with the smaller sample sizes, imbalances may still occur) and permit different doses and schedules to be studied simultaneously by the same group of treating physicians. The goal of such trials is not to substitute for a Phase III comparison with standard treatment but rather to pick a "best bet" to pursue further. There are obvious risks to this approach, including the identification of a "false positive" which is graduated to a Phase III trial and, perhaps equally seriously, the discarding of treatments which may, in fact, be active. The relatively small patient numbers also make it more difficult to discern benefit in subsets of patients, although measurement of surrogate *in vitro* endpoints which can be correlated with response may be productive when evaluating 'targeted' therapies. In a recent article, Estey and Thall discuss these issues in patients with AML in detail and propose an adaptive Bayesian method which allocates the patient randomisation to different treatment arms based on the outcome observed in patients treated earlier on the trial.[30]

OTHER CONSIDERATIONS

Practical problems abound, including the control of most new drugs by pharmaceutical companies with proprietary interests, as well as the absence of an administrative structure permitting rapid multi-institutional implementation of studies in patients with relatively uncommon diseases such as AML and other haematologic cancers. Evaluating combinations of investigational targeted drugs is particularly difficult because of the complexity of the negotiations if the drugs are controlled by different companies. Also, it seems inevitable that any speeding up of the process will raise the probability of making incorrect decisions. The judgement of whether such increased risks are acceptable will be difficult and will obviously depend on the condition under study. In any case, there is an imperative to evaluate new therapies efficiently and with appropriate scientific rigour and it is clear that the current approach to drug development can and needs to be improved.

As an example, possible randomised study designs for trials of new post-remission therapies are shown in Table 12.1 where "conventional" therapy might refer to a few courses of lower dose ara-c. Such therapy has historically yielded median remission durations of 6–10 months and less than 10% long-term DFS,[9] slightly better than observation without treatment which produces very few long-term disease-free survivors

Table 12.1. Possible study designs evaluating new agents as post-remission therapy in older patients with AML

Conventional ± new agent
Observation vs. new agent
Conventional or observation vs. new agent
Conventional followed by ± new agent
Phase II with new therapy alone; new agent in both induction and consolidation

and shorter CR durations. The choice amongst these approaches might vary according to the characteristics of the new agent being studied, although all the designs are subject to many of the problems enumerated above.

Indeed, given the very poor results observed with standard therapy, it could be argued that a Phase II trial in which the new agent is evaluated alone could have merit, although the usual problems with historical controls and patient selection would be issues. A number of anticancer agents have been approved by the FDA in recent years via the accelerated approval process, based solely on Phase II data. In these settings the agents showed benefit in patients with resistant disease and few, if any, other therapeutic options and were felt to be reasonably likely to produce eventual clinical benefit. It is recognised, however, that with the exception of unusually active agents, this is not the ideal means to prove efficacy and randomised confirmatory trials are generally required after initial approval.

SUMMARY

Acute myeloid leukaemia (AML) in the older patient is a common and important, disease which is relatively resistant to current therapies. Careful consideration of the distinctive characteristics of AML are required when designing clinical trials of innovative therapies. There will be a large number of interesting compounds available for evaluation in upcoming years and it is desirable that such studies be conducted using the most efficient and informative designs to rapidly identify therapies which lengthen survival and increase the fraction of patients who are cured.

REFERENCES

1. Cheson BD, Bennett JM, Kopecky KJ, Buchner T, Willman CL, Estey EH, Schiffer CA, Doehner H, Tallman MS, Lister TA, Lo-Coco F, Willemze R, Biondi A, Hiddemann W, Larson RA, Lowenberg B, Sanz MA, Head DR, Ohno R, Bloomfield CD. International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute Myeloid Leukemia. Revised recommendations of the International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute Myeloid Leukemia. *J Clin Oncology* (2003) **21**: 4642–9.
2. Mayer RJ, Davis RB, Schiffer CA *et al.* Intensive postremission chemotherapy in adults with acute myeloid leukemia. Cancer and Leukemia Group B. *New Engl J Med.* (1994) **331**: 896–903.
3. Stone RM, Berg DT, George SL *et al.* Granulocyte-macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myelogenous leukemia. Cancer and Leukemia Group B. *New Engl J Med* (1995) **332**: 1671–77.
4. Rowe JM, Andersen JW, Mazza JJ *et al.* A randomized placebo-controlled phase III study of granulocyte-macrophage colony-stimulating factor in adult patients (>55 to 70 years of age) with acute myelogenous leukemia: a study of the Eastern Cooperative Oncology Group (E1490). *Blood* (1995) **86**: 457–62.
5. Lowenberg B, Suciu S, Archimbaud E *et al.* Use of recombinant granulocyte-macrophage colony-stimulating factor during and after remission induction chemotherapy in patients aged 61 years and older with acute myeloid leukemia (AML): final report of AML-11, a phase III randomized study of the Leukemia Cooperative Group of European Organisation for the Research and Treatment of Cancer (EORTC-LCG) and the Dutch Belgian Hemato-Oncology Cooperative Group (HOVON). *Blood* (1997) **90**: 2952–61.
6. Godwin J, Kopecky K, Head D *et al.* A double-blind placebo-controlled trial of granulocyte colony-stimulating factor in elderly patients with previously untreated acute myeloid leukemia: a Southwest Oncology Group study (9031). *Blood* (1998) **91**: 3607–15.
7. Cassileth PA, Harrington DP, Appelbaum FR, Lazarus HM, Rowe JM, Paietta E, Willman C,

Hurd DD, Bennett JM, Blume KG, Head DR, Wiernik PH. Chemotherapy compared with autologous or allogeneic bone marrow transplantation in the management of acute myeloid leukemia in first remission. *New Engl J Med* (1998) **339**: 1649–56.

8. Baer MR, George SL, Dodge RK, O'Loughlin KL, Minderman H, Caliguiri MA, Powell BL, Kolitz JE, Schiffer CA, Bloomfield CD, Larson RL. Phase III study of the multidrug resistance modulator PSC-833 in previously untreated patients 60 years of age and older with acute myeloid leukemia: Cancer and Leukemia Group B study 9720. *Blood* (2002) **100**: 1224–32.
9. Schiffer CA, Dodge R, Larson RA. Long-term follow-up of Cancer And Leukemia Group B studies in acute myeloid leukemia. *Cancer* (1997) **80**(11, Suppl): 2210–14.
10. Stone RM, Berg DT, George SL *et al*. Postremission therapy in older patients with de novo acute myeloid leukemia: a randomized trial comparing mitoxantrone and intermediate-dose cytarabine with standard-dose cytarabine. *Blood* (2001) **98**: 548–53.
11. Byrd JC, Mrozek K, Dodge RK, Carroll AJ, Edwards CG, Arthur DC, Pettenati MJ, Patil SR, Rao KW, Watson MS, Koduru PR, Moore JO, Stone RM, Mayer RJ, Feldman EJ, Davey FR, Schiffer CA, Larson RA, Bloomfield CD. Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse and overall survival in adult patients with de novo acute myeloid leukemia: results from CALGB 8461. *Blood* (2002) **100**: 4325–36.
12. Tallman MS, Andersen J, Schiffer CA, Appelbaum FR, Feusner J, Ogden A, Shepherd L, Wilman C, Bloomfield CD, Rowe JM, Wiernik PH. All-trans retinoic acid in acute promyelocytic leukemia: long term outcome and prognostic factor analysis from the North American Intergroup Protocol. *Blood* (2002) **100**: 4298–302.
13. Fenaux P, Chastang C, Chevret S *et al*. A randomized comparison of all trans retinoic acid (ATRA) followed by chemotherapy and ATRA plus chemotherapy and role of maintenance therapy in newly diagnosed acute promyelocytic leukemia. *Blood* (1999) **94**: 1192.
14. Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, Capdeville R, Talpaz M. Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *New Engl J Med* (2001) **344**: 1038–4.
15. Bloomfield CD, Lawrence D, Byrd JC, Carroll A, Pettenati MJ, Tantravahi R, Patil SR, Davey FR, Berg DT, Schiffer CA, Mayer RJ. Frequency of prolonged remission duration after high dose cytarabine intensification in acute myeloid leukemia varies by cytogenetic subtype. *Cancer Res* (1998) **58**: 4173–9.
16. Gelber RD, Bonetti M, Cole BF, Gelber S, Goldhirsch A. Quality of life assessment in the adjuvant setting: is it relevant? International Breast Cancer Study Group. *Recent Results Cancer Res* (1998) **152**: 373–89.
17. Gelber RD, Gelber S. Quality-of-life assessment in clinical trials. *Cancer Treat Res* (1995) **75**: 225–46.
18. Cole BF, Gelber RD, Goldhirsch A. Cox regression models for quality adjusted survival analysis. *Stat Med* (1993) **12**: 975–88.
19. Lunceford JK, Davidian M, Tsiatis, AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics* (2002) **58**: 48–57.
20. Peterson B, George SL: Sample size requirements and length of study for testing interaction in a 2 × k factorial design when time-to-failure is the outcome. *Controlled Clin Trials* (1993) **14**: 511–22.
21. Bryant J, Dignam JJ: Semiparametric models for cumulative incidence functions. *Biometrics* (2004) **60**: 182–90.
22. Latouche A, Porcher R, Chevret S. Sample size formula for proportional hazards modelling of competing risks. *Stat Med* (2004) **23**: 3263–74.
23. Bollschweiler E. Benefits and limitations of Kaplan-Meier calculations of survival chance in cancer surgery. *Langenbecks Arch Surg* (2003) **388**: 239–44.
24. Yu B, Tiwari RC, Cronin KA, Feuer EJ. Cure fraction estimation from the mixture cure models for grouped survival data. *Stat Med* (2004) **23**: 1733–47.
25. Sposto R. Cure model analysis in cancer: an application to data from the Children's Cancer Group. *Stat Med* (2002) **21**: 293–312.
26. Sy JP, Taylor JM: Estimation in a Cox proportional hazards cure model. *Biometrics* (2000) **56**: 227–36.
27. Wieand HS. Randomized phase II trials: what does randomization gain? *J Clin Oncol* (2005) **23**: 1794–95.
28. Steinberg SM, Venzon DJ: Early selection in a randomized phase II clinical trial. *Stat Med* (2002) **21**: 1711–26.
29. Simon R, Wittes RE, Ellenberg SS: Randomized phase II clinical trials. *Cancer Treat Rep* (1985) **69**: 1375–81.
30. Estey EH, Thall PF. New designs for phase 2 trials. *Blood* (2003) **102**: 442–8.

CARDIOLOGY

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

13

Acute Stroke

PHILIP BATH
Division of Stroke Medicine, Institute of Neuroscience, University of Nottingham, Queen's Medical Centre,
Nottingham NG7 2UH, UK

INTRODUCTION

Stroke is defined as 'Rapidly developing signs of focal (or global) disturbance of cerebral function, with symptoms lasting 24 hours or longer, or leading to death, with no apparent cause other than vascular origin'.[1] As such, stroke is not a single diagnosis but rather a collection of clinical syndromes resulting from one of several pathological causes including cerebral ischaemia, primary intracerebral haemorrhage (PICH) or subarachnoid haemorrhage. Further, stroke is a diagnosis of exclusion so that mimics (such as abscess, epilepsy, hypoglycaemia or tumour) must be checked for and ruled out.

Stroke is common with a lifetime risk of one in six in the West and is the third commonest cause of death, after ischaemic heart disease and aggregated cancers. Similarly, stroke is devastating, being the commonest cause of long-term adult disability: approximately one-third of patients die by six months *post ictus*, one-third remain disabled and dependent on others (often requiring care in a nursing home), and one-third return to independence (although many of these still have residual impairments and disabilities). Although stroke can occur at any age from uterus to 100+ years, most occur in older people. With ageing Western populations, the incidence of stroke is rising. In parallel, improvements in acute care, rehabilitation and secondary prevention are resulting in an increase in prevalence as more victims survive.

This chapter will focus primarily on drug trials in acute stroke, although limited mention will be made of studies of devices and secondary prevention strategies. Trials assessing the primary prevention of stroke are discussed in the chapter on trials in cardiovascular disease.

REVIEW METHODS

This chapter is based on the author's personal experience as a clinical stroke trialist and his accrued references relating to individual trials and reviews in stroke. Articles and trials were identified following searches of The Cochrane

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

Library and PubMed, some of the author's own systematic reviews, a book listing major stroke trials[2] and reference lists from the resulting articles. A whole book could be written on performing trials in acute stroke. To keep this review to a reasonable length, it focuses on current issues rather than providing a recipe for running an acute stroke trial.

ACUTE ISCHAEMIC STROKE

Acute ischaemic stroke explains about 85% of strokes in the West and follows arterial occlusion, reduced blood flow and cellular ischaemia. In the core, severely ischaemic cells (blood flow < 12 ml/100 mg/min) die within minutes and, hence, are not amenable to salvage. Surrounding the core is a 'penumbral' area of less severe ischaemia (blood flow 12–25 ml/100 mg/min) where neuronal function ceases but cellular death does not immediately follow. A complex cascade of biochemical events follows comprising neuronal excitation, release of free radicals, lipid peroxidation, enzyme inhibition and DNA degradation; these events damage adjacent cells in the penumbra so that many die over the following hours or days if ischaemia persists. A recent review addresses the complex metabolic events that follow acute stroke.[3] Although neurones are the archetypal type of cell affected by ischaemia, other cells are similarly affected, including astrocytes, microglia and vascular cells.

Six overall strategies have been, or are being, tested, for treating stroke and enhancing recovery:

1. Restoring blood flow by lysing the occluding clot (thrombolytics) and preventing clot extension and migration (antiplatelets, anticoagulants).
2. Protecting penumbral neurones (and other cells) by modifying the adverse biochemical milieu and stabilising ischaemic cells (neuroprotection).
3. Regulating homeostatic disequilibrium relating to physiological processes such as blood pressure, glucose, temperature and intracranial pressure.
4. Promoting recovery through administering stems cells or enhancing the activity of endogenous stem cells.
5. Enhancing rehabilitation pharmacologically with drugs that accelerate and enhance recovery.
6. Managing the patient in a dedicated expert environment (acute stroke unit) focusing on rehabilitation (physiotherapy, occupational therapy, speech therapy and psychology) and the prevention of complications such as dehydration, hypoxia, urinary and respiratory infections, venous thromboembolism, pressure sores, contractures, and stroke recurrence.

It is self-evident that management should be started as soon as possible after stroke onset to maximise recovery, i.e. 'lost time is lost brain'.

Since stroke is a syndrome, it is critical to diagnose and distinguish cerebral ischaemia from PICH since their management differs fundamentally. Unfortunately, this is not possible clinically, even using scoring systems such as the Siriraj score,[4] and all patients should be imaged soon after admission to hospital using computerised tomography (or magnetic resonance imaging, MRI).[5]

In spite of numerous randomised controlled trials (189 non-confounded studies by the end of 2001),[6] advances in the management of acute stroke have been slow and limited. The value of acute stroke units, aspirin and alteplase has been recognised and the last is licensed in Europe and North America.[7–10] In contrast, trials of anticoagulants and neuroprotectants have not led to therapeutic change. Numerous studies and reviews have assessed why trials in acute stroke have failed and it is clear that no single explanation exists (Table 13.1).[11–15] Nevertheless, these trial failures provide valuable lessons for the design and practice of future studies. This chapter will not address basic trial methodology but rather will focus on issues specific to stroke, especially those which may have contributed to trial shortcomings or success.

Table 13.1. Possible explanations for the failure of trials in acute stroke. Adapted from Bath P. Failure of acute stroke trials. Presented to the British Association of Stroke Physicians, Stoke-on-Trent (2003)

Issue	Example(s)
Neuroprotection hypothesis	Does human penumbra exist? Is it modifiable?
Laboratory data	Animal data irrelevant to humans,[11–15] and do not fulfil STAIR I criteria (Table 13.2)[16]
Neuroprotectant drugs need to enter brain	Delayed or minimal drug entry into brain, e.g. gavestinel[17]
Timing of move from Phases I to II to III	Rapid move from Phase II to III prevents detailed analysis of Phase II data and adequate design of Phase III trials, e.g. MaxiPost[18]
Dose	Dose too low, e.g. tirilazad (RANNTAS, TESS)[19] Dose too high so adverse events exceeded any beneficial effects, e.g. streptokinase[20]
Sample size	Phase III trial too small, e.g. alteplase (ECASS II)[21,22]
Treatment window	Too long, e.g. magnesium given within 12 rather than 3 hours (IMAGES)[23]
Adverse events equal or exceed any benefit	Anticoagulation prevents thrombus extension and early recurrence but increases symptomatic intracranial haemorrhage by a comparable amount, e.g. heparin (IST)[8]
Primary outcome	Death inappropriate, e.g. lubeluzole (LUB-INT 5 & 9)[24,25]
Statistical analysis	Inappropriate dichotomisation (or 'cut') of modified Rankin Scale, e.g. alteplase (ECASS II)[21]
Subgroups	A positive subgroup in a neutral trial stimulates a second trial which is neutral, e.g. clomethiazole (CLASS, CLASS I)[26,27]
Protocol adherence	Protocol violation rate high, e.g. alteplase (ECASS)[28]
Baseline balance	Baseline characteristics unbalanced, e.g. piracetam (PASS)[29]
Treatment interaction	Negative interaction between investigational and other drugs, e.g. between streptokinase and aspirin, either intended (MAST-I)[30] or unintended (ASK)[31]
Publication	Trials not published, e.g. eliprodil, FISS-bis[32] Publications are of low quality[33]
Completion of development programme	Phase III never performed or completed in spite of favourable findings in Phase II, usually due to expiry of patent life, e.g. pentoxifylline[34]
Funding	Academic trials underfunded[35]

TRANSLATING PRECLINICAL DATA INTO CLINICAL FINDINGS

The classical paradigm for developing novel interventions for stroke (and most other conditions) requires preclinical assessment prior to testing in human volunteers and then patients. Such experimental studies should identify possible mechanisms of action, dosing issues and potential toxicity. Numerous studies of potential treatments for acute stroke have been assessed in experimental models of stroke and many, but not all, have been published. Unfortunately, clinical studies in the last century were often commenced with limited supporting preclinical data. Some drugs which are licensed for other medical conditions, e.g. heparins (as used in venous thromboembolism, ischaemic heart disease and peripheral vascular disease) have been assessed in stroke with minimal laboratory data[36] to support their clinical testing.

In other cases, the preclinical data postdated the clinical data; for example, nimodipine has been assessed in more than 25 clinical randomised controlled trials,[37] the first dating from 1983 and yet the first preclinical studies were only published in 1986 and further studies were reported up to 1997.[38] Unfortunately, systematic reviews of both preclinical and clinical studies of nimodipine were neutral.[37,38] Clearly, much effort, time and money were wasted testing a drug in patients which was ineffective in experimental stroke. Worse, patients were exposed to a potentially toxic intervention with no chance of benefit and were therefore deprived of entering trials of other, possibly effective, drugs.

In an attempt to improve the translation of interventions from the preclinical to clinical arena, an ad hoc group of academic and industry experts, the 'Stroke Trials Academic and Industry Roundtable I' (STAIR I), published guidelines in 1999 (Table 13.2).[16] These make recommendations on the types of stroke models, species of animals, timing and dose of intervention, study design and where research is undertaken. Nevertheless, whether all the recommendations detailed in STAIR I can be followed remains unclear. First, the programme of research which would deliver all these data will take a considerable time thereby threatening the financial viability of the drug due to limitations in the current patent protection period. Second, the complexity of experiments questions whether academic, as opposed to commercial, development is feasible. Last, anti-vivisection groups threaten increasingly the viability of preclinical research, especially that involving primates.

Table 13.2. Guidelines on preclinical studies, updated from the 'Stroke Trials Academic and Industry Roundtable I' (STAIR I)[16]

Factor	Criteria (minimum requirement)
Species	≥2; rodent(s), then primates (e.g. marmosets or squirrel monkeys, then macaques or baboons) if positive results in rodents
Gender	Female and male
Age	Old and young
Models	Permanent and transient focal ischaemia
Intervention, timing	Assess treatment window, ideally up to 4–6 hours after onset for neuroprotectants
Intervention, dose	Assess dose response
Intervention, route	Route depends on whether drug crosses blood–brain barrier
Study design	Randomisation of animals to intervention and placebo
Outcomes, type	Infarct volume (histology, magnetic resonance imaging), motor function (e.g. Rota rod, foot faults), cognitive function (e.g. Morris water maze)
Outcomes, assessment	Blinded to treatment assignment
Outcomes, timing	At least one month after infarction
Lesion	Efficacy in cortex and subcortex
Toxicology	≥2; normal and stroke animals
Laboratories	≥2, including one outside the organisation developing the drug
Meta-analysis	Research synthesis of all studies

Source: Reproduced with permission from Lippincott, Williams and Wilkins

It is also evident that the available preclinical data, whether published or not, should be reviewed *in toto* and systematically before a recommendation is made about moving the drug into humans.[39,40] Few such research syntheses (systematic reviews) have been published (Table 13.3).[38,41–46] These suggest that study quality is poor when judged using multidimensional scoring systems (although this might, in some cases, reflect suboptimal reporting, as has been found for clinical trials[33,47]). Common problems with studies include not performing a prior sample size calculation, not randomising animals to treatment groups, and not performing blinded assessment of outcomes.

Additionally, substantial publication bias may be present with negative or neutral studies being withheld from publication (publication bias); ironically, positive studies may also remain unpublished to protect intellectual property. The use of research synthesis techniques in assessing preclinical data is in its infancy and the appropriate methods for identifying and analysing such studies remain undefined at present.

CLINICAL DEVELOPMENT

Historically, the clinical development of a new intervention has been very variable depending on the nature of the investigational treatment and the type, size and experience of the organisation leading the research. Regulatory considerations

Table 13.3. Examples of published research syntheses of preclinical stroke studies

Intervention	Studies/ animals	Efficacy (95% confidence intervals)	Factors associated with efficacy
Nimodipine[38]	20/450	Infarct size: SMD −1.2 (−1.7, −0.7) Oedema: SMD −0.6 (−1.2, −0.1)	Timing: ≥1 hour? Study quality: ?
FK506[44]	29/1759	Composite:† WMD +0.313 (+0.272, +0.354); heterogeneity present	Species: monkey > rodent Co-morbidity: none > hypoglycaemic, hypertensive animals Anaesthesia: ketamine > others Model: transient > permanent Dose: 0.03–3 mg/kg (inverted 'U' response) Outcome: mortality > lesion size (histology) > lesion size (MRI) Publication: abstract > paper
Melatonin[43]	13/432	Composite:† WMD 0.428 (0.393 – 0.463); no heterogeneity	Gender: female > male Anaesthesia: ketamine > others Dose: multiple > single
Nicotinamide[42]	14/901	Composite:† WMD +0.287 (+0.227, +0.347); heterogeneity present	Species: rats > mice Co-morbidity: none > diabetic, hypertensive animals Anaesthesia: ketamine > others Model: transient > permanent Timing: 90–360 min Administration: iv > ip Dose: 100–750 mg/kg Publication: paper > abstract
Nitric oxide donors[45]	25/1374	Infarct size: SMD −1.21 (−1.69, −0.73); heterogeneity and publication bias present CBF: SMD +0.66 (−0.02, +1.35)	Model: permanent > transient Timing: 0–60 min
Nitric oxide synthase inhibitors[46]	73/2321	Infarct size MCAOp: SMD −0.56 (−0.86, −0.26) Infarct size MCAOt: SMD −0.99 (−1.25, −0.72) CBF MCAOt: SMD −0.57 (−1.26, 0.11)	Species: agyrencephalic > gyrencephalic Model: transient > permanent Timing: no effect Drug: iNOSi = nNOSi > non-selective NOSi

CBF: cerebral blood flow; ip: intraperitoneal; iv: intravenous; MCAOp: middle cerebral artery occlusion permanent; MCAOt: middle cerebral artery occlusion transient; MRI: magnetic resonance imaging; SMD: standardised mean difference; WMD: weight mean difference.
†Across all outcomes, e.g. infarct size, neurological score.

are now forcing consistency and the paradigm of performing studies in normal volunteers (Phase I) and then in patients with the target condition (Phase II – dose, administration, safety, feasibility, tolerability; Phase III – safety and efficacy) is now standard. Indeed, this approach not only applies to commercial drug studies but is equally applicable to academic trials and non-drug interventions such as rehabilitation techniques. Strangely, the regulatory environment for devices is not nearly so rigorous as for drugs, a situation that needs to be addressed to ensure a level playing field.

PHASE I STUDIES

These are not discussed here in detail since they involve the testing of interventions in normal volunteers, usually to determine pharmacokinetic (effect of body on drug) and pharmacodynamic (effect of drug on body) parameters and assess early safety. Studies typically test single and then multiple doses. The results of these studies inform the design of subsequent Phase II studies in the target population, in this case patients with acute stroke, and highlight potential doses, routes of administration and adverse events.

PHASE II STUDIES

Phase II studies assess safety, dose, feasibility, tolerability and mechanisms of action. In the classical model of drug development, the first study (Phase IIa) consists of a dose-escalation trial whereby sequential groups or 'blocks' of patients (often numbering 6–20 patients per block) are randomised to increasing doses of the active intervention versus control (usually a matching placebo). Analysis of pharmacokinetic (e.g. plasma levels of drug) and pharmacodynamic (e.g. blood pressure, fibrinogen, platelet function) and safety (e.g. symptomatic bleeding, liver and renal function) measures after each group of patients allows a decision to be made on whether the next higher dose is tested. Randomisation is often asymmetrical so that more patients receive active drug than control; ratios for active to control of 2:1 or 3:1 are typical.[48,49] It is vital that a parallel control group is included to allow the results to be interpreted in the context of existing data; unfortunately, not all studies have had a control group.[50]

The next study (Phase IIb) will usually assess one or two doses chosen from Phase IIa at several centres with the aims of confirming the first trial's results and of piloting a protocol for use at Phase III.[51] Ultimately, many factors will mitigate in choosing an appropriate dose, but the highest tolerable dose, taking account of adverse effects (such as bruising, headache, nausea and vomiting), should be chosen.[52] A third multicentre Phase IIc trial is often run at multiple centres to further test the proposed Phase III protocol.

Some trials have been labelled as Phase II/III and had the aim of assessing dose, safety and early efficacy. The underlying issue is that conventional-sized Phase IIa/b studies (typically randomising 50–200 patients) are too small to allow sensible decisions to be made with regard to dose. By making such trials larger (1000+ patients), dose–response relationships may be easier to identify. An example includes a trial of MaxiPost (BMS-204352), a maxi-K potassium channel opener; unfortunately, the trial has yet to be published so further details are unavailable.[18] A novel approach in stroke, although previously used in other areas including cancer, is the testing of many doses using a Bayesian design for randomisation; the ASTIN trial of neutrophil inhibitory factor used this method to test 15 different doses.[53]

PHASE III STUDIES

Regulatory authorities expect two positive trials demonstrating safety and efficacy with an overall reduction in poor outcome, usually measured as combined death or disability/dependency (see the section on primary outcome below). Technically, a licence could be obtained with one positive Phase III trial assessing functional outcome and a supporting Phase IIb/c trial positive for a surrogate measure of outcome/efficacy (e.g. reduction in lesion volume on MRI); however, this approach has yet to be used in a successful development programme. Several case studies from stroke drug development emphasise important points for trial design.

Alteplase

Alteplase is a recombinant thrombolytic agent originally developed for treating patients with acute myocardial infarction. As a result, limited preclinical and Phase II work[54–56] was carried out prior to Phase III trials in stroke patients (Table 13.4). A US academic/government/commercial alliance designed, funded and ran the positive NINDS rt-PA trial ($n = 624$). This reported in 1995 that alteplase (0.9 mg/kg) reduced combined death or dependency by an absolute 12–15% in patients treated within 3 hours of ischaemic stroke;[7] regulatory consideration of this two-part study as two separate trials allowed alteplase to be licensed in the

United States and Canada. Running in parallel, a commercial European trial, ECASS-I ($n = 620$), found only a trend to a benefit although the study differed using a higher dose of alteplase (1.1 mg/kg, as used in myocardial infarction) and included patients out to 6 hours post stroke.[28] A second commercial European trial, ECASS-II ($n = 800$), followed the NINDS protocol for dose but recruited patients out to 6 hours post onset;[21] this trial also only found a non-significant trend to benefit. To add to the confusion, a second US trial, ATLANTIS, changed its protocol after the release of the NINDS results and was published piecemeal.[57,58] The study was neutral overall although patients treated after 5 hours tended to fare worse. In each of these trials, CT scanning was performed prior to recruitment so that patients with PICH were excluded.

Two research syntheses, one using summary data and the other individual patient data, reported that alteplase improved functional outcome but at the cost of causing a fivefold increase in symptomatic intracranial haemorrhage.[59,60] Considerable debate (much of it acrimonious) over the design, practice and results of the Phase III trials, and especially NINDS,[61] and the lack of a second positive trial have meant that the European regulators remain uncertain about the safety and efficacy of alteplase. Alteplase currently has a conditional licence in Europe subject to satisfactory results from a third commercial Phase III trial, ECASS-III ($n = 800$, recruiting patients between 3 and 4.5 hours post stroke onset) and a Phase IV post-marketing study, SITS-MOST. Separately, the large academic IST-3 trial ($n = 6000$, time window 0–6 hours) is underway. Overall, it remains unsatisfactory that a positive study of alteplase was published back in 1995[7] and yet its utility remains undefined a decade later.

Aspirin

As with alteplase, aspirin was assessed in acute ischaemic stroke following positive trials in the treatment of acute myocardial infarction. Three academic Phase III trials were performed; in the first, aspirin was tested with streptokinase in a factorial design (MAST-I, Tables 13.4, 13.5)[30] mirroring the design of the landmark ISIS-2 trial in myocardial infarction.[62] MAST-I was stopped prematurely[30] when the negative findings of the MAST-E and ASK streptokinase trials became apparent.[31,63] Two mega-trials of aspirin have also been completed, IST (a factorial trial of aspirin and unfractionated heparin, Tables 13.4, 13.6) and CAST.[8,9] Aspirin appears to have little effect on the index ischaemic event but instead prevents early recurrence; its overall effect on functional outcome is modest (absolute reduction 1.1%) but very worthwhile at the public health level in view of its wide availability, ease of administration and low cost.

Citicoline

Citicoline is an unusual putative neuroprotectant which modifies cell wall structure. Four commercial trials have been completed: a Phase IIa dose comparison study ($n = 259$), a Phase IIb trial ($n = 394$), a further Phase IIb trial assessing the effect of the drug on lesion volume using MRI ($n = 100$),[64] and a Phase III trial ($n = 899$). Although the last study was neutral, a research synthesis reported efficacy across all four trials (Table 13.4).[65] A further and larger Phase III trial is in the planning stage.

Ongoing Development Programmes

Several Phase II and III programmes are underway which involve potentially useful drugs (Table 13.4). The first agent, desmoteplase (another thrombolytic), could be a competitor to alteplase. The second drug, recombinant factor VIIa (initially developed for the treatment of bleeding episodes in von Willebrand’s disease), addresses a neglected area of acute stroke, namely the treatment of primary intracerebral haemorrhage. In both cases, a Phase II trial has been positive on a secondary outcome measuring function.[49,66,67] Phase III trials of each agent are now either underway or planned.

Table 13.4. Selected academic and commercial drug developments in acute stroke. Ongoing trials are shown by their acronym (see www.strokecenter.org for further information)

	Preclinical	Phase I	Phase II	Phase III	Phase IV	Research syntheses	Secondary/ subgroup analyses
Abciximab	Ref. 69	–	Ref. 49	AbESTT-II	–	–	–
Alteplase	Ref. 54	–	Refs 55, 56 DIFFUSE, EPITHET	Refs 7, 21, 28, 57, 58 ECASS-III, IST-3	Ref. 70 SITS-MOST	Ref. 59, 60	Refs 71–75
Aspirin	Ref. 76	–	–	Refs 8, 9, 30	NR	Refs 77, 78	Refs 79–81
Citicoline	Refs 82, 83	–	Refs 64, 84, 85	RICH[86]	–	Ref. 65	–
Desmoteplase	Ref. 87	–	Ref. 66	DIAS-2	–	–	–
Factor VIIa	–	–	Ref. 67	Planned	–	–	–
Gavestinel	Ref. 88	–	Refs 89, 90	Refs 17, 91	NR	Ref. 92	Refs 93–97
Glyceryl trinitrate, nitrates	Ref. 45	Refs 98, 99	Refs 100, 101	ENOS	–	Ref. 102	–
Lubeluzole	Refs 103, 104	Ref. 105	Ref. 51	Refs 24, 25, 106	NR	Ref. 92	Ref. 107
Magnesium	Ref. 108	–	Refs 92, 109, 110	FAST-MAG[23]	–	Ref. 92	–
NXY-059	Refs 111–117	Refs 118, 119	Refs 120, 121	SAINT II[68]	–	–	–
Tirilazad	Refs 122, 123	Refs 124–126	CHANT[48,127]	Refs 128, 129	NR	Ref. 19	Refs 130, 131

NR – not relevant.

Recently, the positive results from a commercial Phase III trial (SAINT I, $n = 1700$) of NXY-059 (a nitrone spin trap free radical scavenger),[68] have been presented. A second trial (SAINT II) is underway. The preclinical testing of NXY-059 addresses most of the STAIR I criteria (as summarised in Table 13.2).[16]

PHASE IV STUDIES

Although not in the remit of this review, it is important to note that the safety (and efficacy) of a new intervention should be monitored following licensing during Phase IV studies (or post-marketing surveillance studies). By example, numerous academic Phase IV studies of alteplase have been performed, as summarised by a published research synthesis.[70] Although the rates of symptomatic intracerebral haemorrhage (principal measure of safety) and 'favourable' outcome (efficacy) were comparable with the NINDS alteplase trial,[7] the mortality rate was associated statistically with the percentage of protocol violations,[70] highlighting the need for clinical usage to mirror trial protocols.

TRIAL DESIGN

There are many differing designs for randomised controlled trials but only a few are relevant to acute stroke.

PARALLEL GROUP TRIALS

The most common trial design has parallel groups whereby patients are assigned to one of two or more groups, usually comprising one or more active groups and a control group. Multiple active groups allow different doses or drugs to be compared with control. Numerous examples of two- group, and some three- group, Phase III trials exist for acute stroke.

FACTORIAL TRIALS

Factorial trials allow two or more interventions to be assessed, each against their own control, within the same trial. In the simplest form, the 2×2 design, patients either receive both interventions A and B, A but not B, B but not A, or neither. Theoretically, the sample size is similar to a parallel group trial studying just one intervention. Several factorial trials have been completed in acute stroke, e.g. the 2×2 factorial trial of streptokinase and aspirin in MAST-I (Table 13.5).[30] A second trial, IST, used a variation of the 2×2 factorial design when assessing unfractionated heparin and aspirin; here, patients randomised to heparin were then randomised to low or medium dose (Table 13.6).[8] The factorial design can be extended to include more than two interventions or dose levels, i.e. amounting to a $N \times N \times N \times \ldots$ design as used in the six-factor trial of interventions for preventing postoperative nausea;[132] no such studies have been completed in stroke to date. Additionally, the factorial design may be nested within a larger parallel group trial (amounting to a partial factorial study), as was done in the ASCOT vascular prevention trial[133] and is being done in the ongoing ENOS trial (Table 13.7).[134]

Table 13.5. Completed 'Multicentre Acute Stroke Trial-Italy' (MAST-I) factorial trial of streptokinase, aspirin, both or neither (total 622 patients)[30]

	Streptokinase	No streptokinase
Aspirin	$N = 156$	$N = 153$
No aspirin	$N = 157$	$N = 156$

Table 13.6. Completed 'International Stroke Trial' (IST) factorial trial of aspirin, heparin, both or neither (total 19 435 patients)[8]

	No heparin	Heparin low dose	Heparin medium dose
No aspirin	$N = 4860$	$N = 2429$	$N = 2426$
Aspirin	$N = 4858$	$N = 2432$	$N = 2430$

Table 13.7. Ongoing 'Efficacy of Nitric Oxide in Stroke' (ENOS) partial factorial trial of glyceryl trinitrate and temporarily stopping or continuing prior antihypertensive therapy (intended sample size 5000 patients)[134]

	No prior anti-hypertensive therapy	Prior antihypertensive therapy	
		Stop	Continue
Glyceryl trinitrate	$N = 2500$	$N = 1250$	$N = 1250$
No glyceryl trinitrate	$N = 2500$	$N = 1250$	$N = 1250$

CROSSOVER TRIALS

Although statistically very efficient, crossover trials, where patients receive each intervention in random order, are not usually relevant in acute stroke since they are based on the premise that there is no underlying change in a subject's status other than that due to the intervention under test; clearly, the natural history for patients to improve with time after a stroke contravenes this assumption. Crossover trials are also not relevant to the assessment of surgical interventions.

FAMILIES OF TRIALS

A fairly recent innovation is the concept of 'families of trials'.[135] A 'family' comprises a series of randomised trials into which patients can be enrolled, either simultaneously or sequentially, with the individual trials sharing common systems for randomisation, data collection and follow-up. Such systems facilitate running parallel large simple trials, reduce research costs, and address concerns regarding co-enrollment.[135] In the ‘FOOD’ family of trials, patients were entered into one, two or three parallel trials. These investigated feed supplements (‘sip-feeds’), the timing of feeding after stroke, and whether enteral feed should be administered via a percutaneous endoscopic gastrostomy or nasogastric tube.[136,137]

CONTROL GROUP

Trials which will be credible to patients, health care professionals, policy makers and licensing agencies must have a control group (although regulatory advice from the European Agency for the Evaluation of Medicinal Products appears relaxed on this issue).[138] For most interventions, the control group consists of patients randomised to receive placebo or no treatment since few effective treatments have been identified to date. In this scenario, the new intervention is compared with no active treatment with all participants receiving standard ‘best care’, including aspirin. Many examples of trials using this approach exist; the following list gives examples of those which changed management, involved novel approaches to trials practice, and/or are examples of high-quality trials practice: ASTIN (neutrophil inhibitory factor, neuroprotectant),[53] CAST (aspirin),[9] ECASS I and II (alteplase),[21,28] GAIN I and A (gavestinel),[17,91] NINDS (alteplase),[7] PROACT II (pro-urokinase)[139] and STAT (ancrod).[140] In each case, active treatment was compared with placebo.

Smaller Phase II trials should also adhere to the principle of having a control group, including those assessing dose (usually defined as Phase IIa studies) such as STIPAS (tirilazad)[48] and AbESTT (abciximab).[49] Unfortunately, not all trials have followed this approach, e.g. TOPAS,[50] which complicates their interpretation.

SUPERIORITY VERSUS EQUIVALENCE STUDIES

Trials in acute stroke are at a relatively early stage of evolution and few proven treatments exist. Hence, most studies aim to show that the new intervention is superior to control. However, if a new intervention could interact with standard treatment or would only be given in its place, then the control group must comprise the standard therapy so as to ensure that patients are not deprived of active treatment. There are a few examples, including trials comparing low-molecular-weight heparins with aspirin: HAEST (dalteparin)[141] and TAIST (tinzaparin)[142]. Such trials can be more complex than conventional parallel group studies in terms of providing placebos for each intervention: for example, in TAIST, patients received tinzaparin and aspirin placebo, or tinzaparin placebo and aspirin (so-called double dummy). Importantly, these trials aimed to show that anticoagulation was superior to aspirin. However, it may be more reasonable to show that a new therapy is only at least equivalent to the best current intervention if, on pharmacological grounds, there is no expectation that it will be superior. Although there are no major examples in acute stroke, it should be noted that the sample size for such studies will be larger than for a similar superiority trial.

BASELINE SEVERITY/IMPAIRMENT

Apart from age and stroke type (ischaemic versus PICH), the most powerful predictor of outcome is baseline stroke severity. Although this may be measured in a single dimension, e.g. level of consciousness as assessed using the Glasgow Coma Scale (Table 13.8), multidimensional assessment using a tool which integrates several aspects of stroke impairment is preferable. A large number of scales have been developed, although most investigators now use the National Institutes of Health Stroke Scale or Scandinavian Neurological Stroke Scale (Table 13.8, 13.9).[143–145] These variably encompass assessments of consciousness, motor power, eye movements, orientation, dysarthria, dysphasia and gait (Table 13.9). An extension of this approach is to add social factors (age, living alone, independence pre-stroke) to a series of impairment measures.[146] In reality, these various approaches include the same types

Table 13.8. Assessment of severity

Scale	Score	Validity	Sensitivity	Problems	Other information
Glasgow Coma Scale[148]	Coma 3–8 Conscious 15	Face, predictive[149]		Developed for use in head injury	Simple to use
National Institutes on Health Stroke Scale (NIHSS)[143]	Severe 42 Normal 0	Inter-rater reliability[150]	Correlates with CT-measured lesion size.[143] Sensitive to treatment effects when used as an impairment outcome measure.[7]	Direction of scale counterintuitive Long-winded Maximum score can vary depending on patient characteristics Relatively insensitive to posterior fossa stroke.	Reliability improved with formal training[147] NIHSS required for administration of alteplase in Europe
Scandinavian Neurological Stroke Scale (SNSS)[144,145]	Severe 0 Normal 58	Inter-rater and intra-rater reliability[151]		Insensitive to posterior fossa stroke	Includes assessment of gait

Table 13.9. Components of the National Institutes of Health Stroke Scale (NIHSS) and Scandinavian Neurological Stroke Scale (SNSS). NIHSS ranges between 0 ('normal') and 42 (although the maximum score varies if certain assessments cannot be made); individual scores are 'good' if low and 'bad' if high. SNSS ranges between 58 ('normal') and 0 ('very severe'); individual scores are 'good' if high and 'bad' if low

Impairment	NIHSS	SNSS
Conscious level	0, 1, 2, 3	6, 4, 2, 0
Orientation	0, 1, 2	6, 4, 2, 0
Follow command	0, 1, 2	
Gaze deviation	0, 1, 2	4, 2, 0
Visual fields	0, 1, 2, 3	–
Face motor	0, 1, 2, 3	2, 0
Arm motor	0, 1, 2, 3, 4, 9	6, 5, 4, 2, 0
Hand motor	–	6, 4, 2, 0
Leg motor	0, 1, 2, 3, 4, 9	6, 5, 4, 2, 0
Limb ataxia	0, 1, 2	–
Gait	–	12, 9, 6, 3, 0
Sensation	0, 1, 2	–
Speech/dysphasia	0, 1, 2, 3	10, 6, 3, 0
Dysarthria	0, 1, 2, 9	–
Extinction/inattention	0, 1, 2	–
Range	0–34 (42)	58–0

of clinical measures and are, therefore, highly correlated, so it probably matters little which scale is used. More importantly, it is vital that any scale should be validated and that investigators are suitably trained to use whichever tool is chosen to minimise inter-rater variance, as epitomised by the formal training and validation required for use of the NIHSS.[147]

OUTCOMES

The choice of outcomes in acute stroke trials, especially at Phase III, should be straightforward since suitable measures need to satisfy a number of criteria, and only a few such validated measures exist. Useful outcomes must be relevant to stroke patients, their carers and society; they must also be validated and sensitive to changes induced by disease and intervention. Many outcome measures have had validation claims made but few, if any, have been exhaustively tested in respect of the main dimensions for assessing a clinical assessment scheme: construct validity, criterion validity, content validity, face validity, ecological validity, intra-rater reliability, inter-rater reliability, test–retest reliability, change sensitivity and simplicity.[149] A detailed description of outcome measures relevant to stroke is given by Wade.[149]

PRIMARY OUTCOME

Trials in acute stroke have used a variety of primary outcomes with most based on ordinal measures of functional outcome, usually composites of 'death or disability' (Barthel index, BI)[152] or 'death or dependency' (modified Rankin Scale, mRS;[153] three questions, 3Q;[154] Glasgow Outcome Scale, GOS)[155] (Tables 13.10, 13.11).[156] Death must be included in the measure since some interventions may have differential effects on dependency and death. For example, intravenous thrombolysis may increase death whilst reducing dependency,[59] so that assessing dependency alone would overestimate overall efficacy and exclude a critical safety measure. Death is usually scored as equal to, or worse than, the lowest score achievable in life, e.g. the mRS scores death as 6, one worse than complete dependency (Table 13.11).

Table 13.11. Modified Rankin Scale, a validated measure of combined death or dependency

0	No symptoms at all
1	No significant disability, despite symptoms; able to carry out all usual duties and activities
2	Slight disability; unable to carry out all previous activities but able to look after own affairs without assistance
3	Moderate disability; requiring some help, but able to walk without assistance
4	Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance
5	Severe disability; bedridden, incontinent and requiring constant nursing care and attention
6	Dead

Historically, most acute stroke trials used the BI as the primary outcome. Nowadays, the mRS is preferred since it is more robust clinically, and powerful statistically (Table 13.12).[96,157] Nevertheless, regulatory authorities do not give guidance on the choice of outcome scale.[158] Importantly, BI, mRS and 3Q have each been validated and are sensitive to therapeutic change (Table 13.10).[7,9,140] Some studies have merged scales in an attempt to avoid their individual

Table 13.10. Information on usual outcome measures in acute stroke trials

Outcome	Assessment	Score	Validity	Sensitivity	Problems	Other information
Barthel index (BI)	Disability, activities of daily living	Dead – 5[†] Disability 0 No disability 100	Construct[170–172] Inter-rater[167]	Outcome[173] Treatment[7,140]	'Floor' (early) and 'ceiling' (late) effects	Can be used over telephone[167,168]
Glasgow Outcome Scale (GOS)[155] (modified)	Impairment/ disability	Dead 1 Full recovery 5 (or 8)	Ref. 155	Treatment[7]	Several versions. Insensitive[149]	Used in head injury research
Rankin Scale (mRS)[153] (Table 13.11)	Impairment/ disability (not handicap as originally intended)	Dead 6 Dependent 5 Independent 0	Inter-rater[174]	Treatment[7]	Multiple versions Limited number of levels Poor content validity[149]	Can be used over telephone
3 questions (3Q)[154]	Dependency	Dead 3 Dependent 2 Independent 1 Normal 0	Ref 154	Treatment[8,9,175]	Very limited number of levels	Essentially the mRS collapsed from 7 to 4 levels

[†] The 20-point BI is the 100 scale divided by 5, i.e. severe 0, normal 20.

Table 13.12. Comparison of measures of dependency (dichotomised modified Rankin Scale, mRS), disability (dichotomised Barthel index, BI), impairment (National Institutes of Health Stroke Scale, NIHSS, or Scandinavian Neurological Stroke Scale, SNSS) and case fatality at three months in positive trials. Significance (*p*) values are shown; the outcome with the smallest *p*-value (i.e. most significant) is highlighted in bold

	mRS	BI	NIHSS or SNSS	Death
Factor VIIa[67]	**0.004**	0.006	0.008	0.02
ECASS II[21]	**0.024**	0.15	0.10	0.82
NINDS rt-PA part 1[7]	<**0.001**	0.012	0.002	–
NINDS rt-PA part 2[7]	**0.019**	0.026	0.033	0.30
PROACT II[139]	**0.04**	0.24	0.30	0.80
SAINT I[68]	**0.038**	0.18	NP	NP
STAT[140]	NR	**0.04**	0.07	0.62

NP: not presented; NR: not recorded.

problems. For example, the ceiling effect commonly observed with the BI can be circumvented, in part, by using the mRS to separate patients who score maximally on both scales (BI = 100, mRS = 0) from those who show minor evidence of dependency on the mRS (BI = 100, mRS = 1).[159] Although this approach may appear appealing, it has not been validated adequately and there is no evidence that it improves sensitivity to treatment effects.

It is common practice, although of limited purpose, to dichotomise these ordinal scales into 'independent' versus 'dead or dependent' although different trials have used varying dichotomies. For example, the mRS has been cut at 0–1 (excellent outcome) versus 2–6,[7] 0–2 (good outcome) versus 3–6,[142] or 0–3 (reasonable outcome) versus 4–6.[137,160] The choice of cut has largely been arbitrary in the past but it is now becoming clear that it should be based on two related principles. First, the cut should be placed between mRS 'levels' with large numbers of subjects on either side since it is these patients who are most available to move under the influence of treatment, i.e. who can move between adjacent strata. Second, around 50% of the patients should lie on each side of the cut since this maximises the statistical power of any binary analysis. Hence, it is essential that the outcome profile of the likely patient population is understood at the trial design stage, from population studies, hospital registers or previous similar trials. The choice of cut is critical since it can influence whether a trial may be considered significant (positive or negative) or neutral. For example, MAST-I (streptokinase in absence of aspirin) and ECASS I were positive with a mRS cut at 1/2 but neutral if the dichotomy was placed at 2/3;[28,30] conversely, ECASS II and PROACT II were positive with a cut of mRS at 2/3 but neutral with the cut at 1/2.[21,139,161] These opposing findings suggest unstable data, most likely due to an inadequate sample size. Nevertheless, an equally important question is whether ordinal variables such as mRS should be dichotomised at all; this statistical issue is dealt with below in the section on analysis.

Death alone has been used as the primary outcome in some stroke trials[162,163] although this cannot be recommended. First, avoiding dependency is more important to stroke patients and the elderly than dying.[164] Second, death is inefficient statistically since it occurs in only a minority of patients in most trials; see Table 13.12 for examples showing that death is a poor outcome as compared with others on statistical grounds. Last, it is questionable whether any treatment has sufficient efficacy to reduce death, and if it does whether this will be more important than improving functional outcome.

The optimal timing of assessment of the primary outcome also remains unclear although most Phase III trials have used three months (as supported by regulatory advice)[158] whilst some used six months.[8,9,142,165] The balance is between measuring outcome too early when significant recovery is still occurring versus measuring it too late when the effect of the intervention becomes confounded or masked by new vascular and non-vascular events which the acute intervention cannot alter or prevent.

It is also important to consider how the primary outcome will be assessed, e.g. at a face-to-face meeting between patient and investigator, over the telephone, or via postal questionnaire.[166] Regulatory authorities favour the former, although blinded central telephone follow-up offers the least biased approach since the local assessor has no prior knowledge about the patient. If central telephone or postal follow-up is chosen, it is vital that the outcome has been validated for this approach, as have the BI and mRS.[167,168] In general, different assessment methods should not be mixed within a trial since telephone, postal and face-to-face follow-up will generate different scorings within the same patient. Whatever form of follow-up is undertaken, it is important to document whether the patient or carer answered the questions since their scorings will differ.

In the context of assessing outcome it is worth considering that many interventions can be unblinded to an extent by investigators due to secondary effects of the drug. For example, prostacyclin causes facial flushing,[169] tirilazad mesylate causes superficial thrombophlebitis at the site of intravenous infusion,[19] alteplase causes superficial bruising (e.g. under blood pressure cuffs), and agents which lower blood pressure may be identified from the patient's ward observation chart. In each case, investigators assessing outcomes may be influenced by their 'guesses' as to whether the intervention was active or placebo. Hence, it is useful to ask the patient or carer at the end of follow-up which treatment was given to assess whether treatment blinding was maintained.

Finally, it is key that assessors are trained in how to administer the outcome measure. Considerable variations exist between assessors in scoring patients from video recordings: for example, scores from the 7-point mRS scale (Table 13.11) often vary about the real estimate by one or more points. Within-observer, between-observer, between-specialty (e.g. nurse versus doctor) and between-country variations each exist. Indeed, some of the between-country differences in functional outcome seen in trials[79] may reflect variations in scoring rather than clinical practice. Assessors should be trained, graded and regraded after several months, using video recordings to minimise inter-rater variance.

Trials should, ideally, only have one primary outcome. This will reduce the potential for misinterpretation and excessive claims if the results vary between outcomes, e.g. one outcome is positive and another neutral. Furthermore, the probability of obtaining false-positive (or false-negative, as opposed to neutral) results increases with the number of primary outcomes. Nevertheless, solutions exist to cope with having several primary outcomes. First, statistical adjustment can be made using multiple comparison procedures, e.g. the Bonferroni approach, so 'smaller' p-values are required to reach significance, i.e. the presence of two primary outcomes would require significance to be set at 0.05/2, namely $p = 0.025$. Second, it is possible to integrate several related dichotomous measures into a single 'global outcome' (see the analysis section),[71,176] as was done in the NINDS alteplase and IMAGES magnesium trials.[7,23]

SECONDARY OUTCOMES

If the situation is relatively straightforward for primary outcomes, it is complex for secondary outcomes. These should assess the intervention across a number of 'dimensions' including safety, recurrent events, impairment, quality of life, cognition and mood.

Safety

All acute stroke trials should assess and report on safety as based on the following key outcomes: early mortality (measured at, or shortly after, the end of treatment, typically at 5–10 days); mortality at end of trial;[158] early deterioration (measured using an impairment/severity stroke scale, e.g. NIHSS or SNSS); and functional outcome at end of trial. Whilst death automatically assesses safety, functional outcome can be the most sensitive measures of hazard (e.g. a toxic drug will increase rather than reduce combined 'death or dependency'), as has been seen in several negative trials.[19,48,177–181] Indeed, these trials were neutral in respect of mortality. Hence, functional outcome should be considered a safety measure in its own right.

Other safety measures will depend on the type of intervention and its mechanism of action. For example, drugs which promote bleeding, such as antiplatelets, anticoagulants and thrombolytics, must assess intracranial haemorrhage,[158] based on both clinical deterioration and neuroimaging. Similarly, conscious level will need to be assessed with centrally depressing drugs (e.g. clomethiazole),[26] white cell count in those which mobilise leucocytes from the bone marrow or reticular endothelial system (e.g. trafermin),[159] and blood pressure in trials of vasoactive agents.

Additionally, a variety of safety events and measures should be recorded in most acute trials, including: brain oedema resulting in herniation/death, seizures, cardiac conduction disturbances, haemostatic effects, blood pressure, temperature, glucose, infection, venous thromboembolism, vomiting and mental disturbances (anxiety, confusion, hallucinations, agitation).[158]

Recurrence

Whilst many treatments are given with the premise of moderating the effects of the presenting stroke, some may alter stroke complications such as recurrence or cerebral oedema. For example, the two aspirin mega-trials (IST, CAST) found that it reduced early ischaemic recurrence.[8,9] Whilst definitions of deterioration, stroke extension and recurrence are weak and overlap, some effort should be made to detect one or more of these using an impairment/severity stroke scale.

Quality of life

Increasing attention is now being paid to measuring function at a 'higher' level than death/disability such as 'quality of life' (QoL). Numerous QoL scales exist[182] and it remains unclear whether disease-specific or generic measures should be used, and whether these are sensitive to therapeutic change. Indirect evidence for the latter comes from the field of hypertension where treatment with angiotensin-modifying drugs has been associated with improved QoL.[183,184] In general, Phase III stroke trials should include a measure of QoL. Some measures can be performed over the telephone or by post, e.g. EuroQOL.[185]

Cognitive Function

In parallel with QoL, many acute stroke trials are starting to include a measure of cognitive function, not least because up to a third of survivors develop cognitive decline making this a worthy therapeutic target. Again, indirect evidence from hypertension trials suggests that cognitive decline can be delayed or slowed.[186–188] Unfortunately, the ideal tool for assessing cognitive function has yet to be identified, but the Mini-Mental State Examination appears sensitive to change and can be measured either in the clinic or over the telephone.[149,189]

Mood

A further therapeutic target is mood since 20% of stroke survivors develop major depression.[190] Antidepressant drugs may have potential non-mood effects, e.g. by possibly promoting recovery.[191] Conversely, it is possible that agents which improve other measures of function will also improve mood or reduce depression. Again, assessment tools for mood remain suboptimal but several have been used in stroke trials. The Zung test can be performed both face to face and over the telephone.[149,166,192]

SURROGATE OUTCOMES

The move from Phase II (assessing safety, dose, feasibility, tolerability) to Phase III (studying safety and efficacy) is fraught because the former trials are usually too small to give meaningful indications about efficacy (unless, perhaps, sequential research syntheses are performed – see the section on analysis). As a result, trialists have searched for surrogate measures of efficacy which might be changed therapeutically and which are statistically associated with functional outcome. Drawing again on an analogy from the field of hypertension, the magnitude of blood pressure lowering is more important in reducing the risk of subsequent stroke than the type of antihypertensive agent.[193,194] Hence, any new agent which has significant antihypertensive effects is likely to reduce the subsequent risk of stroke.

Neuroimaging

MRI techniques have been used to measure stroke lesion size following treatment.[195] This approach attempts to replicate the findings of preclinical studies where putative treatments must be shown, amongst other measures, to reduce lesion size.[16] However, the precise MRI technique (diffusion or T2-weighted imaging) and its timing (5–10 days for diffusion, 5–90 days for T2) have yet to be defined. Several trials have found that interventions can alter both MR and functional outcome measures (Table 13.13) although none has yet resulted in a new licensed treatment. Interestingly, CT scanning appears less useful than MRI; whilst CT measured lesion volume correlated with functional outcome,[130] negative treatment-related effects associated with tirilazad mesylate[19] were not detectable in a subset of patients with CT-measured lesion size.[131] When considering licensing requests, regulatory authorities may be willing to accept surrogate outcome MRI data from a Phase II study when combined with functional outcome data from a Phase III trial (although no successful example exists yet for this strategy); nevertheless, MRI measures cannot replace clinical outcome criteria.[158]

Reperfusion/Recanalisation

Another approach utilising imaging has investigated the effect of thrombolysis on occluded intracerebral arteries, analogous to studies of thrombolysis in acute coronary syndromes.[196] Pro-urokinase, a thrombolytic agent, both improved angiographic recanalisation rates[197] and clinical outcome[139] in a pair of Phase II and III trials involving intra-arterial administration (Table 13.13). Although a further Phase III trial is required for drug registration, this may never happen since the complexity and expense of organising and running a relatively large trial (400+ patients) of intra-arterial therapy is immense. Improved recanalisation (assessed using magnetic

Table 13.13. Examples of the use of surrogate measures of efficacy in stroke trials

Intervention	Trial	Surrogate measure	Effect on surrogate, relative to control	Clinical outcome measure	Effect on clinical outcome, relative to control
Citicoline	Phase II/III	MRI lesion size	Reduced[64]	Function (mRS)	Improved[65]
Desmoteplase	Phase II	MRA recanalisation	Increase[66]	Function (mRS, BI)	Non-significant improvement[66]
Pro-urokinase	Phase II/III	Angiographic recanalisation	Increased[197]	Function (mRS)	Improved[139]
Erythropoietin	Phase II	MRI lesion size	Reduced[206]	Impairment (SNSS)	Improved[206]
		Biomarker S-100ß	Reduced[200,206]		
ONO-2506	Phase II	Biomarker S-100ß	Reduced[207]	Impairment (NIHSS)	Improved[207]
ZK-200775	Phase II	Biomarker S-100ß	Increased[208]	Impairment (NIHSS)	Worsened[208]

resonance angiography MRA), and a trend to improved functional outcome, was also seen in a Phase IIa trial of another thrombolytic, desmoteplase.[66] Guidelines for the further study of intra-arterial thrombolysis in acute ischaemic stroke have been published.[198]

Biomarkers

The third approach is based on estimation of blood biomarker levels. Biomarkers may be defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention".[199] Many proteins and other soluble factors are produced constitutively by the brain and ischaemia leads to either an increase (reflecting cellular activation) or reduction (secondary to cellular failure and death) in circulating levels. Numerous association studies have showed that changes in these factors occur in ischaemic stroke and primary intracerebral haemorrhage and that levels are associated with severity, lesion size and functional outcome.[200] Hence, these factors may be considered to be surrogate measures of outcome. Although most studies, to date, have assessed specific candidate markers such as S-100, neurone-specific enolase and von Willebrand factor (Table 13.14),[201,202] the ongoing proteomics revolution means that large numbers of new markers may be identified.[203] The assay of single biomarkers does not help in the management of individual patients but panels of markers[204,205] are now being developed which may make biomarker measurements more relevant to clinical decision making. Evidence is also accruing that biomarkers may be surrogate measures of efficacy (Table 13.13) although the successful development and clinical use of a new therapeutic intervention based, in part, on the use of biomarkers has yet to be reported.

Table 13.14. Potential blood biomarkers for use in stroke

Cell type	Biomarker
Neurone	Neurone specific enolase[200,202]
Glia	Glial fibrillary acidic protein[200] S-100ß[200,201,204,205]
Endothelial cells	Nitric oxide (NOx)[209,210] von Willebrand factor (vWf)[204,205,211]
Inflammation	Matrix metalloproteinase-9 (MMP-9)[204,205] Monocyte chemotactic protein-1 (MCP-1)[204] Vascular cell adhesion molecule (VCAM)[205]
Haemostasis	Fibrinogen[211] D-dimer[212] P-selectin[211]

SAMPLE SIZE CALCULATION

Most Phase III acute stroke trials have utilised dichotomous outcomes, usually based on a measure of disability (e.g. BI) or dependency (e.g. mRS). In each case, a 'poor' outcome is defined as the combination of 'death or disability' (BI < 60) or 'death or dependency' (mRS > 2). The standard formula for estimating sample size for dichotomous data can be used:

$$N = \frac{\rho_1(100-\rho_1)+\rho_2(100-\rho_2)}{(\rho_2-\rho_1)^2} f(\alpha, ß)$$

where:

ρ_1: probability of a poor outcome in the control group

ρ_2: probability of a poor outcome in the active group

$\rho_2 - \rho_1$: absolute treatment effect

α: significance

$1 - ß$: power

$f(\alpha, ß)$: cumulative distribution function of a standardised normal deviate.[213]

This formula can be adapted to include adjustments for small sample sizes (Yates's correction), multiple active groups, groups of different sizes and multiple comparisons; each of these adaptations result in an increase in sample size.

A variety of parameters have been used in trials in acute stroke.[214] The value used for significance has almost always been 5% ($\alpha = 0.05$) whilst the majority of completed trials used a power of 80% ($1 - ß = 0.80$). This relatively low power can be criticised[214] on the grounds that patients, trialists, funders (whether academic or commercial) and Research Ethics Committees are each unlikely to be willing to miss a positive result by chance one in five times. Increasingly, trials are using a higher power such as 90%.

Using the above definitions of a poor outcome for disability and dependency, the control rate of a poor outcome (ρ_1) is approximately 0.50 in clinical trials; the median value used in acute stroke trials was 0.60.[214] The actual figure occurring in a trial will depend on case mix, which is driven partly by the protocol's inclusion and exclusion criteria (see the next section). It is important that investigators actively manage the outcome rate in the control group during the trial so as to ensure that the assumptions underlying the sample size calculation are achieved as far as possible. Practically, this means monitoring the overall rate of poor outcome ($(\rho_2 - \rho_1)/2$) so as to avoid unblinding of treatment groups. For example, a protocol amendment was made during the 'Tinzaparin in Acute Ischaemic Stroke Trial' (TAIST) to exclude patients with milder stroke;[142] although the trial took longer to complete, the final event rate in the aspirin control group was very close (57.5%) to that assumed (60%) in the protocol's sample size calculation.[142] Sometimes protocol amendments will adversely affect sample size assumptions, as occurred in the 'Trial of ORG 10172 in Acute Stroke Trial' (TOAST).[215] In this study, concerns about bleeding rates led to exclusion of patients with more severe stroke such that the final rate of poor outcome in the control group was much lower than the planned value.[215] Such a deviation from the sample size calculation will significantly reduce the trial's power.

In spite of considerations for significance, power and control outcome rate, the most difficult decision in sample size design usually relates to anticipating a realistic absolute treatment effect, $\rho_2 - \rho_1$. Multiple factors conspire to act in opposite directions. On the one hand, novel patented agents may command a high price (~\$2000 per acute patient) and therefore will need to show significant clinical benefit with a low number-needed-to-treat (NNT), typically less than 25 (equating to an absolute risk reduction of >4%). Of drug interventions, only thrombolysis has achieved this hurdle with alteplase reducing combined 'death or dependency' by 11–13% (NNT 7–9, depending on which of four outcomes are used).[7] In contrast, aspirin reduced combined 'death or dependency' by 1.2% (NNT 83)[77] but is widely prescribed owing to its wide utility (ischaemic stroke within 48 hours of onset), low cost and ease of use (oral, rectal, intravenous preparations). A synthesis of previous stroke trials found the median anticipated absolute treatment effect for combined death or disability was 13%, although the achieved value was low at only 1.8%.[214]

The result of using a lower power of 80% and high absolute treatment effect of >10% is that most Phase III trials in acute stroke were too small to show any benefit. A low sample size in the ECASS II trial of alteplase[21] contributed to its unstable result, i.e. neutral with the primary outcome of mRS > 1 and positive with the more conventional outcome of mRS >2. Had this trial been two to three times in size (i.e. 2000 rather than 800 patients), it is very unlikely that it would have been neutral on its primary outcome.[22] Alteplase might then have received a full European licence in 1999 rather than the conditional licence that was given in 2002. Table 13.15 shows estimated sample sizes, assuming a fixed power of 90%, for many Phase III trials in acute stroke. A comparison of these figures with trial sizes in acute coronary syndromes, where thousands or tens of thousands of patients are recruited, emphasises the shortfall in most stroke trials.

The sample size formula described above is a minimum for designing trials but more sophisticated versions exist which take account

Table 13.15. Estimated sample size parameters for Phase III stroke trials assuming significance = 0.05, power = 0.90 and control event rate = 0.50. The achieved absolute risk reduction assumes a conventional dichotomisation of functional outcome into 'good' versus 'poor'. For further information on the ongoing trials (shown in brackets), see 'The Internet Stroke Center' (www.strokecenter.org/trials)

Absolute risk reduction (ARR %)	Number-needed-to-treat (NNT)	Size	Examples	Achieved ARR (%)
13	7	610	ATLANTIS, EAST, ECASS I, MAST-I, NINDS	Alteplase 12–15
12	8	720	ECASS II, EST, Lub-Int 5 & 9 (ECASS III)	
11	9	860	ANSG, NEST, PASS	
10	10	1100	CLASS & IHT, GAIN-A, IASS-H, TOAST, TRUST (FAST-MAG)	
9	12	1400	TAIST, SAINT I (CLOTS 1)	NXY-059 2.1
8	13	1700	GAIN-I, Lub-Int 13 (AbESTT II)	
7	14	2200	IMAGES (COSSACS, GIST)	
6	17	3000	(CLOTS 2)	
5	20	4200	(ENOS, IST-3)	
4	25	6600		
3	33	11 600	CAST, IST	Aspirin 1.1
2	50	26 000		

of additional parameters such as varying event rates in different subgroups of patients (bearing in mind the heterogeneity of stroke). Further, sample size can be estimated for functional outcomes analysed using ordinal rather than nominal statistical approaches.[216]

INCLUSION/EXCLUSION CRITERIA

A critical design feature in trials is choosing appropriate inclusion and exclusion criteria since these determine which patients enter the trial. Ideally, treatments should be relevant to as many patients as possible, whether from clinical, commercial or licensing perspectives.[158] However, patient heterogeneity may reduce the power of a trial since patients with extreme characteristics may be less sensitive to interventions. For example, patients with very severe or very mild stroke are less likely to respond to most treatments[158] and their inclusion will dilute the effect of the treatment across the whole trial. As a result, selecting patients with a moderate prognosis could reduce sample size by 30% or more.[217] Hence, many stroke trialists believe that limiting the types of patient that enter a trial, especially during Phase II, will enhance its chance of detecting efficacy.

Two sets of data suggest that restricting inclusion to certain subgroups of patients may not be helpful. First, the results of individual trials are discouraging. For example, biological plausibility suggested that patients with cortical rather than subcortical ischaemic stroke might respond to GABA-modifying drugs since subcortical structures contain few GABA receptors. However, the Phase III CLASS-I trial of clomethiazole in patients with large cortical strokes was neutral.[27] Similarly, patients with acute ischaemic stroke and atrial fibrillation, where anticoagulation might be expected to be effective (as it is in primary and secondary stroke prevention),[218] did not benefit from dalteparin, a low-molecular-weight heparin.[141] Second, a systematic review of acute stroke trials found that positive trials had a trend to having fewer rather than more inclusion–exclusion criteria.[219]

Nevertheless, examples exist where patient selection might be useful. First, thrombolysis is likely to be of most benefit to patients with an occluded artery, as identified angiographically in trials such as PROACT II (pro-urokinase) and DIAS (desmoteplase).[66,139] Hence, thrombolysis is being tested in patients with a stroke-treatment time of more than three hours using MRA. Second, putative neuroprotectants are being tested in patients with a perfusion–diffusion deficit on MR scanning, i.e. those with a larger volume of reduced perfusion than the volume of diffusion abnormality (the latter equating to infarcted brain). Patients with a significant mismatch are more likely to have growth of their lesion as their penumbra converts into core. The extreme paradigm is where thrombolysis is only tested in patients with 'an artery to open' (determined with MRA) and 'salvageable brain' (assessed as a MR perfusion–diffusion mismatch) in the same vascular territory.[195,196,220] Recent (DEDAS, DIAS)[66] and ongoing (e.g. DIAS II) trials have utilised these approaches.

Practically, all stroke trials do need to define the entry criteria for certain patient characteristics for pharmacokinetic, pharmacodynamic, legal or other reasons and example criteria are shown in Table 13.16. In contrast, certain exclusion criteria have been used regularly in stroke trials without any obvious reason. For example, many trials excluded older patients although the elderly may have the most to gain from treatment since their absolute risk of a poor outcome or recurrence is higher than that for younger subjects. A minimum age is usually included for legal reasons (Table 13.16).

Table 13.16. Selected inclusion and exclusion criteria and examples from recent or ongoing trials

Factor	Examples	Exclusion criteria
Mechanism of action	Thrombolytic factor for promoting fibrinolysis after ischaemic stroke	Primary or secondary haemorrhage[158]
	Prothrombotic factor to reduce recurrent bleeding/haematoma expansion after intracerebral haemorrhage	Ischaemic stroke[67]
	Neuroprotection	Ischaemic stroke > 6 hours[68]
	Lower an elevated BP	BP < 140 mmHg[134]
Target population	Thrombolysis for rescue of penumbra	No MR diffusion–perfusion mismatch, small perfusion lesion, no diffusion lesion[66]
Metabolism/ excretion	Renal excretion of drugs	Elevated creatinine
	Liver metabolism of drugs	Elevated liver enzymes
Adverse action	Thrombolysis promotes bleeding	Concurrent use of antithrombotics (heparin, aspirin)[7]
Legal	Age ≥ 18 years (legally responsible)	Age < 18 years

RANDOMISATION

All clinical trials should use 'true' randomisation to assign patients to their treatment group. Pseudo-randomisation, where patients are assigned to treatment group by non-random means (e.g. by whether the last digit of the date of admission, date of birth, or hospital registration number is odd or even), does not ensure concealment of allocation and is not acceptable nowadays. Whilst properly executed true randomisation ensures concealment of allocation, it does not guarantee matching of baseline prognostic factors. Hence, it is unsurprising that some prognostic factors may, by chance, be unmatched and influence the trial's result, even in large studies. For example, the PASS trial ($n = 927$) found a non-significant increase in mortality in patients randomised to piracetam, a nootropic agent.[29,221] Much of this increase could be explained by an imbalance in stroke severity at baseline such that adjustment using logistic regression removed any trend for an increase in mortality. Similarly, the IST trial ($n = 19\,435$) was technically neutral for aspirin ($2p = 0.07$) but positive ($2p = 0.03$)

after adjustment for baseline prognostic factors, even though none of these factors differed significantly between patients randomised to aspirin or control.[8]

Historically, 'simple' randomisation was performed by taking a blinded treatment box (or equivalent) in numbered order. Randomisation techniques such as stratification and minimisation help reduce the chance of imbalances in key baseline factors,[158] and improve moderately the statistical power of trials.[222] Modern techniques using telephone or internet randomisation services ensure concealment of allocation and can facilitate stratification and minimisation. Indeed, the internet can also be used for online data collection and verification, supporting adjudication (e.g. of serious adverse events and images), and uploading of data (e.g. images).[223]

CHOICE OF CENTRES

The choice of local investigators (now called principal investigators) and centres is critical to a trial's success. First, investigators should be experienced in taking part in academic and commercial stroke trials (Table 13.17). They should be able to demonstrate that they have suitable types and numbers of patients, and adequate access to investigations, especially neuroimaging and carotid ultrasound. Equally, they should have suitable staff to recruit, manage and follow-up the patients, and deal with the copious amounts of trial paperwork; few principal investigators have the time do this themselves so dedicated research staff are essential. A number of countries (e.g. Australia, Canada, UK) have, or are developing, clinical research networks which offer a portfolio of stroke-experienced centres.

Table 13.17. Key criteria for assessing potential centres

Criteria	Example minima
Experience in stroke trials	Previous involvement in academic and commercial trials (list trials and numbers recruited)
Access to research staff	Research medic, nurse and/or therapist dedicated to research
Access to patients	Ideally >300 per year. Consider annual numbers of specific types of patients relevant to the trial's inclusion criteria: ischaemia, haemorrhage, cardioembolism, cortical, lacunar, proportion presenting within six hours of onset
Access to, and timing of, investigations	Neuroimaging – CT (perfusion CT, xenon CT), MRI (angiography, perfusion, diffusion, spectroscopy); transcranial Doppler; carotid ultrasound; echocardiography

ANALYSIS

The analysis of clinical trial data is an art as well as science and many choices of statistical analysis are available. Crucially, the analysis plan should be chosen prior to the trial end so that 'data-driven analyses' are not performed with the best results being published. Ideally, an analysis plan should be published before or during the trial, and certainly before the database is 'locked'. Implicit in all the following is that the primary analysis is based on the principle of 'intention to treat' (ITT)[158] and not 'per protocol' (PP, or target population). The definition of ITT varies between academic and commercial investigators with the former including all randomised patients whilst the latter tend to include only patients who receive at least one active or control treatment. Results have varied depending on whether ITT or PP analyses are performed; for example, the ECASS I trial was positive by PP and neutral on ITT.[28]

BINARY STATISTICAL APPROACHES

Chi-Square Test/Odds Ratio

The most basic and commonly used approach to analysing acute stroke trial data uses the data in the form of a 2 × 2 table with analysis using a chi-square test or reporting an odds ratio (with confidence intervals). Ordinal functional outcome data are 'collapsed' into good and bad outcomes, e.g. the mRS is dichotomised at 2/3 (or BI at 55/60). This approach matches the basic sample size calculation (as above) and is readily understood by clinicians, e.g. alteplase reduces the proportion of patients who are dead or dependent (mRS) at three months by 13% from 74% to 61%.[7] Some trials have dichotomised at an 'excellent' outcome (mRS 1/2; BI 90/95).[7] Indeed, a retrospective analysis of the neutral ECASS I trial of alteplase revealed that the trial was positive using this approach.[224]

Global Outcome Test

One approach to improving the analysis of a dichotomous outcome is to select several parallel measures and then integrate these using a global statistic such as the Wald test.[71,176] The NINDS trial of alteplase was the first stroke study to use this approach and integrated four outcomes: NIHSS, BI, mRS and GOS.[7] The recent IMAGES trial of magnesium used the same approach for BI and mRS.[23] In a retrospective analysis, the neutral ECASS trial became 'positive' if assessed using the global approach.[224] Although the global test is now well accepted by investigators and regulatory authorities, the meaning of a dimensionless statistic to patients and clinicians[225] and its performance relative to other statistical approaches remain unclear.[226]

Prognosis-Adjusted Outcomes (Patient-Specific Outcomes)

Stroke patients are very heterogeneous in terms of prognostic factors such as age, stroke severity (e.g. NIHSS, SNSS), stroke aetiology (large-artery, small-vessel disease, lacunar),[227] stroke syndrome (using the Oxfordshire Community Stroke Project (OCSP) classification)[228] and comorbid factors (atrial fibrillation, diabetes, hypertension, ischaemic heart disease). As a result, outcomes will vary considerably; for example, the rate of combined death and dependency at one year is 95% in patients with a total anterior circulation syndrome (TACS) presentation but only 35% in those with a lacunar syndrome (LACS).[228] Since the use of a fixed dichotomous outcome, e.g. mRS 2/3, can only detect patients who cross this boundary (e.g. from 3 to 2), patients with a mild (who mostly achieve a mRS of 0 or 1) or severe (who generally obtain mRS 4–6) stroke will not contribute to the trial. Two potential solutions address this conundrum, either to exclude mild and severe patients (as discussed above) or to vary the dichotomi for different patients.

The latter approach is variably called 'prognosis-adjusted outcomes' or 'patient-specific outcomes'. Published mRS and BI dichotomy based on OCSP and NIHSS criteria[143,228] are presented in Table 13.18, although whether these are optimised remains to be established. This approach has been used in two trials. The AbESTT Phase II study of abciximab was neutral on its primary outcome with assessment of death or dependency using a standard mRS dichotomy of 2/3, but positive using the patient-specific outcome approach (based on baseline NIHSS) (Table 13.18).[49] In contrast, the STICH trial of surgery for patients with primary intracerebral haemorrhage was neutral however analysed.[229] Using individual patient data from the GAIN-I trial, a modelling study found that statistical power was increased using the prognosis-adjusted

Table 13.18. Good outcome, judged using the Barthel index or modified Rankin Scale, for patients with different clinical syndromes (Oxfordshire Community Stroke Project classification, OCSP)[228] or stroke severity (National Institutes of Health Stroke Scale, NIHSS).[143] Adapted from Refs 230,231

		OCSP	Category		NIHSS		
	LACI	POCI	PACI	TACS	4–7	8–14	15–22
BI	95 100	95 100	95 100	60–100	–	–	–
mRS	0,1	0,1	0–2	0–2	0	0,1	0–2

outcome approach, e.g. the power of analyses based on the BI increased from 60% to 88%, and that for mRS from 84% to 92%.[230]

Sequential Analysis

An underused approach in trials is the use of a sequential design. This consists of a series of interim analyses which allow the trial to stop when one of the treatment groups becomes superior to the other, or when it becomes futile to continue, i.e. it is unlikely that the treatment groups will ever differ. Since interim analyses reduce a study's power, the maximum possible sample size will be greater than that for a normal non-sequential design, However, the likely sample size is typically smaller, making sequential trials more efficient on average. Sequential designs were used in trials of isradipine[232] and eliprodil (never published) in acute stroke. A modelling exercise suggested that the neutral GAIN-I trial would have been smaller and finished earlier if a sequential design had been used.[94]

ORDINAL (ORDERED CATEGORICAL) APPROACHES

Since functional outcome data based on the mRS and BI are ordinal in nature, i.e. they are categories with an innate order, ordinal statistical approaches are likely to be more efficient statistically than those based on dichotomous tests. The archetypal ordinal analysis uses the Mann–Whitney U/Wilcoxon test,[233] equivalent to a *t* test based on ranked data, and several trials have used this in their primary analysis, e.g. EAST (enlimomab) and ECASS (alteplase).[179,234] Whilst EAST was negative using this approach,[179] ECASS I was neutral[234] but positive in *post hoc* analyses based on dichotomisation for an 'excellent' outcome, or with the global outcome statistic.[224] More recently, computer-intensive approaches such as bootstrapping have been advocated, as illustrated in the *post hoc* analysis of ECASS II.[235]

WHICH APPROACH?

It is not clear which of the above approaches, or indeed others which have not been described here,[236] are most efficient. Table 13.19 illustrates the variability in results seen when different statistical tests are used with summary outcome data from published positive or negative (but not neutral) trials. The ongoing 'Optimising the Analysis of Stroke Trials' (OAST) Project is assessing this issue further with the aim of identifying one or more 'best' approaches;

Table 13.19. Comparison of six statistical approaches for analysing published ordinal outcome data. The resulting significance (*p*) value is given. (The global outcome and patient-specific outcome approaches are not included since their analysis requires access to individual patient data)

	Outcome	Chi-square, 'excellent' outcome†	Chi-square, 'good' outcome†	Chi-square, death†	Mann–Whitney U test‡	Robust ranks test‡	Bootstrap$^{\diamond}$	'Best' analysis
AbESTT[49]	mRS	0.1072	0.3652	0.3328	0.0619	0.0611	0.0593	bootstrap
CAST[9]	3Q	0.0779	0.0865	0.0375	0.0297	0.0296	0.0273	bootstrap
EAST[179]	mRS	0.0953	0.0187	0.0729	0.0041	0.0038	0.0047	robust ranks test
ECASS II[21]	mRS	0.3065	0.0238	0.9877	0.0917	0.0915	0.0913	'Good'
Edaravone[238]	mRS	0.2305	0.3410	0.9762	0.0479	0.0462	0.0420	bootstrap
IST[8]	3Q	0.0778	0.0747	0.1079	0.0290	0.0290	0.0287	bootstrap
MAST-E[165]	mRS	0.5209	0.7072	0.1629	0.6462	0.6479	0.6680	Death
PROACT II[139]	mRS	0.2201	0.0692	0.8778	0.3314	0.3241	0.3200	'Good'

mRS: modified Rankin Scale; 3Q: three questions.

'Excellent' outcome: mRS < 2 or 3Q < 1; 'good' outcome: mRS < 3 or 3Q < 2; death: mRS = 6 or 3Q = 3.

†Corrected for continuity; ‡corrected for ties; $^{\diamond}$3000 cycles/3 iterations.

preliminary findings assessing both individual patient and group data from 20 trials confirm that ordinal approaches (e.g. robust ranks test, Mann–Whitney U test, bootstrap) are, in general, more efficient that binary tests (such as chi-square test).[237]

SUBGROUP ANALYSIS

There is a strong temptation to look for positive subgroups when a trial is neutral on its primary outcome since considerable time, expense and personal emotional energy have gone into completing the study. Occasionally, all prespecified subgroups are neutral[142] but it is usual to find one or more which is positive (or negative). Implicit in any analysis is that findings in subgroups should not be used to drive the overall interpretation of the trial. Similarly, a marketing authorisation will not result from a subgroup analysis where the primary analysis is neutral.[158]

However, the question arises as to whether a follow-on trial is warranted to assess the intervention in a positive subgroup. Several points argue against this approach. First, any analysis in subgroups of patients is likely to be very underpowered (unless the trial was very large and specifically powered to look at subgroups) so obtaining a true positive result is unlikely. This is compounded by the problem, highlighted above in the section on sample size, that most stroke trials have been relatively underpowered across the whole study. Second, the more subgroups that are investigated, the more likely that several will be statistically significant, so a choice of which to study further will be necessary. Last, there must be a negative subgroup (or at least a strong trend to a negative finding) for every positive subgroup (and vice versa) if the trial overall was neutral. By example, if patients with ischaemic cortical stroke benefited, then those with subcortical stroke must have been disadvantaged if the primary result was neutral. Then, it has to be questioned what biological justification could explain why the drug was effective in one subgroup (cortical stroke in this example) but hazardous in another (subcortical stroke). Several examples illustrate the likely futility in chasing subgroups (Table 13.20). In each case, the follow-up trial investigating a positive subgroup was neutral.

This is not to say that chasing subgroups is always doomed to failure. The problem is identifying when the findings are real rather than due to chance. Biological plausibility alone is insufficient to justify a further trial. A research synthesis, preferably based on individual patient data, which suggested efficacy in the subgroup across several studies would be a stronger indication for a further trial. A negative example is the TOAST trial where patients with presumed large-artery ischaemic stroke appeared to respond to intravenous anticoagulation (danaparoid) although the trial was neutral overall and had started out with the expectation that patients with presumed cardioembolic ischaemic stroke might respond

Table 13.20. Examples of positive subgroup analyses in neutral trials which led to further trials, each of which was neutral. A relative risk (RR) or odds ratio (OR) < 1 implies potential benefit, i.e. a reduction in combined 'death or disability/dependency'

Intervention	Index trial(s)	Positive subgroup	Finding in subgroup	Second trial	Finding
Clomethiazole	CLASS[26]	TACS syndrome[228]	RR 0.85	CLASS-I[27]	RR 1.07
Piracetam	PASS[29]	Treatment ≤ 7 hours	RR 0.72	PASS II	Unpublished, neutral
Nimodipine	Research synthesis[239†]	Treatment $\leq 6–12$ hours	OR 0.62	VENUS[160]	RR 1.2 (0.9, 1.6)

[†] A later research synthesis found that nimodipine had no benefit in acute ischaemic stroke, irrespective of timing of administration.[37]

most.[215] Individual trial findings,[141,142] and meta-analyses of them,[240,241] did not confirm efficacy with anticoagulation in either subgroup. A particular dilemma arises for trials with positive subgroups where no other large study data are available, as with the recent IMAGES trial of magnesium (possible efficacy in patients with non-cortical/lacunar syndromes and those with higher BP)[23] and STICH trial of surgery for PICH (possible efficacy in superficial haemorrhages).[229]

SEQUENTIAL RESEARCH SYNTHESES

A further technique for detecting potential efficacy (or hazard) during Phase II is the use of sequential research synthesis. Data are added after each trial for a number of outcomes including early death, death at end of follow-up, and function (usually combined 'death or dependency'). An example where this technique might have highlighted a potential hazard and terminated development and patient exposure earlier than happened relates to tirilazad mesylate. A worse outcome was evident in patients randomised to tirilazad mesylate after a Phase II and two Phase III trials had been analysed. Instead, each trial was interpreted as neutral and a further stage of testing involving three more trials was started investigating higher doses; development was terminated when hazard was identified in the fifth trial. A research synthesis should be commenced after the first Phase II trial and updated with information from each subsequent trial.

A final research synthesis after Phase III has been completed is useful to identify all the available evidence for the intervention and to define the overall efficacy and safety, and effects in major subgroups.

TRIAL COMMITTEES

TRIAL STEERING COMMITTEE

A key determinant in the design and quality of trials is the composition and experience of the Trial Steering Committee (TSC). The TSC should be independent of the sponsor, whether commercial or government/charity, to ensure the trial has relevance to patients, has an optimal design and execution maximising its chance of success and extrinsic validity. These aims differ from those of commercial sponsors who are more interested in patent protection and marketing advantage. Hence, the TSC needs to control publication policy and the availability of study data, and ensure the results, whether positive, neutral or negative, reach the public domain in a timely fashion.[242] In this respect, it is important that the TSC has executive responsibility, i.e. it is a steering committee, rather than having a purely advisory role. Sponsors should be represented on the TSC by members with non-voting status.

A growing tendency is for TSCs to delegate the writing of study reports to a writing committee[242] (which may be a subcommittee of the TSC). This should be resisted since the responsibility for the trial, including its reporting, resides with the entire TSC; nevertheless, it can be challenging to coordinate publications involving a reasonably large number of TSC individuals. Another trend which needs reversing is the removal of executive decisions from the remit of the TSC, e.g. by having a separate 'Trial Executive Committee' (TEC). Parallel TECs and TSCs prevent joined-up decision-making and mean no group can take overall responsibility for the trial.

The composition of the TSC is critical and yet in most stroke trials has largely consisted of stroke neurologists or physicians. Key members of the TSC should include: several stroke physicians (representing the various backgrounds of doctors that deliver stroke care around the world, typically neurology, geriatrics and general medicine) to ensure the trial is relevant to stroke patients, a specialist in the trial's domain (e.g. haematologist for an antithrombotic agent, lipidologist for cholesterol lowering, rehabilitationists if rehabilitation), a clinical pharmacologist for assimilating and interpreting pharmacodynamic and pharmacokinetic data, a methodologist for study design and a statistician for analysis. The practice of 'forcing in' or 'gifting' a representative from each participating country should be resisted; membership of TSCs should depend on expertise, not on geographical politics, and inclusion of numerous TSC members will make the operation of the committee unwieldy. As a rule, the TSC might include 5–15 members although the exact number will depend on the trial's size, complexity and geography. Ideally, the TSC should be chaired by an independent and experienced trialist rather than by the Chief Investigator since the latter may suffer conflicts of interest when difficult decisions have to be made about the trial's future if safety becomes an issue, or when interpreting the results. It is also important that Chief Investigators see their TSC colleagues as having an active role rather than purely being present for the purpose of 'rubber-stamping' decisions. Occasionally, interpretation of the results may vary within a TSC, as occurred in the 'Multicentre Acute Stroke Trial-Italy' (MAST-I),[30,243] and systems for dealing with this need to be addressed in advance.

TRIAL MANAGEMENT COMMITTEE

It is impractical for the TSC to manage the trial on a day-by-day basis and large stroke trials are typically run by a Trial Management Committee (TMC), usually comprising a subset of the TSC and including the Chief Investigator. The TMC must restrict itself to dealing with everyday trial problems rather than infringing on the executive and strategic role of the TSC.

DATA MONITORING COMMITTEE

It is essential that an independent and unblinded look is made at the main efficacy and safety data.[242] The TSC and TMC cannot do this in an unbiased way and, therefore, all multicentre Phase II and III trials should have a Data Monitoring Committee (DMC).[244] Unfortunately, many multicentre Phase II trials, whether academic or commercial, still do not have a DMC.

The DMC members must be experienced in this role although few, to date, have received formal training for this role; training courses for future DMC members are urgently needed. The precise role of the DMC will vary from trial to trial; for example, an additional role in Phase II trials to monitoring safety may be to alter the protocol in response to dose–response issues.[121] However, the principal responsibility of the DMC is to patients within the trial (and those who may be recruited later) by monitoring safety and efficacy. This can only be done through assessment of unblinded data since some safety issues only occur at low rates which remain invisible when assessed in a blinded manner. Further, functional outcome should be included in any safety analysis (see above) irrespective of whether efficacy is itself being assessed. The DMC should not be frightened of recommending that the design of a trial should be changed midway, e.g. by stopping or repeating a dose arm,[120] or that the trial should be stopped altogether if hazard is observed, as happened in AASI, ASK, ASSIST, INWEST, MAST-E, MAST-I and TESS II.[19,30,31,63,178,180,181]

Approvals

Trial bureaucracy has increased to the point of distraction over the last 10 years. Approvals are needed from numerous disparate organisations, including regulatory authorities (e.g. EMEA, FDA), national and local Research Ethics Committees, hospital Research & Development Departments and trial registration bodies. It is to be hoped that the process can be streamlined, as it once was.

SUMMARY

Numerous trials in acute stroke have been performed and yet few have led to advances in treatment. No single reason explains this failure. Figure 13.1 summarises the key components of developing a new intervention and major aspects which need to be considered in the design and execution of component trials. Following this scheme, incorporating STAIR recommendations[16,245,246] and monitoring ongoing advances in trial practice (e.g. design, randomisation, data collection, analysis, research

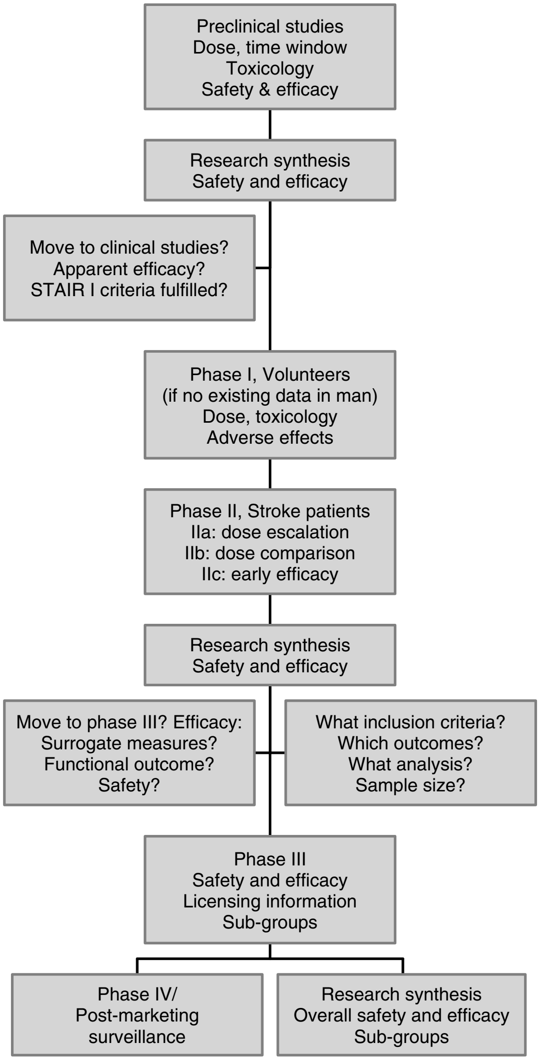


Figure 13.1. Development of a new intervention for acute stroke.

syntheses) will improve the likelihood of developing new and effective interventions for acute stroke.

CONFLICTS OF INTEREST

Apart from being an author of some papers referenced in this review, I took part in the STAIR III workshop,[246] have been the Chief Investigator of several trials and studies (ENOS, OAST, TAIST),[142] have been on the steering committee of several trials,[19,23] chaired the Data Monitoring Committee of two trials, and were or am a member of the Data Monitoring Committee of other trials. I am on the management board of the UK Stroke Research Network, and have consulted for pharmaceutical and diagnostic companies active in stroke.

REFERENCES

1. Hatona S. Experiences from a multicenter stroke register: a preliminary report. *Bull WHO* (1976) **54**: 541–53.
2. Bath P, Albers G. *Controlled Clinical Trials: Stroke*. 1st edn. London: Science Press (2001).
3. Lo EH, Dalkara T, Moskowitz MA. Mechanisms, challenges and opportunities in stroke. *Nature Rev Neurosc* (2003) **4**: 399–412.
4. Poungvarin N, Viriyavejakul A, Komontri C. Siriraj stroke score and validation study to distinguish supratentorial intracerebral haemorrhage from infarction. *Br Med J* (1991) **302**: 1565–7.
5. Wardlaw JM. Is routine computed tomography in strokes unnecessary? Computed tomography provides accurate diagnosis. *Br Med J* (1994) **309**: 1498–1499.
6. Weaver CS, Leonardi-Bee J, Bath-Hexall FJ, Bath PMW. Sample size calculations in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. *Stroke* (2004) **35**: 1216–24.
7. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute stroke. *New Engl J Med* (1995) **333**: 1581–7.
8. International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *Lancet* (1997) **349**: 1569–81.
9. CAST (Chinese Acute Stroke Trial) Collaborative Group. CAST: randomised placebo-controlled trial of early aspirin use in 20 000 patients with acute ischaemic stroke. *Lancet* (1997) **349**: 1641–49.
10. Bath PMW, Lees KR. ABC of arterial and venous disease – acute stroke. *Br Med J* (2000) **320**: 920–3.
11. Grotta J. The current status of neuronal protective therapy: why have all neuronal protective

drugs worked in animals but none so far in stroke patients? *Cerebrovas Dis* (1994) **4**: 115–20.

12. del Zoppo GJ. Why do all drugs work in animals but none in stroke patients? 1 Drugs promoting cerebral blood flow. *J Int Med* (1995) **237**: 79–88.
13. Grotta J. Why do all drugs work in animals but none in stroke patients? 2 Neuroprotective therapy. *J Int Med* (1995) **237**: 89–94.
14. Gladstone D, Black SE, Hakim A. Towards wisdom from failure: lessons from neuroprotective stroke trials and new therapeutic directions. *Stroke* (2002) **33**(8): 2123–36.
15. Lees KR. Neuroprotection is unlikely to be effective in humans using current trial designs: an opposing view. *Stroke* (2002) **33**: 308–09.
16. Stroke Therapy Academic Industry Roundtable (STAIR). Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* (1999) **30**: 2752–58.
17. Lees KR, Asplund K, Carolei A *et al*. Glycine antagonist gavestinel in neuroprotection (GAIN International) in patients with acute stroke: a randomised controlled trial. *Lancet* (2000) **355**: 1949–54.
18. Skanning Jensen B. BMS-204352: a potassium channel opener developed for the treatment of stroke. *CNS Drug Rev* (2002) **8**(4): 353–60.
19. Bath PMW, Blecic S, Bogousslavsky J *et al*. Tirilazad mesylate in acute ischemic stroke – a systematic review. *Stroke* (2000) **31**(9): 2257–65.
20. Cornu C, Boutitie F, Candelise L *et al*. Streptokinase in acute ischemic stroke: an individual patient data meta-analysis – The Thrombolysis in Acute Stroke Pooling Project. *Stroke* (2000) **31**(7): 1555–60.
21. Hacke W, Markku K, Fieschi C *et al*. Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). *Lancet* (1998) **352**: 1245–51.
22. Bath P. Alteplase not yet proven for acute ischaemic stroke. *Lancet* (1998) **352**: 1238–39.
23. Intravenous Magnesium Efficacy in Stroke (IMAGES) Study Investigators. Magnesium for acute stroke (Intravenous Magnesium Efficacy in Stroke trial): randomised controlled trial. *Lancet* (2004) **363**: 439.
24. Grotta J, The US and Canadian Lubeluzole Ischemic Stroke Study Group. Lubeluzole treatment of acute ischemic stroke. *Stroke* (1997) **28**: 2338–46.
25. Diener HC. Multinational randomised controlled trial of lubeluzole in acute ischaemic stroke. *Cerebrovas Dis* (1998) **8**: 172–81.
26. Wahlgren NG, Ranasinha KW, Rosolacci T *et al*. Clomethiazole acute stroke study (CLASS): results of a randomised controlled trial of clomethiazole versus placebo in 1360 acute stroke patients. *Stroke* (1999) **30**: 21–8.
27. Lyden P, Shuaib A, Ng K *et al*. Clomethiazole acute stroke study in ischemic stroke (CLASS-1): final results. *Stroke* (2002) **33**: 122–9.
28. Hacke W, Kaste M, Fieschi C *et al*. Intravenous thrombolysis with tissue plasminogen activator for acute hemispheric stroke. The European Cooperative Acute Stroke Trial (ECASS). *J Am Med Assoc* (1995) **274**: 1017–25.
29. De Deyn P, De Reuck J, Debertdt W, Vlietinck R, Orgogozo JM. Treatment of acute ischemic stroke with piracetam. *Stroke* (1997) **28**: 2347–52.
30. Multicentre Acute Stroke Trial-Italy (MAST-I) Group. Randomised controlled trial of streptokinase, aspirin, and combination of both in treatment of acute ischaemic stroke. *Lancet* (1995) **346**: 1509–14.
31. Donnan GA, Davis SM, Chambers BR *et al*. Streptokinase for acute ischemic stroke with relationship to time of administration: Australian Streptokinase (ASK) Trial Study Group. *J Am Med Assoc* (1996) **276**: 995–6.
32. Hommel M, Group fTFbI. Fraxiparine in ischaemic stroke study (FISS bis). *Cerebrovasc Dis* (1998) **8**: 19.
33. Bath FJ, Owen VE, Bath PMW. Quality of full and final publications reporting acute stroke trials. A systematic review. *Stroke* (1998) **29**: 2203–10.
34. Bath PMW, Bath FJ, Asplund K. Pentoxifylline, propentofylline and pentifylline in acute ischaemic stroke (Cochrane Review). The Cochrane Library, 4th edn. Oxford: Update Software (1998).
35. Rothwell PM. The high cost of not funding stroke research: a comparison with heart disease and cancer. *Lancet* (2001) **357**: 1612–16.
36. Cervera A, Justica C, Reverter JC, Planas AM, Chamorro A. Steady plasma concentration of unfractionated Heparin reduces infarct volume and prevents inflammatory damage after transient focal cerebral ischemia in the rat. *J Neurosci Res* (2004) **77**: 565–72.
37. Horn J, Limburg L, Orgogozo JM. Calcium antagonists for acute ischemic stroke (Cochrane Review). The Cochrane Library, 1st edn. Oxford: Update Software (2001).
38. Horn J, deHaan R, Vermeulen M, Luiten PGM, Limburg M, 2045. Nimodipine in animal model experiments of focal cerebral ischemia. *Stroke* (2001) **32**: 2433–8.
39. Sandercock P, Roberts I. Systematic reviews of animal experiments. *Lancet* (2002) **360**: 586.
40. Pound P, Ebrahim S, Sandercock P, Bracken MB, Roberts I, on behalf of the Reviewing

Animal Trials Systematically (RATS) Group. Where is the evidence that animal research benefits humans? *Br Med J* (2004) **328**: 514–17.

41. Nava-Ocampo AA, Reyes-Perez H, Bello-Ramirez AM, Mansilla-Olivares A, Ponce-Monter H. For ischemic brain damage, is preclinical evidence of neuroprotection by presynaptic blockade of flutamate release enough? *Med Hypotheses* (2000) **54**(1): 77–9.
42. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publications bias. *Stroke* (2004) **35**: 1203–8.
43. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and meta-analysis of the efficacy of melatonin in experimental stroke. *J Pineal Res* (2004) **38**: 35–41.
44. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. *J Blood Flow Metab* (2005) **25**: 713–21.
45. Willmot M, Gray L, Gibson C, Murphy S, Bath PMW. A systematic review of nitric oxide donors and L-arginine in experimental stroke: effects on infarct size and cerebral blood flow. *Nitric Oxide* (2005) **12**: 141–9.
46. Willmot M, Gibson C, Gray L, Murphy S, Bath PMW. Nitric oxide synthase inhibitors in experimental stroke and their effects on infarct size and cerebral blood flow: a systematic review. *Free Radic Biol Med* (2005) **39**: 412–25.
47. Devereaux PJ, Choi PT-L, El-Dika S *et al*. An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. *J Clin Epidemiol* (2004) **57**: 1232–6.
48. The STIPAS Investigators. Safety study of tirilazad mesylate in patients with acute ischemic stroke (STIPAS). *Stroke* (1994) **25**: 418–23.
49. The Abciximab in Ischemic Stroke Investigators. Abciximab in acute ischemic stroke. A randomised, double-blind, placebo-controlled, dose-escalation study. *Stroke* (2000) **31**: 601–9.
50. Diener HC, Ringelstein EB, von Kummer R *et al*. The TOPAS Investigators, 1195. Safety and efficacy of certoparin in patients with acute stroke the TOPAS trial. *Stroke* (2001) 32: 22–29.
51. Diener HC, Hacke W, Hennerici M, Radberg J, Hantson L, De-Keyser J. Lubeluzole in acute ischemic stroke. A double-blind, placebo-controlled Phase II trial. Lubeluzole International Study Group. *Stroke* (1996) **27**: 76–81.
52. Grotta J. Neuroprotection is unlikely to be effective in humans using current trial designs. *Stroke* (2001) **33**: 306–7.
53. Krams M, Lees KR, Hacke W *et al*. Acute stroke therapy by inhibition of neutrophils (ASTIN). An adaptive dose-response study of UK-279,276 in acute ischemic stroke. *Stroke* (2003) **34**: 2543–48.
54. Carter LP, Guthkelch AN, Orozco Jea, 2090. Influence of tissue plasminogen activator and heparin on cerebral ischemia in a rabbit model. *Stroke* (1992) **23**: 883–8.
55. Mori E, Yoneda Y, Tabuchi M *et al*. Intravenous recombinant tissue plasminogen activator in acute carotid artery territory stroke. *Neurology* (1992) **42**: 976–82.
56. Haley EC, Brott TG, Sheppard GL *et al*. Pilot randomized trial of tissue plasminogen activator in acute ischemic stroke. *Stroke.* (1993) **24**: 1000–4.
57. Clark WM, Wissman S, Albers GW, Jhamandas JH, Madden KP, Hamilton S. Recombinant tissue-type plasminogen activator (alteplase) for ischemic stroke 3 to 5 hours after symptom onset. The ATLANTIS study: a randomized controlled trial. *J Am Med Assoc* (1999) **282**: 2019–26.
58. Clark WM, Albers GW, Madden KP, Hamilton S. The rtPA (alteplase) 0–6 hour acute stroke trial, part A (A0276g). Results of a double-blind, placebo-controlled, multicenter study. *Stroke* (2000) **31**: 811–16.
59. Wardlaw JM, Zoppo G, Yamaguchi T, Berge E. Thrombolysis for acute ischaemic stroke. *Cochrane Database Syst Rev* (2003)(3): CD000213.
60. The ATLANTIS ECASS and NINDS rt-PA Study Group Investigators. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. *Lancet* (2004) **363**: 768–813.
61. Lenzer J. Alteplase for stroke: money and optimistic claims buttress the 'brain attack' campaign. *Br Med J* (2002) **324**: 723–9.
62. ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infraction. *Lancet* (1988) **ii**: 349–60.
63. The Multicenter Acute Stroke Trial – Europe Study Group. Thrombolytic therapy with streptokinase in acute ischemic stroke. *New Engl J Med* (1996) **335**: 145–50.
64. Warach S, Pettigrew C, Dashe JF *et al*. Effect of citicoline on ischemic lesions as measured by diffusion-weighted magnetic resonance imaging. *Ann Neurol* (2000) **48**(5): 713–21.

65. Davalos A, Castillo J, Alvarez-Sabin J *et al*. Oral citicoline in acute ischemic stroke. An individual patient data pooling analysis of clinical trials. *Stroke* (2002) **33**: 2850–7.
66. Hacke W, Albers G, Al-Rwi Y *et al*. The Desmoteplase in Acute Ischemic Stroke Trial (DIAS). A Phase II MRI based 9-hour window acute stroke thrombolysis trial with intravenous desmoteplase. *Stroke* (2005) **36**: 66–3.
67. Mayer SA, Brun NC, Begtrup K *et al*. Recombinant activated factor VII for acute intracerebral hemorrhage. *New Engl J Med* (2005) **352**(8): 777–85.
68. Lees KR, Zivin JA, Ashwood T, Davalos A, Davis SM, Diener HC, Grotta J, Lyden P, Shuaib A, Hardemark HG, Wasiewski WW; Stroke-Acute Ischemic NXY Treatment (SAINT I) Trial Investigators. NXY-059 for acute ischemic stroke. *N Engl J Med*. (2006) **354**: 588–600.
69. Zhang L, Zhang ZG, Zhang R *et al*. Adjuvant treatment with glycoprotein IIb/IIIa receptor inhibitor increases the therapeutic window for low low-dose tissue plasminogen activator administration in a rat model of embolic stroke. *Circulation* (2003) **107**: 2837–43.
70. Graham GD. Tissue plasminogen activator for acute ischemic stroke in clinical practice: a meta analysis of safety data. *Stroke* (2003) **34**: 2847–50.
71. Tilley BC, Marler J, Geller NL *et al*. Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA stroke trial. *Stroke* (1996) **27**: 2136–42.
72. The NINDS t-PA Stroke Study Group. Generalized efficacy of t-PA for acute stroke: subgroup analysis of the NINDS t-PA stroke trial. *Stroke* (1997) **28**(11): 2119–25.
73. Brott T, Lu M, Kothari R *et al*. Hypertension and its treatment in the NINDS rt-PA stroke trial. *Stroke* (1998) **29**: 1504–9.
74. Marler JR, Tilley BC, Lu M *et al*. Early stroke treatment associated with better outcome. The NINDS re-PA Stroke Study. *Neurology* (2000) **55**: 1649–55.
75. Broderick JP, Lu M, Kothari R *et al*. Finding the most powerful measures of the effectiveness of tissue plasminogen activator in the NINDS tPA Stroke Trial. *Stroke* (2000) **31**(10): 2335–41.
76. Vartiainen N, Goldsteins G, Keksa-Goldsteine V, Chan PH, Koistinaho J. Aspirin inhibits p44/42 mitogen-activated protein kinase and is protective against hypoxia/reoxygenation neuronal damage. *Stroke* (2003) **34**: 752–7.
77. Chen ZM, Sandercock P, Pan HC *et al*. Indications for early aspirin use in acute ischemic stroke – a combined analysis of 40 000 randomized patients from the Chinese Acute Stroke Trial and the International Stroke Trial. *Stroke* (2000) **31**(6): 1240–9.
78. Gubitz G, Sandercock P, Counsell C. *Antiplatelet therapy for acute ischaemic stroke (Cochrane Review). The Cochrane Library*, 3rd edn. Oxford: Update Software (2002).
79. Weir NU, Sandercock PAG, Lewis SC, Signorini DF, Warlow CP, on behalf of the IST Collaborative Group. Variations between countries in outcome after stroke in the International Stroke trial (IST). *Stroke* (2001) **32**: 1370–77.
80. Saxena R, Lewis S, Berge E, Sandercock PAG, Koudstaal PJ. Risk of Early Death and Recurrent Stroke and Effect of Heparin in 3169 Patients with Acute Ischemic Stroke and Atrial Fibrillation in the International Stroke Trial. *Stroke* (2001) **32**(10): 2333–7.
81. Leonardi-Bee J, Bath PMW, Phillips SJ, Sandercock PAG, for the IST Collaborative Group. Blood pressure and clinical outcomes in the International Stroke Trial. *Stroke* (2002) **33**: 1315–20.
82. Andersen M, Overgaard K, Meden P, Boysen G. Effects of citicoline combined with thrombolytic therapy in a rat embolic stroke model. *Stroke* (1999) **30**: 1464–71.
83. Sobrado M, Lopez MG, Carceller F, Garcia AG, Roda JM. Combined nimodipine and citicoline reduce infarct size, attenuate apoptosis and increase BCL-2 expression after focal cerebral ischemia. *Neuroscience* (2003) **118**: 107–13.
84. Clark WM, Warach SJ, Pettigrew LC, 1202. A randomised dose-response trial of citicoline in acute ischemi stroke patients. *Neurology* (1977) **49**: 671–8.
85. Clark WM, Williams BJ, Selzer KA, Zweifler RM, Sabounjian LA, Gammans RE. A randomised efficacy trial of citicoline in patients with acute ischaemic stroke. *Stroke* (1999) **30**: 2592–7.
86. Clark WM, Wechsler LR, Sabounjian LA, Schwiderski UE. A Phase III randomized efficacy trial of 2000 mg citicoline in acute ischemic stroke patients. *Neurology* (2001) **57**: 1595–1602.
87. Liberatore GT, Samson A, Bladin C, Schleuning W-D, Medcalf RL. Vampire bat salivary plasminogen activator (Desmoteplase). A unique fibrinolytic enzyme that does not promote neurodegeneration. *Stroke* (2003) **34**: 537–43.
88. Bordi F, Pietra C, Ziviani L, Reggiani A. The glycine antagonist GV150526 protects somatosensory evoked potentials and reduces the infarct area in the MCAo model of focal ischemia in the rat. *Exp Neurol* (1997) **145**(2 Pt 1): 425–33.

89. Dyker AG, Lees KR. Safety and tolerability of GV150526 (a glycine site antagonist at the N-Methyl-D-Aspartate receptor) in patients with acute stroke. *Stroke* (1999) **30**: 986–92.
90. Lees K, Lavelle JF, Cunha L *et al.* Glycine antagonist (GV 150526) in acute stroke: a multicentre, double-blind placebo controlled Phase II trial. *Cerebrovasc Dis* (2001) **11**: 20–9.
91. Sacco RL, DeRosa JT, Haley EC *et al.* Glycine antagonist in neuroprotection for patients with acute stroke. GAIN Americas: a randomized controlled trial. *J Am Med Assoc* (2001) **285**: 1719–28.
92. Muir KW, Lees KR. Excitatory amino acid antagonists for acute stroke (Cochrane Review). The Cochrane Library, Issues 3 Oxford: Update Software (2002).
93. Aslayan S, Fazekan F, Weir CJ, Horner S, Lees KR, Investigators ftGISCa. Effect of blood pressure during the acute period of ischemic stroke on stroke outcome. A tertiary analysis of the GAIN international trial. *Stroke* (2003) **34**: 2420–5.
94. Bolland K, Weeks A, Whitehead J, Lees KR, GAIN International Steering Committee and the Investigators. How a sequential design would have affected the GAIN International Study of gavestinel in stroke. *Cerebrovasc Dis* (2004) **17**(2–3): 111–17.
95. Aslayan S, Weir CJ, Johnston C, Lees KR, for the GaIN International Steering Committee and Investigators. Poststroke neurological improvement within 7 days is associated with subsequent deterioration. *Stroke* (2004) **35**: 2165–70.
96. Young FB, Lees KR, Weir CJ, For the Glycine Antagonist in Neuroprotection (GAIN) International Trial Steering Committee and Investigators. Strengthening acute stroke trials through optimal use of disability end points. *Stroke* (2003) **34**: 2676–80.
97. Haley EC, Thompson JL, Levin B *et al.* Gavestinel does not improve outcome after acute intracerebral hemorrhage: an analysis from the GAIN International and GAIN Americas studies. *Stroke* (2005) **36**(5): 1006–10.
98. Dahl A, Russell D, Nyberg-Hansen R, Rootwelt K. Effect of nitroglycerin on cerebral circulation measured by transcranial doppler and SPECT. *Stroke* (1989) **20**: 1733–36.
99. Butterworth RJ, Cluckie A, Jackson SHD, Buxton-Thomas M, Bath PMW. Pathophysiological assessment of nitric oxide (given as sodium nitroprusside) in acute ischaemic stroke. *Cerebrovasc Dis* (1998) **8**: 158–65.
100. Bath PMW, Pathansali R, Iddenden R, Bath FJ. The effect of transdermal glyceryl trinitrate, a nitric oxide donor, on blood pressure and platelet function in acute stroke. *Cerebrovasc Dis* (2001) **11**: 265–72.
101. Rashid P, Weaver C, Leonardi-Bee JA, Fletcher S, Bath FJ, Bath PMW. The effects of transdermal glyceryl trinitrate, a nitric oxide donor on blood pressure, cerebral and cardiac haemodynamics and plasma nitric oxide levels in acute stroke. *J Stroke Cerebrovasc Dis* (2003) **13**: 143–51.
102. Bath PMW, Willmot M, Leonardi-Bee J, Bath-Hextall FJ. Nitric oxide donors (nitrates), *L-arginine, or nitric oxide synthase inhibitors for acute ischaemic stroke (Cochrane Review). The Cochrane Library*, Issues 4. Oxford: Update software.
103. Aronowski J, Strong R, Grotta J. Combined neuroprotection and reperfusion therapy for stroke. *Stroke* (1996) **27**: 1571–77.
104. Haseldonckx M, Van Reempts J, Van de Ven M, Wouters L, Borgers M. Protection with Lubeluzole against delayed ischemic brain damage in rats. *Stroke* (1997) **28**: 428–32.
105. Hantson L, Tritsmans L, Crabbe R, Gheuens J, Van Rooy P. The safety and tolerability of single intravenous doses of lubeluzole (Prosynap) in healthy volunteers. *Int J Clin Pharmacol Ther* (1997) **35**(11): 491–5.
106. Diener H, Cortens M, Ford G *et al.* Lubeluzole in acute ischaemic stroke treatment. *Stroke* (2000) **31**: 2543–51.
107. Hacke W, Lees KR, Timmerhuis T *et al.* Cardiovascular safety of lubeluzole (Prosynap) in patients with ischaemic stroke. *Cerebrovas Dis* (1998) **8**: 247–54.
108. Zhu H-D, Martin R, Meloni B *et al.* Magnesium sulfate fails to reduce infarct volume following transient focal cerebral ischemia in rats. *Neurosci Res* (2004) **49**: 347–53.
109. Muir KW, Lees KR. A randomized, double-blind, placebo-controlled pilot trial of intravenous magnesium sulfate in acute stroke. *Stroke.* (1995) **26**: 1183–8.
110. Muir KW, Lees KR. Dose optimization of intravenous magnesium sulfate after acute stroke. *Stroke* (1998) **29**: 918–23.
111. Marshall JWB, Duffin KJ, Green AR, Ridley RM. NXY-059, a free radical-trapping agent, substantially lessens the functional disability resulting from cerebral ischemia in a primate species. *Stroke* (2001) **32**: 190–8.
112. Lapchak PA, Araugo DM, Song D, Wei J, Zivin JA. Neuroprotective effects of the spin trap agent disodium-[(tert-butylimino)methyl] benzene-1,3-disulfonate N-oxide (generic NXY-059) in rabbit small clot embolic stroke model. Combination studies with the thrombolytic tissue plasminogen activator. *Stroke* (2002) **33**: 1411–15.

113. Lapchak PA, Araugo DM, Song D, Wei J, Purdy R, Zivin JA. Effects of the spin trap agent disodium-[(tert-butylimino)methyl]benzene-1,3-disulfonate N-oxide (generic NXY-059) on intracerebral hemorrhage in a rabbit large clot embolic stroke model. Combination studies with tissue plasminogen activator. *Stroke* (2002)(33): 1665–70.

114. Sydserff SG, Borelli AR, Green AR, Cross AJ. Effect of NXY-059 on infarct volume after transient or permanent middle cerebral artery occlusion in the rat; studies on dose, plasma concentration and therapeutic time window. *Br J Pharmacol* (2002) **135**: 103–12.

115. Marshall JWB, Cummings RM, Bowes LJ, Ridley RM, Green AR. Functional and histological evidence for the protective effect of NXY-059 in a primate model of stroke when given 4 hours after occlusion. *Stroke* (2003)(34): 2228–33.

116. Han M, He QP, Yong G, Siesjo BK, Li PA. NXY-059, a nitrone with free radical trapping properties inhibits release of cytochrome C after focal cerebral ischemia. *Cell Mol Biol* (2003) **49**(8): 1249–52.

117. Lapchak PA, Song D, Wei J, Zivin JA. Coadministration of NXY-059 and tenecteplase six hours following embolic strokes in rabbits improves clinical rating scores. *Exp Neurol* (2004) **188**: 279–85.

118. Edenius C, Strid S, Borga O, Breitholtz-Emanuelsson A, Lanbeck Vallen K, Fransson B. Pharmacokinetics of NXY-059, a nitrone-based free radical trapping agent, in health young and elderly subjects. *J Stroke Cerebrovasc Dis* (2002) **11**(1): 34–43.

119. Strid S, Borga O, Edenius C, Jostell K-G, Odergren T, Weil A. Pharmacokinetics in renally impaired subjects of NXY-059, a nitrone-based, free-radical trapping agent developed for the treatment of acute stroke. *J Clin Pharmacol* (2002) **58**: 409–15.

120. Lees KR, Sharma AK, Barer D *et al*. Tolerability and Pharmacokinetics of the nitrone NXY-059 in patients with acute stroke. *Stroke* (2000) **32**: 675–80.

121. Lees KR, Barer D, Ford GA *et al*. Tolerability of NXY-059 at higher target concentrations in patients with acute stroke. *Stroke* (2003) **34**: 482–7.

122. Hall ED, Smith SL. The 21-aminosteroid antioxidant tirilazad mesylate, U-74006F, blocks cortical hypoperfusion following spreading depression. *Brain Res* (1991) **553**: 243–8.

123. Meden P, Overgaard K, Pedersen H, Boysen G, 1409. Effect of early treatment with tirilazad (U74006F) combined with delayed thrombolytic therapy in rat embolic stroke. *Cerebrovas Dis* (1996) **6**: 141–8.

124. Fleishaker JC, Peters GR, Cathcart BS. Evaluation of the pharmacokinetics and tolerability of tirilazad mesylate, a 21-aminosteroid free radical scavenger: I. Single-dose administration. *J Clin Pharmacol* (1993) **33**: 175–81.

125. Fleishaker JC, Peters GR, Cathcart BS. Evaluation of the pharmacokinetics and tolerability of tirilazad mesylate, a 21-aminosteroid free radical scavenger: II. Multiple-dose administration. *J Clin Pharmacol* (1993) **33**: 182–90.

126. Olsen KS, Videbaek C, Agerlin N *et al*. The effect of tirilazad mesylate (U74006F) on cerebral oxygen consumption, and reactivity of cerebral blood flow to carbon dioxide in healthy volunteers. *Anesthesiology* (1993) **79**(4): 666–71.

127. Unknown. Phase II European trial of high dose Tirilizad. Unpublished (1995).

128. Scott P, Barsan W, Frederiksen S *et al*. A randomized trial of tirilazad mesylate in patients with acute stroke (RANTTAS). *Stroke* (1996) **27**(9): 1453–8.

129. Haley EC. High-dose tirilazad for acute stroke (RANTTAS II). *Stroke* (1998) **29**: 1256–7.

130. Saver JL, Johnston KC, Homer D *et al*. Infarct volume as a surrogate or auxiliary outcome measure in ischemic stroke clinical trials. *Stroke* (1999) **30**: 293–8.

131. van der Worp HB, Kappelle LJ, Algra A, Bar PR, Orgogozo JM, Ringelstein EB. The effect of tirilazad mesylate on infarct volume of patients with acute ischaemic stroke. *J Neurol* (2002) **58**(1): 133–5.

132. Apfel CC, Korttila K, Abdalla M *et al*. A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. *New Engl J Med* (2004) **350**(24): 2441–51.

133. Sever PS, Dahlof B, Poulter NR *et al*. Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian cardiac outcomes trial–lipid lowering arm (ASCOT-LLA): a multicentre randomised controlled trial. *Lancet* (2003) **361**(9364): 1149–58.

134. Bath PM. Major ongoing stroke trials. Efficacy of Nitric Oxide in Stroke (ENOS) trial. *Stroke* (2001) **32**: 2450–51 (abstr).

135. Dennis M, Wardlaw J, Sandercock P, Signorini D, Warlow C. Families of trials: the answers to all our questions? *Cerebrovasc Dis* (1998) **9**: 305–13.

136. The FOOD Trial Collaboration. Routine oral nutritional supplementation for stroke patients in hospital (FOOD): a multicentre randomised controlled trial. *Lancet* (2005) **365**: 755–62.

137. The FOOD Trial Collaboration. Effect of timing and method of enteral tube feeding for dysphagic stroke patients (FOOD): a multicentre randomised controlled trial. *Lancet* (2005) **365**: 764–72.

138. Committee for Proprietary Medicinal Products. Note for guidance on choice of control group in clinical trials (CPMP/ICH/364/96). *ICH Topic E 10. Choice of control groups in clinical trials*. London: The European Agency for the Evaluation of Medicinal Products (2001) 1–29.

139. Furlan A, Higashida R, Wechsler L *et al*. Intraarterial prourokinase for acute ischemic stroke. The PROACT II study: a randomized trial. *J Am Med Assoc* (1999) **282**: 2003–11.

140. Sherman DG, Atkinson RP, Chippendale T *et al*. Intravenous ancrod for treatment of acute ischemic stroke. The STAT study: a randomised controlled trial. *J Am Med Assoc* (2000) **283**: 2395–402.

141. Berge E, Abdelnoor M, Nakstad PH, Sandset PM. Low molecular-weight heparin versus aspirin in patients with acute ischaemic stroke and atrial fibrillation: a double-blind randomised study. *Lancet* (2000) **355**: 1205–10.

142. Bath P, Lindenstrom E, Boysen G *et al*. Tinzaparin in acute ischaemic stroke (TAIST): a randomised aspirin-controlled trial. *Lancet* (2001) **358**: 702–10.

143. Brott T, Adams HP, Olinger CP *et al*. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke* (1989) **20**: 864–70.

144. Scandinavian Stroke Study Group. Multicenter trial of hemodilution in ischemic stroke – background and study protocol. *Stroke.* (1985) **16**: 885–90.

145. Scandinavian Stroke Study Group. Multicenter trial of hemodilution in acute ischemic stroke. I. Results in the total patient population. Scandinavian Stroke Study Group. *Stroke* (1987) **18**: 691–9.

146. Counsell C, Dennis M, McDowall M. Predicting functional outcome in acute stroke: comparison of a simple six variable model with other predictive systems and informal clinical prediction. *J Neurol Neurosurg Psychiatr* (2004) **75**: 401–5.

147. Lyden P, Brott T, Tilley B *et al*. Improved reliability of the NIH stroke scale using video training. *Stroke* (1994) **25**: 2220–26.

148. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* (1974) **2**: 81–3.

149. Wade DT. Measurement in neurological rehabilitation. Oxford: Oxford University Press (1992).

150. Goldstein LB, Bertels C, Davis JN. Interrater reliability of the NIH stroke scale. *Arch Neurol* (1989) **46**: 660–2.

151. Lindenstrom E, Boysen G, Christiansen LW, Rogvi Hansen B, Nielsen PW, 1514. Reliability of Scandinavian neurological stroke scale. *Cerebrovas Dis* (1991) **1**: 103–7.

152. Mahoney FI, Barthel DW. Functional evaluation: The Barthel index. *Maryland State Med J* (1965): 61–5.

153. Rankin J. Cerebral vascular accidents in patients over the age of 60. 2. Prognosis. *Scott Med J* (1957) **2**: 200–15.

154. Lindley RI, Waddell F, Livingstone M *et al*. Can simple questions assess outcome after stroke? *Cerebrovasc Dis* (1994) **4**: 314–24.

155. Jennett B, Bond M. Assessment of outcome after severe brain damage. A practical scale. *Lancet* (1975) **i**: 480–4.

156. Sulter G, Steen C, de Keyser J, 1540. Use of the Barthel Index and modified Rankin Scale in acute stroke trials. *Stroke* (1999) **30**: 1538–41.

157. Weimar C, Kurth T, Kraywinkel K *et al*. Assessment of functioning and disability after ischemic stroke. *Stroke* (2002) **33**: 2053–9.

158. Committee for Proprietary Medicinal Products. Points to consider on clinical investigation of medicinal products for the treatment of acute stroke. London: The European Agency for the Evaluation of Medicinal Products (2001): 1–7.

159. Bogousslavsky J, Donnan GA, Fieschi C *et al*. Fiblast (trafermin) in acute stroke: results of the European -Australian Phase II/III safety and efficacy trial. *Cerebrovas Dis* (2000) **10** (Suppl 2): 106.

160. Horn J, deHaan R, Vermeulen M, Limburg M. Very early nimodipine use in stroke (VENUS). A randomised, double-blind, placebo-controlled trial. *Stroke* (2001) **32**: 461–5.

161. Wardlaw J, Sandercock P, Warlow C, Lindley RI. Trials of thrombolysis in acute ischemic stroke. Does the choice of primary outcome measure really matter? *Stroke* (2000) **31**: 1133–35.

162. Diener HC, Kaste M, Hacke W *et al*. *Lubeluzole in acute ischemic stroke*. Stroke (1997) **28**: 271.

163. Grotta J, Hantson L, Wessel T, Group TL-I-LS. The efficacy and safety of lubeluzole in patients with acute ischemic stroke. Stroke (1997) **28**: 271.

164. Hanger HC, Fogarty B, Wilkinson TJ, Sainsbury R. Stroke patients' views on stroke outcomes: death versus disability. *Clin Rehabil* (2000) **14**: 417–24.

165. Hommel M, Boiseel JP, Cornu C *et al*. Termination of trial of streptokinase in severe acute ischaemic stroke. *Lancet* (1994) **345**: 57.

166. Korner BN, Wood DS, Siemiatycki J, Shapiro S, Becker R, 1479. Health-related information post

discharge: telephone versus face-to-face interviewing. *Arch Phys Med Rehabil* (1994) **75**: 1287–96.

167. Shinar D, Gross CR, Bronstein KS, *et al.*, 1481. Reliability of the activities of daily living scale and its use in telephone interview. *Arch Phys Med Rehabil* (1987) **68**: 723–8.
168. Lyden PD, Broderick J, Mascha E. Reliability of the Barthel Index outcome measure selected for the NINDS t-PA stroke trial, 3rd edn. Tokyo: Springer (1995).
169. Hsu CY, Faught RE Jr, Furlan AJ *et al.* Intravenous prostacyclin in acute nonhemorrhagic stroke: a placebo-controlled double-blind trial. *Stroke.* (1987) **18**: 352–8.
170. Wade DT, Langton Hewer R, 1506. Functional abilities after stroke measurement, natural history and prognosis. *J Neurol* (1987) **50**: 177–82.
171. Wade DT, Collin C, 1504. The Barthel ADL index: a standard measure of physical disability? *Int Disab Stud* (1987) **10**: 64–7.
172. Brown EB, Tietjen GE, Deveshwar RK *et al.* Clinical Stroke Scales: an intra- and interscale evaluation. *Neurology* (1990) **40** (Suppl 1): 352.
173. Granger CV, Hamilton BB, Gresham GE, Kramer AA, 1509. The stroke rehabilitation outcome study: part II. Relative merits of the total Barthel Index score and a four item subscore in predicting patient outcomes. *Arch Phys Med Rehabil* (1989) **70**: 100–3.
174. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJA, van Gijn J, 1519. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* (1988) **19**: 604–7.
175. Kay R, Wong KS, Yu YL *et al.* Low-molecular-weight heparin for the treatment of acute ischemic stroke. *New Engl J Med* (1995) **333**: 1588–93.
176. Lefkopoulou M, Ryan L. Global tests for multiple binary outcomes. *Biometrics* (1993) **49**: 975–88.
177. Saxena R, Wijnhoud AD, Carton H *et al.* Controlled safety study of a hemoglobin-based oxygen carrier, DCLHb, in acute ischemic stroke. *Stroke* (1999) **30**: 993–6.
178. Davis SM, Lees KR, Albers GW *et al.* Selfotel in acute ischemic stroke. Possible neurotoxic effects of an NMDA antagonist. *Stroke* (2000) **31**: 347–54.
179. Enlimomab acute stroke trial investigators. Use of anti-ICAM-1 therapy in ischemic stroke: results of the Enlimomab acute stroke trial. *Neurology* (2001) **57**(8): 1428–34.
180. Wahlgren NG, MacMahon DG, de Keyser J, Indredavik B, Ryman T, INWEST Study Group. Intravenous Nimodipine West European Stroke Trial (INWEST) of nimodipine in the treatment of acute ischaemic stroke. *Cerebrovasc Dis* (1994) **4**: 204–10.
181. Albers G, Goldstein LB, Hall D, Lesko LM. Aptiganel hydrochloride in acute ischemic stroke. *J Am Med Assoc* (2001) **286**: 2673–82.
182. De Haan R, Horn J, Limburg M, van der Meulen J, Bossuyt P. A comparison of five stroke scales with measures of disability, handicap, and quality of life. *Stroke* (1993) **24**: 1178–81.
183. Croog SH, Levine SR, Testa MA *et al.* The effects of antihypertensive therapy on the quality of life. *New Eng J Med* (1986) **314**(26): 1657–64.
184. Hansson L, Lithell H, Skoog I *et al.* Study on COgnition and Prognosis in the Elderly (SCOPE). *Blood Press* (1999) **8**(3): 177–83.
185. Dorman PJ, Slattery J, Farrell B, Dennis MS, Sandercock PAG, United Kingdom Collaborators in the International Stroke Trial. A randomised comparison of the EuroQol and short form-36 after stroke. *Br Med J* (1997) **315**: 461.
186. Forette F, Seux ML, Staessen JA *et al.* Prevention of dementia in randomised double-blind placebo-controlled Systolic Hypertension in Europe (Syst-Eur) trial. *Lancet* (1998) **352**: 1347–51.
187. Chalmers J, MacMahon S, Committee obotPM. PROGRESS – perindopril protection against recurrent stroke study: main results. *J Hypertens* (2001) **19** (Suppl 2): S260.
188. Lithell H, Hansson L, Skoog I *et al.* The study on cognition and prognosis in the elderly (SCOPE): principal results of a randomized double-blind intervention trial. *J Hypertens* (2003) **21**(5): 875–86.
189. Galasko D, Klauber MR, Hofstetter R *et al.* The Mini-Mental State Examination in the early diagnosis of Alzheimer's disease. *Arch Neurol* (1990) **47**: 49–52.
190. Robinson RG. Poststroke depression: prevalence, diagnosis, treatment and disease progression. *Biol Psychiatr* (2003) **54**(3): 376–87.
191. Dam M, Tonin P, De Boni A *et al.* Effects of fluoxetine and maprotiline on functional recovery in poststroke hemiplegic patients undergoing rehabilitation therapy. *Stroke* (1996) **27**: 1211–14.
192. Tucker MA, Ogle SJ, Davison JG, Eilenberg MD, 1740. Validation of a brief screening test for depression in the elderly. *Age Ageing* (1987) **16**: 139–44.
193. Staessen JA, Wang J-G, Thijs L. Cardiovascular protection and blood pressure reduction: a meta-analysis. *Lancet* (2001) **358**: 1305–15.
194. Rashid P, Leonardi-Bee J, Bath P. Blood pressure reduction and the secondary prevention of stroke and other vascular events: a systematic review. *Stroke* (2003) **34**: 2741–49.

195. Warach S. New imaging strategies for patient selection for thrombolytic and neuroprotective therapies. *Neurology* (2001) **57**(5 (Suppl 2)): 48–52.
196. Warach S. Thrombolysis in stroke beyond three hours: targeting patients with diffusion and perfusion MRI. *Ann Neurol* (2002) **51**(1): 11–13.
197. del Zoppo GJ, Higashida RT, Furlan AJ *et al.* PROACT: a Phase 2 randomised trial of recombinant pro-urokinase by direct arterial delivery in acute middle cerebral artery stroke. *Stroke* (1998) **29**: 4–11.
198. Higashida RT, Furlan AJ, Technology Assessment Committees of the American Society of Interventional and Therapeutic Neuroradiology and the Society of Interventional Radiology. Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke. *Stroke* (2003) **34**: e109–37.
199. De Gruttalo VG, Clax P, DeMets DL *et al.* Consideration in the evaluation or surrogate endpoints in clinical trials: summary of a national institutes of health workshop. *Controlled Clin Trials* (2001) **22**: 485–502.
200. Herrmann M, Enhrenreich H. Brain derived proteins as markers of acute stroke: their relation to pathophysiology, outcome prediction and neuroprotective drug monitoring. *Restorative Neurol Neurosci* (2003) **21**: 177–90.
201. Abraha HD, Butterworth RJ, Bath PMW, Wassif WS, Garthwaite J, Sherwood RA. Serum S-100 protein, a prognostic marker of clinical outcome in acute stroke. *Ann Clin Biochem* (1997) **34**: 366–70.
202. Butterworth RJ, Wassif WS, Sherwood RA *et al.* Serum neurone specific enolase, carnosinase and their ratio in acute stroke: an enzymatic test for predicting outcome? *Stroke* (1996) **27**: 2064–68.
203. Prados J, Kalousis A, Sanches J-C, Allard L, Carrette O, Hilario M. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics* (2004) **4**: 2320–32.
204. Reynolds MA, Kirchick HJ, Dahlen JR *et al.* Early biomarkers of stroke. *Clin Chem* (2003) **49**(10): 1733–9.
205. Lynch JR, Blessing R, White WD, Grocott HP, Newman MF, Laskowitz DT. Novel diagnostic test for acute stroke. *Stroke* (2004) **35**: 57–63.
206. Ehrenreich H, Hasselblatt M, Dembowski C *et al.* Erythropoietin therapy for acute stroke is both safe and beneficial. *Mol Med* (2002) **8**(8): 495–505.
207. Pettigrew LC, Kasnor SE, Gorman M, Atkinson RP, Dietrich D, Ishibashi H. ONO-2506 suppresses S-100B in ischemic stroke patients. *Stroke* (2004) **35**(1): 288.
208. Etling JW, Sulter GA, Kaste M *et al.* AMPA antagonist ZK200775 in patients with acute ischemic stroke: possible glial cell toxicity detected by monitoring of S-100B serum levels. *Stroke* (2002) **33**: 2813–18.
209. Ferlito S, Gallina M, Pitari GM, Bianchi A. Nitric oxide plasma levels in patients with chronic and acute cerebrovascular disorders. *Panminerva Med* (1998) **40**: 51–4.
210. Rashid PA, Whitehurst A, Lawson N, Bath PMW. Plasma nitric oxide (nitrate/nitrite) levels in acute stroke and their relationship with severity and outcome. *J Stroke Cerebrovasc Dis* (2003) **12**(2): 82–7.
211. Bath PMW, Blann A, Smith N, Butterworth RJ. Von Willebrand factor, p-selectin and fibrinogen levels in patients with acute ischaemic and haemorrhagic stroke, and their relationship with stroke sub-type and functional outcome. *Platelets* (1998) **9**: 155–9.
212. Barber M, Langhorne P, Rumley A, Lowe GDO, Stott DJ. Hemostatic function and progressing ischemic stroke D-dimer predicts early clinical progression. *Stroke* (2004) **35**: 1421–5.
213. Pocock SJ. *Clinical trials: a practical approach.* Chichester: John Wiley & Sons, Ltd (1983).
214. Weaver CS, Leonardi-Bee J, Bath F, Bath PMW. Sample size calculations in acute stroke trials: a systematic review of their reporting characteristics, and relationship with outcome. *Stroke* (2004) **35**: 1216–24.
215. The Publications Committee for the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) Investigators. Low molecular weight heparinoid, ORG 10172 (danaparoid), and outcome after acute ischemic stroke. *J Am Med Assoc* (1998) **279**: 1265–72.
216. Whitehead J. Sample-Size Calculations for Ordered Categorical-Data. *Stat Med* (1993) **12**: 2257–71.
217. Weir CJ, Kaste M, Lees KR, for the Glycine Antagonist in Neuroprotection (GAIN) International Trial Steering Committee and Investigators. Targeting neuroprotection clinical trials to ischemic stroke patients with potential to benefit from therapy. *Stroke* (2004) **35**: 2111–6.
218. Hart RG, Benavente O, McBride R, Pearce LA. Antithrombotic therapy to prevent stroke in patients with atrial fibrillation: a meta-analysis. *Ann Intern Med* (1999) **131**(7): 492–501.
219. Weaver CS, Leonardi-Bee J, Bath-Hextall FJ, Bath PMW. Inclusion and exclusion criteria in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. *Cerebrovasc Dis* (2004) **17**: 124 (abstr).
220. Fisher M. Characterizing the target of acute stroke therapy. *Stroke* (1997) **28**: 866–72.

221. Ricci S, Celani MG, Cantisani AT, Righetti E. Piracetam for acute ischaemic stroke (Cochrane Review). The Cochrane Library, 4th edn. Oxford: Update Software (1999).
222. Weir CJ, Lees KR. Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *StatMed* (2003) **22**: 705–26.
223. Bath PM, Young S, Watson G, Weaver C. Improving trials in acute stroke using internet randomisation and data collection. *Br Assoc Stroke Physicians* (2003).
224. Hacke W, Bluhmki E, Steiner T *et al*. Dichotomized efficacy end points and global end-point analysis applied to the ECASS intention to treat data set. Post hoc analysis of ECASS 1. *Stroke* (1998) **29**: 2073–75.
225. Bath PMW, LETTER. Re: Stroke therapy academic industry roundtable II (STAIR II). *Stroke* (2002) **33**(2): 639.
226. Gray LJ, Bath PMW. Do global outcomes increase efficiency in stroke clinical trials? *Cerebrovas Dis* (2004) **17**(Suppl 5): 125.
227. Adams HP, Bendixen BH, Kappelle LJ *et al*. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. *Stroke* (1993) **24**: 35–41.
228. Bamford J, Sandercock P, Dennis M, Burn J, Warlow C. Classification and natural history of clinically identifiable subtypes of cerebral infarction. *Lancet* (1991) **337**: 1521–26.
229. Mendelow AD, Gregson BA, Fernades HM *et al*. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial. *Lancet* (2005) **365**: 387–97.
230. Young FB, Lees KR, Weir CJ, for the GAIN International Steering Committee and Investigators. Improving trial power through use of prognosis-adjusted end points. *Stroke* (2005) **36**: 597–601.
231. Adams HP, Leclerc JR, Bluhmki E, Clarke WR, Hansen MD, Hacke W. Measuring outcomes as a function of baseline severity of ischemic stroke. *Cerebrovasc Dis* (2004) **18**: 124–9.
232. Whitehead J. Application of sequential methods to a Phase III clinical trial in stroke (ASCLEPIOS). Unpublished (1995).
233. Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences, 2nd edn. Singapore: McGraw-Hill (1988).
234. Hacke W, Kaste M, Fieschi C. Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke. *J Am Med Assoc* (1995) **274**: 1017–25.
235. Stingele R, Bluhmki E, Hacke W, (1786). Bootstrap statistics of ECASS II data: just another post hoc analysis of a negative stroke trial? *Cerebrovas Dis* (2001) **11**: 30–3.
236. Koch GG, Edwards S. Clinical efficacy trials with categorical data. Biopharmaceutical statistics for drug development. New York: Marcel Dekker (1998): 403–57.
237. Optimising Analysis of Stroke Trials (OAST) Collaborators, Bath PMW. Optimising the statistical analysis of functional outcome in stroke trials. *Stroke* (2003) **34**: 316 (abstr).
238. The Edaravone Acute Brain Infarction Study Group (Chair: Eiichi Otomo MD). Effect of a novel free radical scavenger, Edaravone (MCI-186), on acute brain infarction. *Cerebrovasc Dis* (2003) **15**: 222–9.
239. Mohr JP, Orgogozo JM, Harrison MJG *et al*. Meta-analysis of oral nimodipine trials in acute ischaemic stroke. *Cerebrovas Dis* (1994) **4**: 197–203.
240. Bath PMW, Iddenden R, Bath FJ. Low molecular weight heparins and heparinoids in acute ischaemic stroke: a systematic review. *Stroke* (2000) **31**: 311–14.
241. Bath P, Leonardi-Bee J, Bath F. Low molecular weight heparin versus aspirin for acute ischemic stroke: a systematic review. *J Stroke Cerebrovasc Dis* (2002) **11**(2): 55–62.
242. Lees KR. Stroke: trial by jury. *Eur Neurol* (2003) **49**: 106–8.
243. Tognoni G, Roncaglioni MC. Dissent: an alternative interpretation of MAST-1. *Lancet* (1995) **346**(8989): 1515.
244. DAMCLES study group. A proposed charger for clinical trial data monitoring committees: helping them to do their job well. *Lancet* (2005) **365**: 711–22.
245. Stroke Therapy Academic Industry Roundtable II (STAIR-II). Recommendations for clinical trial evaluation of acute stroke therapies. *Stroke* (2001) **32**: 1598–1606.
246. Fisher M, for the Stroke Therapy Academic Industry Roundtable. Recommendations for advancing development of acute stroke therapies. Stroke Therapy Academic Industry Roundtable 3. *Stroke* (2003) **34**: 1539–1546.

14

Cardiovascular Disease

KRISHNA PRASAD
Medicines and Healthcare products Regulatory Agency, London, UK & Hon Consultant Cardiologist,
St Thomas' Hospital

INTRODUCTION

Clinical trials in cardiovascular disease have gained a high profile over the last couple of decades in parallel to those relating to the field of oncology. The primary reasons appear to be the high prevalence of cardiovascular disease in the Western world, the advances in cardiovascular therapeutics and, lastly, the ability to demonstrate significant clinical event reduction (mainly mortality) in the medium term (3–5 years) with the currently available therapeutic agents. The factor that significantly influences the cardiovascular clinical trials is therefore the prevalence of the cardiac disease in the developed and developing nations. The majority of the cardiovascular disease is attributable to the following three conditions: atherosclerosis including coronary heart disease, hypertension and, in the elderly, heart failure.

FACTS TO CONSIDER IN CARDIOVASCULAR CLINICAL TRIALS

The principles governing development, management and analysis of trials in cardiovascular diseases are similar to other conditions discussed elsewhere in this book. There is one aspect, however, that has changed over the years. The incidence and prevalence of *premature* atherosclerosis has made cardiovascular disease a headline feature both in terms of actual public health impact and as a topic of considerable scientific or political debate. Whilst atherosclerosis and hypertension are primary factors influencing cardiovascular disease, surrogate markers for these factors, such as cholesterol levels, and other risk factors (e.g. diabetes, obesity) have attained importance because of two reasons: the proportion of the population that is affected and the relative importance of these risk factors in determining outcome. In very few other situations will one find so many 'modifiable' risk factors that determine the outcome! Unsurprisingly therefore, a number of the cardiovascular trials in the last few decades have been in lipid lowering, control of hypertension, prevention or treatment of myocardial infarction and improving mortality and morbidity in heart failure. Whilst this represents a large part of the clinical cardiovascular load, other areas such as congenital heart disease or peripheral vascular disease have received less attention. Indeed these aspects have been

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

the subject of few clinical trials. Similarly, valve disease arising from rheumatic heart disease is fairly common in developing countries but has not been a major subject for multinational clinical trials. This chapter will mainly consider the issues that are common in both developed and developing nations, such as hypertension, atherosclerosis and heart failure. The following few paragraphs will consider factors important in determining the nature of a clinical trial and those that might have a bearing on the outcome of such trials.

An important consideration in any trial is the chronicity of the condition to be investigated. Atherosclerosis and hypertensive cardiovascular diseases are chronic conditions, often taking decades to develop, lasting many years after being diagnosed. Clinical trails may thus be initiated before manifestation/development of risk factors (primordial prevention), after manifestation of risk factors but before organ damage (primary prevention) or after organ damage (secondary prevention). Interventions may be applicable in all three settings, or be specific to a particular situation. Specific and invasive interventions are likely to be particularly expensive and are therefore more often restricted to secondary prevention strategy. The relative importance of risk factors is likely to differ in the above-mentioned three situations affecting the likely impact of an intervention. For example, after a heart attack (myocardial infarction), the function of the surviving myocardium will be more important in determining longevity than cholesterol level alone, which is related to recurrent infarctions and death. One important aspect of such chronic conditions is that treatments or interventions are frequently not curative of the underlying condition, but rather reduce the likelihood of clinical sequelae.

TRIAL SETTINGS – IVORY TOWER OR COMMON GARDEN

Clinical trials have often been the preserve of academic/large institutions, purely because of the ability to designate specific trial personnel and provide the necessary infrastructure. This may, however, be changing with more trials being conducted in the community setting for many reasons. In certain situations, because of the high prevalence, small treatment benefits such as 'use of aspirin in the prevention of myocardial infarction' are likely to have significant public health benefits. Similarly, statins have changed the outlook of the management of cardiovascular risk factors and atherosclerosis in particular. General use of statins (prescribed or otherwise) has increased along lines that aspirin use followed in the 1970s and 1980s. It is therefore not surprising to see many trials involving aspirin or statins being organised and conducted outside of the academic centres (or centres of academic excellence!), which were the main theme some years ago. We may yet see many trials conducted in the community, or 'general practice setting' in the future, and it is conceivable that these could influence the quality of the data collected, the impact of the clinical trial results on the standard of care in the community and the general perception of a clinical trial. In the community setting, the participants are more likely to have less complicated or less advanced disease, the trial personnel are likely to be few and less specialised, and the process less invasive. Furthermore, in recent years the public awareness of the existence of clinical trials and their results has increased considerably. Therefore, the complexity or simplicity of the trial design will be determined by the setting, complex trials being set in academic institutions and simple trials often in the community setting.

GENOTYPE/PHENOTYPE INTERACTIONS

The common causes or risk factors of cardiovascular diseases (atherosclerosis, hypertension, diabetes) are multifactorial in origin. Both atherosclerosis and hypertension may be influenced by lipid levels, smoking (active or passive), obesity, physical activity, age, gender, diabetes mellitus and, in some cases, genetic factors. A number of cardiovascular diseases or their risk factors are believed to be a result of 'many interactions between many genes' and not influenced by single genes/factors. One example is the relation of ACE gene polymorphism (I or D allele) with hypertension, ventricular hypertrophy and clinical events. Those with DD polymorphism appear to be at greater risk for all the above. Similarly, the RAS (Renin–Angiotensin System) inhibitors (ACE inhibitors or A-II receptor blockers) elicit poor response in low renin populations, e.g. blacks in whom there is higher prevalence of DD genotype in contrast to high renin populations (Caucasians – II or ID genotype). These, however, are not considered consistent or hard relationships and the environmental influences on the genes are considered relevant in a number of these. Because of such genetic and environmental influences on multiple risk factors, interventions (in clinical trial or practice) that alter individual factors may yield only modest reduction in clinical outcomes, especially myocardial infarction or cardiovascular death. A case in point is the example of antihypertensive drugs that lower blood pressure. These produce impressive reductions in stroke, but lead only to modest reductions in heart disease.[1] The reason may probably have been that other risk factors, such as atherosclerosis or smoking, remained unchanged in these studies. Similarly, gender may influence the results; in most lipid-lowering trials, men derived greater benefit or women 'appeared to derive less' benefit. Such a finding may have been due to the small number of women included in such trials or some other factor. This multifactorial nature of cardiovascular disease has led to trial designs that attempt intervention at several levels or attempt to address several risk factors simultaneously. Some examples of such trials are the ALLHAT trial with its hypertension arm plus the lipid arm or the ASCOT trial (hypertension and lipid arms-LLA), both of which used a statin in addition to antihypertensive agents, pravastatin in ALLHAT and atorvastatin in ASCOT-LLA.

SIMPLE OR COMPLEX TRIAL?

Typically, clinicians are prone to perceive cardiovascular disease as affecting the heart, e.g. coronary heart disease or brain as in stroke. Other parts of the body such as kidneys, or limbs (legs – intermittent claudication) are often affected. Even within the heart, spectrum cardiovascular disease may assume various forms or presentations such as angina, infarction, cardiac arrhythmias, heart failure and sudden death. When multiple interventions aimed at reducing all such presentations or multiple risk factors become part of the clinical trial, these can affect protocol adherence, cause drug interactions, or even affect end results. These effects may be positive (potentiation) or negative (dilution) of the primary endpoints. Such events should be taken into consideration in designing complex trials or complex endpoints. The MACH-1 trial is a good example to demonstrate unexpected drug interactions seen during a complex trial. Mibefradil, a vasodilatory T-channel calcium blocker with negative chronotropic properties (heart rate lowering), was compared with placebo in addition to standard treatment in heart failure. The standard treatment included ACE inhibitors, diuretics and digitalis. In those with pre-existing coronary disease, statins, aspirin and antiarrhythmics were also permitted. Whilst the trial did not show additional reductions in mortality or other endpoints at the end of three years, a significant interaction of Mibefradil with statins and amiodarone forced complete withdrawal of this agent.

CHOICE OF ENDPOINTS

In designing and interpreting results of trials in cardiovascular diseases, the choice of the endpoint is an extremely important issue. As noted, interventions may affect one aspect of the disease and the investigators often prefer to use the most likely aspect modifiable by treatment as the 'outcome of interest'. This may be difficult for endpoints such as cause-specific mortality, and may lead to oversimplification of trial results or could encourage choice of an inappropriate endpoint. One such example is the WHO Clofibrate trial that showed reduction in cholesterol but over the time period increased overall mortality and cause-related mortality for several diseases.[2,3] A number of these, however, were unrelated (such as increased respiratory infections) to the mechanism of action of clofibrate and hence the results remain unexplained. Even the adoptions of broader outcomes such as death due to cardiovascular disease have limitations as many deaths are unwitnessed and autopsies are not common. Any attempts at refinement of filters (or endpoints), such as death due to arrhythmia or myocardial infarction, may cause the difficulties to increase.[4] As an example, diagnosis of myocardial infarction involves an electrocardiogram and laboratory testing as confirmatory evidence, introducing new factors for identifying the endpoint. The laboratory tests are currently very well defined and so specific that even tests considered gold standard a few years ago, such as the CK-Mb fraction for diagnosing myocardial infarction, have been superseded by other more specific tests. Thus defining and adopting such an endpoint may increase the trial complexity. Death due to arrhythmia is another such area. Whilst arrhythmic death may be dramatic, many deaths are unwitnessed and arrhythmia may be assumed, thus diluting the impact of the results. A trial design assessing cardiovascular mortality would need to consider these carefully and plan for inclusion of appropriate definitions in the protocol. Alternatively, such difficulties may not be reduced by use of nonfatal events. A case in point is the Framingham Heart Study, where more than 25% of myocardial infarctions were 'silent', occurring without symptoms and could only be recognised on electrocardiographic examination.[5] Such examinations or efforts would have to be specified in the trial protocol in relevant studies, because, even if they were asymptomatic, these events convey considerable risk of death.

In most recent cardiovascular trials, all-cause mortality and/or cardiovascular mortality have been the most frequently used endpoints. Most European regulatory agencies expect to see an analysis of 'all-cause mortality' and find use of 'cardiovascular mortality' alone as an endpoint perhaps a little restrictive and specific. This distinction has clinical relevance. This divarication or distinction was highlighted by the WHO Clofibrate trial, which predated most guidelines on clinical trials. Cholesterol reduced with Clofibrate treatment (now an accepted surrogate endpoint based on statin data), but all-cause mortality worsened, while cardiovascular mortality did not. Therefore, the debate 'whether all-cause mortality or cardiovascular mortality' as an endpoint is likely to continue, but suffice it to say that, on occasion, a clear benefit in all-cause mortality may not be achievable in a trial.

SURROGATE MARKER OR CLINICAL EVENT

Aspects of a clinical trial such as participant numbers, surrogate endpoints, need for specialist investigators and stratification to ensure balance are often not major considerations in cardiovascular disease, unlike other diseases in view of the high prevalence or commonness of cardiovascular disease. In general, for cardiovascular disease, usually there are sufficient numbers of participants available in multicentre or multinational studies such that clinical outcome studies are feasible. Hence surrogate endpoint trials are not often necessary. Classic examples of these are ACE inhibitors in heart failure or after myocardial infarction. In both these situations, most trials have included a large number of participants, running to several thousands. Another example would be betablockers in heart failure where surrogate endpoints have not been necessary. On the contrary, when a trial involves specific or special subtypes of cardiovascular disease, participant availability may become an issue and use of surrogate endpoints may become necessary. For example, in diseases causing sudden death, hypertrophic cardiomyopathy is considered high on the list after ischaemic heart disease. However, its prevalence in the general population and the frequency (relative infrequency) of sudden cardiac death in this condition dictate use of other markers of sudden death than death itself. Similarly, in those with ventricular arrhythmia inducible by electrophysiological testing, a record of a serious arrhythmia and defibrillation are used as surrogate markers of sudden death. Even when participant numbers are sizeable, one example that provides arguments for and against use of surrogate markers would be a trial involving a new statin or other pharmacotherapy for lipid lowering. On the one hand, a primary endpoint of cholesterol reduction alone may not be considered an adequate surrogate endpoint for reduction in mortality or clinical events (e.g. Clofibrate trial). On the other hand, it is well accepted that 'the lower the lipid level, the better it is' in terms of derived mortality or morbidity benefit based on epidemiological data, and therefore this could be considered the ideal surrogate marker.[6]

In some situations, popular endpoints may not be appropriate surrogate markers of benefit. For example, in acute heart failure, reduction of pulmonary capillary wedge pressure is considered an extremely important clinical endpoint for any pharmacotherapy. A number of trials used this as a surrogate marker (endpoint) of clinical outcome (survival), but this has frequently failed to influence mortality for several drugs such as milrinone,[7,8] more recently nesiritide,[9] or even inotropes. Similarly, ejection fraction is a good prognostic marker in those with failing hearts. Positive inotropes improved left ventricular ejection fraction in the short term, but failed to improve long-term outcome, thus highlighting the dichotomy between this surrogate marker (ejection fraction) and actual event reduction.

ROLE FOR SPECIALIST INVESTIGATORS

The advantages derived from the large population available for study should be considered carefully as cardiovascular disease has a myriad of manifestations. In order to avoid dilution or inappropriate recruitment of participants, specialist investigators may need to be involved in some of these cardiovascular clinical trials. Another additional factor that will influence inclusion of specialist investigators or physicians is use of interventions such as percutaneous coronary intervention, stent insertion, etc., that are accepted standard treatments for a number of cardiovascular conditions. Thus, despite the availability of large number of participants, the potential confounding factors are many, and involvement of specialist physicians is therefore recommended in order to ensure sufficient participant recruitment.

TO STRATIFY OR NOT TO STRATIFY

Stratification is usually unnecessary because of the large size of the trials in cardiovascular diseases. However, the following key factors lend themselves for stratification: study sites are always a stratification variable; in heart failure trials, the left ventricular ejection fraction has been an important variable for stratification; and for arrhythmias, the type of arrhythmia and left ventricular function are often used for stratification as the clinical impact of these variables are considered significant. In primary prevention trials, age or sex may be often needed for stratification because the prevalence of cardiac disease differs between genders and, moreover, affects outcome considerably. An example is the recent ASCOT Study that showed greater benefit of lipid lowering in men with atorvastatin and only a third of such benefit in women.[10] This may have been influenced by the small proportion of women in the ASCOT-LLA trial population (~20%) and the few events overall (17 and 19, placebo versus atorvastatin). Interestingly, a subsequent meta-analysis of lipid-lowering trials that included women appeared to support this.[11] These data would need to be confirmed in a truly comparative prospective study. These results are in contrast to the CARE Study results where the benefit in women was greater than men, although the trial included only 13.8% women (see under statins). Traditionally, in most clinical trials, the proportion of female participants has been low to modest and very few indeed hold the ratio of 1:1 as regards gender. This does have an impact on the trial results as seen above and begs the question '*Is there a rationale for insisting that in a cardiovascular trial, 50% of participants should be women?*' Such a question would be extremely difficult to answer; insistence on this distribution may influence the outcome because, for many cardiovascular events, male gender in itself is a risk factor (especially over 55 years of age). Alternatively, we should consider if we could extend observations in a trial of predominantly male participants to women as regards benefit and risk. Opinions differ and the debate rages on. This certainly is not an issue of parity between sexes but an important determinant of whether to adopt a gender-based stratification scheme in clinical trials.

Age has a similar influence in cardiovascular diseases and, by corollary, the clinical trials. Incidence and prevalence of left ventricular dysfunction and heart failure increase with age,[12] especially above age 70. Most heart failure trials include a large proportion of subjects with ages lower than these, and frequently those over 75 years are only represented in a minority. Whether this influences the conclusions drawn and whether we can extrapolate results from the lower age groups to the elderly should be considered carefully. Another example is the WOSCOPS Trial that specified age an as inclusion criteria (45–64 years; see under statins). Age itself might be an important risk factor for clinical events as discussed previously (see the HOPE Study). Stratification by age will help answer this conundrum, at least partially.

Prior use of therapeutic classes of drugs such as antihypertensives could also serve as a stratification variable. Care should be taken to ensure that such stratification does not compound confusion while reporting endpoints. For example, in the ALLHAT trial, the results may have been affected by the fact that a number of participants were receiving diuretics pre-randomisation. It is claimed that diuretic withdrawal at randomisation may have increased reporting of heart failure in the non-diuretic arms. So, if pre-randomisation diuretic use was employed as a stratification variable, with the inclusion of an arm that continued diuretic treatment during the trial, the results might have been projected or seen differently.

PLACEBO OR ACTIVE CONTROL

In recent years, with very many interventions being made available for the same condition and the complexity of the treatments needed, the question of whether the comparator should be placebo or an available active treatment has become a point of debate. Inclusion of a placebo arm in any trial has the advantage of limiting the 'placebo effect' that on occasion may confound the comparison of benefits obtained. This should therefore be considered carefully prior to finalisation of the design. The placebo effect, however, has not been a major issue in trials on cardiovascular mortality and morbidity and, inclusion of a standard treatment arm should limit such an impact further. The debate on whether placebo use is ethical in clinical trials has also been raised, which is beyond the scope of this chapter. For certain types of trials where active treatments are available, inclusion of a placebo arm would be inappropriate when lack of treatment could expose participants to unacceptable risks. One particular scenario is in situations where the standard treatment has not been firmly established, so a comparison could lead to the new agent being considered better than the existing agent. This is particularly so where the existing active 'comparator' was only marginally better than placebo in efficacy, but may have greater adverse events. On the other hand, if the existing active 'comparator' was the only, or most effective, treatment available, such a comparison may be unfavourable to the newer agent. This may leave the condition with no alternative treatment for those intolerant to the original agent. Inclusion of a placebo arm permits further analysis between groups and therefore has the advantage of *not* stifling innovation.

While placebo comparisons are an important step and provide valuable information, adequate attention needs to be paid in designing the trials to include the standard treatments adopted for the particular condition being investigated. The standard treatments may influence the outcome even when the placebo arm is included in the trial. As an example, in heart failure trials, it would be unimaginable to randomise participants into groups without the standard treatments such as ACE inhibitors, diuretics with or without betablockers, as the combination of these agents has greatly influenced outcome in such patients. This reduction of expected events in the so-called standard treatment arm or control arm increases the sample size required to find a difference between 'new' and 'standard or control' treatments. This has led in recent times to large complex trials such as Val-Heft (Valsartan in heart failure)[13], Valiant (valsartan in POST MI) trial[14] and RALES trial etc.,[15] where ACE inhibitors and diuretics were included as standard treatment in all the arms. Similarly, the use of statins in coronary heart disease may have to be considered standard treatment for future trials.

It is also important to consider the impact of current treatments (standard clinical practise) on public health and its perception on the endpoints chosen. For example, the public health impact of cholesterol reduction using statins has been significant, even if mainly relevant to secondary prevention. Whilst this may be a very happy state from the public health and patient standpoint, it means that clinical trials must be designed with the expectation that event rates may be considerably less than before and that the benefit will have to be over and above the current standard treatments and not merely placebo. The trials therefore either need to be much larger or may need to include a combination of endpoints such as death or other clinical events. Thus, large sample size running even to several thousand subjects is a common feature of cardiovascular trials and this is likely to continue.

PHASE I AND PHASE II TRIALS

IMPORTANT CONSIDERATIONS IN PHASE I OR PHASE II TRIALS

Phase I and II trials in cardiology are limited primarily to demonstration of effect on short-term endpoints such as physiological effects, pharmacodynamic effects in man and some pharmacokinetic effects. Examples include blood pressure reduction, lipid or cholesterol reduction, anticoagulant or antithrombotic effect, and inotropic effect in heart failure or antagonism of renin–angiotensin system. Whilst these surrogate endpoints have thus far served the pharma industry well as regards development of pharmacological agents, it should be noted that more agents are now expected to pass the main hurdle of clinical event reductions, and these are likely to be addressed only in Phase III trials.

The investigational agents could belong to several different classes (RAS inhibitors, betablockers, catecholamine inotropes, etc.) and all these could be grouped as neurohumoral response modulators or ‘biological response modifiers’. The second term is usually reserved for agents that are chemotherapeutic agents but the distinction is becoming hazy. They may also act by specific enzyme inhibition such as statins (HMG Co-A reductase) or pure simple agonists such as dopamine or dobutamine. Phase I trials in conjunction with preclinical data should help categorise the agent pharmacodynamically.

Phase I trials in cardiovascular medicine are likely to be specific and aim to determine the demonstration of physiological basis of pharmacological actions and to determine the dose response that will enable further studies. These trials could also serve to elicit any specific limitations of these investigational agents.

AIMS OF PHASE I AND II TRIALS

Issues for Phase I or Phase II trials include:

1. Identification of the target organ within the scope of cardiovascular medicine and defining the mechanism of action. For example, Inotropes should aim to improve myocardial contractility or function, while antiarrhythmics should suppress automaticity, conduction velocity and alter action potential duration.
2. Limiting the scope of action of the agent to the target organ only; the majority of agents used in cardiovascular medicine are likely to have a wide distribution and wide-ranging effects. Most vasodilators are effective on systemic vasculature and on pulmonary vasculature. This could be an advantage or a disadvantage depending on the aim or the trial. As an example, in subjects/situations where left ventricular output is dependent on filling pressure, pulmonary vasodilatation is likely to be hazardous, while in right heart failure due predominantly to pulmonary vascular disease such as corpulmonale, this is likely to be of benefit. Similarly, a number of antiarrhythmics suppress conduction not only in cardiac

tissue but also act in other tissues through a number of mechanisms. Such effects may initiate adverse events that are entirely unacceptable. A classic example is amiodarone which not only suppressed cardiac arrhythmias, but affected several other systems such as peripheral nervous tissue inducing neuropathy. It also affected the hepatic function. Such effects may have been specific to amiodarone because of the cellular mechanism of action, i.e. altering the Cy AMP and Cy GMP ratios within the conducting tissue and elsewhere (neurons, myelin and hepatocytes).

3. Assessing the boundary line for risk–benefit transitions in neurohumoral modulation; ACE inhibitors and angiotensin receptor blockers, whilst producing peripheral vasodilatation that is expected, also reduce renal blood flow and may induce renal failure.

The Phase I and II trials should provide sufficient data from such scenarios in order that Phase III trials aim to provide outcome data.

The Phase I and II trial designs should therefore take into account all the following factors;

1. Identification of appropriate participants to include in a clinical trial – appropriate subjects need to be included to obtain the desired benefit.
2. Phase I and II trials play a major part in determining the suitability of drugs with narrow therapeutic index or those that are extremely relevant to emergency situations such as antiarrhythmics.
3. Similar situations are applicable to agents in the treatment of myocardial infarction (MI) or percutaneous intervention.
4. Often, comparative efficacy and safety with other agents may be necessary before Phase III trials can be considered.
5. Last but not the least, it is expected that Phase I and II trials of all cardiovascular agents specifically investigate the effect of the pharmacologic agent on QT interval, QT dynamics and potential interaction with other agents that alter QT interval. If these data are not available, a very sound justification would be expected.

These points are not exhaustive but should act as a guide to Phase I and II clinical trials in the cardiovascular field. The points relevant to Phase III trials are discussed in the subsequent sections.

We shall now consider the issues in specific interventional modalities and specific trials. The modalities of importance are ‘pharmacological interventions/agents’, ‘devices and surgical procedures’, life-style changes (behavioural change) and lastly ‘nurse (health worker) led interventions’. The chapter also considers specific trials in short, as detailed analysis of each trial is beyond the scope of this chapter.

PHARMACOLOGICAL AGENTS/INTERVENTIONS

Most trials involving drugs in cardiovascular (CV) diseases are similar to other fields of medicine. They could be preventive (primary, e.g. lipid lowering), secondary prevention (lipid lowering post-MI), secondary prevention of heart failure after MI, etc. There are few primordial prevention trials thus far. A number of significant points that may be specific to cardiovascular disease should, however, be noted. The myriad manifestations of CV disease is preceded by a long period of development that has virtually no symptoms and may extend to decades. Conditions such as atherosclerosis, a risk factor for CV disease, is influenced or even accelerated by the presence of other ‘risk factors’ such as hypertension, hyperlipidaemia and diabetes. This may also be influenced by family history, advanced age and the sophisticated imaging facilities available.

As the diseases take a number of years to develop, preventive strategies are especially applicable and these are often primary in nature. Secondary prevention in relatively asymptomatic individuals is also an important consideration, such as in diabetics. All caveats regarding *primary prevention* trials therefore apply to these

CV trials specifically; serious or troublesome adverse events would be unacceptable (note relatively healthy population), drugs/agents need to be extremely well characterised, the rate of clinical events may be low even if surrogate endpoints are used, definition of appropriate surrogate endpoints could prove difficult, and lastly, compliance may be an issue in healthy individuals who do not see obvious benefit immediately. The *sample size* is likely to be enormous (thousands of subjects) because of the above factors. Preventive strategies do have a role even in established diseases such as heart failure; sequelae such as premature death may be prevented by use of agents (betablockers, antiarrhythmics) or devices, either alone or in combination. Such combination therapy is also influenced by additional risk factors and co-morbid conditions.

LIPID-LOWERING TRIALS

Pre-Statin Era

The story of cholesterol reduction over the decades highlights a number of points raised in this chapter relevant to clinical trials and hence these are discussed in some detail. The importance of lipid lowering was identified nearly 40 years ago in those with coronary heart disease, and this was investigated by two landmark projects, 'The Coronary Drug Project' (CDP)[16,17] and 'the WHO Clofibrate trial'.[2,3] The CDP began in the 1960s and tested five interventions (clofibrate, nicotinic acid, dextrothyroxine, and two doses of oestrogen) in men with a history of myocardial infarction.[16,17] There were major adverse events and only modest benefits compelling early termination of the trial. The degree of lipid lowering was relatively small ($\sim 9\%$ reduction in LDL or total cholesterol) and hence the lack of benefit in mortality was perhaps not surprising in hindsight. Nicotinic acid, however, showed a reduction in non-fatal reinfarction at the end of the trial and subsequent longer follow-up showed a reduction in mortality. The WHO clofibrate trial, however, yielded different results. Indeed the reduction in cholesterol was modest with clofibrate (10%) but there was no significant reduction in CV death or cause-related mortality. Indeed overall mortality was higher in the treated group (nearly 47%) during the treatment phase (five years)[2] but post-treatment, this difference in mortality abated (eight year follow-up).[3] This result has confounded the scientific community in several ways: Was this result specific to clofibrate or a class effect of all PPARα agonists? Is cholesterol reduction therefore a valid surrogate endpoint? Or is there a numerical level for cholesterol reduction that needs to be achieved before benefit in mortality is observed? Some of these issues may never be addressed until another PPARα agonist is tested in a large trial, and in this trial the control arm will need to include statins because of their established status in mortality reduction.

There were a few other important, additional key findings of the above trials. One was the lower mortality rate than expected (pre-trial projections) in the control group in the Coronary Drug Project. It also became obvious that the length of follow-up needed to observe mortality benefit was considerable (during and after treatment) as shown by the nicotinic acid arm. The WHO clofibrate trial[2,3] also provided important epidemiological information; the untreated control group (lower third of cholesterol distribution) had the lowest mortality of all groups, proving in a large population that 'the lower the cholesterol, the better the outcome'. The higher all-cause mortality during the trial phase in the treated group was neutralised during the follow-up primarily due to an increase in the ischaemic heart disease (IHD) mortality in the first control group (upper third of cholesterol distribution), supporting the above statement. This would have been confirmatory had an increase in cholesterol been demonstrated during follow-up in the WHO clofibrate trial. It is interesting to note that other fibrates have been tested for secondary prevention in much smaller trials (gemfibrozil in the Helsinki Heart Study 1987 and bezafibrate in the BIP trial 1997). Neither of these has shown clear mortality benefit emphasising the importance of sample size calculations, even for secondary prevention trials.

Other methods (or agents) for cholesterol reduction have been through the trial phase but yielded variable results. Comparison of Cholestyramine in the lipid research Clinics Coronary Prevention trial to placebo showed some significance (<0.05, one-sided, 155 events in the intervention arm vs. 187 in the placebo arm) for primary prevention in 3806 men without prior MI but with hyperlipoproteinaemia. The significance level has been questioned, primarily as this was a one-sided test. The arrival of statins made the use of Cholestyramine powder rather obsolete.

Statins

The appearance of statins (HMG Co-A reductase inhibitors) heralded a new era in lipid lowering for two reasons: the reduction in cholesterol levels far exceeded any other previous experience and use of these agents was associated with definitive benefit in clinical outcome. Whilst the initial trials with statins only demonstrated cholesterol reduction in the late 1980s, the definitive outcome trials were all conducted in the 1990s. These trials provide a clue to the issues faced by most trialists for design, for sample size, problems of recruitment and of developing public interest in clinical trials.

SECONDARY PREVENTION TRIALS

The 4S trial (Scandinavian Simvastatin Survival Study; 4S trial study group, 1994) and CARE (Cholesterol And Recurrent Events) were the two initial studies that highlighted the potential benefit of use of statins in those with coronary heart disease (4S) or prior coronary events (CARE). The baseline cholesterol levels differed in these two studies, being lower in CARE than 4S (5.5 to 8.0 mmol/L).

4S Study

Simvastatin was compared against placebo in 4444 participants with known CHD and elevated cholesterol. Over 5.4 years of follow-up, simvastatin reduced LDL cholesterol by 35% and relative mortality by 30%, 12% (256) deaths in the placebo arm and 8% (182) in the simvastatin arm with a relative risk of death in the simvastatin group of 0.70 (95% CI, 0.58 to 0.85). This impressive reduction in mortality was associated with 33% reduction for major coronary events and 37% reduction in risk of revascularisation procedures. All subjects were on a low-fat diet and thus the influence of diet balanced.

CARE Study

Unlike the 4S study, CARE included 4159 participants (3583 men and 576 women) with average cholesterol but after a recent MI.[18] The effect of 40 mg pravastatin was compared against placebo over a four-year follow-up period. There was a 24% relative risk reduction of coronary events with pravastatin (3% absolute risk reduction), a 25% reduction in bypass grafting (7.5 vs. 10%) and 31% reduction in the occurrence of strokes. Interestingly, no difference in overall mortality or mortality from non-cardiovascular causes was observed. Not surprisingly, those with higher pre-treatment cholesterol levels derived greater benefit. An important paradox of this study (in comparison with other similar studies) was that coronary event risk reduction was greater in women than men, albeit only 13.8% of the participants were women.

These studies provided the impetus for primary prevention trials, some of which were already in flow, such as WOSCOPS.

OTHER SECONDARY PREVENTION TRIALS

LIPID Study

In this large mortality study, 9014 subjects with CHD and variable cholesterol levels (155 mg to 270 mg) received pravastatin or placebo and were followed up for six years. The primary endpoint was mortality from CHD.[19] There was a 24% relative risk reduction (2.1% absolute risk reduction) in the pravastatin group overall for the primary endpoint. Overall mortality reduced by 3% (absolute reduction) with a relative risk reduction of 22%. All other CV endpoints (MI

(29% reduction), death from CHD or non-fatal MI (24% reduction), stroke (19% reduction in risk) and coronary revascularization (20% reduction)) showed benefits with 40 mg pravastatin.

IMPLICATIONS OF STATINS IN SECONDARY PREVENTION

Several statins have shown similar clinical event reductions although the magnitude of the effect on overall mortality varied. Whilst the absolute risk reduction was modest and relative risk reduction impressive, the trials had large sample sizes, thereby increasing the impact in terms of both public health and clinical practice. From a clinical trial standpoint, the following are some of the lessons learnt from statin trials. Most statins have shown benefit in CV clinical event reduction. However, it is a far cry from being heralded as a 'class effect'. These trials demonstrated the importance of appropriate surrogate endpoints in a limited fashion as other modalities of lipid reduction did not show the same effect. From a regulatory perspective, statins were authorised based on their ability to reduce cholesterol and not based on long-term mortality. The data on mortality became available only subsequently. Other issues highlighted from the experience with statins are: a long duration is needed for event reduction; large populations and selected high-risk groups are needed often to demonstrate a clear benefit that there is a possible relation between level of LDL cholesterol and event reduction; and lastly, early evidence for the concept of 'the lower, the better' in terms of plasma lipids.

PRIMARY PREVENTION TRIALS

The impetus for primary prevention arose from the lessons learnt during the secondary prevention trials with statins. Naturally the population under investigation needed to be extended from those with CHD to those without and consideration of other risk factors such as diabetes assumed importance. The most important development appears to be the lowering of LDL cholesterol level as an inclusion criterion in the trials.

Woscops

The WOSCOPS (West of Scotland Coronary Prevention Study)[20] was the first primary prevention trial in 6595 men with hypercholesterolaemia (7.0 ± 0.6 mmol/L) but without other evidence of CV disease. Over an average follow-up period of 4.9 years, LDL cholesterol reduced by 26% with a relative risk reduction of 31% ($n = 171$ vs. 248) for coronary events (non-fatal MI or death) with the use of pravastatin. Death from all cardiac causes reduced by 32% (relative risk) and, interestingly, death from any cause reduced by 22%. Despite the impressive results, there are certain limitations of this study that should be kept in mind: it included only males, excluded a proportion of the high-risk population (>65 years of age) by limiting inclusion (45–64 years) and the dietary habits of these participants were legendary for their high-fat content. Other confounding factors such as hypertension or diabetes were not analysed in detail.

The MRC/BHF Heart Protection Study[21,22] This large randomised study of simvastatin investigated the benefit offered by statins to those with or without CHD, but were otherwise at high risk of coronary events. It included participants with risk factors such as coronary disease, other occlusive arterial disease, or diabetes, and found 40 mg simvastatin beneficial regardless of the baseline cholesterol levels. The average reduction of LDL cholesterol was 1.0 mmol/L with significant reductions in coronary deaths (18% risk reduction), vascular deaths, non-fatal infarctions or revascularisation procedures. The overall mortality was significantly reduced (14.7 vs. 12.9%)[21]. Interestingly, a subgroup analysis of 5963 diabetics (40–80 years) demonstrated that diabetics had highly significant reduction in the first occurrence of coronary events (~22% in comparison with placebo), and this was irrespective of the baseline arterial disease or LDL cholesterol level.[22] This specific issue of diabetics deriving greater benefit from such preventive measures, even in the absence of manifest CV disease, is extremely important in terms of public health and is anticipated to occupy the scientific headlines for some considerable time to come.

These results, viewed along with those of the UKPDS studies, which evaluated aspects of diabetes control on vascular complications, provide major insight into impact on preventive or treatment measures on public health.

TEXCAPS/AFFCAPS Studies

TEXCAPS is the first primary prevention trial in CV disease that included women (15% of 6605; 997 post-menopausal women and 5608 men) without evidence of CVD and average LDL cholesterol levels.[23] However, it had an additional inclusion criterion of low HDL. Lovastatin of 20–40 mg reduced the primary endpoint of acute major coronary event (including fatal or non-fatal MI) by 37%, after an average follow-up of 5.2 years. The genders did not differ in prespecified analyses of the endpoint (ACMEs). Specifically, LDL cholesterol reduced by 25% in women but did not detect a treatment difference between groups as events were infrequent and therefore the power of the study was insufficient (7 of 499 vs. 13 of 498 for lovastatin or placebo respectively). In contrast to some of the other statin trials, this did not show benefit in mortality (all causes) with lovastatin treatment.[23,24] One of the objectives behind this study was to mimic clinical situations and it succeeded to the extent that 8 million Americans fitted the inclusion–exclusion criteria used in this study.

Several other trials using statins (both primary and secondary) have now been completed and reported with variable results. Some examples are the ASCOT-LLA trial[10] and the TNT trial, both with atorvastatin, rosuvastatin trials and ALLHAT-LLA trials.[25,26] In some of these an effort to replicate normal clinical practise was made in the trial design. This has lead to surprising outcomes and more questions on designs of the trials.

TRIALS OF ANTIHYPERTENSIVE AGENTS

The antihypertensive and lipid-lowering treatment to prevent heart attack trial, or ALLHAT, compared in a blinded fashion three different antihypertensive agents against control in more than 40 000 people aged 55 years or over with hypertension and at least one other risk factor.[25,26] Fatal CHD or non-fatal MI was the primary outcome. Interestingly, ALLHAT also included a lipid-lowering arm (pravastatin) that recruited ~10 000 subjects. The primary endpoint for this open, non-blinded part of the study was 'all-cause mortality'.

The trial did not show any differences in the primary outcome for either the hypertension part or the lipid-lowering part for three of the antihypertensives. The doxazosin arm of the hypertension trial was stopped early, as the incidence of heart failure, a secondary endpoint, was significantly higher with this alphablocker. The diuretic, chlorthalidone, proved to be superior to other antihypertensives (amlodipine or ACE inhibitor, Lisinopril) for the endpoint of heart failure. Whether the fact that 'a large number of participants were receiving the diuretic pre-trial, which was withdrawn around randomisation, contributed to these findings is debatable. Thus, ALLHAT emphasised several important caveats in relation to trial conduct, choice of comparators and sample size. These were: for hypertension, a large sample size is required to achieve adequate power even with combination endpoints; withdrawal rates may contribute to lack of power (30% during FU in ALLHAT); for participants already on an established antihypertensive pre-trial, the withdrawal effect of this agent needs to accounted for, and such an event will influence the choice of the comparators. For the lipid component of the trial,[25,26] several reasons have been ascribed for lack of effect/difference with pravastatin which had proved effective in several previous trials: there was only a modest difference in LDL cholesterol between groups; nearly 30% of the control group were receiving non-trial statins by the end of the trial; the non-blinded design further complicated this non-trial statin use; and the public campaigns to reduce cholesterol. Hypertension was a requirement in all participants. Participants thought to require definite statin therapy were excluded from the trial. *Such factors are likely to play major roles in most future* *clinical trials as the design becomes more complex in order to achieve the primary endpoint.*

Stepped Care Approach

This approach, often used in clinical practice, was also applied in some clinical trials and is predominantly useful in assessing antihypertensive agents. When a single agent is either insufficient or not well tolerated by the participant, use of second-, third- or fourth-choice drugs may be built into the protocol. This could also be achieved by incremental dosing of the same agent. However, this is more related to assessment of a strategy or approach rather than trial-assessing risk–benefit of a single agent. The Systolic Hypertension in the Elderly Program used this approach in a multicentre, randomised, double-blind, community-based trial. The trial started with 12.5 mg Chlorthalidone or placebo, stepped up to 25 mg or placebo after two months. If poor response was noted, 25 and subsequently 50 mg of atenolol were added and, as a last step, 0.05 to 0.1 mg of reserpine was used. Use of reserpine is no longer attractive even in those with resistant isolated systolic hypertension because of the side effects.

The 'non-responder trial' design is utilised frequently for combination therapies and could be considered a modified version of the stepped care approach, but particularly useful for investigation of combination therapy with specific agents. Those subjects not responding to monotherapy with predefined criteria of response are then randomised in the second phase of the trial to either addition of the second agent or placebo whilst the first drug is continued. It is anticipated that the combination produces a greater reduction in blood pressure than the *single agent + placebo*. Variations of this theme are numerous with an increased dose of the first agent forming the control arm instead of the placebo. Such an approach became popular in treating hypertension primarily, or hyperlipidaemia (statins plus other agents such as fibrates) as the tendency to use combination therapy earlier in the course of the disease became the vogue.

Many newer antihypertensives or combination therapies are emerging continuously. Most trials involving these rarely include mortality as a primary endpoint, as reduction in systolic or diastolic blood pressure at the end of a blinded period (usually 12 weeks) is considered an adequate surrogate endpoint for this condition. The scientific basis for this has been the epidemiological finding that CVD (MI) and stroke incidence are related to level of BP. Moreover, guidance documents such as the Joint National Committee reports and NANHC reports have provided claims based on meta-analyses that every 1 mm reduction in BP leads to a relevant reduction in risk of cerebrovascular accident (CVA) or MI. Prior to designing the trial, whether these guidelines and the meta-analyses are relevant in the current context of many facets of CVD/CVA prevention should be considered carefully. There are only a few trials purely assessing mortality/morbidity in hypertensive populations, primarily because of the influence of co-existing risk factors such as diabetes or hyperlipidaemia. Moreover, risk factors for CVD are multifactorial. The ASCOT trial comparing the amlodipine/perindopril combination to atenolol/bendrofluazide is perhaps one of the first few trials to include CV clinical events as endpoints (non-fatal MI or fatal CHD) in treatment of hypertension after ALLHAT. However, as the trial was stopped 12 months early, the primary endpoint did not reach statistical significance.

Role of Guidelines in Designing Trials

It is also important to note that guidelines on treatment of hypertension have undergone a paradigm shift: out with the old and in with the new. Combination therapies are recommended by most guidelines as an appropriate step in the course of management of hypertension at an earlier stage than some years ago. This could be attributed to changing definitions of 'control of hypertension', the acceptable level of systolic BP being 140 mmHg rather than 160 mmHg, and the additive benefit derived from combinations such as ACE inhibitors and diuretics. These facts are likely to govern the choice and definition of 'standard treatment' for the control arm in most forthcoming trials.

As discussed earlier, antihypertensive agents and lipid-lowering drugs have generally been approved by regulatory authorities/agencies on the basis of their effects on BP and cholesterol rather than effects demonstrated in trials on clinical outcomes. However, their effects on clinical outcomes have been tested on death, MI and stroke, leading to subsequent modification of the indication for these agents (e.g. see 4S, CARE, HPS Studies).

TRIALS AFTER MI

Trials in patients with MI or for prevention of MI present a different dimension for either primary or secondary prevention. Rates of death in people who had MI used to be very high. Modern therapy (advent of thrombolysis, use of aspirin and betablockers) has reduced this considerably and remarkably. Therefore, trial with mortality alone as an endpoint may no longer be feasible even in survivors of MI, without the identification of a very high-risk group. This has led to the use of combination endpoints that include CV mortality, non-fatal MI or even stroke. From a preventive aspect, even in those with established CHD or with multiple risk factors for such disease, combination endpoints have been used frequently in recent years and there are many examples of such studies. The Heart Outcomes Prevention Evaluation (HOPE)[27] Study compared ACE inhibitor ramipril against placebo in 9297 high-risk subjects with known vascular disease or diabetes plus another risk factor. The results of this study with a composite endpoint of MI, stroke or death from CV causes *are of landmark significance*. Even in this high-risk group, a sizeable sample was needed to achieve adequate power and the results showed a highly significant and impressive reduction in the primary endpoint.

The PEACE (Prevention of Events with Angiotensin Converting Enzyme inhibitor therapy)[28] Study emphasised the need for an appropriate choice of endpoint and study population by achieving neutral results in comparison with the HOPE trial. PEACE compared an ACE inhibitor, trandolapril, against standard treatment in 8290 subjects with documented CHD and normal left ventricular function (ejection fraction of > 40%). ACE inhibitors have been shown to benefit those with heart failure or reduced ejection fraction (SOLVD[29,30] and SAVE[31] trials, amongst others) and hence the limit of 40% for ejection fraction in the PEACE trial. This placed the participants in the PEACE trial in the lower risk group. The original sample size was 14 000, but needed to be reduced to 8200, and the primary outcome expanded to include coronary revascularisation. Over a mean period of 4.8 years, the incidence of the primary endpoint (CV death, MI or revascularisation) was 22.5% in the standard therapy group and 21.9% in the ACE inhibitor group, leading the authors of the report to conclude that ACE inhibitors offer no additional benefit in those with stable CHD and preserved ventricular function. Understandably, the event rate in the control group was lower than in those with impaired ventricular function and despite the combination endpoint, the study did not show benefit with ACE inhibitors in this group. Inclusion of 'need for revascularisation' in the endpoint is often debatable and may be subject to considerable bias if the trial is not blinded.

COUNTER-INTUITIVE OR 'SURPRISE' RESULTS

Clinical trials often show effects that appear counter-intuitive and some classic examples are discussed below. Improvement of left ventricular function and outcome in those with either acute MI or heart failure of other aetiology has been attempted with various agents including ACE inhibitors,[32] combination of vasodilators (hydralazine + nitrates),[33] or inotropic agents.[8,34,35] Several interesting phenomena were noted over the years in a number of clinical trials involving the above agents. Inotropic agents, whether catecholamines (e.g. dobutamine or dopamine in shocked patients, or oral ibopamine in heart failure)[34,35] or other classes such as phophodiesterase inhibitors (e.g.

milrinone or amrinone)[7,8] uniformly worsened mortality in contrast to their perceived effect of improving ejection fraction and the anticipated long-term benefit. On the contrary, vasodilators such as ACE inhibitors or angiotensin receptor blockers have shown consistent benefit in those with heart failure in improving mortality and morbidity without demonstrable short-term benefit on ejection fraction. A natural corollary was to assume that other vasodilators such as hydralazine and nitrates would provide similar benefit, but this proved inconsistent in V-Heft trials.[33] The above assumptions proved further away from the truth, when it was noted that there were racial differences in response or benefit. The African/Caribbean populations (black races) appear to derive less benefit with ACE inhibitor therapy than the Caucasian populations based on subgroup analysis.[36] Interestingly, the same subpopulations (blacks) derived greater benefit with Hydralazine + nitrate combination in V-Heft trials in contrast to Caucasians. Both these observations were, however, limited by the small size of the subgroups, especially the black population, and these were *post hoc* observations. A plausible but not wholly satisfactory explanation that these subgroups were low renin populations and hence derive less benefit either in BP reduction or heart failure control with ACE inhibitors still exists. This has not been conclusively proven. The issue of sample size was resolved by the most recent African American heart failure trial that has shown significant benefit with the use of Hydralazine + nitrate combination in comparison with placebo.[37] This trial specifically in 1050 blacks with NYHA class III or IV heart failure showed clear benefit with reduction in mortality in comparison with placebo (with standard treatment). The primary endpoint was a composite score of weighted values for death from any cause, a first hospitalisation for heart failure and a change in quality of life. This study was terminated early because of the significantly higher mortality in those receiving placebo (10.2% vs. 6.2%). Several explanations and several questions arise from these observations that are beyond the scope of this chapter, but suffice it to say that clinical trialists would do well to bear in mind that potential for racial differences in response with the same agent does exist. The other important lesson from the inotrope trials in heart failure is that short-term benefit does not necessarily reflect long-term gain in survival and this may apply to a number of surrogate markers as discussed in earlier parts of this chapter.

ANTIARRHYTHMIC AGENT TRIALS

After MI or in those with CHD, arrhythmia, especially ventricular arrhythmia, is a major cause of mortality and use of antiarrhythmic agents in such subjects is an important aspect of management. Cardiac ventricular arrhythmias with or without coronary artery disease are known to correlate with death or sudden cardiac death. A number of agents were used to control these arrhythmias and these agents were expected to improve mortality or sudden death in all patients with ventricular arrhythmias. This assumption stood up to logic until trials of antiarrhythmic agents showed that suppression of arrhythmia was not a surrogate for reduction in mortality. On the flip side, atrial arrhythmia, especially atrial fibrillation, was not expected to increase mortality, although it contributed to morbidity. Whilst there is some evidence that onset of atrial fibrillation may influence mortality after MI or in those with heart failure, recent trials have shown that maintenance of sinus rhythm in preference to rate control of atrial fibrillation offers no advantage in terms of clinical outcomes.[38,39]

The Cardiac Arrhythmia Suppression Trial (CAST)[29,40,41] compared three antiarrhythmic drugs to test the hypothesis that suppression of ventricular arrhythmias reduced sudden cardiac death. In a phased trial, 1700 patients whose arrhythmias were suppressed by encainide, flecainide or moricizine were randomly assigned to the drug that was most effective or a matching placebo. Early in the trial (after 10 months), both flecainide and encainide were seen to increase mortality from either arrhythmia or nonfatal cardiac arrests. These two agents were discontinued with the investigators concluding that neither encainide nor flecainide should be used in the treatment of patients with asymptomatic or minimally symptomatic ventricular arrhythmia after MI, whilst the moricizine part of the trial (CAST-II) was continued. The CAST-II trial had approximately 1325 patients and in the short term (14-day post-MI phase) itself showed significant mortality with moricizine compared with placebo. The long-term part of the trial also showed similar results,[29,42] confirming the view that antiarrhythmics should not be used for pure suppression of ventricular arrhythmias, especially when asymptomatic. As all subjects were post-MI with ventricular dysfunction, whether ischaemia contributed to the increased mortality or was worsened by the agents used remains unknown.

Several similar examples are available and the significant messages appear to be: short-term control of arrhythmia does not necessarily imply long-term survival (as in inotropes and heart failure); the presence of asymptomatic arrhythmias does not imply poor long-term survival; the population included in the trial should be appropriately assessed for risk of events and this could be influenced by a number of other factors.

TRIALS ON TREATMENT STRATEGIES

Some trials compare treatment strategies rather than individual agents. One cardiac condition that lends itself to variation in treatment strategies is atrial fibrillation, where rate or rhythm control has always been a debate between the conservative and the aggressive physicians. The AFFIRM (Atrial Fibrillation Follow-up Investigation of Rhythm Management)[38] Study evaluated these two strategies in 4060 participants who were over 65 years (or less than 65 with high risk of stroke), and had atrial fibrillation. Investigators selected treatment from options on an approved list of the pharmacological and non-pharmacological therapies. Rhythm control was achieved by cardioversion or antiarrhythmic drugs. Digitalis, calcium-channel blockers, betablockers or A–V nodal ablation and pacemaker implantation were options for rate control (ventricular response). Interestingly, while the all-cause mortality was similar, the rate control group had fewer adverse events and a better mortality trend. Similar results were seen in the recently concluded UK-Pace study that compared single and dual chamber pacing modalities in the elderly.[43–45] Dual chamber pacing as a treatment strategy was also favoured in preference to continued medical treatment for heart failure in a multicentre, randomised study.[44] These are excellent examples of strategies that compared drugs or devices. This brings us to the interesting but complex issue of trials involving devices and other procedures.

TRIALS INVOLVING DEVICES OR SURGICAL PROCEDURES

Devices and surgical procedures are commonly used in cardiological patients. Devices are unique to these patients as cardiologists and not surgeons now implant many of the devices. The devices may include coronary stents, cardiac pacemakers and defibrillators, or even replacement heart valves. Examples of common surgical procedures are coronary bypass graft (CABG), aneurysm resection and correction of congenital abnormalities. The comparisons usually are of the surgical procedure against the use of medical treatment or a device (e.g. CABG vs. stent insertion in patients with stable angina). Comparisons of one surgical procedure against another are less frequent, whilst devices are compared frequently against other devices such as single chamber vs. dual chamber pacemakers. As with trials in pharmacological agents, the question raised in these clinical trials needs to be relevant and important. The studies should be appropriately designed and randomised and data properly analysed. A number of factors influence the outcome of these trials.

BLINDING AND POTENTIAL FOR BIAS

First, there is obvious scope for bias when comparing a surgical procedure against a device or medical therapy or when comparing a device against medical therapy as blinding may not be feasible and is often impractical. Occasionally blinding may be possible but may also involve a second implantation/extraction procedure to upgrade the device. In order to avoid unblinding or bias, a crossover design is often employed. The difficulties with such trials/designs are highlighted by the two trials discussed below. The MOST (Mode Selection Trial in Sinus-Node Dysfunction[45] trial showed that blinding might be possible when implanted devices are similar but the functional modality is alterable. The Canadian trial of Physiologic Pacing (CPP)[46] showed that if it involves reimplantation or upgrade of an existing device, the rate of crossover might differ significantly (31% vs. 2.7% respectively between the two trials, MOST and CPP). The final conclusions of the two trials comparing similar devices are therefore likely to be different!

OPERATOR-RELATED ISSUES

The second issue with devices concerns operator competence or skill. Unless the investigator has considerable surgical experience or skill in implanting a device, an intervention might be claimed to be not beneficial or even harmful. This was highlighted by the 'Department of Veterans Affairs trial' comparing surgical and medical management of angina pectoris.[47] Of the 13 hospitals participating in this trial, 3 had surgical mortality considerably greater than the other 10. The results comparing surgery against medical therapy were in favour of surgery among the high-risk group even when all 13 were included in the analysis. Invariably, for lower risk groups, the results from the 10 best centres showed benefit from surgery. Important caveats in this trial are that the medical therapy was pre-thrombolysis or other advances in anti-anginal therapy. Also, the results for high- and low-risk patients between centres may reflect normal variation but the issue of skill and competence of the operator and experience are raised. This issue was re-emphasised in the recent AVID trial (Antiarrhythmics Versus Implantable Defibrillators trial) that specified the experience required of investigators and the establishment of minimum standards for device and lead systems.[48,49] Whilst this approach does not guarantee that only highly skilled operators will be involved, it permits the trial to be a better test of performance of the devices. Additional questions that would arise from involvement of highly skilled investigators are 'How broadly could the results be generalised?' and 'could the skills (surgical technique or device implantation ability) be transferred to other practitioners?' such that similar results are achieved in the general population. These are not straightforward and cannot be easily addressed. The implication of this is, if in a trial a device is found to be beneficial, would this benefit be achievable when the device is implanted by less well-trained operators? They may not be(!) and this has to be considered seriously. Another issue with devices or surgical trials is that modifications or changes in technique could affect the results or cause difficulties in interpreting them unless care is taken to ensure that the groups remain comparable. The AVID[49] and CIDS (Canadian Implantable Defibrillator Study)[50] trials both compared defibrillators against medical therapy, and were enrolling patients when thoracotomy for defibrillator implantation was replaced by transvenous implantation. In these studies, as both types of defibrillators performed similarly and were shown to be significantly better than antiarrhythmic drugs in reducing mortality, combining the results for the two types of defibrillators was not an issue. Such changes in technique could have a bearing on the skill of the operators and hence influence the results.

DURATION OF TRIALS

These trials with devices or surgical techniques are also impacted by early adverse. In these invasive trials, an early adverse experience associated with the procedure is likely and expected and these include: the trauma, the consequences of anaesthesia, risk of infection, and finally the effect on blinded randomisation of the treatment arms. In order to elicit true benefit (especially long term), the length of the trial should be sufficient to overcome this early unfavourable morbidity and possibly mortality. In such a situation, considerations regarding early termination of the study will need to be made. Sometimes, this expected benefit might be seen only after a considerable length of time, even several years. The Program on the Surgical Control of Hyperlipidemia (POSCH) trial that compared ileal bypass surgery against medical therapy in patients with prior MI is a prime example.[51,52] The aim was to decrease lipid absorption and thus serum cholesterol. The trial failed to show a difference in the primary outcome and all-cause mortality in the first two years, with the surgical group faring slightly worse than the control group.[51] Three years after the surgery, the curves crossed and, by the scheduled end of the trial period, a non-significant trend in favour of surgery was noted. Five years after the formal end of the trial the mortality difference was statistically significant in favour of surgery.[52] Adverse events (diarrhoea and increased rate of kidney and gall stones in POSCH) combined with a lack of benefit might have led to a decision to stop the study prematurely and the long-term benefit may not have been seen at all.

OTHER ISSUES INCLUDING RISK LEVEL OF PARTICIPANTS

Several other aspects specific to devices are unusual and particular to CV trials that are uncommon in other kinds of trials. These are: engineering issues of functionality of the device, the level of risk of the participants enrolled in the studies, and the ethical aspects of withholding an effective treatment in the medical therapy group. The engineering issues regarding the device would have to be clearly addressed before the clinical trial. The clinical trial should be designed to answer specifically questions posed by the clinicians. As to the risk level of the participants, trials have looked at various ways of identifying patients at sufficiently high risk to determine if the devices are beneficial and this specifically applies to defibrillators. However, the risk should not be unacceptably high that it would become unethical to randomise participants. MADIT-I and MADIT-II provide particular examples of the above strategy. MADIT-I involved patients with post-MI heart failure and non-suppressible ventricular tachycardia on electrophysiological testing, recruiting 196 patients. There were highly significant reductions in mortality (all-cause and cardiac) with defibrillator use in this very high-risk group.[53] The subsequent MADIT-II assessed use of defibrillators in patients with prior MI and ejection fraction less than or equal to 30%, but electrophysiological testing was not used as an inclusion criterion.[54] Whilst the defibrillators proved superior to medical therapy even here, the study randomised 742 patients, as risk of mortality was expected to be lower in this group. While the two trials came to similar conclusions (that defibrillators were superior to antiarrhythmic drugs), the risk level of the participant groups was different and hence the required sample size differed.

ETHICAL CONSIDERATIONS

Ethical aspects and considerations may be very important in trials involving devices or surgical techniques. For example, the REMATCH trial in patients with congestive heart failure raised a number of ethical issues from both a randomisation and alternative therapy aspects. The study compared medical therapy versus left ventricular assist devices in 129 participants between 1997 and 2001. The differences in survival were highly significant (25% vs. 52% in favour of assist devices at the end of one year; 8% vs. 23% at the end of two years). There were, however, higher rates of infection, bleeding and device malfunction in the device group. As these assist devices were meant as a bridge to transplantation, their long-term use in this study raised many questions. The randomisation itself was considered unethical by many, as the survival rates in subjects with end-stage heart failure are so low.

END OF TRIAL ISSUES

Last but not the least, unlike most pharmaceutical agent trials that are withdrawn at the end of the trial, explantation of the device is a tricky issue. If the trial has shown benefit with the device, it might be easier to implant devices into the control group, but if the trial turns out not to show benefit, explantation of the device would seriously need to be considered and might even be obligatory. Such a situation arose in the CABG patch trial, which compared transthoracic implantation of defibrillators against control in patients undergoing CABG. At the end of the trial (32 months), the two groups did not differ and all the participants were given the trial results and therapy individualised. Nearly 40% of the patients in the intervention groups elected to have the devices explanted or turned off.[55] A similar situation is imaginable when a surgical technique that is not shown to be beneficial cannot be undone or reversed at the end of the trial.

TRIALS OF BEHAVIOURAL CHANGE

For a number of years, life-style modifications have been advocated and used commonly in cardiological practice. Unsurprisingly, trials of behavioural change have become fairly common. They take various forms: diet change for weight loss, cholesterol reduction or BP control; exercise to reduce cardiac risk; and better outcomes after heart attack or heart failure. These trials differ from most other trials in that there is likely to be considerable crossover by the participants; they are quite resource intensive; often the study duration is considerably shorter than other types of trials and, because of the limited duration, surrogate markers are more frequently used. Two other issues are also significant in that they may impact on the results of these trials: standardisation of intervention and measurement of degree of compliance are very complicated. Peer pressure and societal changes often influence these. For example, getting volunteers to adhere to an exercise programme consistently even for a few months let alone years is extremely difficult. This is complicated by volunteers interested in exercise unable to remain in the 'non-exercise arm' and therefore tend to cross over after randomisation. Similarly, in assessing a dietary intervention, if maintenance of caloric intake over a period of time is an objective of the trial, weight change will not serve as a marker of diet alteration for reasons detailed before. Another such situation arises in trials of smoking cessation. Regulations are likely to affect trials of smoking cessation if workplace or public place restrictions against smoking come into force during the trial as an example of societal changes or peer pressure.

Despite all the unusual aspects above, there have been some highly successful behavioural change trials. The DASH (Dietary Approaches to Stop Hypertension) trial[56] and the subsequent trial on sodium intake[57] were prime examples of dietary influences on hypertension. The DASH trial randomised 459 adults with hypertension (160 mmHg systolic and 95 mmHg diastolic) to one of three groups: a control diet (standard American diet), a combination diet (fruit + veg + low-fat), and lastly, a diet rich in fruits and vegetables. Both interventions (the combination diet and the fruit + veg diet) reduced BP at the end of eight weeks with greater benefit (BP reduction) in the low-fat group. In the second of the DASH studies, there were 412 participants in a factorial design that included the control diet and the combination diet with each group being classed at three levels of sodium intake: low, moderate and control. The moderate- and low-sodium intake groups reduced BP in the DASH diet and control groups, with the DASH diet leading to lower BP for the same level of salt intake. In both the DASH studies, the food was prepared and provided to the trial participants for purposes of consistency and to avoid significant lapses. The subsequent PRIMIER trial, with 810 mildly hypertensive participants, assessed whether benefits of diet (based on DASH) were sustainable over time.[58] The participants obtained their food in the usual way, unlike the DASH trials (prepared and provided). Participants were grouped (randomised) into three arms: advice only; comprehensive life-style intervention using behavioural approaches; and a combined life-style intervention plus DASH diet. Behavioural approaches consisted of 18 counselling sessions over a six-month period and the primary outcome was systolic BP six months after randomisation. The results at the end of six months were striking: reductions in systolic BP of 11.1 mmHg in the combined group, 10.5 mmHg in the behavioural intervention group and 6.6 mmHg in the advice-only group. The combined group also showed the largest reductions in diastolic BP and in the percentage of participants with hypertension at six months. Whilst the combination of behavioural change plus diet provided significant and meaningful reduction in BP, unlike the DASH trials the DASH diet did not appear to contribute beyond the life-style intervention.

The ENRICHD (Enhancing Recovery in Coronary Heart Disease)[59] trial assessed a different kind of behavioural intervention at 73 hospitals in 2481 patients who had a recent MI within the previous 28 days. All participants were noted to have depression, low social support, or both, and it was deemed that intervening on these factors might improve survival. Those randomised to intervention received counselling whilst the control group received the usual medical care. Both groups were provided information on heart disease risk factors. Whilst depression decreased in the treatment group, there was no difference at three years in the primary endpoint of death or recurrent MI (24.1% vs. 24.2%).[59] The results highlight many important issues: despite their association, depression may not be a causative factor for mortality; the observed improvement in depression may not have been of sufficient magnitude to alter mortality; and the measures of depression and social support particularly after a major event may not reflect the true baseline. Nevertheless, the apparent improvement in depression was not a trivial finding and is an important clinical outcome.

There have been several other behavioural intervention trials aimed at changing the risk factors for heart disease. Whether these will be effective on a community-wide basis remains to be seen. There is one area that has shown variable improvements with community-based intervention in terms of outcome and this leads us to a discussion of 'Nurse-led intervention trials' in CV disease.

TRIALS OF 'NURSE-LED INTERVENTION'

A number of areas and approaches have assessed the influence of 'Nurse-led programmes/interventions' on outcome. As this area of clinical trials has little return to the pharmaceutical or devices industry, the funding for any such trial would have to come from either charitable sources or from the NIH in the United States, the Medical Research Council or Department of Health in the UK, and similar bodies in other countries. This may lead to a limited number of trials or trials with limited sample size influencing the ability of the trial to find differences in mortality. Nevertheless, these trials provide valuable information regarding the cost-effectiveness of nurse-led intervention for health planners. Some such trials have indeed found that intervention led to a reduction in mortality. The two studies from Australia in patients with congestive heart failure[60,61] examined the effect of home-based intervention (HBI). One study consisted primarily of home visits to 200 subjects on unplanned readmission or out-of-hospital deaths within six months of discharge from hospital after the index admission. The intervention consisted of a single home visit by a cardiac nurse 7–14 days after discharge. In the six months of follow-up there were 129 primary endpoint events in the usual care group while the HBI group had 77 events ($p = 0.02$). There were fewer unplanned readmissions (68 vs. 118; $p =$ 003). The hospital-based costs were \$490 300 vs. \$922 600 in the intervention vs. control groups. In the other study, there were fewer out-of-hospital deaths (1 vs. 5; $p = 0.11$) and fewer admissions (36 vs. 63; $p = 0.03$). Whilst the difference in number of deaths was not statistically significant, this only enumerated out-of-hospital deaths and did not take into account deaths following readmission. Similar results were found in a more recent study[62] from two English hospitals. A hybrid programme of clinic + HBI (C + HBI, $n = 58$) was compared with the usual post-discharge care ($n = 48$) and used a cluster randomisation method. Death and hospital readmission were co-primary endpoints. Whilst the hospital readmissions were significantly reduced (44% vs. 22%; $p = 0.019$), the deaths showed a non-significant trend in favour of intervention (7 (15%) vs. 5 (9%); $p = \text{ns}$). It is more than likely that the above studies were not powered to detect differences in mortality and an alternative endpoint might have highlighted the impact more clearly. However, the differences in hospital readmissions and cost of care appear impressive even for a condition such as heart failure. A community-based secondary prevention study for CHD in a primary care setting[63] in middle England noted the cost-effectiveness of such nurse-led intervention: costs of clinics; overall costs to health service; and cost per life year and per quality adjusted life year (QALY) gained – all expressed as incremental gain in intervention group compared with control group as the main outcome measures. Whilst the cost of the intervention (clinics and drugs) and the difference in other NHS costs were not statistically significant, overall, 28 fewer deaths occurred in the intervention group leading to a gain in mean life years per patient of 0.110 and of 0.124 QALYs. The incremental cost per life year saved was £1236 and that per QALY was £1097. Such cost-effective analyses may be unattractive to the physician, the pharma industry or the regulator, but do have a significant bearing on the overall health delivery.

In such nurse-led trials, the choice of the condition being investigated and the patient population do have an impact on the outcome as shown by the following examples. In a study[64] assessing the impact of nurse-led intervention/advice on smoking cessation for hospitalised smokers ($n = 1422$), in a tertiary hospital setting in Australia, no differences were noted at 3 or 12 months follow-up in two groups of smokers ($n = 711$ each). The endpoint was self-reported success in smoking cessation. The brief nurse-delivered intervention incorporated the following: tailored information: assessment of withdrawal, offer of nicotine replacement therapy, booklets and a discharge letter. The reasons for the ‘lack of impact’ may have been multifactorial as smoking cessation is dependent on a number of individual and social factors. A total quality improvement model of outcome-focused treatment, incorporating assertive physician-led pharmacotherapy, routine assessment and recording of nicotine dependence, and engagement from multidisciplinary teams of health professionals, may be required. Another study,[65] whilst not directly related to cardiovascular disease, assessed the effect of regular standardised telephone contact by a diabetes nurse educator (DNE) on metabolic control, treatment compliance and quality of life in 46 adolescents (13–17 year old) with poorly controlled type-I diabetes. A the end of six months of follow-up, regular telephone contact did not lead to immediate improvements in metabolic control in contrast to adults with type-II diabetes. Nevertheless, the knowledge and skills gained during the intervention may have had a delayed beneficial effect in these high-risk adolescents. Whether such efforts have an impact on long-term vascular complications of diabetes is debatable.

Thus, the less than outstanding results of trials of behavioural change and nurse-led intervention from these community-based efforts illustrate the difficulties in achieving either the behaviour change or the problems in the primary care/community-based care management as opposed to individual, specific, pharmacological or device interventions.

SUMMARY

Trials in cardiovascular diseases are overall designed, conducted and analysed in ways similar to trials in other conditions. There are, however, some specific features that should be considered carefully and appear specific. These apply whether they are primary prevention or secondary prevention trials. Cardiovascular diseases are chronic in nature and the pathophysiology precedes clinical manifestations by several years. A number of multiple risk factors are responsible for these diseases and they are likely to affect outcome whether short-term or long-term outcome is considered. They are often multifactorial and may have a genetic predisposition, but single genes are not thought to be major determinants. This may not be true in all cases as mutations may cause certain disease. A huge number of people develop cardiovascular disease of various forms and thus they have a large public health impact. Consequently, any impact on risk factor modification is likely to affect the outcome of the trials. Moreover, because of the high prevalence (large population affected), these trials usually need to have a sufficiently large sample size to encompass the variety of manifestations and risk factors.

This chapter describes the above factors and others, including choice of endpoints, role of specialist investigators, and trials in certain specific clinical situations such as hypercholesterolaemia, heart failure and myocardial infarction. Points specific to devices or surgical procedures have been touched upon. Important aspects of different types of trials such as stepped care or non-responder trials are highlighted. The remaining sections deal with situations or behavioural and other interventions, including nurse-led interventions.

Future trials may include targeted and optimised interventions based on a number of factors including genotype, genotype/phenotype interactions and even gene therapy. The main governing aspects of these trials are likely to include the points raised above to a large extent.

REFERENCES

1. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr, Jones DW, Materson BJ, Oparil S, Wright JT Jr, Roccella EJ. Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. National Heart, Lung, and Blood Institute; National High Blood Pressure Education Program Coordinating Committee. *Hypertension* **42**: 1206–52.
2. WHO cooperative trial on primary prevention of ischaemic heart disease using clofibrate to lower serum cholesterol: mortality follow-up. Report of the Committee of Principal Investigators). *Lancet* (1980) **2**(8191): 379–85.
3. WHO cooperative trial on primary prevention of ischaemic heart disease with clofibrate to lower serum cholesterol: final mortality follow-up. Report of the Committee of Principal Investigators. *Lancet* (1984) **2**(8403): 600–4.
4. Greene HL, Richardson DW, Barker AH, Roden DM, Capone RJ, Echt DS, Friedman LM, Gillespie MJ, Hallstrom AP, Verter J. Classification of deaths after myocardial infarction as arrhythmic or non-arrhythmic (the Cardiac Arrhythmia Pilot Study). *Am J Cardiol* (1989) **63**: 1–6.
5. Kannel WB. Silent myocardial ischaemia and infarction: insights from the Framingham study. *Cardiol Clin* (1986) **4**: 583–91.
6. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* (2002) **106**(25): 3143–421.
7. Cuffe MS, Califf RM, Adams KF Jr, Benza R, Bourge R, Colucci WS, Massie BM, O'Connor CM, Pina I, Quigg R, Silver MA, Gheorghiade M. Short-term intravenous milrinone for acute exacerbation of chronic heart failure. *J Am Med Assoc* (2002) **287**: 1541–7.
8. Gheorghiade M, Gattis WA, Klein L. OPTIME in CHF trial: rethinking the use of inotropes in the management of worsening chronic heart failure resulting in hospitalization. *Eur J Heart Fail* (2003) **5**(1): 9–12.
9. Sackner-Bernstein JD, Kowalski M, Fox M, Aaronson K. Short-term risk of death after treatment with nesiritide for decompensated heart failure. *J Am Med Assoc* (2005) **293**: 1900–5.
10. Sever PS, Dahlof B, Poulter NR *et al.* Ascot investigators. Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial–Lipid Lowering Arm (ASCOT-LLA): a multicentre randomised controlled trial. *Drugs* (2004) **64** (Suppl 2): 43–60.
11. Walsh JM, Pignone M. Drug treatment of hyperlipidemia in women. *J Am Med Assoc* (2004) **291**(18): 2243–52.
12. Ho KK, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham Study. *J Am Coll Cardiol* (1993) **22**(4 Suppl A): 6A–13A.
13. Cohn JN, Tognoni G. Valsartan Heart Failure Trial Investigators. A randomized trial of the angiotensin-receptor blocker valsartan in chronic heart failure. *New Engl J Med* (2001) **345**(23): 1667–75.
14. Pfeffer MA, McMurray JJ, Velazquez EJ, Rouleau JL, Kober L, Maggioni AP, Solomon SD, Swedberg K, Van de Werf F, White H, Leimberger JD, Henis M, Edwards S, Zelenkofske S, Sellers MA, Califf RM. Valsartan in Acute Myocardial Infarction Trial Investigators. Valsartan, captopril, or both in myocardial infarction complicated by heart

failure, left ventricular dysfunction, or both. *New Engl J Med* (2003) **349**(20): 1893–906.

15. Pitt B, Zannad F, Remme WJ, Cody R, Castaigne A, Perez A, Palensky J, Wittes J. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaluation Study Investigators. *New Engl J Med* (1999); **341**(10): 709–17.
16. Canner PL, Klimt CR. The Coronary Drug Project. Experimental design features. *Controlled Clin Trials* (1983) **4**: 313–32.
17. The Coronary Drug Project Research Group. Clofibrate and Niacin in coronary heart disease. *J Am Med Assoc* (1975) **231**: 360–81.
18. Sacks FM, Pfeffer MA, Moye LA, Rouleau JL, Rutherford JD, Cole TG, Brown L, Warnica JW, Arnold JM, Wun CC, Davis BR, Braunwald E. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. Cholesterol and Recurrent Events Trial investigators *(CARE). New Engl J Med* (1996) **335**(14): 1001–9.
19. Lipid Study Group (1998). Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. *New Engl J Med* (1998) **339**(19): 1349–57.
20. Shepherd J, Cobbe SM, Ford I, Isles CG, Lorimer AR, MacFarlane PW, McKillop JH, Packard CJ. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. West of Scotland Coronary Prevention Study Group. *New Engl J Med* (1995) **333**(20): 1301–7.
21. Heart Protection Study Collaborative Group (2002). MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* (2002) **360**(9326): 7–22.
22. Collins R, Armitage J, Parish S, Sleigh P, Peto R. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 5963 people with diabetes: a randomised placebo-controlled trial. *Lancet* (2003) **361**(9374): 2005–16.
23. Downs JR, Clearfield M, Weis S, Whitney E, Shapiro DR, Beere PA, Langendorfer A, Stein EA, Kruyer W, Gotto AM Jr. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. Air Force/Texas Coronary Atherosclerosis Prevention Study. *J Am Med Assoc* (1998) **279**(20): 1615–22.
24. Downs JR, Clearfield M, Tyroler HA, Whitney EJ, Kruyer W, Langendorfer A, Zagrebelsky V, Weis S, Shapiro DR, Beere PA, Gotto AM. Air Force/Texas Coronary Atherosclerosis Prevention Study (AFCAPS/TEXCAPS): additional perspectives on tolerability of long-term treatment with lovastatin. *Am J Cardiol* (2001) **87**(9): 1074–9.
25. ALLHAT Writing Group. The ALLHAT Officers and Coordinators for ALLHAT Collaborative Research Group. Major outcomes in moderately hypertensive patients randomised to angiotensin converting enzyme inhibitor, calcium channel blocker vs diuretic. The antihypertensive and lipid lowering treatment to prevent heart attack trial. *J Am Med Assoc* (2002) **288**: 2981–97.
26. ALLHAT Writing Group. The ALLHAT Officers and Coordinators for ALLHAT Collaborative Research Group. Major outcomes in moderately hypertensive patients randomised to pravastatin vs usual care. The antihypertensive and lipid lowering treatment to prevent heart attack trial. *J Am Med Assoc* (2002) **288**: 2998–3007.
27. Yusuf S, Sleight P, Pogue J, Bosch J, Davies R, Dagenais G. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. The Heart Outcomes Prevention Evaluation Study Investigators. *New Engl J Med* (2000) **342**(3): 145–53.
28. Braunwald E, Domanski MJ, Fowler SE, Geller NL, Gersh BJ, Hsia J, Pfeffer MA, Rice MM, Rosenberg YD, Rouleau JL. PEACE Trial Investigators. Angiotensin-converting-enzyme inhibition in stable coronary artery disease. *New Engl J Med* (2004) **351**(20): 2058–68.
29. Anonymous. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. The Cardiac Arrhythmia Suppression Trial II Investigators. *New Engl J Med* (1992) **327**(4): 227–33.
30. SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. The SOLVD Investigators. *New Engl J Med* (1991) **325**: 293–302.
31. Pfeffer MA, Braunwald E, Moye LA, Basta L, Brown EJ, Cuddy TE, Davis BR, Geltman EM, Goldman S, Flaker GC *et al*. Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. Results of the survival and ventricular enlargement trial. The SAVE Investigators. *New Engl J Med* (1992) **327**: 669–677.
32. Swedberg K, Kjekshus J, Snapinn S. Long-term survival in severe heart failure in patients treated with enalapril. Ten year follow-up of CONSENSUS I. *Eur Heart J* (1999) **20**(2): 136–9.
33. Loeb HS, Johnson G, Henrick A, Smith R, Wilson J, Cremo R, Cohn JN. Effect of enalapril, hydralazine plus isosorbide dinitrate, and prazosin

on hospitalization in patients with chronic congestive heart failure. The V-HeFT VA Cooperative Studies Group. *Circulation* (1993) **87**(6 Suppl): VI78–87.

34. O'Connor CM, Gattis WA, Uretsky BF, Adams KF Jr, McNulty SE, Grossman SH, McKenna WJ, Zannad F, Swedberg K, Gheorghiade M, Califf RM. Continuous intravenous dobutamine is associated with an increased risk of death in patients with advanced heart failure: insights from the Flolan International Randomized Survival Trial (FIRST). *Am Heart J* (1999) **138**(1 Pt 1): 78–86.
35. Hampton JR, van Veldhuisen DJ, Kleber FX, Cowley AJ, Ardia A, Block P, Cortina A, Cserhalmi L, Follath F, Jensen G, Kayanakis J, Lie KI, Mancia G, Skene AM. Randomised study of effect of ibopamine on survival in patients with advanced severe heart failure. Second Prospective Randomised Study of Ibopamine on Mortality and Efficacy (PRIME II) Investigators. *Lancet* (1997) **349**(9057): 971–7.
36. Carson P, Ziesche S, Johnson G, Cohn JN. Racial differences in response to therapy for heart failure: analysis of the vasodilator-heart failure trials. Vasodilator-Heart Failure Trial Study Group. *J Card Fail* (1999) **5**(3): 178–87.
37. Taylor AL, Ziesche S, Yancy C, Carson P, D'Agostino R Jr, Ferdinand K, Taylor M, Adams K, Sabolinski M, Worcel M, Cohn JN. The African-American Heart Failure Trial Investigators. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *New Engl J Med* (2004) **351**: 2049–57.
38. Wyse DG, Waldo AL, DiMarco JP, Domanski MJ, Rosenberg Y, Schron EB, Kellen JC, Greene HL, Mickel MC, Dalquist JE, Corley SD. Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) Investigators. A comparison of rate control and rhythm control in patients with atrial fibrillation. *New Engl J Med* (2002) **347**(23): 1825–33.
39. Carlsson J, Miketic S, Windeler J, Cuneo A, Haun S, Micus S, Walter S, Tebbe U. STAF Investigators. Randomised trial of rate-control versus rhythm-control in persistent atrial fibrillation: the Strategies of Treatment of Atrial Fibrillation (STAF) study. *J Am Coll Cardiol* (2003) **41**(10): 1690–6.
40. Anonymous. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *New Engl J Med* (1989) **321**(6): 406–12.
41. Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH, Arensberg D, Baker A, Friedman L, Greene HL *et al*. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *New Engl J Med* (1991) **324**(12): 781–8.
42. Epstein AE, Hallstrom AP, Rogers WJ, Liebson PR, Seals AA, Anderson JL, Cohen JD, Capone RJ, Wyse DG. Mortality following ventricular arrhythmia suppression by encainide, flecainide, and moricizine after myocardial infarction. The original design concept of the Cardiac Arrhythmia Suppression Trial (CAST). *J Am Med Assoc* (1993) **270**(20): 2451–5.
43. Toff WD, Camm AJ, Skehan JD. United Kingdom Pacing and Cardiovascular Events Trial Investigators. Single-chamber versus dual-chamber pacing for high-grade atrioventricular block. *New Engl J Med* (2005) **353**(2): 145–55.
44. Bristow MR, Saxon LA, Boehmer J, Krueger S, Kass DA, De Marco T, Carson P, DiCarlo L, DeMets D, White BG, DeVries DW, Feldman AM. Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure (COMPANION) Investigators. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. *New Engl J Med* (2004) **350**(21): 2140–50.
45. Lamas GA, Lee KL, Sweeney MO, Silverman R, Leon A, Yee R, Marinchak RA, Flaker G, Schron E, Orav EJ, Hellkamp AS, Greer S, McAnulty J, Ellenbogen K, Ehlert F, Freedman RA, Estes NA III, Greenspon A, Goldman L. Mode Selection Trial in Sinus-Node Dysfunction. Ventricular pacing or dual chamber pacing for sinus node dysfunction. *New Engl J Med* (2002) **346**: 1854–62.
46. Connolly SJ, Kerr CR, Gent M, Roberts RS, Yusuf S, Gillis AM, Sami MH, Talajic M, Tang AS, Klein GJ, Lau C, Newman DM. Effects of physiologic pacing versus ventricular pacing on the risk of stroke and death due to cardiovascular causes. *New Engl J Med* (2000) **342**: 1385–91.
47. Detre K, Peduzzi P, Murphy M, Hultgren H, Thomsen J, Oberman A, Takaro T. Effect of bypass surgery on survival in patients in low-and high-risk subgroups delineated by the use of simple clinical variables. *Circulation* (1981) **63**: 1329–38.
48. The AVID Investigators. Antiarrhythmics versus Implantable Defibrillators (AVID) - rationale, design and methods. *Am J Cardiol* (1995) **75**: 470–5.
49. The Antiarrhythmics versus Implantable Defibrillators (AVID) Investigators. A comparison of antiarrhythmic-drug therapy with implantable defibrillators in patients resuscitated from near-fatal ventricular arrhythmias. *New Engl J Med* (1997) **337**: 1576–84.

50. Bokhari F, Newman D, Greene M, Korley V, Mangat I, Dorian P. Long-term comparison of the implantable cardioverter defibrillator versus amiodarone: eleven-year follow-up of a subset of patients in the Canadian Implantable Defibrillator Study (CIDS). *Circulation* (2004) **110**(2): 112–16.
51. Buchwald H, Varco RL, Matts JP, Long JM, Fitch LL, Campbell GS *et al*. Effect of partial ileal bypass surgery on mortality and morbidity from coronary heart disease in patients with hypercholesterolemia. Report of the Program on Surgical Control of Hyperlipidemia (POSCH). *New Engl J Med* (1990) **323**(14): 946–55.
52. Buchwald H, Varco RL, Boen JR, Williams SE, Hansen BJ, Campos CT *et al*. Effective lipid modification by partial ileal bypass reduced long-term coronary heart disease mortality and morbidity: five year post trial follow-up report from the POSCH. Program on the Surgical Control of Hyperlipidemias. *Arch Intern Med* (1998) **158**: 1253–61.
53. Moss AJ. Background, outcome, and clinical implications of the Multicentre Automatic Defibrillator Implantation Trial (MADIT). *Am J Cardiol* (1997) **80**: 28F–32F.
54. Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS, Daubert JP, Higgins SL, Brown MW, Andrews ML. Multicenter Automatic Defibrillator Implantation Trial II Investigators. Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *New Engl J Med* (2002) **346**: 877–83.
55. Bigger JT Jr. Prophylactic use of implanted cardiac defibrillators in patients at high risk for ventricular arrhythmias after CABG surgery. Coronary Artery Bypass Graft (CABG) Patch trail investigators. *New Engl J Med* (1997) **337**: 1569–75.
56. Appel LJ, Moore TJ, Obarzanek E, Vollmer WM, Svetkey LP, Sacks FM, Bray GA, Vogt TM, Cutler JA, Windhauser MM, Lin PH, Karanja N. A clinical trial of effects of dietary patterns on blood pressure. DASH collaborative Research Group. *New Engl J Med* (1997) **336**: 1117–24.
57. Sacks FM, Svetkey LP, Vollmer WM, Appel LJ, Bray GA, Harsha D, Obarzanek E, Conlin PR, Miller ER III, Simons-Morton DG, Karanja N, Lin PH. DASH-Sodium Collaborative Research Group. Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. *New Engl J Med* (2001) **344**: 3–10.
58. Writing Group of Premier Collaborative Research Group. Effects of comprehensive lifestyle modification on blood pressure control. Main results of PREMEIR clinical trial. *J Am Med Assoc* (2003) **289**: 2083–93.
59. Writing Committee for ENRCHD Investigators. Effects of treating depression and low perceived social support on clinical events after myocardial infarction. The Enhancing Recovery in Coronary heart disease (ENRCHD) randomised trial. *J Am Med Assoc* (2003) **289**: 3106–16.
60. Stewart S, Marley JE, Horowitz JD. Effects of a multidisciplinary, home-based intervention on unplanned readmissions and survival among patients with chronic congestive heart failure: a randomised controlled study. *Lancet* (1999) **354**(9184): 1077–83.
61. Stewart S, Pearson S, Horowitz JD. Effects of a home-based intervention among patients with congestive heart failure discharged from acute hospital care. *Arch Intern Med* (1998) **158**(10): 1067–72.
62. Thompson DR, Roebuck A, Stewart S. Effects of a nurse-led, clinic and home-based intervention on recurrent hospital use in chronic heart failure. *Eur J Heart Fail* (2005) **7**(3): 377–84.
63. Raftery JP, Yao GL, Murchie P, Campbell NC, Ritchie LD. Cost effectiveness of nurse led secondary prevention clinics for coronary heart disease in primary care: follow up of a randomised controlled trial. *Br Med J* (2005) **330**(7493): 707.
64. Nagle AL, Hensley MJ, Schofield MJ, Koschel AJ. A randomised controlled trial to evaluate the efficacy of nurse-provided intervention for hospitalised smokers. *Aust NZ J Public Health* (2005) **29**(3): 285–91.
65. Lawson ML, Cohen N, Richardson C, Orrbine E, Pharm B. A randomised trial of regular standardised telephone contact by a diabetes nurse educator in adolescents with poor diabetes control. *Paediatr Diabetes* (2005) **6**(1): 32–40.

DENTISTRY AND ORAL HEALTH CARE

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

15

Dentistry and Oral Health Care

MAY C.M. WONG, COLMAN MCGRATH AND EDWARD C.M. LO
Dental Public Health, Faculty of Dentistry, The University of Hong Kong, Hong Kong

INTRODUCTION

SCOPE OF THE CHAPTER

Dentistry is concerned with the prevention and treatment of diseases and disorders of the teeth, gums (periodontium) and oral cavity. The two most common oral diseases are dental caries (tooth decay) and periodontal (gum) diseases. Data from the Global Oral Health Databank of the World Health Organization (http://www.whocollab.od.mah.se/index.html) report that at least half of the children and nearly all of the adults in most countries throughout the world have been affected by dental caries. In addition, findings from epidemiological surveys throughout the world have reported that less than 10% of the adult population have no periodontal disease (completely healthy gums). One of the more life-threatening diseases of the oral cavity is oral cancer, primarily cancer involving the oral mucosa (lining of the oral cavity). The prevalence of oral cancer varies form country to country. In most countries it accounts for less than 1% of the total cancer incidence whereas in the Indian subcontinent it can account for 30–50% of the total cancer incidence.[1]

Apart from oral diseases, there are a number of conditions or disorders of the oral cavity that are of concern. Malalignment an malocclusion of teeth (crooked teeth) is prevalent and severe in many countries and most report a growing demand for orthodontic treatment to correct the malocclusion. In the United States, it is estimated that around half of the population are in need of some kind of orthodontic treatment to improve their occlusion.[2] Another problem is the need for replacement of missing teeth, congenitally absent or lost for various reasons, including caries or periodontal disease. Removable prosthesis (dentures or false teeth) and fixed prosthesis (bridges) as well as implants (screw-in teeth) have been used to address these problems.

As the scope of dentistry is very wide, it is not possible to include all kinds of clinical trials in dental research in this chapter. Instead, the discussion here will focus on the more common oral diseases and conditions. First, the disease aetiology and measurements of dental caries and periodontal disease will be presented. Second, clinical trials methods used in dentistry will then be outlined and illustrated with examples. Third, the designs of clinical trials conducted in the areas of dental caries, oral rehabilitation,

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

periodontal disease and orthodontics will be discussed. Finally, the current issues of evidence-based dentistry and hierarchical data analysis will also be discussed. Last, the impact of clinical trials on dental practice will be summarized.

DISEASE AETIOLOGY AND MEASUREMENTS

Evidence from animal and epidemiological studies shows that dental caries arise from demineralisation of tooth hard tissues due to organic acids produced by plaque bacteria on the tooth surface.[3] Frequent intake of fermentable carbohydrates, especially sugars, has been shown to be related to the development of dental caries.[4] The most common measure used in clinical trials to quantify the severity of dental caries is to count the number of decayed, filled and missing (due to caries) teeth or tooth surfaces, the DMFT or DMFS index.[5] The current treatment approach for dental caries emphasises prevention of the disease by strengthening the tooth, such as the use of fluorides and fissure sealants, modification of the diet, such as the use of sugar substitutes, and appropriate health behaviours. Cavities produced by the caries process can be filled by various methods and restorative materials.

Periodontal disease is characterised by the inflammatory response of the gums and its sequel to the toxic substances produced by the plaque bacteria.[6] The current treatment approach for periodontal diseases emphasises primary and secondary prevention of the disease through the removal of plaque by mechanical and chemical means, e.g. toothbrushing and the use of mouthrinses. There are also various surgical and non-surgical ways to treat the periodontal pockets that are formed in more advanced periodontal disease stages. Many indices have been used in clinical trials to quantify the amount of plaque on the tooth surfaces, ranging from a simple dichotomous scale of presence or absence of visible plaque[7] to recording the thickness of the plaque at the gum margin[8] or the area of tooth surface covered by plaque.[9] Gingivitis is usually measured by the presence or absence of bleeding after gentle probing[7] or in an ordinal scale using various clinical signs.[10] For more advanced periodontal diseases, usually the depth of the periodontal pocket and/or the attachment loss are measured in millimetres using marked probes.

Following the development in medical research, patient-based outcome measures have also been developed and employed in dental research. These have focused primarily on the concept of patient satisfaction and health-related quality of life measures. A number of questionnaires have been developed recently to measure the oral-health-related quality of life of people, e.g. the Oral Health Impact Profile (OHIP) and the General Oral Health Assessment Index (GOHAI).[11] Some of these have been used in clinical studies to assess the outcome of dental treatment[12–14] to supplement the clinical assessments.

CLINICAL TRIALS METHODS

In dental research, the clinical trials methods used mainly follow those developed and adopted in medical research. The basic design principles and considerations are very similar, thus the following discussion on the clinical trials methods is kept short for the general areas, while for the specific areas unique to dentistry more details are given. General issues in conducting clinical trials as discussed in Chapter 2 of this book and a few references on clinical trials methodology from the dental literature[15–17] are recommended for general reading.

RANDOMISED CONTROLLED TRIALS

As with the development in medical research, randomised controlled trials (RCTs) have become the gold standard in conducting clinical trials in dentistry. The key features of RCTs are treatment modalities being assigned randomly to the subjects and the existence of a control group. The controls can be either concurrent controls as in parallel studies or self-controls as in crossover studies. The treatment received in the control group can be placebo or standard treatment. In the perfect setting, RCTs should also be double-blinded, which requires that both the subjects and the examiners/observers involved in the trials are not aware of the assignment of the treatment modalities to the subjects, thus reducing biases besides randomisation when the groups are compared.

Informed consent, ethical consideration, data monitoring and pre-study sample size calculation are also important issues in conducting RCTs. Subjects should be informed about the research protocols, their roles as participants in the studies, the different treatment modalities, the possibility of any side effects or risks arising from receiving the treatments, and their right to withdraw from the trial. After the explanation of the above details to the subjects, they should be given ample opportunity and time to ask questions and to discuss the trial with their families and friends. Written consent is normally required, but under special circumstances verbal consent may be employed.[18] It is good practice that clinical trials can only be conducted after approvals from ethical committees are obtained in order to protect human rights.

Data monitoring is especially important in large-scale, multicentre RCTs and usually a data monitoring committee is established to monitor the data collected during the study. The committee needs to monitor the data for patient safety and statistical significance, while keeping their findings confidential to prevent the introduction of bias.[15]

PARALLEL AND CROSSOVER STUDY DESIGNS

In the parallel study design, concurrent groups of subjects are involved in the study and the comparison of the different modalities is the comparison of between-subject variation. When the number of treatment modalities increases, the corresponding sample size required in order to achieve a particular level of power and significance needs to be increased considerably. Crossover study design is a self-controlled study design, in which subjects serve as their own controls. Thus, the comparison of the different modalities is the comparison of within-subject variation. The use of subjects as their own controls prevents confounding by many characteristics that may influence the outcome. In a crossover study each subject is given the different treatments (or treatment and placebo) one after another. Each subject is his or her own control. The sequence of assignment is generally randomised, so that this is in effect a type of RCT. A wash-out period may be required between treatments, to permit the effects of the previous treatment to disappear. However, since subjects who participate in clinical trials with crossover design need to receive all treatment regimens and undergo wash-out periods between the treatments, it can make the investigation periods of the clinical trials very long and not feasible.[19]

Crossover trials are frequently employed in oral hygiene studies where treatment effects are reversible. An example is a single-blinded, short-term crossover clinical trial where the plaque removal performance of three commercially available manual toothbrushes was compared.[20] A sample of 25 dental hygiene students (aged 19 to 42 years old) participated in this trial. The participants were instructed to refrain from toothbrushing or flossing for 24 hours before the trial. The pre-brushing plaque level of the subjects was recorded. One of the three test brushes was then randomly assigned to each participant, and the subjects were allowed to brush for 90 seconds without the aid of a mirror. The post-brushing plaque level was recorded on each participant. This procedure was repeated twice at two-week intervals so that each participant had used all three types of toothbrushes.

Crossover design will not be applicable if the treatment has protracted ‘carry-over’ effect, i.e. the effect of the treatment is not easily reversible. In this situation, either the parallel or the split-mouth design should be adopted. For example, in the case of dental caries, since most of the carious lesions occur in the pit and fissure on the occlusal surfaces of the posterior teeth, the effectiveness of sealing these pits and fissures in order to prevent dental caries is studied. In these studies the number of new caries in the sealed teeth is compared with that of non-sealed teeth. The crossover design is not applicable in these studies because once the teeth are sealed, the process is not reversible. Thus either parallel design or split-mouth design would be used. In the setting of parallel design, subjects are assigned to different concurrent test groups, while in the setting of split-mouth design, different teeth of a subject are assigned to different test groups in the same study period.[21]

SPLIT-MOUTH DESIGN

Split-mouth design is one of the self-controlled study designs that is unique in dentistry. This design is characterised by subdividing the mouth of the subjects into homogeneous within-patient experimental units such as quadrants (upper left, upper right, lower left and lower right sides of the mouth), sextants (upper left posterior, upper anterior, upper right posterior, lower left posterior, lower anterior and lower right posterior), contra-lateral (left and right) or ipsi-lateral (upper and lower) quadrants or sextants or a symmetrical combination of these. With these within-patient experimental units, a range of two to six different treatment modalities can be randomly assigned to the experimental units.[22] The number of treatment modalities usually equals the number of within-patient experimental units. For instance, in a study where two treatment modalities are compared, the within-patient experimental units would usually be the left or right sides of the mouth. In a study where four treatment modalities are compared, the within-patient experimental units could be the four quadrants of the mouth. The split-mouth design has been the principal research design used in periodontal clinical trials to compare different treatment modalities. In the periodontal literature, at least 11 different types of split-mouth design have been described.[22]

The major advantage of using the split-mouth design, like the crossover design, is that subjects serve as their own controls, thus this design may be more efficient than designs that use between-subject comparisons. However, in contrast to the crossover design, since the subjects are concurrently receiving all the treatment modalities in the different parts of the mouth, the study period of the investigation will be shorter. The study period can then be the same duration as if the parallel design is used, but the number of subjects required will be reduced. In an investigation of the efficiency of split-mouth design compared with whole-mouth design (with the use of parallel study design), it was concluded that when disease characteristics were symmetrically distributed over the within-patient experimental units, the split-mouth design could provide moderate to large gains in relative efficiency. For periodontal disease, division of the mouth into two experimental units, either left and right or upper and lower sides, provided the greatest similarities of the disease characteristics.[22]

Besides the distribution of the disease, one should also consider the carry-across effects arising from the split-mouth designs. Carry-across effects occur where treatment performed in one experimental unit can affect the treatment response in other experimental units of the mouth. These effects cannot be estimated from the split-mouth data. Therefore, unless prior knowledge indicates that no carry-across effects exist, reported estimates of treatment efficacy are potentially biased. When deciding on whether a split-mouth design should be used in a clinical trial, researchers should weigh the potential gain in precision against a potential decrease in validity.[23] In a study to compare the effectiveness of two electric toothbrushes in plaque removal, a split-mouth design was used in which either the first (upper right) and third (lower left) quadrants or the second (upper left) and fourth (lower right) quadrants of 84 subjects were brushed with one or the other toothbrush in a random assignment.[24] In this situation, since the distribution of plaque inside the mouth is symmetrical and the use of different toothbrushes to brush different parts of the mouth should not affect the other parts, i.e. no carry-across effect, the use of a split-mouth design is appropriate. In studies that compare the effectiveness of fluoride toothpaste in preventing dental caries, split-mouth design is not advisable as the fluorides from the toothpaste can move freely within the mouth, i.e. with a carry-over effect. For these studies, parallel design should be more appropriate.

BLINDING

In order to achieve double-blinding, the subjects and the examiners/observers should not be able to tell which treatment modalities have been assigned to the subjects. This can be done, for example, in a mouthrinse study comparing the test mouthrinse with placebo. In this type of study, the placebo mouthrinse is made to the same appearance and taste as the test mouthrinse, so that the subjects are not be able to distinguish the two by sight, and they are not told which mouthrinse they have been assigned. On the other hand, the examiners/observers are also not able to access the information on the assignment of the mouthrinse to the subjects.[25] However, in many other situations, one can only blind the examiners/observers but not the subjects (single binding). In the study mentioned above concerning the comparison of the effectiveness of two electric toothbrushes in plaque removal, it is inevitable that the subjects will know which toothbrushes they were using. Thus, in this situation, the best that one can achieve is to blind the investigator from knowing which toothbrushes have been assigned to which quadrants of the subjects.[24] There are situations where even single-blinding is not feasible. In a study comparing the performance of two dental filling materials, amalgam (metal) versus resin-modified glass ionomer cement (tooth colour), it would be very difficult to blind the investigator as one can distinguish the two by their appearance.[26] In conducting clinical trials, one should use the maximum degree of blindness that is possible. In studying the prevalence of caries and fluorosis of children from a water-fluoridated site and a nearby non-fluoridated site, instead of having the investigators examining the children at the sites and then recognising the fluoride content in the water, children were transported to a common examination site so as to blind the investigators from knowing the residence of the children.[27] This method demonstrates an innovative way for the researchers to maximise the degree of blindness.

CLINICAL TRIALS IN DENTISTRY

Although RCTs have become the gold standard in conducting clinical trials in dentistry, most of the clinical trials conducted are of Phase II type. In the following discussion, only examples from RCTs are illustrated.

CARIES PREVENTION AND TREATMENT STUDIES

The aims for these prevention studies are to investigate the effectiveness of different dental caries prevention methods. These include different methods of strengthening the teeth (such as the use of fluorides in different forms), modification of diet (such as the use of sugar substitutes) or modification of health behaviours (such as toothbrushing techniques and habits, and oral education programmes). The target populations for these studies are mainly children, the elderly and special needs groups. For those studies investigating the effectiveness of different forms of fluorides (in forms of toothpaste, varnish, gel or foam), randomisation of the assignment of groups with different regimes (including the control group) can be done at the individual level with parallel design. In a study to compare the effectiveness of two types of toothpaste with different concentration of fluoride to arrest root carious lesions, 201 subjects with at least one root carious lesion were recruited from dental school patients.[28] They were randomly assigned to use either Prevident 5000 Plus (5000 ppm F) or Colgate Winterfresh Gel (1100 ppm F), both containing sodium fluoride in the same silica base. Measurements of lesion hardness, area, distance from the gingival margin, cavitation and plaque were recorded at baseline and three months later by a single examiner.

For those studies involving the modification of diet and behaviour, field (Phase III) studies were used more often because randomisation of the assignment of groups with different regimes would be easier to achieve at the school or community level. In a three-year community intervention trial to determine the caries preventive effect of sugar-substituted chewing gum among Lithuanian school children,[29] a total of 602 children, aged 9–14 years, from five secondary schools in Kaunas, Lithuania, were recruited. Baseline clinical and radiographic caries examinations were given. The schools were randomly allocated to receive one of the five interventions: sorbitol/carbamide gum; sorbitol gum; xylitol gum; control gum; and no gum. Children in the four active intervention groups were asked to chew at least five pieces of gum per day, preferably after meals. The children were re-examined clinically after one, two and three years, and radiographically after three years to find out the number of new dental caries lesions developed during the study period.

In both the above examples, parallel design was adopted. However, studies like the comparison of two different fissure sealant materials can be carried out using the split-mouth design. In a study to compare the retention and the caries preventive effect of a glass ionomer developed for fissure sealing (Fuji III) and a chemically polymerised resin-based fissure sealant (Delton), 179 7-year-old children with at least one pair of permanent first molars that were caries-free or only had incipient lesions were recruited from schools. A split-mouth design was adopted by assigning the two sealing materials randomly to the contralateral teeth. Follow-up examinations for sealant retention and caries development were done after six months, one year, two years and three years.[30]

In order to determine the level of fluoride that should be used (dose finding), some studies have focused on comparing the effectiveness of different concentrations of fluorides in caries prevention. For example, in a randomised, double-blinded study comparing the anti-caries effectiveness of sodium fluoride dentifrices containing 1700 ppm, 2200 ppm and 2800 ppm fluoride ion relative to an 1100 ppm fluoride ion control, a population of 5439 schoolchildren, aged 6–15 years, was recruited from an urban area in Ohio, USA with a low fluoride content in the water supply.[31] Subjects were stratified according to gender, age and caries experience (DMFS scores) derived from the visual–tactile baseline examination. Random allocation of subjects into four treatment groups was done: 0.243% sodium fluoride (1100 ppm fluoride ion) 0.376% sodium fluoride (1700 ppm fluoride ion), 0.486% sodium fluoride (2200 ppm fluoride ion), and 0.619% sodium fluoride (2800 ppm fluoride ion). All products were formulated with the same fluoride-compatible silica abrasive. Subjects were examined by visual–tactile and radiographic examination at baseline and after one, two, and three years to assess their caries development.

The aims of the caries treatment studies were to investigate the performance of the different materials or different approaches used to fill up the decayed cavities in the mouth in terms of bond strength and retention of the materials. As the treatments being delivered cannot be reversed, crossover design is not applicable for these studies. Among these studies, the use of a split-mouth design was more common. In a study to compare the clinical performance of two glass-ionomer cements, ChemFlex (Dentsply DeTrey) and Fuji IX GP (GC), when used with the atraumatic restorative treatment (ART) approach in China, 89 school children aged between 6 and 14 years who had bilateral matched pairs of carious posterior teeth were included.[32] A split-mouth design was used in which the two materials were randomly placed on contra-lateral sides. The performance of the restorations was assessed directly and also indirectly from die-stone replicas at baseline and after 6, 12 and 24 months.

ORAL REHABILITATION STUDIES

One objective of oral rehabilitation studies is to compare a range of treatment modalities to replace missing teeth, including removable and fixed dental prosthesis, and the use of implants. Depending on the treatment modalities, parallel, split-mouth and crossover designs have been applied in these studies. In a five-year parallel study[33] to compare implant-retained mandibular overdentures (IRO) with complete dentures (CD), 61 and 60 patients were randomly assigned to the IRO and CD groups. The clinical aspects and patient satisfaction were measured.

In a study that used a split-mouth design,[34] sandblasted and acid-etched (SLA) implants (recently introduced to reduce the healing period between surgery and prosthesis) were compared with titanium plasma-sprayed (TPS) implants under loaded conditions. Thirty-two healthy patients with comparable bilateral edentulous sites and no discrepancies in the opposing dentition were recruited. The surgical procedure was performed by the same operator and was identical for all the test and control sites. Abutment connection was carried out at 35 N cm six weeks post-surgery for test sites and 12 weeks for the controls by the same dentist who was blinded to the type of surface of the implant. Provisional restoration was fabricated and a new tightening was performed after six weeks. Similar gold–ceramic restorations were cemented on the same type of solid abutments on both sites. Clinical measures and radiographic changes were recorded by the same operator who was blinded to the type of surface of the implant, one year after the surgery.

In a study to compare two designs of maxillary implant overdentures,[35] a crossover design was used. In this study, differences in patient satisfaction with maxillary long-bar implant overdentures with and without palatal coverage opposed by a fixed mandibular implant-supported prosthesis were measured. A mandibular fixed prosthesis was inserted in 13 total edentulous participants, who were then divided into two groups. One group ($n = 7$) received maxillary long-bar overdentures with palate coverage, then long-bar overdentures without palate coverage. The other group ($n = 6$) received the same treatments in the reverse order. For each overdenture design, mastication tests and patient satisfaction were assessed two months after the fitting of the overdenture to allow for adaptation.

Besides clinical outcomes, very often patient-based outcomes such as patient satisfaction and quality of life measures were also measured in oral rehabilitation studies. As in the above quoted example, patients were asked to rate (1) their general satisfaction with the upper prosthesis; (2) satisfaction with the physical functions of the prosthesis such as retention, stability, comfort, ease of cleaning, etc.; and (3) satisfaction with the psychosocial functions such as aesthetics, self-confidence, etc., using both the visual analogue scale and the category scale.[35] In another study by Allen *et al.*,[13] an oral-specific quality of life measure, the Oral Health Impact Profile,[36] which is one of the most comprehensive instruments available in evaluating oral health-related quality of life, was used to measure the impact of the clinical intervention on quality of life. Three groups of subjects were compared: edentulous/edentate subjects who requested and received implant-stabilised complete oral prostheses (IG, $n = 26$), edentulous/edentate subjects who requested implants but received conventional dentures (CDG1, $n = 22$), and edentulous subjects who had new conventional complete dentures (CDG2, $n = 35$).

TRIALS RELATED TO PERIODONTAL DISEASE

In order to prevent periodontal disease, plaque removal and prevention of calculus deposit was one of the key concerns. Thus ways to improve toothbrushing (mechanical means to remove plaque) or the use of mouthrinse or toothpaste with active ingredients like chlorhexidine and pyrophosphate ion (chemical means to remove plaque and to prevent calculus deposition) were investigated. Two main streams of therapy existed: non-surgical and surgical. The aims of these studies were to investigate the effectiveness of the treatment modalities in reducing the depth of the periodontal pocket or improving periodontal attachment level.

Similar to caries research, different study designs in RCT have been applied in periodontal research; one particular design issue that has been discussed more in periodontal research compared with other areas in dental research is the consideration of therapeutic equivalency. This is important because clinical trials for testing superiority and for testing equivalency should have different designs and there has been a tendency by the clinicians in periodontal research to confuse the difference between the two.[37] Clinical trials whose purposes are to show equivalence of two or more treatments have traditionally utilised methods for demonstrating superiority. Thus, if no statistical differences are found, this only demonstrates that the treatments are not superior to the others but they may not be equivalent.[19] The sample size requirements for both equivalence and superiority studies investigating products used in periodontal regeneration have been investigated. It was found that since equivalence clinical trials require much smaller differences between groups, much larger sample sizes are required.[38]

ORTHODONTIC STUDIES

In orthodontic treatments, two main concerns are the effectiveness of early orthodontic treatment for Class II malocclusion and the value of maxillary arch expansion for the treatment of posterior crossbite. Relatively speaking, fewer RCTs have been done in this area. Some researchers have argued that even though RCT has become the gold standard of conducting medical research, it can only apply to a very narrow spectrum of orthodontic questions. One quoted example is 'it would be nearly impossible to enrol fully-informed subjects into any study whose alternatives are of markedly different morbidity: extraction versus non-extraction or orthodontics versus surgery'.[39] Three confusions (or inertia) have been summarised in explaining why the move to conduct RCTs in orthodontic research has been slow. First, there is a remarkable level of non-acceptance that the highest level of evidence is derived from RCTs and many researchers regard retrospective investigation as more useful. Second, there is the argument that it is not ethical to subject patients to a random allocation of treatments, if it is already known that one of the treatments is superior. Third, there is a feeling that RCTs are very large and difficult to manage, are expensive and require a large amount of funding.[40] Solutions to the above confusions have been suggested by O'Brien *et al.*:[40] (1) one should accept that RCTs derive the highest level of evidence and retrospective investigation still has great value in generating questions for RCTs; (2) one should not simply believe that most treatments are superior to others without being tested in an unbiased manner, so it is actually unethical to provide treatment that has not been properly evaluated; (3) careful planning and monitoring of the trials are all that is needed for conducting RCTs. Journal editors should also promote the publication of RCTs. It is worth noting that in the past few years, an increasing number of RCTs have been conducted in this field.[41,42] Hopefully, this indicates a paradigm shift in conducting RCTs in orthodontic research.

CURRENT ISSUES

EVIDENCE-BASED DENTISTRY

One of the major implications of conducting clinical trials is to provide scientific evidence for the clinicians when they are making clinical decisions. Evidence-based medicine is defined as 'the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients'.[43] Dental researchers started addressing the issues of evidence-based dentistry (EBD) in the 1990s and a series of articles has been published to address the concepts of EBD, the misunderstandings of EBD, the barriers to EBD and the processes of finding, evaluating and applying the evidence.[44–50] Various studies making use of the EBD approach have been applied to different areas in dental research and the journal *Evidence-Based Dentistry* has also been published since November 1998.

Systematic literature review is the foundation of the EBD approach. It differs from narrative review as in the latter case the authors or experts do not use standardised ways to retrieve articles and summarise information. Thus, different authors may arrive at different conclusions for the same question. In systematic review, standards in finding, evaluating and synthesising evidences are reported and thus the conclusion is more reliable.

The Cochrane Collaboration is an international organisation that has developed for systematic, up to date reviews of all relevant RCTs of health care (http://www.cochrane.org). In an influential book, Cochrane[51] drew attention to the importance of organising critical summaries of all relevant RCTs by specialty or subspecialty areas of health care, so that people can make more informed decisions. The first ‘Cochrane Centre’ was opened and funded in the UK in October 1992, to facilitate systematic reviews of RCTs across all areas of health care. Currently, there are 12 Cochrane Centres around the world. However, the Cochrane Centres are not directly responsible for preparing and maintaining systematic reviews. This is the responsibility of international collaborative review groups, which also maintain registers of systematic reviews currently being prepared or planned, so that unnecessary duplication of effort can be minimised and collaboration promoted. Currently there are about 50 review groups covering all the important areas of health care and dentistry has been included in the Cochrane Oral Health Group (http://www.ohg.cochrane.org). The principal output of the Cochrane Collaboration is Cochrane Reviews which are published electronically in successive issues of *The Cochrane Database of Systematic Reviews*. Over 50 Cochrane Reviews have been completed in the Cochrane Oral Health Group (as of August 2006). These include reviews of topical fluoride in preventing caries in children and adolescents; different interventions for replacing missing teeth; guided tissue regeneration for periodontal infrabony defects; interventions for preventing oral mucositis or oral candidiasis for patients with cancer receiving chemotherapy; interventions for treating burning mouth syndrome, oral mucositis, oral leukoplakia, oral candidiasis and oral lichen planus; orthodontic treatment for posterior crossbites; and the treatment of dentine hypersensitivity. Over 40 protocols are registered currently including ones that aim to review various fluoride products in preventing caries, various regimes of interventions for replacing missing teeth, and various orthodontic treatment protocols.

HIERARCHICAL AND MULTILEVEL DATA ANALYSIS

Hierarchical data (or clustered data) are common in dental research, as people may have up to 32 teeth and taking measurements from multiple teeth of the same individual is very typical in the data collection of clinical trials in dentistry. For example, in caries prevention studies, the individual teeth of the subjects will be examined and in periodontal research usually six sites of each tooth will be examined, and all these measurements are possibly correlated or clustered. More examples were given by Macfarlane and Worthington[52] and Gilthorpe *et al.*[53] With these correlated or clustered data, conventional statistical methods, which assume observations being independent, are not appropriate for the analysis. Special statistical analysis is required when data have a hierarchical structure. Analysing data without recognising the hierarchical structure and treating the observations as independent will lead to a spurious increase in the sample size, and a corresponding spurious significant relationship.[54] For example, in periodontal research, one might take up to six site measurements for each tooth and with 28 teeth (not including the wisdom teeth) for an individual, the total number of observations for one individual could then be 168. Thus, it is easy to have thousands of observations with only a relatively small number of subjects. In a study by Axtelius *et al.*,[55] the total number of sites assessed was 2236 distributed on 559 teeth in 22 subjects. One way to overcome this problem is to aggregate the raw observations to the highest level of the data structure; for example, site/teeth-level measurements can be aggregated to the level of an individual subject like the DMFT/DMFS index (number of decayed, missing and filled teeth) and mean probing pocket depth. A single aggregated measurement is then obtained from each individual and conventional statistical analysis can be applied. However, aggregation has several drawbacks, the most obvious of which is the loss of detailed information. Therefore, the aggregated measure differentiates poorly both trait and severity of the problems or diseases made at the disaggregated level. Furthermore, aggregation is of little value when the focus of interest lies at a lower level of the data hierarchy, e.g. sites with periodontal pockets.[56]

'Multilevel modelling' (MLM)[57] or equivalently 'hierarchical linear modelling'[58] is a class of techniques developed to analyse hierarchical data. It was first adopted to analyse educational data. These techniques are the modified version of statistical methods available for the analysis of single-level data structures (e.g. multiple regression, logistic regression, log–linear modelling) for the analysis of data with hierarchical structure. One can carry out the multilevel data analysis through the use of software specially written for MLM, such as MLwiN (http://multilevel.ioe.ac.uk), or one can write macros in other statistical software such as SAS and S-plus.[59] In carrying out MLM, one should specify the number of levels in the model and then the variables incorporated at each of the levels. Returning to the example of the periodontal study in which the number of sites assessed was 2236 distributed on 559 teeth in 22 subjects,[55] one of the analyses performed by the researchers was a multilevel analysis of the factors affecting the change in probing periodontal pocket depth at the sites over the course of the treatment. A three-level model was built with site as level 1, tooth as level 2 and subject as level 3. At the site level, 12 variables were constructed (e.g. presence of plaque at the site, treatment received at the site); at the tooth level, three variables were included (e.g. tooth type); and at the subject level, 19 variables were constructed (e.g. age, gender, smoking habit). The above analysis was performed using the software MLn (the non-windows version of MLwiN). With the use of MLM, it is possible to investigate the change in probing periodontal pocket depth at the site in relation to the effects from the site itself, the tooth that it belongs to and the subject overall. In order to evaluate whether a three-level model is necessary, one should test the 'null model' in which no independent variables are included, check the significance of the variance at each level, and then fit the model accordingly. Several other studies using multilevel modelling in analysing dental caries, periodontal and orthodontics data have also been published.[56,59–63,101,102]

IMPACT OF TRIALS ON DENTAL PRACTICE

DIET AND DENTAL CARIES

Evidence of the role of diet, particularly sugars, in relation to dental caries has largely been collated from animal experiments or *in vitro* studies. Human studies have largely been of the observational type: worldwide epidemiological studies, 'before and after' studies, and studies among people with both high and low sugar consumption. Very few interventional studies on human subjects have been conducted[4] and are unlikely to be undertaken in the future given the difficulties of placing groups of people on rigid dietary regimes for long periods of time and because of ethical issues. The main conclusion of studies relating to sugar and dental caries has been that (1) consumption of sugar, even at high levels, is associated with only a small increase in caries increment if the sugar is taken up to four times a day and none between meals; (2) consumption of sugar both between meals and at meals is associated with a marked increase in caries increment.[64,65] These conclusions have shaped key dental education messages of oral health promotion campaigns relating to diet and dental health around the world and also formed the basis of dentist–patient dental health education relating to diet and dental caries.[66] Other trials have provided evidence of variations in caries incidence with different types of sugars, notably the low caries rate associated with the use of sugar alcohols like xylitol.[67] This has led to more widespread use of non-carcinogenic sugar alternatives in drinks and foodstuffs.[68] However, this may be more attributed to their low caloric value than their low carciogenicity.

WATER FLUORIDATION

Evidence of the effectiveness of water fluoridation has largely been based on cross-sectional and ecological studies, 'before and after' studies, and a few cohort or case–control studies. No RCTs have been reported in the dental literature. Systematic reviews of the effectiveness of water fluoridation have concluded that it is an effective, efficient and safe method of preventing dental caries and possibly promotes equity in oral health in society.[69] The studies have examined the relationship between dental caries experience and the fluoride content of the water supply and have shown clearly the association between an increase in fluoride concentration in the drinking water and a decrease in dental caries experience in the population. However, the studies have suggested that there is little benefit where water fluoride concentrations exceed 1 ppm. These findings have resulted in the implementation of water fluoridation in many industrialised and developing countries where central water supplies have made it feasible to do so. It remains a key measure promoted by the World Health Organization for improvement of oral health.

However, there are also a number of studies reporting on the negative influences of water fluoridation on dental and general health, primarily on the effects of water fluoridation in producing dental fluorosis (tooth mottling) among the population.[70] Fluoride at a concentration of 1 ppm is likely to produce a small increase in dental mottling. However, such mottling is unlikely to be of aesthetic concern. Despite strong evidence of the effectiveness and safety of water fluoridation, some communities have ceased to fluoridate their water supplies because of legal issues, social acceptance and concern about the additional benefits of fluoridated water where other sources of fluoride are readily accessible.

ALTERNATIVES TO WATER FLUORIDATION

A wide range of alternative methods for administering systemic fluoride have been suggested in the literature, particularly milk fluoridation, salt fluoridation[71] and recently sugar fluoridation.[72] Extensive literature describing the study of fluoride compounds administered with calcium-rich food, as well as clinical trials and laboratory experiments with fluoridated milk, have demonstrated the effectiveness of milk fluoridation in caries prevention.[73] However, the criticism of decreased bioavailability of the fluoride, the cost and administrative burden involved, and conflicting evidence of efficacy have resulted in few community milk-fluoridated programmes.

Salt fluoridation has also been advocated as an alternative to water fluoridation. Evidence of the effectiveness of salt fluoridation has largely come from test and control community studies in several countries.[74] Despite the fact that salt fluoridation offers the consumer a choice to use fluoride supplements or not, there are only a handful of countries where it is widely available and consumed. Concerns about the appropriate dosage (a minimum of 200 mg/l F is recommended) and safety for general health may impede its widespread implementation.[75]

Fluoride supplements in the form of tablets or drops have long been considered an alternative to water fluoridation. Although the effectiveness of fluoride supplements was endorsed by many small clinical studies, closer examination of the experimental conditions of these, their methods and the analysis of their results undermined confidence in their findings. More modern, well-conducted clinical trials of supplements suggest that today, in children who also exposed to fluoride from other sources such as toothpaste, the marginal effect of fluoride supplements is very small and there is substantial risk of fluorosis if supplements are used by young children.[76] This has resulted in changes to the recommended fluoride dosage schedules, and deferment of the age commencing the use of supplements, being implemented in many countries. Overall, poor compliance and potential risks of fluorosis make fluoride supplements a poor public health measure and they are infrequently prescribed in dental practice.[77]

FLUORIDE TOOTHPASTE

The daily use of fluoride toothpaste is a highly effective method in delivering fluoride to the tooth surface and has proved to be a major aid to caries prevention.[78] Concentration of fluoride at 1000 ppm F has been suggested as a safe and effective means of preventing caries.[79] Although evidence suggests that toothpastes with higher fluoride concentrations are more effective in preventing dental caries,[80] because of safety concerns dentifrices exceeding 1500 ppm F are only sold by prescription in many countries. However, a few clinical trials have suggested that a lower concentration of fluoride in dentifrices (250–500 ppm F) can be used for children and that only a minimum amount (less than 5 mm) should be placed on the toothbrush to minimise risk.[81] Some trials have suggested that combining more than one fluoride agent is more effective than using one source of fluoride agent in preventing dental caries. However, different formulations of toothpaste appear to have similar effectiveness.[82] To some extent, the use of dentifrice has removed the need for professionally applied fluoride agents, except in special circumstances.

OTHER FORMS OF TOPICAL FLUORIDE

Many forms of professionally applied fluoride have been studied, including solutions, gels or foams of sodium fluoride, stannous fluoride, organic amine fluoride, acidulated phosphate fluoride and non-aqueous fluoride varnishes in an alcoholic solution of natural resins and difluorosilane agents covered by a polyurethane coating. All of these professionally applied topical agents have anticaries benefits, although their benefits and the ease of application vary.[83] However, a fairly recent systematic review of the scientific literature undertaken to determine the strength of the evidence for the efficacy of professional caries preventive methods applied to high-risk individuals, and the efficacy of professionally applied methods to arrest or reverse non-cavitated carious lesions, concluded that the strength of the evidence was judged to be fair for fluoride varnishes and insufficient for all other methods.[84] In dental practice, professionally applied fluoride is infrequently employed owing to the more widespread use of other fluoride sources, reports of inconclusive evidence, and because of health care reimbursement for such preventive procedures.

FISSURE SEALANTS

Most carious lesions occur in the pit and fissure on the occlusal surface of posterior teeth. Over the years clinical trials have demonstrated the effectiveness of sealing these fissure and pits in preventing dental caries.[21] Light-curing and autopolymerising sealants are equally effective. However, the cost-effectiveness of fissure sealants remains questionable.[85] Thus, fissure sealants should be employed on clinical grounds on patients with special needs, a history of extensive caries in the primary dentition or caries involving one or more molars.[86] It is important that they are reviewed at regular intervals.

TREATMENT OF CARIES LESIONS

A key focus of research has been the performance of direct posterior restorations (fillings): the longevity and reasons for failure of direct resin-based composite (RBC), amalgam and glass-ionomer cement (GIC) restorations in stress-bearing posterior cavities. A majority of the studies are either of the longitudinal or the retrospective cross-sectional type, with few controlled clinical studies. GIC performs significantly worse than amalgam and RBC.[87] However, the reasons for placement and replacement of direct restorations in dental practice relate to many factors, and aesthetic and safety concerns have resulted in an increased use of RBC or GIC restorations in posterior teeth.[88] The handling and fluoride leaching properties of GIC have made it popular in general practice.[89]

REPLACEMENT OF MISSING TEETH

A range of treatment modalities to replace missing teeth have been studied; including removable and fixed dental prosthesis, and the use of implants. Increasingly these studies have incorporated patients' perceptions of outcomes. Evidence has largely been collated from longitudinal or case–control studies with relatively few RCTs. Implant-retained overdentures are reported to be superior to complete dentures.[33] In addition, implants are useful in the treatment of partial edentulism. However, the widespread use of implants in practice has been limited by a number of factors including health care cover and costs, operator experience and appropriateness for individual cases.

Another contentious issue has been the use of resin-bonded bridges (RBBs) which provide a greater degree of conservation of tooth structure of abutments compared with designs of conventional fixed prostheses in the treatment of partial edentulism.[90] A key concern has been the longevity of RBBs. However, studies suggest that with appropriate case selection, preparation design and cementation, RBBs are a viable treatment option compared with conventional bridges. Increasingly RBBs are being employed in dental practice in the treatment of short edentulous spaces.

TRIALS RELATING TO PERIODONTAL DISEASE

A key focus of periodontal trials has been the need for plaque control to prevent periodontal diseases and for the maintenance of periodontal health.[91] Primarily, studies have focused on mechanical methods of plaque control. Studies have shown that the most important plaque control method is toothbrushing; precise technique is less important than the result, which is removal of plaque without causing damage to the teeth or gums.[92] It is widely promoted to establish toothbrushing practice as a daily routine from childhood.

Additional methods of mechanical plaque control include the use of interdental cleaning aids such as dental floss. While such aids have been shown to be effective in plaque control with minimum damage if used correctly, they are generally prescribed depending on the individuals' periodontal health and their ability to use them appropriately.[93]

Chemical antimicrobial agents may be a useful adjunct measure to managing periodontal health. The use of chlorhexidene in the chemical control of plaques has been widely advocated, particularly in acute phases and in preventing post-surgical infection.[94] However, with the long-term use of chlorhexidene, there is a tendency for it to stain (extrinsic) teeth. In more recent times, the additions of antimicrobial agents to dentifrices to aid plaque control have become commonplace.[95] The use of chemical agents in the removal of plaque, while effective, is not recommended over the use of mechanical agents.[96]

In the treatment of periodontal disease many trials have concluded that non-surgical periodontal therapy is more appropriate than surgical periodontal therapy, and that surgical therapy should only be considered when sites fail to respond to non-surgical methods despite adequate oral hygiene.[97] State-funded and third-party payers of oral health care usually require detailed justification for surgical periodontal care.

ORTHODONTICS

While there has been considerable growth in the practice of orthodontics, there is a dearth of EBR, particularly RCTs.[40] A contentious issue has been the timing of orthodontics and the need for early orthodontic intervention. The evidence relating to early orthodontic treatment is inconclusive, with the result that many clinicians decide, on a case-by-case basis, when to provide orthodontic treatment.[98] Another key concern has been the value of maxillary arch expansion for the treatment of posterior crossbite. A Cochrane Review on the subject was unable to propose recommendations based on inadequate trials.[99] Lack of evidence relating to the value of orthognathic surgery versus orthodontic camouflage in the treatment of mandibular deficiency, and also as to the need for extraction of teeth for orthodontic purposes, has resulted in clinicians deciding on a case-by-case basis without any clinical guidelines.[100] With an increase in the number of RCTs conducted recently, hopefully the practice of orthodontics will be more evidence based.

CONCLUSION

Currently the lack of a sufficient number of high-quality research studies in certain areas within dentistry, namely the lack of randomised controlled trials, has impeded the identification of the best dental practice and the implementation of evidence-based dentistry in those areas. There is, however, widespread recognition of the problem and concerted efforts have been undertaken in recent years through more collaborative high-quality research that can inform policies and develop best guidelines to be implemented in dental practice.

REFERENCES

1. Stjernsward J, Stanley K, Eddy D, Tsechkovski M, Sobin L, Koza I, Notaney KH. National cancer control programs and setting priorities. *Cancer Detect Prevent* (1986) **9**: 113–24.
2. Proffit WR, Fields HW, Morav LJ. Prevalence of malocclusion and orthodontic treatment need in the United States: estimates from the NHANES III survey. *Int J Adult Orthodont Orthognat Surg* (1998) **13**: 97–106.
3. Bowden GHW, Edwardsson S. Oral ecology and dental caries. In: *Textbook of Clinical Cardiology*, Thylstrup A and Fejerskov O, eds, Copenhagen: Munksgaard (1994) 45–70.
4. Gustafsson BE, Quensel CE, Lanke LKS. The Vipeholm dental caries study. The effect of different levels of carbohydrate intake on caries activity in 436 individuals observed for five years (Sweden). *Acta Odontol Scand* (1954) **11**: 232–64.
5. World Health Organization. *Oral Health Surveys: basic methods*, 4th edn. Geneva: WHO (1997).
6. Zambon JJ. Periodontal diseases: microbial factors. *Ann Periodontol* (1996) **1**: 879–925.
7. Ainamo J, Bay I. Problems and proposals for recording gingivitis and plaque. *Int Dent J* (1975) **25**: 229–35.
8. Silness J, Loe H. Periodontal disease in pregnancy. II. Correlation between oral hygiene and periodontal condition. *Acta Odontol Scand* (1964) **22**: 112–35.
9. Turesky S, Gilmore ND, Glickman I. Reduced plaque formation by the chloromethyl analogue of Victamine C. *J Periodontol* (1970) **41**: 41–3.
10. Loe H, Silness J. Periodontal disease in pregnancy. I. Prevalence and severity. *Acta Odontol Scand* (1963) **21**: 533–51.
11. Slade GD, ed. *Measuring oral health and quality of life*. Chapel Hill, NC: University of North Carolina Press (1997).
12. Dolan TA. The sensitivity of the geriatric oral health assessment index to dental care. *J Dent Educ* (1997) **61**: 37–46.
13. Allen PF, McMillan AS, Locker D. An assessment of sensitivity to change of the Oral Health Impact Profile in a clinical trial. *Commun Dent Oral Epidemiol* (2001) **29**: 175–82.
14. McGrath C, Comfort MB, Lo ECM, Luo Y. Can third molar surgery improve quality of life? A 6-month cohort study. *J Oral Maxillofac Surg* (2003) **61**: 759–63.
15. Dennison DK. Components of a randomized clinical trial. *J Periodont Res* (1997) **32**: 430–8.
16. Koch GG, Paquette DW. Design principles and statistical considerations in periodontal clinical trials. *Ann Periodontol* (1997) **2**: 42–63.
17. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of clinical trials*, 3rd edn. New York: Springer (1998).
18. Lau G. Ethics. In: Kalberg J, Tsang K, eds, *Introduction to Clinical Trials.* Hong Kong Press: The University of Hong Kong (1998) 73–112.
19. Burns DR, Elswick Jr RK. Equivalence testing with dental clinical trials. *J Dent Res* (2001) **80**: 1513–17.
20. McDaniel TF, Miller DL, Jones RM, Davis MS, Russell CM. Effects of toothbrush design and brushing proficiency on plaque removal. *Compend Contin Educ Dent* (1997) **18**: 572–7.
21. Morphis TL, Toumba KJ, Lygidakis NA. Fluoride pit and fissure sealants: a review. *Int J Paediatr Dent* (2000) **10**: 90–8.
22. Hujoel PP, Loesche WJ. Efficiency of split-mouth designs. *J Clin Periodontol* (1990) **17**: 722–8.
23. Hujoel PP, DeRouen TA. Validity issues in split-mouth trials. *J Clin Periodontol* (1992) **19**: 625–7.
24. Dörfer CE, Berbig B, von Bethlenfalvy ER, Staehle HJ, Pioch T. A clinical study to compare the efficacy of 2 electric toothbrushes in plaque removal. *J Clin Periodontol* (2001) **28**: 987–94.
25. Corbet EF, Tam JO, Zee KY, Wong MC, Lo EC, Mombelli AW, Lang NP Therapeutic effects of supervised chlorhexidine mouthrinses on untreated gingivitis. *Oral Dis* (1997) **3**: 9–18.
26. Donly KJ, Segura A, Kanellis M, Erickson RL. Clinical performance and caries inhibition of resin-modified glass ionomer cement and amalgam restorations. *J Am Dent Assoc* (1999) **130**: 1459–66.

27. Stephen KW, Macpherson LM, Gilmour WH, Stuart RA, Merrett MC. A blind caries and fluorosis prevalence study of school-children in naturally fluoridated and nonfluoridated townships of Morayshire, Scotland. *Commun Dent Oral Epidemiol* (2002) **30**: 70–9.
28. Lynch E, Baysan A, Ellwood R, Davies R, Petersson L, Borsboom P. Effectiveness of two fluoride dentifrices to arrest root carious lesions. *Am J Dent* (2000) **13**: 218–20.
29. Machiulskiene V, Nyvad B, Baelum V. Caries preventive effect of sugar-substituted chewing gum. *Commun Dent Oral Epidemiol* (2001) **29**: 278–88.
30. Poulsen S, Beiruti N, Sadat N. A comparison of retention and the effect on caries of fissure sealing with a glass-ionomer and a resin-based sealant. *Commun Dent Oral Epidemiol* (2001) **29**: 298–301.
31. Biesbrock AR, Gerlach RW, Bollmer BW, Faller RV, Jacobs SA, Bartizek RD. Relative anti-caries efficacy of 1100, 1700, 2200, and 2800 ppm fluoride ion in a sodium fluoride dentifrice over 1 year. *Commun Dent Oral Epidemiol* (2001) **29**: 382–9.
32. Lo EC, Luo Y, Fan MW, Wei SH. Clinical investigation of two glass-ionomer restoratives used with the atraumatic restorative treatment approach in China: two-years results. *Caries Res* (2001) **35**: 458–63.
33. Meijer HJ, Raghoebar GM, Van't Hof MA, Geertman ME, Van Oort RP. Implant-retained mandibular overdentures compared with complete dentures; a 5-years' follow-up study of clinical aspects and patient satisfaction. *Clin Oral Implants Res* (1999) **10**: 238–44.
34. Roccuzzo M, Bunino M, Prioglio F, Bianchi SD. Early loading of sandblasted and acid-etched (SLA) implants: a prospective split-mouth comparative study. *Clin Oral Implants Res* (2001) **12**: 572–8.
35. de Albuquerque Junior RF, Lund JP, Tang L, Larivee J, de Grandmont P, Gauthier G, Feine JS. Within-subject comparison of maxillary long-bar implant-retained prostheses with and without palatal coverage: patient-based outcomes. *Clin Oral Implants Res* (2000) **11**: 555–65.
36. Slade GD, Spencer AJ. Development and evaluation of the Oral Health Impact Profile. *Commun Dent Health* (1994) **11**: 3–11.
37. Duke SP, Garrett S. Equivalence in periodontal trials: a description for the clinician. *J Periodontol* (1998) **69**: 650–4.
38. Gunsolley JC, Elswick RK, Davenport JM. Equivalence and superiority testing in regeneration clinical trials. *J Periodontol* (1998) **69**: 521–7.
39. Johnston LE Jr. Let them eat cake: the struggle between form and substance in orthodontic clinical investigation. *Clin Orthodont Res* (1998) **1**: 88–93.
40. O'Brien K. Editorial: Is evidence-based orthodontics a pipedream? *J Orthodont* (2001) **28**: 313.
41. O'Brien K, Wright J, Conboy F, Sanjie Y, Mandall N, Chadwick S, Connolly I, Cook P, Birnie D, Hammond M, Harradine N, Lewis D, McDade C, Mitchell L, Murray A, O'Neill J, Read M, Robinson S, Roberts-Harry D, Sandler J, Shaw I. Effectiveness of treatment for Class II malocclusion with the Herbst or twin-block appliances: a randomized, controlled trial. *Am J Orthod Dentofacial Orthop* (2003) **124**: 128–37.
42. Kau CH, Durning P, Richmond S, Miotti FA, Harzer W. Extractions as a form of interception in the developing dentition: a randomized controlled trial. *J Orthodont* (2004) **31**: 107–14.
43. Sackett DL, Rosenberg WM, Gary JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *Br Med J* (1996) **312**: 71–2.
44. Sutherland SE. The building blocks of evidence-based dentistry. *J Can Dent Assoc* (2000) **66**: 241–4.
45. Sutherland SE. Evidence-based dentistry: Part I. Getting started. *J Can Dent Assoc* (2001) **67**: 204–6.
46. Sutherland SE. Evidence-based dentistry: Part II. Searching for answers to clinical questions: how to use MEDLINE. *J Can Dent Assoc* (2001) **67**: 277–80.
47. Sutherland SE. Evidence-based dentistry: Part IV. Research design and levels of evidence. *J Can Dent Assoc* (2001) **67**: 375–8.
48. Sutherland SE. Evidence-based dentistry: Part V. Critical appraisal of the dental literature: papers about therapy. *J Can Dent Assoc* (2001) **67**: 442–5.
49. Sutherland SE. Evidence-based dentistry: Part VI. Critical appraisal of the dental literature: papers about diagnosis, etiology and prognosis. *J Can Dent Assoc* (2001) **67**: 582–5.
50. Sutherland SE, Walker S. Evidence-based dentistry: Part III. Searching for answers to clinical questions: finding evidence on the Internet. *J Can Dent Assoc* (2001) **67**: 320–3.
51. Crochane AL. *Effectiveness and Efficiency. Random Reflections on Health Service.* London: Nuffield Provincial Hospital Trust (1972).
52. Macfarlane TV, Worthington HV. Some aspects of data analysis in dentistry. *Commun Dent Health* (1999) **16**: 216–19.
53. Gilthorpe MS, Maddick IH, Petrie A. Introduction to multilevel modelling in dental research. *Commun Dent Health* (2000) **17**: 222–6.

54. Altman DG, Bland JM. Units of analysis. *Br Med J* (1997) **314**: 1874.
55. Axtelius B, Söderfeldt B, Attström R. A multilevel analysis of factors affecting pocket probing depth in patients responding differently to periodontal treatment. *J Clin Periodontol* (1999) **26**: 67–76.
56. Gilthorpe MS, Griffiths GS, Maddick IH, Zamzuri AT. The application of multilevel modelling to periodontal research. *Commun Dent Health* (2000) **17**: 227–35.
57. Goldstein H. *Multilevel statistical models* 2nd edn. London: Edward Arnold (1995).
58. Bryk AS, Raudenbush S. *Hierarchical linear models: applications and data analysis methods,* Newbury Park, CA: Sage (1992).
59. Hannigan A, O'Mullane DM, Barry D, Schäfer F, Robers AJ. A re-analysis of a caries clinical trial by survival analysis. *J Dent Res* (2001) **80**: 427–31.
60. Albandar JM, Goldstein H. Multi-level statistical models in studies of periodontal diseases. *J Periodontol* (1992) **63**: 690–5.
61. Gilthorpe MS, Cunningham SJ. The application of multilevel, multivariate modelling to orthodontic research data. *Commun Dent Health* (2000) **17**: 236–42.
62. Tu YK, Gilthorpe MS, Griffiths GS, Maddick JH, Eaton KA, Johnson NW. The application of multilevel modeling in the analysis of longitudinal periodontal data – part I: absolute level of disease. *J Periodontol* (2004) **75**: 127–36.
63. Tu YK, Gilthorpe MS, Griffiths GS, Maddick JH, Eaton KA, Johnson NW. The application of multilevel modeling in the analysis of longitudinal periodontal data – part II: changes in disease levels over time. *J Periodontol* (2004) **75**: 137–45.
64. Rugg-Gunn AJ. Nutrition, diet and dental public health. *Commun Dent Health* (1993) **10**: 47–56.
65. Sheiham A. Dietary effects on dental diseases. *Public Health Nutr* (2001) **4**: 569–91.
66. Duggal MS, van Loveren C. Dental considerations for dietary counselling. *Int Dent J* (2001) **51**: 408–12.
67. Scheinin A, Makinen KK. Turku sugar studies. An overview. *Acta Odontol Scand* (1976) **34**: 405–8.
68. Linke HA. Sugar alcohols and dental health. *World Rev Nutr Diet* (1986) **47**: 134–62.
69. Tresaure ET, Chestnutt IG, Whiting P, McDonagh M, Kleijnen J. The York review – a systematic review of public water fluoridation. *Br Dent J* (2002) **192**: 495–7.
70. McDonagh MS, Whiting PF, Wilson PM, Sutton AJ, Chestnutt I, Cooper J, Misso K, Bradley M, Treasure E, Kleijnen J. Systematic review of water fluoridation. *Br Med J* (2000) **321**: 855–9.
71. O'Mullane DM. Systemic fluorides. *Adv Dent Res* (1994) **8**: 181–4.
72. Mulyani D, McIntyre J. Caries inhibitory effect of fluoridated sugar in a trial in Indonesia. *Aust Dental J* (2002) **47**: 314–20.
73. Marino R. Should we use milk fluoridation? A review. *Bull Pan Am Health Org* (1995) **29**: 287–98.
74. Stephen KW. Fluoride prospects for the new millennium – community and individual patient aspects. *Acta Odontol Scand* (1999) **57**: 352–5.
75. Bergmann KE, Bergmann RL. Salt fluoridation and general health. *Adv Dent Res* (1995) **9**: 138–43.
76. Riordan PJ. Fluoride supplements for young children: an analysis of the literature focusing on benefits and risks. *Commun Dent Oral Epidemiol* (1999) **27**: 72–83.
77. Burt BA. The case for eliminating the use of dietary fluoride supplements for young children. *J Public Health Dent* (1999) **59**: 269–74.
78. Clarkson JJ, McLoughlin J. Role of fluoride in oral health promotion. *Int Dent J* (2000) **50**: 119–28.
79. Stephen KW. Dentifrices: recent clinical findings and implications for use. *Int Dent J* (1993) **43**: 549–53.
80. Ripa LW. Clinical studies of high-potency fluoride dentifrices: a review. *J Am Dent Assoc* (1989) **118**: 85–91.
81. Warren JJ, Levy SM. A review of fluoride dentifrice related to dental fluorosis. *Paediatr Dent* (1999) **21**: 265–71.
82. Holloway PJ, Wothington HV. Sodium fluoride or sodium monofluorophosphate? A critical view of a meta-analysis on their relative effectiveness in dentifrices. *Am J Dent* (1993) **6**: S55–8.
83. Newbrun E. Topical fluorides in caries prevention and management: a North American perspective. *J Dent Educ* (2001) **65**: 1078–83.
84. Bader JD, Shugars DA, Bonito AJ. A systematic review of selected caries prevention and management methods. *Commun Dent Oral Epidemiol* (2001) **29**: 399–411.
85. Deery C. The economic evaluation of pit and fissure sealants. *Int J Paediatr Dent* (1999) **9**: 235–41.
86. Smallridge J. UK National Clinical Guidelines in Paediatric Dentistry. Management of the stained fissure in the first permanent molar. *Int J Paediatr Dent* (2000) **10**: 79–83.
87. Hickel R, Manhart J, Gracia-Godoy F. Clinical results and new developments of direct posterior restorations. *Am J Dent* (2000) **13**: 41D–54D.

88. Wilson NH, Burke FJ, Mjor IA. Reasons for placement and replacement of restorations of direct restorative materials by a selected group of practitioners in the United Kingdom. *Quintessence Int* (1997) **28**: 245–8.
89. Yip Hk and Smales RJ. Glass ionomer cements used as fissure sealants with the atraumatic restorative treatment (ART) approach: review of literature. *Int Dent J* (2002) **52**: 67–70.
90. Botelho M. Resin-bonded prostheses: the current state of development. *Quintessence Int* (1999) **30**: 525–34.
91. Loe H. Oral hygiene in the prevention of caries and periodontal disease. *Int Dent J* (2000) **50**: 129–39.
92. Westfelt E. Rationale of mechanical plaque control. *J Clin Periodontol* (1996) **23**: 263–7.
93. Iacono VJ, Aldrege WA, Luckus H, Schwartzstein S. Modern supragingival plaque control. *Int Dent J* (1998) **48**: 290–7.
94. Jones CG. Chlorhexidine: is it still the gold standard? *Periodontology* (2000) **15**: 55–62.
95. Sheen S, Pontefract H, Moran J. The benefits of toothpaste – real or imagined? The effectiveness of toothpaste in the control of plaque, gingivitis, periodontitis, calculus and oral malodour. *Dent Update* (2001) **28**: 144–7.
96. Sheiham A. Is the chemical prevention of gingivitis necessary to prevent severe periodontitis? *Periodontology 2000* (1997) **15**: 15–24.
97. Greenstein G. Nonsurgical periodontal therapy in 2000: a literature review. *J Am Dent Assoc* (2000) **131**: 1580–92.
98. Kluemper GT, Beeman CS, Hicks EP. Early orthodontic treatment: what are the imperatives? *J Am Dent Assoc* (2000) **131**: 613–20.
99. Harrison JE, Ashby D. Orthodontic treatment for posterior crossbites. *Cochrane Database Syst Rev* (2001) **1**: CD000979.
100. Berg R. Orthodontic treatment – yes or no? A difficult decision in some cases. *J Orofac Orthop* (2001) **62**: 410–21.
101. Wong MCM, Lam KF, Lo ECM. Bayesian Analysis of Clustered Interval-censored Data. *J Dent Res* (2005) **84**: 817–21.
102. Wong MCM, Lam KF, Lo ECM. Multilevel modeling of clustered grouped survival data using Cox regression model: an application to ART dental restorations. *Stat Med* (2006) **25**: 447–57.

DERMATOLOGY

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green

16

Dermatology

LUIGI NALDI[1] AND COSETTA MINELLI[2]
[1]Study Centre of the Italian Group for Epidemiologic Research in Dermatology (GISED), Department of Dermatology, Ospedali Riuniti di Bergamo, Bergamo, Italy
[2]Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, Leicester, UK

WHAT IS DERMATOLOGY?

Dermatology deals with disorders affecting the skin and associated specialised structures such as hair and nails. The skin is a biological barrier between ourselves and the outside world consisting of a stratified epithelium, an underlying connective tissue, i.e. dermis, and a fatty layer usually designated as 'subcutaneous'. The skin is not a simple inert covering of the body but a sensitive dynamic boundary. It offers protection against infections, ultraviolet radiation and trauma. It is essential for controlling water and heat loss and contributes to the synthesis of substances such as vitamin D. The skin is also an important organ of social and sexual contact. Body perception, which is deeply rooted within the culture of any given social group, is largely affected by the appearance of the skin and its associated structures.[1]

Extensive disorders affecting the skin may disrupt its homeostatic functions resulting in a properly speaking 'skin failure,' needing intensive care with hydration, maintenance of caloric balance and temperature. However, this is a rare event occurring with conditions such as extensive bullous disorders or exfoliative dermatitis. The most usual health consequence of skin disorders is connected with the discomfort of symptoms, such as itching and burning or pain, which frequently accompany skin lesions and interfere with everyday life and sleeping. Moreover, visible lesions may result in a loss of confidence and disrupt social relations. Feelings of stigmatisation and major changes in lifestyle caused by a chronic skin disorder such as psoriasis have been repeatedly documented in population surveys.[2,3] Additional problems may arise under diverse circumstances: the exudation or loss of substances that interfere locally with the barrier function (and dressing); the shedding of scales whenever excessive desquamation occurs; the need to prevent contact dissemination in the case of transmissible diseases.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

A LARGE NUMBER OF DIFFERENT SKIN DISEASES

Unlike most other organs that usually have around 50 to 100 diseases, the skin has a complement of 1000 to 2000 conditions and over 3000 dermatological categories can be found in the International Classification for Disease version 9 (ICD-9). This is partly justified by the skin being a large and visible organ. Beside disorders primarily affecting the skin, there are cutaneous manifestations with most of the major systemic diseases (e.g. vascular and connective tissue diseases). The classification of skin disorders is far from satisfactory (Table 16.1). Currently, there is a widespread use of symptom-based or purely descriptive terms, such as parapsoriasis or pytiriasis rosea, which reflects our limited understanding of the causes and pathogenetic mechanisms of a large number of skin disorders. For the dermatologist the International Classification of Disease (ICD) coding system is a crude and cumbersome instrument full of bizarre inconsistencies. The recognition of the deficiencies prompted a number of initiatives including the Diagnostic Index of the British Association of Dermatologists and a new Diagnostic Index commissioned by the US National Institute for Arthritis and Musculoskeletal and Skin Diseases (NIAMS). These instruments may improve classification and communication.[4]

Skin diseases as a whole are very common in the general population. A limited number of prevalence surveys have documented that skin disorders may affect 20% to 30% of the general population at any one time.[5] The most common diseases are also the most trivial ones. They include such conditions as mild eczematous lesions, mild to moderate acne, benign tumours and angiomatous lesions. More severe skin disorders, which may have an impact in terms of physical disability or even mortality, are rare or very rare. They include, among others, autoimmune bullous diseases, such as pemphigus, severe pustular and erythrodermic psoriasis, generalised eczematous reactions, and such malignant tumours as malignant melanoma and lymphoma. The disease frequency may show

Table 16.1. Operational classification of skin diseases

	Anatomical distribution	Morphology	Pathology	Pathogenetic process	Aetiology
Genodermatoses		X			X
Nevi and other development defects		X	X		
Mechanical and thermal injuries					X
Photodermatoses		X		X	
Eczemas		X	X	X	X
Lichenoid disorders		X	X		
Disorders of keratinisation		X	X		
Psoriasis		X	X		
Infections and infestations		X			X
Disorders of skin colour		X			
Bullous eruptions		X	X	X	
Disorders of sebaceous glands	X	X		X	
Disorders of sweat glands	X	X			
Immune-related diseases		X		X	
Urticaria		X		X	X
Vascular disorders		X		X	X
Disorders of hair	X	X	X		
Disorders of nails	X	X	X		
Disorders of subcutaneous fat	X	X	X		
Tumours	X	X	X		

variations according to age, sex and geographic area. Eczema is common at any age while acne is decidedly more frequent among male adolescents. Skin tumours are particularly frequent in aged white populations. Infestations and infections such as scabies, pyoderma and dermatophytosis predominate in developing countries and some developed countries' urban pockets. In many cases, skin diseases are minor health problems, which may be trivialised in comparison with other, more serious medical conditions. However, as mentioned above, skin manifestations are visible and may cause more distress to the public than more serious medical problems. The issue is complicated by the fact that many skin disorders are not present in the population as a yes or no phenomenon but as a spectrum of severity. The public's perception of what constitutes a 'disease' requiring medical advice may vary according to cultural issues, the social context, resources and time. Minor changes in health policy may have a large health and financial impact simply because a large number of people may be concerned. For example, most of the campaigns conducted to increase the public awareness of skin cancer have led to a large increase in the number of benign skin conditions such as benign melanocytic nevi being evaluated and excised.

Large variations can be documented among different countries in terms of health service organisation for treating skin disorders. A rough indicator of these variations is the density of dermatologists ranging, in Europe, from about 1 in 20 000 in Italy and France to 1 in 150 000 in the United Kingdom. It has been repeatedly pointed out that public response to skin disease is not necessarily one of understanding and empathy but rather at best disinterest in, and disregard for, its implications for those who have it – if not prejudice and stigmatising negative judgements.[3] The origins of such a response are to be found deeply rooted in history and culture. Negative connotations and moral evaluation attach to skin manifestations. For centuries in many different cultures, skin diseases have been associated with disgrace and danger. A notable component of the process is the connotation of dirtiness attaching to them, bound up with fears of infection or contagion, but not by any means wholly understandable and explicable in these terms.[3]

In general, only a fraction of those with skin diseases are expected to seek medical help, while an estimated large proportion opt for self-medication. Pharmacists occupy a key role in advising the public on the use of over-the-counter products. Primary care physicians seems to treat the majority of people among those seeking medical advice. Primary care of dermatological problems seem to be imprecisely defined with a large overlap with specialist activity. Most of the dermatologist's workload around the world is concentrated in the outpatient department. In spite of the vast number of dermatological diseases, it has been documented that just a few categories account for about 70% of all dermatological consultations. Brief, more detailed descriptions of the most frequent skin categories are given below while skin cancer is dealt with in a later section.

Generally speaking, dermatology requires low-technology clinical practice. Clinical expertise is mainly dependent on the ability to recognise a skin disorder quickly and reliably, which, in turn, depends to a large extent on the awareness of a given clinical pattern, based on previous experience and on the exercised eye of a visually literate physician.[6] Complementary diagnostic procedures include skin biopsy, patch testing and immunopathology.

A peculiar aspect of dermatology is the possible option for topical treatment. This treatment modality is ideally suited to localised lesions, the main advantage being the restriction of the effect to the site of application and the limitation of systemic side effects. A topical agent is usually described as a vehicle and an active substance, the vehicle being classified as powder, grease, liquid or combinations such as pastes and creams.

Much traditional topical therapy in dermatology has been developed empirically with so-called magistral formulations. Most of these products seem to rely on physical rather than chemical properties for their effects and it may be an arbitrary decision to appoint one specific ingredient as the 'active' one. Physical effects of topical agents may include detersion, hydration and removal of keratotic scales. The border between pharmacological and cosmetic effects may be imprecisely defined and the term 'cosmeceuticals' is sometimes employed.[7] It should be noted that the evaluation of even the most recent cosmetic products is far from being satisfactory. In addition to pharmacological treatment, a number of non-pharmacological treatment modalities exists including phototherapy or photochemotherapy and minor surgical procedures such as electroessication and criotherapy. Large variations in treatment modalities for the same condition have been documented, which mainly reflect local traditions and preferences.[8,9]

ACNE

The term 'acne' refers to a group of disorders characterised by abnormalities of the sebaceous glands. *Acne vulgaris* is the most common condition and is characterised by polymorphous lesions, including comedones (blackheads), inflammatory lesions such as papules or pustules, and scars, affecting the face and less frequently the back and shoulders. A combination of factors are considered as pathogenetic, including the hormonal influence of androgens, seborrhea, abnormalities in the bacterial flora with overgrowth of *Propionibacterium acnei*, and plugging of pilosebaceous openings. Mild degrees of acne are extremely common amongst teenagers (more than 80%) and decrease in later life. The prevalence of moderate to severe acne has been estimated at about 14% in 15–24 year olds, 3% of 25 to 34 year olds and about 1% of those aged 35–54 years. It is likely that the vast majority of sufferers of mild acne do not seek medical advice. Around 70% of those affected with acne experience shame and embarrassment because of it. Criteria for treatment include clinical severity, as judged by the extension and presence of inflammatory lesions, and the degree of psychological distress from the disease. The aim of treatment is to prevent scarring, limit disease duration and reduce psychological stress. Mild acne is usually treated by topical modalities such as benzoyl peroxide and/or tretinoin, while moderate severity acne is treated by systemic antibiotics or antiandrogens in women. Oral isotretinoin is used under specialist supervision for severe unresponsive disease. There are a number of published systems for measuring the severity of acne.[10] These vary from sophisticated systems with up to 100 potential grades to simple systems with 4 or 5 grades. A specially designed acne disability index has been also devised to assess the psychological impact of the disease and disability, and has been found to correlate well with severity as measured by an objective grading system, even if a small group experiences disability which is out of proportion with severity.[11]

ATOPIC DERMATITIS

Typically, this condition is characterised by itching, dry skin and inflammatory lesions especially involving skin creases. Patients suffering from atopic dermatitis may also develop IgE-mediated allergic diseases such as bronchial asthma or allergic rhinitis. An overall cumulative prevalence of between 5% and 20% has been suggested by the age of 13–14. Around 60% to 70% of children are clear of significant disease by their mid-teens. Even if genetic factors seem to play a major role, environmental factors such as allergens and irritants are important and there is reasonable evidence to suggest that the prevalence has increased two to threefold over the last 30 years. There is some evidence that it may be possible to prevent atopic dermatitis in high-risk children born to parents with atopic disease by restricting maternal allergens and reducing house dust mite levels.[12,13] Moreover, the role of parents' education should not be underestimated. No treatment has been shown to alter the natural history of established eczema and the mainstay of treatment is emollients, which moisturise the skin and topical steroids.

PSORIASIS

This is a chronic inflammatory disorder characterised by red scaly areas, which tend to affect extensory surfaces of the body and scalp. Its overall prevalence is about 1% to 3%. Several varieties have been described including guttate, pustular and erythrodermic psoriasis. In about 3% of cases it may associate with a peculiar arthritis. Significant disability has been documented with psoriasis. Multifactorial heredity is usually considered for disease causation. This implies interaction between a genetic predisposition and environmental factors. Heritability, a measure that quantifies the overall role of genetic factors, ranges from 0.5 to 0.9. Acute infections, physical trauma, selected medications and psychological stress are usually viewed as triggers. The risk of psoriasis has been linked with smoking (especially pustular varieties) and increased body mass index.[14] Sun exposure usually temporarily improves the disease. Altered kinetics of epidermal cells has been repeatedly documented. The lesions are visible and may itch, sting and bleed easily. The aim of treatment is to achieve short-term suppression of symptoms and long-term modulation of disease severity improving the quality of life with minimal side effects. Topical agents such as vitamin D derivatives, dithranol and steroids can be used for short-term control of the disease. Ultraviolet B phototherapy, psoralen plus ultraviolet A phototherapy (PUVA) and systemic agents such as methotrexate or cyclosporine are employed to control extensive lesions that fail to respond to topical agents. Relapse is common upon withdrawal. After decades of limited advancements, a remarkable number of new systemic agents, collectively termed ‘biologicals’, including tumour necrosis factor inhibitors and drugs interfering with the function of T cells, are entering the market and being made available This has been paralleled by a revitalisation of the immunopathogenesis of the disease with a new emphasis on T cell and antigen presenting cell (APC) interaction. To date, only short-term placebo-controlled randomised control trials have been conducted with biological agents, only surrogate outcome measures have been employed and study results have not provided any clue concerning the role of these agents as compared with traditional ones in the *long term* management of psoriasis.[15] In addition, recent data seem to cast some doubt about the central role of immunological mechanisms on psoriasis.[16]

Outcomes that matter to the patient include disease suppression and duration of remission, patient satisfaction and autonomy, and disease-related quality of life. A number of clinical activity scores have been developed, the most popular being the Psoriasis Area and Severity Index – PASI.[17] There is no documented evidence that such indexes are reliable proxies for the above-mentioned outcomes. In the long term, a simple measure such as the number of patients reaching complete or nearly complete stable remission appears to be the most relevant outcome variable.

LEG ULCERS

Venous and arterial leg ulcers are recognised as the most common chronic wounds in Western populations. A skin ulcer has been defined as a loss of dermis and epidermis produced by sloughing of necrotic tissue. Ulcers persisting for four weeks or more have been rather arbitrarily classified as chronic ulcers. Based on population surveys, the point prevalence of leg ulcers ranges from 0.1% to 1.0% and increases with age. Venous ulcers are the end result of superficial or deep venous insufficiency and a venous origin is diagnosed in about 80% of cases. Arterial ulceration may be regarded as a multistep process, starting, in general, with a systemic vascular derangement such as atherosclerosis. The prognosis of leg ulcers is less than satisfactory with about one-quarter of subjects not healing in over two years and the majority of patients having recurrence. In a large-scale clinical study, the healing time varied according to the dimension of the ulcers, their duration and the mobility of the patient. The quality of life of ulcer patients may be severely affected. Social isolation, depression and negative self-image have been associated with ulcers in a high percentage of patients.[18] A number of studies point to the less than satisfactory management of ulcer patients in the community, including the lack of any clinical assessment leading to long periods of ineffective or inappropriate treatment and delays in instituting effective pain-relieving strategies. Ulcer clinics in vascular surgical services in the UK proved to offer advantages over home treatment.[19] The overwhelming rates of recurrence clearly suggest that more attention should be paid to prevention.

SKIN DISORDERS AND CLINICAL TRIAL METHODS: ADAPTING STUDY DESIGN TO SETTING AND DISEASE

As for other disciplines, the last few decades have seen an impressive increase in the number of clinical trials carried out in dermatology. However, there are indications that the upsurge of clinical research has not been paralleled by a refinement in clinical trial methodology and the quality of randomised control trials (RCTs) in dermatology falls well below the usually accepted standards.[20–22] In this section we would like to mention some issues which deserve special attention when designing a randomised clinical trial in this speciality area. There is a need for innovative thinking in dermatology to make clinical research address the important issues and not simply ape the scientific design.

RANDOMISATION

It can be estimated that there are at least 1000 rare or very rare skin conditions where no single randomised trial has been conducted. These conditions are also those which carry a higher burden in terms of physical disability and mortality. The annual incidence rate of many of them is lower than 1 case per 100 000 and frequently less than 1 case per 1 000 000. International collaboration and institutional support are clearly needed. There are no examples of such an effort.

For quite different reasons, there are also common skin conditions where RCTs have been rarely performed. These conditions include several varieties of eczematous dermatitis (e.g. nummular eczema), psoriasis (e.g. guttate psoriasis) and urticaria (e.g. pressure urticaria), a number of exanthematic reactions (e.g. pytiriasis rosea), rosacea and common infections such as warts and molluscum contagious. One alleged difficulty with mounting RCTs in dermatology is the visibility of skin lesions and the consideration that, much more than in other areas, patients self-monitor their disease and may have preconceptions and preferences about specific treatment modalities.[23] The decision to treat is usually dictated by subjective issues and personal feelings. As we will consider below, there is a need to educate physicians and the public about the value of randomised trials to assess interventions in dermatology. The need to evaluate the attitudes of patients and to educate should be clearly considered when planning a study and developing modalities to obtain the informed consent of the patient.

RCTs are usually designed in dermatology with an expected large effect from the test treatment and most trials do not recruit more than a few dozen patients. In small trials there may be substantial differences in group sizes that will reduce the precision of the estimated differences in treatment effect and hence the efficiency of the study. As a consequence, block randomisation may be preferable. On the other hand, a substantial imbalance may persist in prognostic characteristics, and minimisation can be used to make small groups more similar with respect to major prognostic variables.[24] There is some evidence that the group sizes of clinical trials, apparently based on simple randomisation and published in a number of leading dermatological journals, tend to be much too similar than expected by chance (unpublished data from the European Dermatoepidemiology Network Psoriasis Project). The cluster around equal sample sizes may be due to publication bias, failure to report blocking, or even to the rectification of an unsatisfactory imbalance by adding extra patients to one treatment.

In many instances, the management of a chronic skin disorder is a multiple step process where different phases can be identified. For example, at least two phases are usually considered when treating psoriasis: a clearance phase, which involves a more intensive treatment approach with the aim of clearing existing psoriasis lesions, and a maintenance phase, with the main aim of preventing disease relapse. The different phases are not necessarily well separated in time. Long-term disease-modifying strategies can be adopted at the same time when a treatment modality for reaching clearance has been started. An example is the treatment of atopic dermatitis by topical steroids and diet. Most RCTs in dermatology use a simplified approach to evaluate treatment effects and most of them analyse the effect of a single manoeuvre over a limited time span. One as yet not fully explored issue is the potential for combining different treatment approaches in a simultaneous or subsequent order. There are examples of combinations of such treatment modalities such as calcipotriol and ultraviolet B radiation in psoriasis treatment, but other rationale combinations are not fully explored. A way of addressing the issue is by relying on factorial design. An example of such a design would be a randomised clinical trial of the effect of a low allergen diet compared with an unrestricted diet in atopic women during pregnancy and breast-feeding on the subsequent development of atopic disorders in children where women are randomised to all the possible combinations of restricted and unrestricted dietetic measures during the periods examined.

BLINDING

There are several reasons for considering blindness as a major bias-reducing procedure in RCTs of skin disorders. First, it is expected that physicians and patients are subject to strong, though difficult to document, hopes and prejudices about the optimal care of skin disorders. This is reflected, for example, in the large variations of treatment procedures for the same condition which have been repeatedly documented in different areas of dermatology. Second, most outcome measures are soft endpoints involving subjective judgement, which may be influenced to a significant extent by the previous knowledge of the treatment adopted. Third, the visibility of lesions may influence the decision to rely or not on a given treatment to a larger extent as compared with situations where disease variables are not so obvious. On the other hand, there may be problems with blinding which may be difficult or impossible to solve, as with trials comparing complex procedures such as ultraviolet light radiation and drug regimens. An issue which warrants more attention than it is often given in randomised trials is the possibility that certain ‘marker variables’ occur, together with obvious side effects. These variables, observable during treatment, may in part unblind the trial, even at a subliminal level.[25] This is an issue with the use of topical retinoids and the associated mild cutaneous irritation, which may be noticed but not reported at all as a ‘side effect’. It is quite common to find RCTs where the authors claim blindness in situations where treatment modalities are responsible for frequent and obvious side effects. In 1982, for example, a trial was published examining three different therapeutic strategies for psoriasis: oral etretinate associated with topically applied betamethasone, oral etretinate associated with topically applied placebo, and oral placebo associated with topical betamethasone.[26] Systemic retinoids such as etretinate are responsible for common side effects which are reminiscent of vitamin A overdosage including dryness of the skin and mucous membranes, while topical steroids commonly produce a transitory blanching effect. It is difficult to accept blindness in the trial when there is no additional information on how blinding was actually assured. It is worth mentioning that the drop-out rates showed large variations among the different trial arms because of alleged side effects of treatment.

One way to overcome the problem of an unachievable blindness and avoid the influence of the researcher’s subjective judgement is to plan the study so that the clinician who treats the patient is not the same one who judges the effect of the therapy. This way the second clinician can be blind to the treatment assigned to the patient.

STANDARDISATION OF TREATMENT MODALITIES AND ACCESSORY CARE

Independently of the ‘active’ intervention administered, accessory non-pharmacological treatment and skin care seem to play a significant role in the outcome of most skin disorders. It is common sense that emollients may improve dry skin and wet soaks may help to dry exudating lesions. As a consequence, accessory care requires careful standardisation. However, while it is relatively easy to ensure that the pharmacological treatment is conducted in an appropriate way (particularly timing and administration route), non-pharmacological accessory care is prone to a larger variability that is affected by social and cultural factors among others. To a greater extent, variability may affect topical treatment as compared with systemic treatment. Topical treatment is usually more cumbersome in comparison with systemic treatment and may well depend on the physician’s and patient’s consistency. As recently documented in RCTs of the retinoid derivative tazarotene in psoriasis, the modalities of application may play a significant role in tolerability and side effects.[27] The variations which have been documented in the placebo arms of RCTs for psoriasis also point to the need for standardisation and strict entry criteria.[28]

Once again a well-informed patient as well as an active and supportive clinician are vitally important. The issue of standardisation is also important for assessing compliance when the treatment is self-applied by the patient. If indeed there are limitations with such methods as tablet count for assessing compliance with systemic agents, the limitations are even greater when similar methods are used to monitor the consumption of topical agents in the absence of strict rules to define a ‘single dose’. The amount consumed cannot be monitored if patients are not carefully instructed on how to apply the topical agent. The observed compliance behaviour may range over compliant, overusing, erratic using and dropping out.

DIFFERENT STUDY SETTINGS AND DISPARATE DISEASE SEVERITY

We have already mentioned that the contents of primary care for many skin disorders are imprecisely defined as opposed to specialist clinical practice, with possibly large overlapping areas. In addition, it has been noted that there may be wide variations in terms of severity within any given diagnostic category, with conditions ranging from subclinical manifestations, e.g. psoriatic ‘markers’, to skin failure, e.g. erythrodermic or generalised pustular psoriasis. Moreover, it should also be noted that for any given disease there might be clinical variants, which may have peculiar prognostic features and responses to treatment, e.g. guttate psoriasis versus plaque psoriasis. As a consequence, it is of the outmost importance that entry criteria in RCTs of skin disorders are defined as precisely as possible. This should include as a major requirement the definition of the study setting, clinical variety, disease severity and duration, previous treatments, and concomitant systemic disorders.

It should be stated that the severity assessment of most skin disorders implies an understanding of the many influences of the disease on the patient’s life, including disease-associated discomfort, level of disability and social disruption. Most of these influences are better expressed as a continuum of severity rather than a yes or no phenomenon. On the other hand, there are practical advantages in trying to translate the continuum into a limited number of workable severity categories. The main advantage is a better compliance with the discrete nature of most clinical decisions where thresholds are usually required for implementing interventions (examples of categorical classifications of a severity continuum are tumour staging and arterial hypertension definition). Unfortunately, for many inflammatory skin disorders no reliable severity criteria have been developed. Even when such criteria are available, there is uncertainty about severity thresholds. Consequently, large variations are expected to occur among different RCTs and, in fact, have been documented on several occasions. A rather common attitude in published RCTs is the lack of entry criteria and severity definitions, so that the patient population appears to be recruited in a vacuum.[21,29] One habit which should be discouraged is to include broad diagnostic categories that lack specificity, such as the category of 'steroid responsive dermatoses' or 'itching disorders'.

OUTCOME MEASURES

There are obvious analogies between the problems implied in the development of severity criteria and those implied in outcome measures. They both consist of measures that must have the properties of validity and reliability. In addition, outcome measures must be responsive to change, i.e. they must have the ability to identify what may be small but nevertheless clinically important changes. On the other hand, with severity criteria the intent is more to discriminate between individuals. If an instrument is useful to discriminate between severity levels of psoriasis, it does not necessarily mean that it will be able to detect changes which are important as a result of treatment within these categories. The distinction between the two aims (i.e. describing variations between individuals and assessing variations within individuals) is frequently blurred when developing measurement systems for skin disorders. In spite of the fact that, from a clinical point of view, distinguishing between disease severity levels may represent a different issue as compared with assessing clinically important changes in individual patients, the two issues are usually dealt with by relying on identical scale systems in dermatology.

There are indications that many score systems employed in dermatology lack the basic requirements for reliability and validity. Even a simple measure such as the approximate percentage of area involved by a skin disease is prone to wide inter- and intra-observer variations if the evaluation methods are not clearly specified.[30] In spite of their lacking basic requirements, a large number of different scales have been developed for such common disorders as psoriasis or atopic dermatitis (Table 16.2). One example is the 'Psoriasis Area Severity Index' (PASI).[17] This index is obtained by summing up the scores concerning three features of psoriasis, namely the body district affected, the severity of the condition (judged by the degree of erythema, infiltration and desquamation) and the extension of the disease. The last two are judged according to the body district analysed. Although the PASI score has been widely used, it is largely unsatisfactory.[30] It has never been standardised and there is limited testing for inter- and intra-rater reliability. Validity is another issue. It has never been demonstrated that the weights arbitrarily attributed to each item in the PASI score actually reflect the clinical severity of lesions. PASI is only relying on the dermatologist's judgement of a few clinical features of psoriasis and there is increasing awareness that the patient's judgement is equally important. An additional drawback of PASI is that similar scores can be attributed to varieties of psoriasis which differ clinically and in terms of response to treatment.[40] The 'Self-Administered PASI' (SAPASI), which asks the patient to make the same evaluation as the physician for PASI, does not escape the limitations we have pointed out for PASI as an outcome measure.[41]

To overcome the problems arising from subjective judgement, more 'objective' measures have been repeatedly advocated, such as the use of ultrasound to evaluate the thickness of psoriasis plaques.[40] In fact, any measurement is fully justified only when it represents a good surrogate for clinically important outcomes, such as the patient disability and quality of life.[42] The notion of responsiveness to change expresses the idea that any measure used in a trial should be sensitive to 'clinically important changes' in response to therapy.[43] A conceptual difficulty arises in specifying what a clinically important change is. With most scales developed in dermatology the issue remains fraught since no 'gold standard' has gained wide acceptance. It should be considered that the 'outcome' of the treatment refers to 'all the possible results that stem from preventive or therapeutic interventions' and consists of several separate dimensions (e.g. discomfort and disability), which may be broken down into components

Table 16.2. Measures used in the outcome evaluation of selected skin diseases

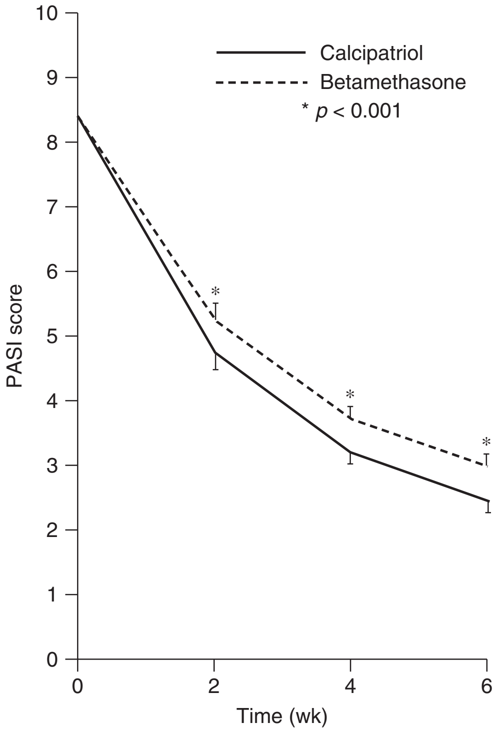
Disease	Clinical scales	Disease-specific quality of life measures
Acne[10,31,32]	• Lesion counting (papule, pustule and comedone counts) • Plewing and Kligman grading system • Cunliffe score (Leeds technique) • Cook's photonumeric method • American Academy of Dermatology (*AAD*) classification • Allen and Smith photographic method • Fluorescence photography • *GAGS* (Global Acne Grading System)	• *ADI* (Acne Disability Index) • *CADI* (Cardiff Acne Disability Index) • *APSEA (*Assessment of the Psychological and Social Effects of Acne) • *AQL* (Acne Quality of Life) index
Atopic dermatitis[32–34]	• *SCORAD* (severity Scoring of Atopic Dermatitis) • *SASSAD* (Six-Area, Six-Sign Atopic Dermatitis) severity index • *ADASI* (Atopic Dermatitis Area and Severity Index) • *EASI* (Eczema Area and Severity Index) • Rajka and Langerland scoring system • *SSS* (Simple Scoring System) • *BCSS* (Basic Clinical Scoring System) • *ADSI* (Atopic Dermatitis Severity Index) • *SIS* (Skin Intensity Score) • *ADAM* (Assessment Measure for Atopic Dermatitis) • Nottingham Eczema Severity Score	• *EDI* (Eczema Disability Index)
Psoriasis[30,35]	• *Severity scores* based on *individual signs* (involved body surface area, erythema, induration, desquamation) • *PASI* (Psoriasis Area and Severity Index) • *SAPASI* (Self-administered PASI) • *Ultrasound* evaluation of the thickness of psoriasis	• *PDI* (Psoriasis Disability Index) • *PLSI* (Psoriasis Life Stress Inventory)
Leg ulcers[36,37]	• Clinical skin score • *Simple* wound measurements • *Planimetric* wound area measurements	
Dermatological diseases as a class[32,38,39]	• *DIDS* (Dermatology Index of Disease Severity)	• *DLQI* (Dermatology Life Quality Index) • *CDLQI* (Children's Dermatology Life Quality Index) • *IMPACT* (Impact of Skin Disease Scale) • *SKINDEX*

and simple measurable items. Any given measure achieves its value only to the extent that it serves as a proxy for an outcome component. For example, if the PASI accurately quantifies disability or discomfort, then it may be of value as a surrogate outcome measure for psoriasis. What may be a relevant outcome variable is a matter of judgement, based on knowledge of the disease, the patients' requirements, and the values of that society. The outcome of skin disorders that affect the quality rather than the quantity of life is expected to be largely culture-dependent. Very recently, some statistical refinements of existing quality of life scales have been made by the use of models relying on an item response theory (IRT) such as the Rasch model, rather than classical test theory. These refinements allow us to address some issues like threshold order, item fit and differential item functioning, and to ensure the transition from the representational (nominal, ordinal, interval and ratio scales) to a sort of more fundamental measurement of quality of life.[44] Rasch analysis has been mainly applied in dermatology to the Skindex scale.[45] In chronic disease, one issue is to provide a long-term assessment of quality of life measures. Only one paper attempted such an assessment, highlighting the complexity of factors accounting for long-term changes in quality of life measures.[46] Self-reported general health and age were identified as confounding factors to be considered when comparing quality of life changes among groups over time. Such an assessment pointed to changes which were largely independent of physical disease severity and treatment.

Papers assessing long-term correlates of changes in terms of coping strategies and accommodation to the disease are needed.

It is our conviction that the development of a 'gold standard' requires a deep understanding of patients' requirements and expectations from treatment. In the lack of reliable scales, trials with the simplest and most objective outcome variables are preferable. Such measures as complete remission or recurrence should be preferred, provided that these categories are clearly defined. Clearly, remission or recurrence are events which occur with a lower frequency as compared with less dramatic variations in disease activity measured by clinical scales. This, in turn, affects the sample size calculation.

There are at least two different choices when analysing outcome measures expressed by any given score system. We might compare the difference between the initial and final scores in the treatment and control groups, or, alternatively, ignore the pre-test scores and simply analyse the scores after treatment. There are two important analytical reasons to consider in the use of change scores.[47] The first is that the subtraction of scores before treatment has the effect of removing stable individual differences between subjects, thereby increasing the power of the statistical test. The second reason, which is of relatively minor importance in a randomised trial, is that there may be overall differences between the two groups at the baseline, and the use of change scores can potentially correct for these differences. The usual presentation of score data over time (e.g. PASI score) is to build up a curve based on the mean score values of the treatment and control groups. A common but inappropriate analysis of these curves is to apply separate sample tests on mean score values at several time points (Figure 16.1). The means may not represent a good descriptor of a typical curve for an individual and the separate analyses of different time points does not convey information on how individual subjects respond over time. Moreover, this practice can be criticised on statistical grounds because of multiple potentially data-driven statistical tests and because the values over time are not independent and one time point is likely to influence successive time points.[49] It should be noted that the information from each patient might be diluted when comparing the mean, or better the median, of indexes such as PASI in different treatment groups. In addition, the score of patients who leave the study prematurely and are lost to follow-up cannot be evaluated and the intention-to-treat analysis may be difficult to perform. In this respect, the use of simple clinical variables (e.g. the number of total or partial remissions) could be more informative.



Source: Reprinted with permission from Elsevier.

Figure 16.1. Problematic analysis of PASI score over time (from Kragballe K *et al.*[48]). The figure shows a common but unsatisfactory modality of analysing PASI scores collected serially in a RCT. The mean score is calculated at different time points and a graph is presented with lines joining the means at the different time points for the experimental and control group. In the graph, 'errors' bars are attached at each time point and an indicator of statistical significance is placed by each time point to summarise the results of separate significance tests. The curve joining the means may not be a good descriptor of a typical curve for an individual and no account in the analysis is taken of the fact that measurements at different time points are from the same subjects and are likely to be correlated. The number of statistical tests performed and the choice of time points to be tested are additional problematic issues. Further, dividing the results into 'significant' and 'not significant' introduces an artificial dichotomy into serial data.

A remedy has been proposed for the analysis of serial measurements. In the first stage, a suitable summary of the response in each individual, such as a rate of change or an area under the curve, is identified and calculated. Subsequently, these summary measures are analysed by simple statistical techniques.[49]

PHASE I AND PHASE II STUDIES

The ordered development of treatment modalities according to well-identifiable phases[50] is the exception rather than the rule in dermatology. There are several reasons for this situation. Many treatments are non-pharmacological interventions (e.g. ultraviolet phototherapy) which do not need to comply with the regulatory requirements for drug development, and there are no strict guidelines on how to assess them at an early clinical phase. Second, in spite of their being so common, with few exceptions, e.g. psoriasis, the resources allocated to the study of skin disorders are limited as compared with other clinical areas. As a consequence, our understanding of pathomechanisms is limited, as it concern the development of disease-specific therapy. Until the causation of the main skin disorders is unravelled, disparate therapies with imprecisely defined biological activities will continue to be available and many treatments will enter the therapeutic arena serendipitously. This was the case with a renal-transplant recipient with psoriasis whose skin lesions cleared with cyclosporine that led to studies demonstrating the efficacy of that drug.[51] Similar considerations can be made for such treatment modalities as topical vitamin D in psoriasis or the use of minoxidil in androgenetic alopecia. In more recent years, some advancements have been made with our understanding of pathomechanisms for important diseases like psoriasis and atopic dermatitis. The introduction of 'biological agents' in the therapeutic armamentarium of psoriasis partly reflects these advancements. Patent activity around psoriasis continues to accelerate and approaches based on suppression of chemokine activity considered as playing some crucial role in the pathogenesis of the disease are being developed. About 30% of late stage (Phase II or Phase III) development drugs for the treatment of psoriasis are biologics. One aspect to consider in this area is the acceleration of the developmental process with collapsed phases, such as the combination of dose finding phase with efficacy Phase III studies into a single study. The typical design is one where patients are randomised to placebo and different active drug dosages and where the active treatment arms are initially taken separately and then combined together and compared with the placebo arm.[52] Such a practice should be discouraged. It is frequently unclear if the analyses of combined treatment arms were planned in advance. Moreover, the precision of the estimates for efficacy obtained by pooling active treatment arms together is questionable.

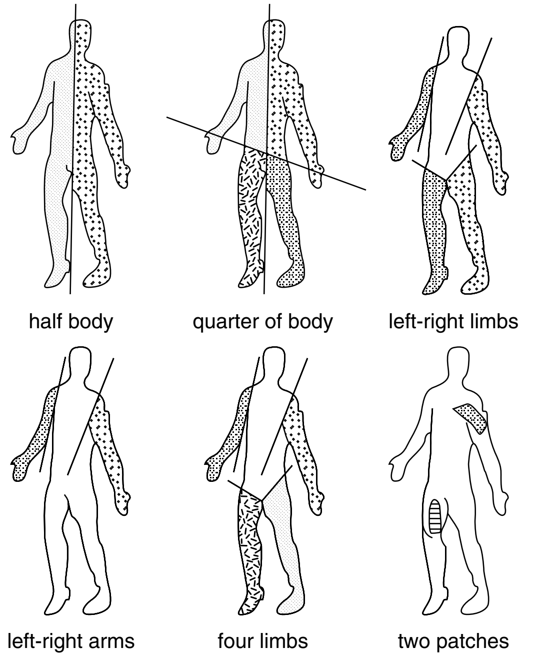
It is widely accepted that a Phase I study is one that examines the initial introduction of a drug in human beings with the treatment tested either in normal volunteers or in patients. The main issues are the pharmacokinetics, pharmacodynamics and tolerability of the drug being tested with a focus on assessing inter-patient variability. While problems with systemic drugs in dermatology do not differ from those usually encountered in other speciality areas, some peculiarities exist with the assessment of topical drugs. Penetration within the deep epidermal layers and dermis is a parameter of particular interest since it clearly affects the local activity of the drug itself. On the other hand, pharmacokinetic parameters describing such a penetration are less stringent as compared with systemic drugs. The assessment can be performed on normal or diseased skin. Relevant methods are those which allow measurements of the concentration of the drug in the skin, in a given time, after topical application, while concentration gradients are formed. Such profiles are usually obtained by direct invasive techniques (e.g. skin biopsy) using topically applied radiolabelled drugs. In some instances, a close correlation has been documented between the barrier function of the horny layer, its reservoir function and the resulting penetration into the skin. Penetration into human skin can thus be predicted from drug quantification in horny layer strippings. This allows non-radioactive methods of drug dosage, like high-performance liquid chromatography, to be applied. Indirect measurements such as urinary excretion or blood levels are also analysed as parameters indicative of the systemic adsorption of the drug and possible toxicity. In many instances, it may be of interest to perform penetration studies in the same patient with the drug being applied on the involved versus the uninvolved skin. Whenever the horny layer barrier is disrupted, penetration within the diseased area is usually facilitated. In addition to adsorption, tolerability of a locally applied drug may be of interest. This is usually evaluated by studying local reactions with increasing concentrations of the drug. All the above-mentioned studies are usually conducted on a few healthy subjects or voluntary patients and in an uncontrolled way. Measurement error is a crucial issue, which needs standardisation and careful evaluation at the design level.

For a limited number of topical drugs, pharmacodynamic parameters have been developed. An example is the blanching or vasoconstriction assay, which has been employed to screen new topical steroids for clinical efficacy. The bioavailability of steroids from topical formulation has been rather improperly defined as the relative absorption efficiency of a drug, as determined by the release of the steroid from its formulation. Its subsequent penetration through the stratum corneum and viable epidermis into the dermis would produce the characteristic blanching effect. This effect is measured through scores that have a subjective component and need careful standardisation. There have also been some attempts to identify biologic pharmacodynamic markers of some chronic skin disorders like psoriasis to be used at an early stage of drug development.[53] However, these indexes are based on cross-sectional studies and there is still limited information on their modifications with disease activity.

According to the Food and Drug Administration (FDA) regulations, a Phase II study is the first controlled clinical study that evaluates the effectiveness of a drug for a given specific therapeutic use in patients. It is also the first study to evaluate the risks of a drug's side effects. Such a study is typically a well-controlled, very closely monitored trial that tests a relatively small, narrowly defined patient population, usually numbering no more than a few hundred. If the criterion is the number of patients recruited, then most RCTs in dermatology would come under this definition.

Study designs that are frequently employed at a preliminary stage in drug development are within-patient control studies, i.e. crossover and self-controlled studies or simultaneous within-patient control studies. In dermatology they are also used, albeit improperly, at a more advanced stage. In a survey of more than 350 published RCTs of psoriasis (unpublished data), a self-controlled design accounted for one-third of all the studies examined and was relied on at any stage in drug development. Crossover studies are studies where patients are randomly allocated to study arms, where each arm consists of a sequence of two or more treatments given consecutively. These trials allow the response of a subject to a given treatment A to be contrasted with the same subject's response to treatment B. There are some inconsistencies with the definition of self-controlled studies provided by different authors. We consider self-controlled studies to be those clinical trials where patients act simultaneously as their own control. A prerequisite for these kinds of studies is the existence of pair organs, e.g. eyes, which can be treated by a locally applied drug in the lack of any significant systemic effect. From our definition we exclude, those studies where a single treatment is administered to patients and a 'before–after' comparison is carried out, and the so-called '*N*-of-one' RCTs, where different time periods are randomised in a single patient to different treatment.

The main advantage of a within-patient study over a parallel concurrent study is a statistical one. A within-patient study obtains the same statistical power with far fewer patients while reducing problems of variability between the populations confronted. Within-patient studies may be useful when studying conditions that are uncommon or show a high degree of patient-to-patient variability. On the other hand, within-patient studies impose restrictions and artificial conditions, which may undermine validity and generalisability of results and may also raise some ethical concern. The wash-out period of a crossover trial as well as the treatment schemes of a self-controlled design, which entails applying different treatments to various parts of the body, do not seem to be fully justifiable from an ethical point of view. In fact, they do not satisfy the principle of providing the patients enrolled in clinical trials with the best-proven diagnostic and therapeutic method. By necessity these studies are restricted to the evaluation of short-term outcomes. A higher degree of collaboration from the patients is requested as compared with other study designs. Clearly, the impractical treatment modalities in self-controlled studies or wash-out period in crossover studies may be difficult for the patient to accept. In these kinds of studies the number of drop-outs may be higher when compared with parallel group designs. In a survey of 26 self-controlled trials on short-contact dithranol in psoriasis (Figure 16.2), which had a median number of 16 patients (range 5 to 63), half of the trials experienced drop-out.[54] Drop-outs may have more pronounced effects in a within-patient study as compared with other study designs because each patient contributes a large proportion of the total information, and the design is sensitive to departure from the ideal plan. The situation is compounded in self-controlled studies where the dropping out from the study may be caused by observing a difference in treatment effect between the parts in which the patient has been 'split up'. In this case, given that drop-outs are related to a difference in treatment effect between interventions, the estimate of the effect of the intervention could be incorrect and falsely equalised. There are several more problems to be considered. Contamination of treated areas and systemic absorption may complicate the interpretation of self-controlled studies, while crossover studies require that the disease lasts long enough to allow the investigator



Source: Reproduced by permission of Blackwell Publishing Ltd.

Figure 16.2. Design of intra-patient comparison in 26 trials concerning dithranol short-contact therapy of psoriasis (from Naldi L *et al.*[54]).

to expose patients to each of the experimental treatments and measure the response. Also the treatment must be one that does not permanently alter the disease or process under study. Carry-over and period effects may clearly compound the analyses.[55] Generalisability is an issue of concern in within-patient controls. Not only are entry criteria usually greatly restricted, e.g. symmetrical lesions, but also outcome measures need to be selected among those reflecting short-term changes in disease activity. Such issues as patient satisfaction and quality of life are obviously beyond the scope of a self-controlled design. It is surprising that self-controlled designs have been the preferred design in situations like topical immunotherapy of alopecia areata or short-contact therapy of psoriasis where patient satisfaction and maintenance of effects over time (e.g. maintenance of hair restoration to an acceptable extent) are vitally important.

PHASE III TRIALS

From Phase III studies we request randomised trials that gather additional information regarding the effectiveness and safety of a treatment, under conditions which are closer to the usual clinical practice as compared with Phase II trials. They should study those clinical outcomes that are of major interest to physicians and patients (as opposed to those driven by surrogate endpoints) and last longer than Phase II trials. The distinction between Phase II and Phase III trials is blurred in dermatology, where most randomised trials are small and, being short-term, employ surrogate measures in well-selected groups of patients. A few points are worth mentioning when discussing the design of Phase III studies in the area of dermatology.

PATIENT MOTIVATION AND PREFERENCE

It has already been mentioned that one of the main concerns of patients suffering from a skin disorder is the visibility of lesions and, much more than in other areas, the patient self-monitoring his on her disease. Patients' motivations and previous experience are obvious crucial points when entering a trial. Motivations and expectations are likely to influence the clinical outcome of all treatments, but they may have a more crucial role in situations where 'soft endpoints' are of concern, as in dermatology. Commonly, more than 20% of the patients entering RCTs of psoriasis experience improvement on placebo independently of the initial disease extension. Motivations are equally important in pragmatic trials where different packages of management are evaluated, such as in the comparison of a self-administered topical product for psoriasis with hospital-based therapy like phototherapy. Traditionally, motivation is seen as a characteristic of the patient that is assumed not to change with the nature of the intervention. However, it has been argued that it is more realistic to view motivation in terms of the 'fit' between the nature of the treatment and the patient's wishes and perceptions, especially with complex interventions that require the patient's active participation. We have already mentioned that the boundary between disease and non-disease is particularly shady in dermatology. On the other hand, the public is confronted with a great deal of uncontrolled and sometimes misleading or unrealistic messages on how to improve the body's appearance. All in all, there is a need to ensure that patient information and motivations are taken into proper consideration when designing and analysing clinical trials on skin disorders. The issue is not only a matter of 'informed consent.' There is a need to study the influences that determine patients' preferences and to understand how these may affect the outcome of clinical trials. A distinction should be drawn between an informed choice based on factual data – such as a reliable estimate of the risks and benefits of interventions – and attitudes towards treatment based on emotional aspects and preconceptions. In recent years, a number of design variants on the traditional randomised trial have been proposed to take into account the patient's preferences. They include the partially randomised patient preference design and the so-called randomised consent or Zelen design.[56] These designs have never gained wide acceptance and none have ever been used in dermatology. The shift from a paternalistic attitude, whereby enrolment decisions are made by doctors, to the choice freely exercised by individual patients is likely to affect the composition of populations in clinical trials. However, when agreement to enrol is based on patients' preferences for individual treatments, as in the Zelen design, the group assembled is unlikely to mirror the target population of all the eligible patients. There is a need to study the influences that determine patients' preferences and understand how these may influence the final outcome of a trial. In a relatively recent survey, Dutch patients affected by psoriasis considered the safety issue and long-term management as more important than fast clearing.[23] It was also important to them to have a vote in the selection of the treatment. It is worth mentioning that the large majority of RCTs in psoriasis are short-term studies dealing with short-term clearance rates that are assessed by the treating physicians. There is room for testing study designs that allow for different preference assessment strategies.

ENTRY CRITERIA

The definition of the study population is of particular importance in dermatology where large variations in disease severity and different clinical subgroups may exist, e.g. plaque versus guttate psoriasis. In addition, there may be problems with variations in disease severity over time. This is commonly observed with chronic inflammatory skin diseases characterised by a relapsing course such as atopic dermatitis or psoriasis. In situations where a variable time-course of the clinical condition is expected, it may be advisable to proceed with sequential evaluations using standardised criteria to judge the stability of the disease over time. Quite surprisingly, information about the stability of the clinical condition is often neglected in clinical trial reports. A review that focused on the selection of patients with psoriasis examined more than 60 clinical trials between 1988 and 1989[29] and documented that information about the stability of the condition was missing in more than 70% of the studies.

Exclusion criteria have the function of selecting the 'more suitable' patients among all possible candidates (e.g. excluding patients in whom the treatment under investigation is contraindicated). This selection also has the aim of reducing factors of variability in the study population, in order to maximise the chance of detecting and quantifying the treatment effect (e.g. excluding patients who are too young or too old). The best way to provide an account of the selection process is a log that lists the included and excluded patients and specifies the reasons for exclusion. This is rarely found in clinical trial reports concerning skin disorders. An example of how far exclusion criteria may operate and limit the possibility of generalising the study results is offered by a clinical trial on the effectiveness of a Chinese herbal extract called 'Dabao' in the treatment of alopecia androgenetica.[57] Among the 3000 patients available to take part in the trial, only 396 were eventually selected to be randomised in the treatment or placebo group. Such exclusion must be a warning when interpreting the actual effectiveness of Dabao on males affected by alopecia androgenetica. It is quite plausible that a similar selection process may operate in many RCTs concerning skin disorders.

PLACEBO USE

There are still controversies about the use of placebos in RCTs. It is widely accepted that 'in any medical study every patient should be assured of the best proven diagnostic and therapeutic method.' As a consequence, the use of placebos should be proscribed when a 'proven' therapeutic method exists. In spite of these principles, studies which breach the ethical principle are still commonly conducted with the approval of regulatory agencies and institutional review boards. It is widely accepted that placebo-controlled trials have high internal validity, but they may be difficult to apply to clinical practice in situations where alternative interventions of proven efficacy already exist. In these circumstances, the information of clinical value is the effect size of the new intervention as compared with the alternative treatment strategy. The use of placebo may sometimes undermine the validity of the study if the treatment falls short of the patient's expectations, resulting in reduced compliance and a large drop-out rate. Published some years ago was a placebo-controlled trial on the effect of ebastine, an H_1 receptor antagonist, in chronic urticaria.[58] A number of other non-sedative antihistamine drugs of proven efficacy were available when the trial was conducted. One might argue that it is unethical to deprive the patients in the placebo group of any effective therapy, even if only for a limited time (14 days in this study). As a matter of fact, the authors reported a high number of drop-outs due to the lack of effect in the placebo group. A remark on the possible misinterpretation of the results of placebo-controlled trials comes from this study. The authors' conclusion that 'ebastine represents an effective and well tolerated alternative to other non-sedative antihistamine drugs in the treatment of chronic urticaria' is likely to be true but far from proven.

Researchers may have a number of different options for their choice of placebo or comparison intervention in RCTs but, in practice, many regulatory agencies still consider placebo controls as the 'gold standard.' Placebo controls are usually required for the evaluation of symptom relief or short-term modification of disorders of moderate severity even when an alternative treatment is available. The usual but questionable claim that justifies this practice from an ethical point of view is that withholding the active therapy does not necessarily affect the long-term prognosis. The above-mentioned issues of symptomatic relief and moderate severity disorders are commonly encountered in dermatology and, in fact, a large number of placebo-controlled RCTs are conducted in this area even when alternative therapies exist. The results of delaying or withholding the treatment may not be straightforward in dermatology. However, there is no question that an extraordinary large number of similar molecules employed for the same clinical indications can be found in this area. These molecules are mostly assessed in placebo-controlled RCTs rather than in comparative RCTs. Examples include topical steroids, oral antihistamines, antifungal drugs, topical antibacterial drugs. More than 200 treatment modalities were identified in a recent survey of published clinical trials of psoriasis with only a few comparative trials. There is a need to establish criteria for the use of placebo in dermatology. They should be developed with the active and informed participation of the public and should be considered by review boards and regulatory agencies. Pragmatic randomised trials contrasting alternative therapeutic regimens are urgently needed to inform clinical decisions. In many instances, traditional medications have not been fully exploited. For example, the first short-term comparative randomised trial of methotrexate versus cyclosporin in psoriasis was only published in 2003 after several years of use of the tested drugs.[59] Given these delays with development of old medications it is difficult to make reasonable statements regarding the comparative effectiveness profile of new versus old-fashioned medications.

TIME FRAME FOR EVALUATION AND OUTCOME MEASURES IN CONTEXT

This discussion will focus on chronic inflammatory skin disorders like psoriasis or atopic dermatitis. There is a necessary link between the time frame for evaluation and the measures adopted to assess clinical response; therefore the two issues should be dealt with together. Many chronic inflammatory skin diseases do not necessarily have a progressive deteriorating course, but they may vary in severity over time causing problems that are similar to those encountered with many psychiatric disorders and some rheumatic diseases. Whenever a definite cure is not reasonably attainable, it is common to distinguish between short-, intermediate- (usually measurable within months) and long-term outcomes. We have already mentioned that clearing the disease in the short term is different from maintaining clearance over time, and long-term results are not simply predictable from short-term outcomes. Most of the score systems available for skin disorders seem to fit best with the clearance issue. On the other hand, it is not easy to define what represents a clinically significant long-term change in the disease status. This is an even more difficult task than defining outcome for other clinical conditions, such as cancer or ischemic heart disease, where mortality or major hard clinical endpoints (e.g. myocardial infarction) are of particular interest. In the long term, the way the disease is controlled and the treatment side effects are vitally important. It has been documented that compliance with the duration of the treatment is limited and is worst with topical treatments.[23] Measures of the quality of life appear rather attractive. However, what represents an important change for most quality of life measures is imprecisely defined, especially if one considers a long-term time frame for evaluation. Clearly, treatment effects can be seen from different perspectives and several dimensions can be taken into account. However, in view of the limitations of the available measures in the long term, simple and cheap outcome measures applicable in all patients seem to be preferable. These may include the number of patients in remission, the number of hospital admissions or ambulatory consultations, major disease flare-ups. Drop-outs merit special attention. In chronic inflammatory skin diseases that lack hard endpoints, they may strongly reflect dissatisfaction with treatment. Whatever the outcome measure adopted, drop-outs cannot simply be ignored because the patients who do not provide PASI, disability or quality of life scores might be different from those who do. Analysis by randomised group irrespective of subsequent changes is the method recommended for the analysis of clinical trials. This analysis poses special problems when relying on quantitative scores. It is suggested that every effort should be made to ensure that patients have a complete assessment at withdrawal and are followed up. If some categorical outcome variable is also considered, e.g. hospital admission, the relation between the score value at withdrawal and the final outcome may be explored.

OTHER ISSUES

The most precise definition of the profile of an intervention requires a perspective on the risks and benefits, which is wider than the one usually provided by any single RCT. For many chronic skin diseases, efficacy data are derived from short-term RCTs, whereas patients tend to be treated over years. The main issues of safety and long-term effectiveness are usually addressed in the context of observational studies, i.e. Phase IV studies. One of the best examples of such a study is the PUVA follow-up study, a cohort study of more than 1400 patients who had received a first course of PUVA-treatment in 1977. These patients are still being followed up and provide information on disease associations and prognostic factors. The study pointed to a dose-related increased risk of non-melanoma skin cancer in PUVA treated patients. We lack similar studies for many other systemic treatments of

Table 16.3. List of the systematic reviews on skin conditions already available, or in an advanced stage of development, in the Cochrane Library (*Cochrane Skin Group, January 2006*)

Completed reviews
Oral treatments for fungal infections of the skin of the foot
Interventions for guttate psoriasis
Systemic treatments for metastatic cutaneous melanoma
Topical treatments for fungal infections of the skin and nails of the foot
Minocycline for acne vulgaris: efficacy and safety
Interventions for toxic epidermal necrolysis (TEN)
Surgical treatments for ingrowing toenails
Interventions for photodamaged skin
Local treatments for cutaneous warts
Drugs for discoid lupus erythematosus
Laser resurfacing for the improvement of facial acne scar
Antistreptococcal interventions in the treatment of guttate and plaque psoriasis
Chinese herbal medicine for atopic eczema
Interventions for bullous pemphigoid
Interventions for impetigo
Interventions for mucous membrane pemphigoid and epidermolysis bullosa acquisita
Interventions for Rosacea
Interventions for basal cell carcinoma of the skin
Statins and fibrates for preventing melanoma
Topical Vitamin A or its derivatives for treating and preventing napkin dermatitis in infants
Protocols under conversion to reviews
Interventions for chronic palmoplantar pustular psoriasis
Antihistamines for atopic eczema
Complementary therapies for acne
Topical 5 fluorouracil for non-melanoma skin cancer
Interventions for localised excessive sweating
Interventions for Melasma
Interventions for vitiligo
Interventions for preventing occupational hand dermatitis
Interventions for hand eczema
Interventions for alopecia areata
Dietary exclusions for established atopic eczema
Oral isotretinoin for acne
Interventions for actinic keratoses
Excision margins for localised cutaneous melanoma
Antimycobacterials for Buruli ulcer
Interventions for the prevention of non-melanoma skin cancers in high-risk groups
Bacillus Calmette – Guerin (BCG) vaccine for preventing Buruli ulcer
Photodynamic therapy for localised squamous cell carcinoma of the skin

(*continued overleaf*)

Table 16.3. (*continued*)

Interventions to reduce Staphylococcus Aureus for atopic eczema
Psychological and educational interventions for atopic eczema in children
Systemic antifungal therapy for tinea capitis in children
Disposable nappies for the prevention of napkin dermatitis in infants
Educational programmes for skin cancer prevention
Oral intake of evening primrose oil and borage oil for atopic eczema
Interventions for morphea
Laser and photoepilation for unwanted hairgrowth
Interventions for cellulitis and erysipelas
Traditional Chinese herbs for psoriasis
Treatments for molluscum contagiosum in children
Chemoimmunotherapy versus chemotherapy for metastatic malignant melanoma
Interventions for skin changes caused by nerve damage in leprosy
Interventions for polymorphic light eruption
Interventions for mucocutaneous leishmaniasis
Topical treatments for chronic plaque psoriasis
Oral treatments for onychomycosis
Topical pimecrolimus for atopic dermatitis
Methotrexate for psoriasis
Interventions for solitary or limited cutaneous leishmaniasis
Interventions for pityriasis rosacea
Fumaric acid esters for psoriasis
Dietary supplements for established atopic eczema

psoriasis, including methotrexate, retinoids and cyclosporin. The safety profiles of most systemic antihistamines is also imprecisely defined. Observational studies may represent the most feasible way to study the usefulness of long-term treatment strategies for chronic inflammatory skin diseases, when disease modification rather than symptom control becomes a desired outcome. As has been proposed for some rheumatologic disorders, e.g. rheumatoid arthritis, drug survival, e.g. the interval during which individual patients remain on an agent may offer an indication for long-term acceptability that takes into account adverse effects, lack or loss of effect and patients' preference. The introduction of biological agents has prompted the development of some new registries to assess long-term safety and effectiveness, like the one being started by the British Association of Dermatologists[60] and the Psocare initiative in Italy (www.psocare.it).

A final mention should be made for those activities that aim at summarising the results of several RCTs on the same issue. There is a large burden of small RCTs[61] addressing disparate clinical questions, as well as a lack of consensus on the management of many skin disorders. This creates an increasing emphasis on systematic reviews, and a Cochrane Skin Group has been established within the Cochrane Collaboration in 1997. A list of systematic reviews already available within the Cochrane Library is reported in Table 16.3.

In the light of the increasing role systematic reviews may play with informing clinical practice, special care should be devoted to setting priorities so that the most important questions are addressed first. Otherwise, they would risk amplifying the irrelevant issues. The importance of the involvement of consumers cannot be underestimated. The first analyses from systematic assessment of published RCTs point to some 'peculiarities' of dermatology that have already been discussed in the previous sections, and include among others:

1. The 'moving boundary' between cosmetology and medicine.
2. The need to develop study designs that address questions posed by chronic recurrent diseases.
3. The limitations of available outcome measures that neglect patients' needs and expectations.
4. Problems with external generalisability like the lack of adequate description of the study populations and study settings.
5. The lack of comparative RCTs.
6. The overwhelming role of pharmaceutical industries with defining priorities.

REFERENCES

1. Jobling R. Therapeutic research into psoriasis: patients' perspectives, priorities and interests. In: Rawlins M, Littlejohns P, eds, *Delivering Quality in the NHS 2005*. Oxford: Radcliffe Publishing (2005) 53–6.
2. McKenna KE, Stern RS. The outcomes movement and new measures of the severity of psoriasis. *J Am Acad Dermatol* (1996) **34**: 534–8.
3. Jobling R. Psychosocial issues in dermatology. In: Hughes E, van Onselen J, eds, *Dermatology Nursing. A practical guide*. Edinburgh: Churchill Livingstone (2001) 93–101.
4. Papier A, Chalmers RJG, Byrnes JA, Goldsmith LA. Framework for improved communication: The Dermatology Lexicon Project. *J Am Acad Dermatol* (2004) **50**: 630–4.
5. Wolkenstein P, Grob JJ, Bastuji Garin S, Ruszczynski S, Roujeau JC, Revuz J. French people and skin disease: results of a survey using a representative sample. *Arch Dermatol* (2003) **139**: 1614–19.
6. Webster GF. Is dermatology slipping into its anecdotage? *Arch Dermatol* (1995) **131**: 149–50.
7. Vermeer BJ, Gilchrest BA. Cosmeceuticals – a proposal for rational definition, evaluation, and regulation. *Arch Dermatol* (1996) **132**: 337–40.
8. Peckham PE, Weinstein GD, McCullough JL. The treatment of severe psoriasis: a national survey. *Arch Dermatol* (1987) **123**: 1303–7.
9. Farr PM, Diffey BL. PUVA treatment of psoriasis in the United Kingdom. *Br J Dermatol* (1991) **124**: 365–7.
10. Doshi A, Zaheer A, Stiller MJ. A comparison of current acne grading systems and proposal of a novel system. *Int J Dermatol* (1997) **36**: 416–8.
11. Motley RJ, Finlay AY. Practical use of a disability index in the routine management of acne. *Clin Exp Dermatol* (1992) **17**: 1–3.
12. Zeigher RS, Heller S, Mellon MH, Forsythe AB, Hamburger RN, Schatz M. Effect of combined maternal and infant food-allergen avoidance on development of atopy in early infancy: a randomized study. *J Allergy Clin Immunol* (1989) **84**: 72–89.
13. Arshad SH, Matthews S, Gant C, Hide DW. Effect of allergen avoidance on development of allergic disorders in infancy. *Lancet* (1992) **339**: 1493–7.
14. Naldi L, Chatenoud L, Linder D, Belloni A, Peserico A, Virgili AR, Bruni PL, Ingordo V *et al*. Cigarette smoking, body mass index, and stressful life events as risk factors for psoriasis: results from an Italian case-control study. *J Invest Dermatol* (2005) **125**: 61–7.
15. Stern RS. A promising step forward in psoriasis therapy. *J Am Med Assoc* (2003) **290**: 3133–5.
16. Zenz R, Eferl R, Kenner L, Florin L, Hummerich L, Mehic D, Scheuch H, Angel P, Tschachler E, Wagner EF. Psoriasis–like skin disease and arthritis caused by inducible epidermal deletion of Jun proteins. *Nature* (2005) **437**: 369–75.
17. Frederiksson AJ, Peterssonn DC. Severe psoriasis: oral therapy with a new retinoid. *Dermatologica* (1978) **157**: 238–44.
18. Phillips T, Stanton B, Provan A, Lew R. A study of the impact of leg ulcers on quality of life: financial, social, and psychological implications. *J Am Acad Dermatol* (1994) **31**: 49–53.
19. Moffatt CJ, Franks PJ, Oldroyd M, Bosanquet N, Brown P, Greenhalgh RM, McCollum CN. Community clinics for leg ulcer and impact on healing. *Br Med J* (1992) **305**: 1389–92.
20. Lehmann HP, Robinson KA, Andrews JS, Holloway V, Goodman SN. Acne therapy: a methodologic review. *J Am Acad Dermatol* (2000) **47**: 231–40.
21. Naldi L, Svensson A, Diepgen T, Elsner P *et al*. Randomized clinical trials for psoriasis 1977–2000. The EDEN survey. *J Invest Dermatol* (2003) **120**: 738–41.
22. Van Coevorden AM, Coenraads PJ, Svensson A, Bavinck JN, Diepgen TL, Naldi L, Elsner P, Williams HC. European Dermato-Epidemiology Network (Eden) (2004). Overview of studies of treatments for hand eczema – the EDEN hand eczema survey. *Br J Dermatol* (2006) **151**: 446–51.
23. Van de Kerkhof PCM, De Hoop D, De Korte J, Cobelens SA, Kuipers MV. Patient Compliance

and disease management in the treatment of psoriasis in the Netherlands. *Dermatology* (2000) **200**: 292–8.

24. Day S. Commentary: treatment allocation by the method of minimisation. *Br Med J* (1999) **319**: 947–8.
25. Boateng F, Sampson A, Schwab B. Statistical analysis of possible bias of clinical judgements due to observing an on-therapy marker variable. *Stat Med* (1996) **15**: 1747–55.
26. Christiansen JV, Holm P, Moller R, Reymann F and Schmidt H. Etretinate (Tigason) and betamethasone valerate (Celeston valerate) in the treatment of psoriasis. A double-blind, randomized, multicenter trial. *Dermatologica* (1982) **165**: 204–7.
27. Weinstein GD. Safety, efficacy and duration of therapeutic effect of tazarotene in the treatment of plaque psoriasis. *Br J Dermatol* (1996) **135**(suppl): 32–36.
28. Spuls PL, Witkamp L, Bossuyt PM, Bos JD. The course of chronic plaque psoriasis in placebo groups of randomised clinical trials. *Arch Dermatol* (2004) **140**: 338–44.
29. Petersen LJ, Kristensen JK. Selection of patients for psoriasis clinical trials: a survey of the recent dermatological literature. *J Dermatol Treat* (1992) **3**: 171–6.
30. Ashcroft DM, Li Wan Po A, Williams HC, Griffiths CEM. Clinical measures of severity and outcome in psoriasis: a critical appraisal of their quality. *Br J Dermatol* (1999) **141**: 185–91.
31. Gupta MA, Johnson AM, Gupta AK. The development of an Acne Quality of Life scale: reliability, validity, and relation to subjective acne severity in mild to moderate acne vulgaris. *Acta Dermato-Venereol* (1998) **78**: 451–6.
32. Finlay AY. Quality of life measurement in dermatology: a practical guide. *Br J Dermatol* (1997) **136**: 305–14.
33. Charman C, Williams H. Outcome measures of disease severity in atopic eczema. *Arch Dermatol* (2000) **136**: 763–9.
34. Emerson RM, Charman CR, Williams HC. The Nottingham Eczema Severity Score: preliminary refinement of the Rajka and Langeland grading. *Br J Dermatol* (2000) **142**: 288–97.
35. Ashcroft DM, Li Wan Po A, Williams HC, Griffiths CEM. Quality of life measures in psoriasis: a critical appraisal of their quality. *J Clin Pharm Ther* (1998) **23**: 391–8.
36. Kantor J, Margolis DJ. Efficacy and prognostic value of simple wound measurements. *Arch Dermatol* (1998) **134**: 1571–4.
37. Lindholm C, Bjellerup M, Christensen OB, Zederfeldt B. Quality of life in chronic leg ulcer patients. An assessment according to the Nottingham Health Profile. *Acta Dermato-Venereol* (1993) **73**: 440–3.
38. Faust HB, Gonin R, Chuang TY, Lewis CW, Melfi CA, Farmer ER. Reliability testing of the dermatology index of disease severity (DIDS). An index for staging the severity of cutaneous inflammatory disease. *Arch Dermatol* (1997) **133**: 1443–8.
39. Chren MM, Lasek RJ, Quinn LM, Mostow EN, Zyzanski SJ. Skindex, a quality of life measure for patients with skin disease: reliability, validity, and responsiveness. *J Invest Dermatol* (1996) **107**: 707–13.
40. Marks R, Barton SP, Shuttleworth D, Finlay AY. Assessment of disease progress in psoriasis. *Arc Dermatol* (1989) **125**: 235–40.
41. Feldman SR, Fleischer AB Jr, Reboussin DM, Rapp SR, Exum ML, Clark DR, Nune L. The self-administered psoriasis area and severity index is valid and reliable. *J Invest Dermatol* (1996) **106**: 183–6.
42. Krueger GG, Feldman SR, Camisa C, Duvic M, Elder JT, Gottlieb AB *et al*. Two considerations for patients with psoriasis and their clinicians: what defines mild, moderate and severe psoriasis? What constitutes a clinically significant improvement when treating psoriasis? *J Am Acad Dermatol* (2000) **43**: 281–3.
43. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol* (2000) **53**: 459–68.
44. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* (2004) **7**: S22–6.
45. Nijsten T, Sampogna F, Chren MM, Abeni D. Testing and reducing the Skindex-29 using Rasch analysis: Skindex-17. *J Invest Dermatol* (2006) **126**(6): 1244–50.
46. Unaeze J, Nijsten T, Murphy A, Ravichandran C, Stern RS. Impact of psoriasis on health-related quality of life decreases over time: an 11-year prospective study. *J Invest Dermatol* (2006) **126**: 1480–9.
47. Norman GR. Issues in the use of change scores in randomized trials. *J Clin Epidemiol* (1989) **42**: 1097–1105.
48. Kragballe K, Gjertsen BT, De Hoop D, Karlsmark T, van de Kerkhof PC, Larko O, Nieboer C, Roed-Petersen J, Strand A, Tikjob G. Double-blind, right/left comparison of calcipotriol and betamethasone valerate in treatment of psoriasis vulgaris. *Lancet* (1991) **337**: 193–6.

49. Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *Br Med J* (1990) **300**: 230–5.
50. Temple R. Current definitions of phases of investigation and the role of the FDA in the conduct of clinical trials. *Am Heart J* (2000) **139**: S133–5.
51. Krueger GG. Psoriasis therapy – observational or rational? *New Eng J Med* (1993) **328**: 1845–6.
52. Leonardi CL, Powers JL, Matheson RT, Goffe BS, Zitnik R, Wang A, Gottlieb AB. Etanercept Psoriasis Study Group. Etanercept as monotherapy in patients with psoriasis. *New Engl J Med* (2003) **349**: 2014–22.
53. Gottlieb AB. Product development for psoriasis: clinical challenges and opportunities. In: Roenigk HH, Maibach HI, eds, *Psoriasis*, 3rd ed. New York: Marcel Dekker (1998) 421–33.
54. Naldi L, Carrel CF, Parazzini F, Cavalieri d'Oro L, Cainelli T. Development of anthralin short-contact therapy in psoriasis: survey of published clinical trials. *Int J Dermatol* (1992) **31**: 126–30.
55. Louis TA, Lavori PW, Bailar JC, Polansky M. Crossover and self-controlled designs in clinical research. *New Engl J Med* (1984) **310**: 24–31.
56. Lambert MF, Wood J. Incorporating patient preferences into randomized trials. *J Clin Epidemiol* (2000) **53**: 163–6.
57. Kessels AG, Cardynaals RL, Borger RL, Go MJ, Lambers JC, Knottnerus JA, Knipschild PG. The effectiveness of the hair-restorer 'Dabao' in males with alopecia androgenetica. A clinical experiment. *J Clin Epidemiol* (1991) **44**: 439–47.
58. Peyri J. Ebastine in chronic urticaria: a double-blind placebo-controlled study. *J Dermatol Treat* (1991) **2**: 51–3.
59. Heydendael VM, Spuls PI, Opmeer BC, de Borgie CA, Reitsma JB, Goldschmidt WF, Bossuyt PM, Bos JD, de Rie MA. Methotrexate versus cyclosporine in moderate-to-severe chronic plaque psoriasis. *New Engl J Med* (2003) **349**: 658–65.
60. Smith CH, Anstey AV, Barker JNWN, Burden AD, Chalmers RJ, Chandler D *et al*. British Association of Dermatologists guidelines for use of biological interventions in psoriasis. *Br J Dermatol* (2005) **153**: 486–97.
61. Williams HC, Seed P. Inadequate size of 'negative' clinical trials in dermatology. *Br J Dermatol* (1993) **128**: 317–26.

INFECTIOUS DISEASES

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

17

HIV Infection

MICHAEL D. HUGHES
Center for Biostatistics in AIDS Research, Harvard School of Public Health, Boston, Massachusetts, USA

INTRODUCTION

The epidemic of human immunodeficiency virus (HIV) infection was estimated to affect about 38 million people worldwide in 2003, including about 5 million newly infected people in that year alone.[1] About 25 million people living in sub-Saharan Africa were HIV-infected. In the same year, almost 3 million people died from the acquired immunodeficiency syndrome (AIDS), the later stage of HIV infection characterised by advanced immunodeficiency and significant clinical disease. Over 20 million people have died from the disease since the first cases of AIDS were identified in 1981. In the face of this increasing epidemic, considerable research has been undertaken to develop and evaluate interventions for both prevention of transmission and treatment of infection. Clinical trials continue to be critically important in this research and considerable progress has been made, particularly in the development of treatments. In this chapter, the focus is on trials that evaluate antiretroviral drugs including Phase I to III trials and treatment management trials, with a brief discussion of other issues related to antiretroviral trials including the use of antiretroviral drugs to prevent transmission of HIV. It should, however, be appreciated that many trials, particularly Phase I and II trials, have been conducted to evaluate other treatments for HIV-infected people such as immune-based therapies and therapeutic vaccines, as well as treatments and prophylaxes for the opportunistic diseases that characterise AIDS, and treatments for the adverse effects associated with the use of antiretroviral drugs. In addition, clinical trials have been important in evaluating other approaches for prevention of transmission of HIV including, for example, candidate vaccines,[2] microbicides[3,4] and behavioural interventions[5–7].

HIV AND ITS TREATMENT

In this section, some background relevant to the design of trials involving HIV-infected people is provided. The virus and the mode of action of the multiple treatments available for HIV infection are briefly described. Then the primary marker of disease status used in trials, the level of HIV RNA in the host subject, is described and critiqued as a surrogate endpoint. Finally, there is a brief description of one of the major challenges

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

to treatment, specifically the development of viral resistance to drugs.

HIV is a retrovirus – its genetic material is stored as a single strand of RNA which is converted to DNA in a process known as reverse transcription after the viral RNA has been released into one of a human host's CD4+ T lymphocytes (described hereafter, for short, as a CD4 cell). Subsequent steps of the replication process within the CD4 cell take advantage of the host's genetic machinery to produce new virus. Antiretroviral drugs for the treatment of HIV attempt to inhibit different steps in the replication cycle. The first two classes of drug that became available, nucleoside and non-nucleoside reverse transcriptase inhibitors (NRTIs and NNRTIs), and the third class, protease inhibitors (PIs), target steps in the cycle which occur within the CD4 cell. More recently developed classes, including fusion inhibitors (FIs) and CCR5 inhibitors, include drugs that attempt to inhibit entry of the virus into the CD4 cell. In the late 1980s, zidovudine (ZDV, also known as AZT), an NRTI, was the first antiretroviral drug shown to delay progression of AIDS. However, the standard of care since about 1996 has been to use three (or sometimes more) drugs, from at least two drug classes, in combination. Most commonly, such 'highly active antiretroviral therapy' (HAART) includes two NRTIs and either an NNRTI or a PI. By mid-2005, the Food and Drug Administration of the United States had approved eight NRTIs, three NNRTIs, eight PIs and one FI for the treatment of HIV, as well as additional drug products that combined some of these drugs. With so many drugs available, there are potentially numerous HAART combinations available. This diversity of regimens is an important issue for trial design.

Since the mid-1990s, assays have been available which allow measurement of viral levels in various body fluids. In clinical trials, it is most commonly measured in plasma and expressed as the number of copies of HIV RNA per millilitre of plasma (copies/ml), and changes in HIV RNA are usually expressed in terms of log reductions ($\log_{10}$ copies/ml). There are a number of assays available with different sensitivities at low levels of viral load. The most widely used assays have lower limits of quantification of 50 or 400 copies/ml. Below these thresholds, virus may not be detectable or, if detectable, the precision of quantification is poor. In HIV-infected subjects who have not been previously treated with antiretroviral therapy, the proportion of subjects who achieve viral loads below these thresholds is often in excess of 80%, thus making it difficult to quantify treatment effects on HIV RNA levels in this population.

It is well established that higher levels of HIV RNA and lower CD4 cell counts are predictive of increased risk of clinical disease progression (e.g. to AIDS) and of death in both the absence and the presence of treatment.[8,9] It is also clear that the reductions in viral load, when sustained, and the subsequent improvements in CD4 cell count achieved with HAART have resulted in substantial reductions in the risk of progression to AIDS and of death which can be seen readily in national surveillance data.[10,11] These facts, coupled with a biological rationale for the mechanisms by which treatment effects on these markers impact disease progression, the very low incidence of clinical endpoints in trials (i.e. new AIDS-defining events and deaths), and the widespread use of these markers for treatment management in clinical practice, have led to the almost exclusive use of HIV RNA level as a primary endpoint in clinical trials (including Phase III trials), with CD4 cell count being important as a secondary endpoint. Drug approval of antiretroviral drugs is also based on establishing effects on these parameters.[12] However, it is important to appreciate that these are being used as surrogate endpoints for clinical endpoints. Thus, while it is reasonably clear that the substantial effects that HAART has on these markers is associated with substantial effects on the incidence of clinical events, data from a large meta-analysis show that small differences between treatments in these marker levels are less reliable as indicators of the magnitude of the corresponding difference in the risk of progression to AIDS or death.[10,11]

A major issue in the development and use of an antiretroviral drug concerns the development of viral resistance to that drug, which may also lead to cross-resistance to some or possibly all other drugs in the same class. Resistance arises when an error in viral replication produces a mutation in the viral genome such that the mutant form of the virus is less susceptible to the drug. For some drugs, notably NNRTIs, a mutation at a single codon may lead to a high level of resistance, whereas for other drugs, successive mutations may confer increasing levels of resistance. Because the turnover of virus is rapid – about one-half of the viral population in a host is cleared and replaced daily – and the replicative process is highly error prone, resistance is a common problem. The risk of development of resistance increases with higher levels of virus in the host and so is exacerbated by factors such as partial adherence to therapy, suboptimal dosing, and the combination of a new drug with drugs to which the host's viral population already shows resistance.

PHASE I/II CLINICAL TRIALS

The goal of Phase I and II clinical trials of a new antiretroviral drug is, ultimately, to identify a dose of a drug which shows adequate activity against HIV and acceptable safety to be taken forward to longer-term randomised evaluation in Phase III testing. The labelling of trials as Phase I or Phase II is, however, not well standardised. In general, initial trials ('Phase I') of an antiretroviral drug are often short-term monotherapy studies involving HIV-infected subjects. These are designed to evaluate the short-term antiviral effects and pharmacokinetics of different doses of the drug, while also providing a preliminary assessment of safety parameters.[13] These are followed by longer-term ('Phase II') trials which evaluate a drug in combination with other antiretroviral drugs.

Although a drug would be used as part of a combination therapy in clinical practice, the purpose of studying a drug as a monotherapy is to establish that it does indeed have adequate antiviral activity. Because antiviral effects are rapid, this can readily be demonstrated in trials of one to two weeks' duration. Commonly, these trials will be of the dose-finding or dose escalation type whereby successive cohorts of subjects are assigned increasing doses of the drug.[13,14] Typical cohort sizes are 4 to 10 subjects. They may explore different schedules of dosing, such as once versus twice daily, as well as different dose weights or routes of administration. The lowest dose studied is chosen based on *in vitro* studies such that the dose is expected to have adequate antiviral activity and so minimise the risk of exposing a subject to the development of viral resistance to the drug. This is a key issue as such resistance may rule out not only a future treatment option for the patient using this drug but also future treatment options using other drugs in the same class because of cross-resistance. Increases in dose occur only when a sufficient number of subjects have completed treatment at the lower dose without unacceptable levels of toxicity. However, these trials are not generally designed to detect a maximum tolerated dose (as, for example, in the cancer setting) because the dose–response curves for antiviral effects of available antiretroviral drugs have tended to plateau at levels of toxicity which are acceptable for a life-threatening disease such as HIV infection. The duration of treatment in these monotherapy studies is deliberately short to minimise the risk of resistance as the antiviral activity of monotherapy will be markedly less than that of the combination therapies used in clinical practice. An adaptation of this type of design is to add the new drug to an existing combination antiretroviral therapy that is maintaining a stable level of virus in a patient but is not achieving suppression of virus to the levels detectable using current assays. Dynamic designs which aim to identify a dose which maximises the proportion of subjects who achieve adequate viral suppression without toxicity have also been developed.[15]

Phase II trials of an antiretroviral drug are longer-term evaluations typically involving 48 weeks of treatment in combination with other antiretroviral drugs. Most Phase II trials are randomised and often are used to evaluate two or three doses of the new drug. Two examples illustrate common designs. The first involved a comparison of three doses of a drug, enfuvirtide (the first approved fusion inhibitor), added to a 'background' antiretroviral regimen comprising of four specified drugs, compared with a control group which received only the same background regimen.[16] This design allowed evaluation of the additional antiviral effect of enfuvirtide over and above the standard background regimen. It is therefore a superiority design in the hope that it will show that there is an additional antiviral effect. The second example involved a comparison of two doses of a new PI, atazanavir, versus another drug from the same drug class, nelfinavir, when both atazanavir and nelfinavir were used in combination with two specified NRTIs.[17] In this particular study, the design was an equivalence design in that the aim was to show that the effects of atazanavir and nelfinavir were similar when used in combination with the same two NRTIs. The same type of design could be used in the superiority setting if the intent was to establish that atazanavir improved viral suppression compared with nelfinavir when both drugs are used in combination with the same two NRTIs.

Hybrid Phase I/II designs are also used. An example involved randomising subjects who had not been previously treated with antiretroviral therapy to one of three doses of atazanavir or to an established dose of nelfinavir, taken as monotherapy for two weeks. After two weeks, all subjects then added the same two NRTIs and continued treatment for a further 46 weeks.[18] The monotherapy period provides a direct assessment of the similarity of the antiviral effects of atazanavir and nelfinavir, while the longer-term follow-up provides a comparison of effects in a combination regimen as would be used in clinical practice (with the caveat that all three drugs in the regimen would be initiated simultaneously in practice). This study also used a two-stage design. In Stage 1, akin to a Phase I trial, 87 subjects were randomised to the four arms and then enrolment to the trial was put on hold until all subjects had completed at least four weeks of treatment and the data had been analysed. This provided an opportunity to confirm the safety and antiviral effects of atazanavir before proceeding to a larger Stage 2, akin to a Phase II trial, in which a further 272 subjects were randomised. A similar hybrid design, without the staging, has also been used to evaluate a new CCR5 inhibitor in subjects who were already on antiretroviral therapy but for whom the regimen was failing. In this design, subjects were randomised to receive one of three doses of the new CCR5 inhibitor or a matching placebo to be added to their existing antiretroviral regimen for two weeks. This two-week 'add-on' period provides a direct assessment of the antiviral activity of the new CCR5 inhibitor. However, in contrast to the atazanavir study for subjects with no prior treatment, after two weeks, all subjects changed to a new background antiretroviral regimen while continuing their randomised CCR5 inhibitor dose or placebo. This change in background regimen was done in recognition that, in clinical practice, when drugs need to be changed because a subject is on a failing regimen, it is desirable to maximise the chance of achieving good viral suppression and hence minimise the risk of development of further viral resistance to drugs by changing at least two drugs.

PHASE III CLINICAL TRIALS

Phase III clinical trials involve randomised comparisons of combinations of antiretroviral drugs and are undertaken at multiple clinical centres often in many countries. They are therefore active-controlled trials. Regimens may be partially blinded through the use of placebos. As an example, a recent trial compared two three-drug combinations, zidovudine plus lamivudine plus abacavir (ZDV+3TC + ABC) and zidovudine plus lamivudine plus efavirenz (ZDV+3TC+EFV), with a four-drug combination, ZDV+3TC + ABC + EFV.[19] Both

ZDV+3TC + ABC and ZDV+3TC are used in clinical practice as combinations of drugs in a single pill. Patients in all three arms therefore took placebos: for ZDV+3TC and EFV, for ZDV+3TC + ABC and for ZDV+3TC in the three arms, respectively. This achieves blinding over the drugs that differ between the three arms and hence reduces the possibility for bias due to differences in treatment management, outcome assessment or patient preferences that affect retention on randomised treatment that may arise in the absence of blinding. These are significant issues in HIV trials. However, the disadvantage is that the comparison is less 'real-life' in that there is an additional pill burden in all three arms, which also differs between the three arms, compared with how the drugs will be used in clinical practice. This is an important consideration in some HIV trials as a potential benefit of some regimens may be achieved by lower pill numbers or reduced dosing (e.g. daily versus twice daily), leading to improved adherence and viral suppression and hence also lower risk of development of resistance.

Besides defining the interventions to be evaluated, two other key considerations in randomised trials are the definitions of the eligible population and the outcome measures for comparing interventions. For HIV trials, a major factor in defining the eligible population is whether or not subjects have previously used antiretroviral therapy. Most trials will restrict entry either to treatment-naive subjects or to treatment-experienced subjects. This is because viral resistance to drugs in treatment-experienced subjects may impose a substantial restriction on the drugs which are likely to be efficacious for any particular subject, whereas this issue is much less important when considering treatment-naive subjects (though it is possible that a treatment-naive subject could have been infected with a virus that shows resistance). This distinction according to prior treatment also impacts the definition of the primary outcome measure used in trials. Consideration of these two distinct populations is therefore given below in more detail. Other key eligibility criteria used in trials generally reflect disease status including HIV RNA level, CD4 count and clinical status such as having previously been affected by an AIDS-defining condition.

TRIALS ENROLLING TREATMENT-NAIVE PATIENTS

A major issue in trials of treatment-naive patients concerns the choice of primary efficacy outcome measure. HAART regimens recommended for initial treatment of HIV-infected patients achieve long-term sustained suppression of HIV RNA below 400 copies/ml in a large proportion of subjects while taking the initial treatment. For example, in one pivotal trial (the GS-903 Study), about 80%, 75% and 70% of patients were 'responders' after 48, 96 and 144 weeks, respectively.[20] In the United States, regulatory approval for an antiretroviral drug requires efficacy results over 48 weeks of treatment.[12] The approximately 20% of subjects who were 'non-responders' within the first 48 weeks of treatment in the aforementioned study included about 5% who were virologic failures (never achieving HIV RNA < 400 copies/ml or experiencing rebound from <400 copies/ml), about 6% who discontinued randomised treatment due to adverse events, <1% who died, and about 8% who discontinued treatment due to loss to follow-up, patient withdrawal and other reasons.[21] Thus, only about one-quarter of the non-responders within the first 48 weeks of treatment were due to virologic failure. Losses to follow-up, patient withdrawals and other reasons which often reflect patient preferences or decisions leading to treatment discontinuation constituted the major category of non-response. This is typical of other trials.[22] Caution therefore needs to be taken when reviewing results of trials as these results are often described in terms of 'time-to-loss-of-virologic-response' (TLOVR; terminology used in an FDA Guidance document)[12] but the 'losses of virologic response' are often dominated by non-virologic reasons. For this reason, 'treatment failure' might be better terminology, particularly if the reasons described as patient preference primarily reflect issues such as intolerance of treatment due to low-grade side effects such as nausea etc.

So-called 'as-treated' analyses may also be important for understanding the strength of conclusions from a study concerning virologic efficacy when the primary outcome measure is treatment failure. One form of as-treated analysis evaluates time to virologic failure with censoring of follow-up for subjects who discontinue randomised treatment prior to virologic failure. Although such an analysis may be biased when discontinuation might be related to virologic outcome, discordance of results from this analysis and an analysis of treatment failure should be explained.[22] An example of this is a trial that compared tenofovir plus emtricitabine with zidovudine plus lamivudine (ZDV+3TC) when both were used in combination with efavirenz. Preliminary 24-week results showed a significant difference in the percentage of patients without treatment failure (73% versus 65%) but an as-treated analysis showed no difference between the groups in virologic suppression.[23] The explanation was a higher rate of adverse events and an associated higher rate of treatment discontinuation in the ZDV+3TC arm.

One reason that the treatment failure composite endpoint is advocated is that it provides for a valid intention-to-treat (ITT) analysis because it is well defined for all randomised patients, whether or not they have met the endpoint, including patients who discontinue from the trial early. An alternative approach is to evaluate the time to virologic failure irrespective of whether randomised treatment is being taken at the time of failure. This endpoint provides for a valid ITT analysis (except that deaths and losses to follow-up need consideration in the analysis), but is sometimes criticised because some of the virologic failures will occur after discontinuation of randomised treatment (for reasons such as toxicity and tolerability as well as patient preferences) and hence a differential between randomised treatments may reflect the potency of subsequent treatments received rather than the potency of the randomised treatments. However, use of this endpoint answers an important treatment management question concerning which antiretroviral treatment should be preferred as initial treatment of HIV infection for maximising time until first virologic failure, allowing for the fact (as in clinical practice) that some subjects will switch treatments for a variety of reasons prior to failure. Losses to follow-up and deaths prior to virologic failure complicate the interpretation of results of analyses using this endpoint, particularly if there is a difference in the conclusions that might be drawn from an analysis in which follow-up is censored at the time of loss to follow-up or death versus those that might be drawn from an analysis in which a composite endpoint of time to virologic failure, loss to follow-up or death is used. In practice, irrespective of the definition of a virologic failure endpoint used as a study's primary endpoint, it is recommended that trials follow subjects after discontinuation of study treatment so that analyses of both 'treatment failure' and 'virologic failure' endpoints can be undertaken.[24] Consistency between the conclusions that can be drawn from a study about an initial treatment based upon the different endpoints makes the interpretation more straightforward. Inconsistency in the conclusions requires further exploration to understand, if possible, how the relative numbers of patients discontinuing the different randomised treatments according to the types of reason for discontinuation might explain the inconsistency.

Regardless of whether a treatment failure or a pure virologic failure endpoint is used, the low rates of failure over the typical durations of trials (one to three years) mean that it is difficult to anticipate substantial improvements in the failure rate. Trials of new antiretroviral drugs are, therefore, typically designed as non-inferiority trials with the aim of demonstrating that the rate of non-response is not inferior within some defined margin to a standard of care regimen. In the aforementioned GS-903 Study, a margin of 10% was used.[20] However, it is arguable whether a regimen with a true response rate of 70% at 48 weeks would be considered non-inferior to one with a true rate of 80% and so smaller margins, e.g. about 5% at 48 weeks, might be used. For example, in a study of the AIDS Clinical Trials Group scheduled to open in

2005, the margin (expressed formally in terms of a hazard ratio for a pure virologic failure endpoint) corresponds to differences of about 5% and 10% at 48 and 96 weeks, with anticipated rates of failure of 17% and 32%, respectively. A further important consideration in non-inferiority (or equivalence) HIV trials is the extent to which treatment switches and discontinuations affect the conclusions. When a treatment failure endpoint is being used, a conclusion of non-inferiority or equivalence might be driven by a high rate of subjective reasons for discontinuation and the concern is whether these reflect what might be found in clinical practice. Alternatively, such a conclusion might be obtained in the face of quite different distributions of types of endpoints, e.g. a high rate of virologic failures and a low rate of treatment-related adverse events for one treatment with the reverse occurring for another treatment.

TRIALS ENROLING TREATMENT-EXPERIENCED SUBJECTS

The major difference between trials enrolling treatment-experienced subjects and those enrolling treatment-naive subjects is that many treatment-experienced subjects will have experienced virologic failure due to the development of viral resistance to one or more drugs often involving multiple drug classes. This restricts new treatment options that are effective in suppressing virus, a problem which increases as patients experience virologic failure on each successive treatment regimen that they take. For this reason, trials typically use a quantitative change in HIV RNA from pre-treatment levels as the primary efficacy outcome measure. As noted earlier, however, this has limitations as it is unclear as to the extent to which modest differences in HIV RNA, particular over short periods of time such as 24 weeks, are a surrogate for longer-term differences in patient-relevant clinical outcomes.[11] In trials including patients with less extensive treatment experience, as in trials including treatment-naive patients, sustained suppression of HIV RNA to below limits of assay detection is likely to be a more appropriate outcome measure.

Given the large number of antiretroviral drugs available and the numerous combinations of these drugs that are possible, it is difficult to conduct trials involving treatment-experienced subjects which compare specific drug combinations. To overcome this, trials have used a design in which a test drug is added to an ‘optimised background regimen’ (OBR) and compared with OBR alone (sometimes with a placebo for the test drug). The OBR is a combination of three or more drugs selected prior to randomisation and individualised for each specific subject based on their treatment history and using the results of viral genotyping and/or phenotypic testing to evaluate viral resistance to available drugs for that subject. This design was used, for example, in the TORO 1 and TORO 2 trials which evaluated enfuvirtide.[25,26] Because antiretroviral therapy is expected to be less effective in treatment-experienced patients, these trials used a primary efficacy endpoint of the quantitative change in HIV RNA from baseline to 24 weeks. To encourage enrolment, other features of the design of these two trials were a 2:1 randomisation in favour of receiving enfuvirtide, and the availability of enfuvirtide to patients in the control arm who experienced virologic failure after at least eight weeks of treatment. One criticism of the latter design feature is that it makes it more difficult to evaluate the primary efficacy endpoint as well as the long-term immunological changes and safety of the test drug in randomised comparisons, as the use of enfuvirtide in both randomised arms will tend to dilute differences between the two arms. An alternative but similar design randomises subjects to the test drug or to drug from a specified set of alternatives from the same drug class as the test drug, with the addition of an OBR to both randomised arms. This was used in RESIST-1 which compared a new PI, tipranavir, added to ritonavir and an OBR, with another protease inhibitor added to ritonavir and an OBR.[27]

A general concern about these types of design is that patients are at risk of developing resistance to the new drug, particularly when it is added to an OBR which has limited activity against the virus.[28,29] For this reason, some patients will choose not to enter trials using these types of design, preferring instead to delay taking the new drug until it can be used in combination with some other new drug that may become available, hence increasing the chance of achieving greater viral suppression from the combined effects of the two new drugs and so reducing the risk of developing resistance to one or both drugs. One way of overcoming this issue when there are multiple investigational agents available at the same time (as is not uncommon for HIV infection) is to design trials in which a regimen which includes at least two investigational agents is compared with the OBR. Alternatively, a modified factorial design has been proposed when three investigational agents (labelled A, B and C) are available: comparing OBR+A+B with OBR+A+C with OBR+B+C with OBR+A+B+C.[29] Use of such designs obviously may carry the significant practical challenge of requiring collaboration between pharmaceutical companies including early evaluation of the pharmacokinetics and safety of combinations of investigational drugs.

In trials involving treatment-experienced patients, evaluating genetic characteristics of the virus at the time of screening and entry to a trial is important for allowing retrospective identification of mutations present prior to the start of treatment for which suppression of the virus by the new drug is reduced. For tenofovir, for example, patients having three or more mutations that are associated with another NRTI, zidovudine, including two specific mutations, had their level of HIV RNA suppressed by a factor of four less than patients with no zidovudine-associated mutations.[21,29] Such cross-resistance between drugs may sometimes be identified during Phase I/II trials. In this case, it may be lead to the exclusion of subjects from Phase III trials as occurred in the development of tipranavir.[29] A similar issue arises in trials that may be targeted for patients carrying virus with specific characteristics. For example, the CCR5 co-receptor inhibitors attempt to inhibit entry into the CD4 cell of virus which attaches to the CCR5 co-receptor on the cell. However, some viruses instead use the CXCR4 co-receptor. From a mechanistic perspective, it is therefore important to exclude subjects hosting virus with the CXCR4 phenotype from trials of these drugs.

TREATMENT MANAGEMENT TRIALS

With the large number of drugs approved for the treatment of HIV infection, there have been numerous trials that have evaluated various aspects of treatment management. Some examples are provided here for illustration. Many of these concern the identification of specific drug combinations that might be most efficacious in particular settings and the design issues are very similar to those discussed above for Phase III trials. This type of question extends to the issue as to what sequence of drug combinations might be most efficacious. For example, one study, ACTG 384, used a factorial design which included a comparison of two different pairs of NRTIs (ddI+d4T or ZDV+3TC) combined with either an NNRTI (EFV) or a PI (NFV) as initial treatment for treatment-naive patients, with defined switches after failure of this initial regimen to a second-line regimen (e.g. patients initially receiving ddI+d4T+EFV switched to ZDV+3TC+NFV).[30] The primary endpoint was failure of the second regimen so providing an evaluation of the four different sequences of three-drug regimens. Other trials address questions such as whether it is possible to simplify a three-drug regimen by discontinuing a drug after adequate virologic suppression has been successfully achieved. The potential benefit of this is a reduction in risk of adverse effects, some of which are life-threatening and cumulate with long-term exposure (e.g. lipid abnormalities leading to increased risk of cardiovascular morbidity).

Trials have also been undertaken to evaluate other aspects of treatment management. One example was trials that established that the use of genotypic antiretroviral resistance testing prior to selection of a new drug regimen for treatment-experienced subjects improved disease status versus standard clinical practice.[31,32] Other trials have shown that planned interruptions of antiretroviral therapy have led to a worsening of disease status versus continuous treatment.[33,34] There are important management questions being addressed by trials which are ongoing at the time of writing (in 2005). Two trials, ACTG 5115 and PENPACT-1, address the question as to when to change an antiretroviral regimen that is failing. Specifically, both involve a randomised comparison of switching at low HIV RNA levels (e.g. shortly after the level becomes detectable by standard assays) versus allowing the virus to rebound to higher levels before switching. Another question is being addressed by the SMART Study: what are the long-term consequences of a strategy which aims to suppress viral loads as much as possible irrespective of a subject's CD4 count versus a 'drug conservation' strategy which defers treatment until the subject's CD4 count is below a defined threshold (associated with a more marked immediate risk of AIDS-defining events) and then also discontinues treatment whenever it rises to a sufficiently high level? This trial is unusual in that its primary efficacy outcome measure is, appropriately, clinical disease progression or death. This has meant that the study is large (involving 6000 patients) and will require long-term follow-up (the average duration is expected to be about 7.5 years). The SMART Study should answer an adaptation of a key treatment management question concerning when first to start antiretroviral treatment in an HIV-infected patient. There are, however, significant challenges in undertaking such a trial (e.g. concerning rate of accrual if insufficient patients are willing to be randomised to the two competing strategies), and the potential difficulties in interpreting the results when considerable progress is still being made in developing new drugs and classes of drugs which would become available while such a trial is ongoing.

OTHER ISSUES

TREATMENT OF CHILDREN AND ADOLESCENTS

The HIV epidemic impacts children and adolescents significantly, particularly in countries with more limited resources. Vertical transmission of HIV from mothers to their babies occurs during pregnancy, labour and delivery or through breastfeeding, and adolescents may acquire HIV infection through high-risk behaviour. For adolescents, clinical trials are often focused on the evaluation of approaches to treatment management that promote adherence to treatment as this is a major challenge in this patient group. For infants and children, most current clinical trials are relatively small studies that aim to find a dose of a drug that has an acceptable safety profile and which has pharmacokinetic properties that are similar to those obtained for the dose used in adults. This has been challenging for some drugs, particularly in infants, because of substantial variability in pharmacokinetic parameters and difficulties in formulating products for the paediatric population.

In addition, there are some aspects of HIV infection that arise in children which do not arise in adults. One concerns the negative impact of the disease on growth and neurodevelopment. Another concerns the fact that there is a very high risk of morbidity and mortality in children with perinatally acquired infection during the first few months of life, at a time when CD4 counts and HIV RNA levels are less reliable as markers of prognosis. This has led to substantial differences in treatment management guidelines internationally. An important question being considered in trials in development therefore concerns how to treat HIV-infected infants with higher CD4 levels within the first two to three months of life. These trials will involve a randomisation between, for example, immediate therapy (to be continued indefinitely or, alternatively, to be discontinued at 1 or 2 years of age when risk of morbidity and mortality is lower and more reliably predictable in children with higher CD4 counts) versus deferring treatment until there is a defined deterioration in CD4 or clinical disease status.

THE INTERFACE OF PREVENTION AND TREATMENT

Clinical trials are important in evaluating potential approaches for the prevention of transmission of HIV. The greatest success has been achieved in preventing transmission of HIV from a mother to her baby. The pivotal trial in the United States was ACTG 076 which showed that a course of zidovudine used during pregnancy, labour and delivery by the mother and during the first six weeks of life by the infant could reduce the risk of transmission from in excess of 25% to about 8%.[35] Use of more potent regimens has led to reductions in risk to less than 2%. In the resource-limited setting, a single dose of nevirapine (NVP) given to the mother during labour and to the infant during the first three days of life reduced the transmission rate from 20% to 12% at 6–8 weeks of age, and from 26% to 16% by 18 months of age despite breastfeeding (which can lead to transmission of the virus).[36] Despite this success, it has become clear that the virus in a large proportion of the mothers and their infants (when infected) who have taken this NVP regimen has developed resistance to NVP.[37] A major concern, therefore, is that this resistant virus might be archived in the mother or infant thus reducing the effectiveness of NVP-based and other NNRTI-based regimens when subsequently needed for their own health. These regimens are often the standard of care provided in national treatment programmes for initial treatment in resource-limited countries. To evaluate this, a study (ACTG 5208) that incorporates parallel randomised trials, one in a population of women with no prior single dose NVP exposure and one in a population with such exposure, has been developed to evaluate whether this reduction in effectiveness actually occurs. In each population, women are randomised to a NVP-based regimen or a PI-based regimen. The trial involving women with no prior exposure is designed as an equivalence trial as the effects of the two regimens are anticipated to be similar. The trial involving women with prior exposure is designed as a superiority trial to establish whether PI-based therapy is superior to NVP-based therapy as might be expected if archived resistant virus is a concern. The trial involving women with no prior exposure then provides a ‘control’ comparison for the comparison in women with prior exposure.

Antiretroviral drugs might also be used by infected people to prevent horizontal transmission of HIV to an uninfected person. This is being evaluated in a randomised comparison of immediate versus deferred use of antiretroviral therapy by an infected person for whom treatment is not immediately indicated (HPTN 052). The randomisation involves couples, one of whom is HIV-positive and one of whom is HIV-negative. The primary efficacy outcome measure is the proportion of HIV-negative subjects who become infected. A secondary objective of this study concerns the ‘when to start’ antiretroviral treatment (in the HIV-positive partner) discussed earlier.

ETHICAL ISSUES

In the 20 years or so since the HIV epidemic was first identified, there have been, and continue to be, significant ethical and political challenges in the areas of treatment and prevention of HIV infection, and in the associated research. There has been intense scrutiny of research by both the scientific community and the public, accompanied by a significant level of patient activism. Addressing some of these challenges has had an impact on trial design, conduct and analysis, as well as the drug approval process, not only in HIV but also in other areas of medicine. Early in the epidemic, a major issue concerned access by patients to promising new drugs.[38] One significant development concerned the provision of expanded access programmes whereby patients who were unable to participate in clinical trials could access a drug in parallel with the drug’s evaluation in Phase II and III clinical trials. Most commonly, these programmes offered access to patients with later-stage (symptomatic) disease who had no satisfactory alternative treatment options (e.g. due to viral resistance to, or adverse effects from, alternative therapies). Such programmes became more generally available for patients with other serious or immediately life-threatening diseases. The programmes for HIV-infected subjects have been used to provide additional safety information to support drug approval, sometimes from cohorts of several thousand patients.[39,40] However, these programmes are not without controversy. In particular, limitations on drug availability due to manufacturing restrictions for new products meant that access to many programmes involved ballots among eligible patients, and there were issues concerning availability or programmes in different geographical locations.[41]

A second and related development was that regulatory agencies developed approaches for expedited (or accelerated) approval of drugs. This was achieved in part by allowing approval based upon shorter-term studies that evaluated effects on surrogate markers. Earlier, this was largely based on effects on CD4 cell count pending clinical outcome studies for full traditional approval. More recently, it has been based on shorter-term effects on HIV-1 RNA levels (e.g. over 24 weeks) with traditional approval based on longer-term effects (e.g. over 48 weeks or more). These developments had a similar impact on drug approval for other life-threatening diseases.

The international side of the epidemic also led to major ethical issues concerning the design and conduct of clinical trials in resource-limited countries, including drug access to study participants at the conclusion of a trial. The most significant issue concerned the choice of control group to be used in randomised trials; specifically, whether the intervention used in the control group should reflect the best standard of care available anywhere in the world, or whether it is acceptable to use the best standard of care available in the country in which the trial is conducted. Discussion centred on trials in which antiretroviral drugs were used to prevent mother-to-child transmission of HIV. As described above, ACTG 076 was a placebo-controlled randomised trial conducted in the United States and France that established that maternal use of zidovudine initiated between 14 and 32 weeks of gestation and continued through labour and delivery, coupled with six weeks of zidovudine use by the infant, reduced transmission from over 25% to 8%.[35] The question of interest in the resource-limited setting was whether shorter, less expensive and less complex regimens would also be effective in reducing transmission. The debate centred on whether a control group should receive the zidovudine regimen provided in ACTG 076 or whether it was acceptable to use a placebo control on the basis that the zidovudine regimen was not available in the countries in which the trials were to be conducted to address a question of considerable local importance, or whether the ACTG 076 regimen should be provided. The debate was extensive[42–45] and led to a revision in 2000 to the World Medical Association Declaration of Helsinki on ethical principles for medical research involving human subjects.[46] However, the debate has continued resulting in a note of clarification to the Declaration of Helsinki in 2002 and ongoing controversy with multiple national and international bioethics committees providing alternative opinions and recommendations.[47]

REFERENCES

1. UNAIDS. *2004 Report on the Global AIDS Epidemic. Executive Summary*. Geneva: UNAIDS (2004) 5–6.
2. The rgp120 HIV Vaccine Study Group. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *J Infect Dis* (2005) **191**: 654–65.
3. Hoffman IF, Taha TE, Padain NS *et al.* Nonoxynol-9 100 mg gel: multi-site safety study from sub-Saharan Africa. *AIDS* (2004) **18**: 2191–5.
4. Fleming TR, Richardson BA. Some design issues in trials of microbicides for the prevention of HIV infection. *J Infect Dis* (2004) **190**: 666–74.
5. Dilley DW, Woods WJ, Sabatino J *et al.* Changing sexual behavior among gay male repeat testers for HIV: a randomized, controlled trial of a single-session intervention. *J Acquir Immune Defic Syndr* (2002) **30**: 177–86.
6. Chesney MA, Koblin BA, Barresi PJ *et al.* An individually tailored intervention for HIV prevention: baseline data from the EXPLORE study. *Am J Public Health* (2003) **93**: 933–8.

7. DiClemente RJ, Wingood GM, Harrington KF *et al*. Efficacy of an HIV prevention intervention for African American adolescent girls. A randomized controlled trial. *JAMA* (2004) **292**: 171–9.
8. Mellors JW, Munoz A, Giorgi JV *et al*. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Ann Intern Med* (1997) **126**: 946–54.
9. Phillips A, CASCADE Collaboration. Short-term risk of AIDS according to current CD4 cell count and viralload in antiretroviral drug-naïve individuals and those treated in the monotherapy era. *AIDS* (2004) **18**: 51–8.
10. HIV Surrogate Marker Collaborative Group. Human immunodeficiency virus type I (HIV-1) RNA level and CD4 count as prognostic markers and surrogate endpoints: a meta-analysis. *AIDS Res Hum Retroviruses* (2000) **16**: 1123–33.
11. Hughes MD. The evaluation of surrogate endpoints in practice: experience in HIV. In: Burzykowski T, Molenberghs G, Buyse M, eds, *The Evaluation of Surrogate Endpoints*. New York: Springer (2005).
12. Food and Drug Administration. *Guidance for industry. Antiretroviral drugs using plasma HIV RNA measurements – clinical considerations for accelerated and traditional approval*. Rockville, MD: Food and Drug Administration (2002).
13. Rousseau FS, Kahn JO, Thompson M *et al*. Prototype trial design for rapid dose selection of antiretroviral drugs: an example using emtricitabine (Coviracil). *J Antimicrob Chemother* (2001) **48**: 507–13.
14. Kilby JM, Hopkins S, Venetta TM *et al*. Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. *Nat Med* (1998) **4**: 1302–7.
15. O'Quigley J, Hughes M, Fenton T. Dose-finding designs for HIV studies. *Biometrics* (2001) **57**: 1018–29.
16. Lalezari JP, DeJesus E, Northfelt DW *et al*. A controlled Phase II trial assessing three doses of enfuvirtide (T-20) in combination with abacavir, amprenavir, ritonavir and efavirenz in non-nucleoside reverse transcriptase inhibitor-naïve HIV-infected adults. *Antivir Ther* (2003) **8**: 279–87.
17. Murphy RL, Sanne I, Cahn P *et al*. Dose-ranging, randomized, clinical trial of atazanavir with lamivudine and stavudine in antiretroviral-naïve subjects: 48-week results. *AIDS* (2003) **17**: 2603–14.
18. Sanne I, Piliero P, Squires K *et al*. Results of a Phase 2 clinical trial at 48 weeks (AI424-007): a dose-ranging, safety, and efficacy comparative trial of atazanavir at three doses in combination with didanosine and stavudine in antiretroviral-naive subjects. *J Acquir Immune Defic Syndr Hum Retrovirol* (2003) **32**: 18–29.
19. Gulick RM, Ribaudo HJ, Shikuma CM *et al*. Triple-nucleoside regimens versus efavirenz-containing regimens for the initial treatment of HIV-1 infection. *New Engl J Med* (2004) **350**: 1850–61.
20. Gallant JE, Staszewski S, Pozniak AL *et al*. Efficacy and safety of tenofovir DF vs stavudine in combination therapy in antiretroviral-naive patients. A 3-year randomized trial. *JAMA* (2004) **292**: 191–201.
21. Gilead Sciences, Inc. T*enofovir (Viread) U.S. Prescribing Information*. (May 2005). Accessed at www.viread.com/pdf/pi.pdf, 10 August 2005.
22. Hill A, DeMasi R. Discordant conclusions from HIV clinical trials – an evaluation of efficacy endpoints. *Antivir Ther* (2005) **10**: 367–74.
23. Gazzard B, DeJesus E, Campo R *et al*. The combination of tenofovir DF (TDF), emtricitabine (FTC), and efavirenz (EFV) has significantly greater response vs fixed-dose zidovudine/lamivudine (CBV) and EFV in antiretroviral-naïve patients: a 24-week preliminary analysis. *44th Interscience Conference on Antimicrobial Agents and Chemotherapy*, October 30–November 2 2004, Washington, DC, USA. Abstr H-1137c.
24. Phillips AN, Walker AS. Drug switching and virologic-based endpoints in trials of antiretroviral drugs for HIV infection. *AIDS* (2004) **18**: 365–70.
25. Lalesari JP, Henry K, O'Hearn M *et al*. Enfuvirtide, an HIV-1 fusion inhibitor, for drug-resistant HIV infection in North and South America. *New Engl J Med* (2003) **348**: 2175–85.
26. Lazzarin A, Clotet B, Cooper D *et al*. Efficacy of enfuvirtide in patients infected with drug-resistant HIV-1 in Europe and Australia. *New Engl J Med* (2003) **348**: 2186–95.
27. Hicks C. RESIST-1: a phase 3, randomized, controlled, open label, multicenter trial comparing tipranavir/ritonavir (TPV/r) to an optimized comparator protease inhibitor/r(CPI/r) regimen in antiretroviral(ARV) experienced patients: 24 week data. *44th Interscience Conference on Antimicrobial Agents and Chemotherapy*, October 30–November 2 2004, Washington, DC, USA. Abstr H-1137a.
28. Youle M, Cohen C, Katlama C *et al*. Codevelopment of new antiretrovirals in very treatment-experienced HIV-infected individuals (letter). *Lancet* (2004) **364**: 1036–7.
29. Struble K, Murray J, Cheng B *et al*. Antiretroviral therapies for treatment-experienced patients: current status and research challenges. *AIDS* (2005) **19**: 747–56.

30. Robbins GK, De Gruttola V, Shafer RW *et al*. Comparison of sequential three-drug regimens as initial therapy in HIV-1 infection. *New Engl J Med* (2003) **349**: 2293–303.
31. Durant J, Clevenbergh P, Halfon P *et al*. Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomized controlled trial. *Lancet* (1999) **353**: 2195–9.
32. Baxter JD, Mayers DL, Wentworth DN *et al*. A randomized study of antiretroviral management based on plasma genotypic antiretroviral resistance testing in patients failing therapy. *AIDS* (2000) **14**: F83–F93.
33. Lawrence J, Mayers DL, Hullsiek KH *et al*. Structured treatment interruption in patients with multidrug-resistant human immunodeficiency virus. *New Engl J Med* (2003) **349**: 837–46.
34. Ruiz L, Ribera E, Bonjoch A *et al*. Role of structured treatment interruption before a 5-drug salvage antiretroviral regimen: the Retrogene Study. *J Infect Dis* (2003) **188**: 977–85.
35. Connor EM, Sperling RS, Gelber R *et al*. Reduction of maternal-infant transmission of human immunodeficiency virus type I with zidovudine treatment. Pediatric AIDS Clinical Trials Group Protocol 076 Study Group. *New Engl J Med* (1994) **331**: 1173–80.
36. Jackson JB, Musoke P, Fleming T *et al*. Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kamapala, Uganda: 18-month follow-up of HIVNET 012 randomized trial. *Lancet* (2003) **362**: 859–68.
37. Eshleman SH, Mracna M, Guay LA *et al*. Selection and fading of resistance mutations in women and infants receiving nevirapine to prevent HIV-1 vertical transmission (HIVNET 012). *AIDS* (2001) **15**: 1951–7.
38. Hirsch MS. Chemotherapy of human immunodeficiency virus infections: current practice and future prospects. *J Infect Dis* (1990) **161**: 845–57.
39. Pike IM, Nicaise C. The didanosine Expanded Access program: safety analysis. *Clin Infect Dis* (1993) **16**(Suppl 1): S63–8.
40. Kessler HA, Johnson J, Follansbee S *et al*. Abacavir expanded access program for adults infected with human immunodeficiency virus type I. *Clin Infect Dis* (2002) **34**: 535–42.
41. Alcorn K, Nary G, Pernet AG *et al*. Europeans outraged over expanded access. *J Int Assoc Physicians AIDS Care* (1996) **2**: 48–51.
42. Lurie P, Wolfe SM. Unethical trials of interventions to reduce perinatal transmission of the human immunodeficiency virus in developing countries. *New Engl J Med* (1997) **337**: 853–6.
43. Varmus H, Satcher D. Ethical complexities of conducting research in developing countries. *New Engl J Med* (1997) **337**: 1003–5.
44. Brennan T. Proposed revisions to the Declaration of Helsinki – will they weaken the ethical principles underlying human research? *New Engl J Med* (1999) **341**: 527–31.
45. Levine RJ. The need to revise the Declaration of Helsinki. *New Engl J Med* (1999) **341**: 531–4.
46. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. Adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964, and amended by the 48th WMA General Assembly, Edinburgh, Scotland, October 2000, with Note of Clarification on Paragraph 29 added by the WMA General Assembly, Washington, 2002 and Note of Clarification on Paragraph 30 added by the WMA General Assembly, Tokyo, 2004. Accessed at www.wma.net/e/policy/pdf/17c.pdf, 26 September 2005.
47. Lie RK, Emanuel E, Grady C, Wendler D. The standard of care debate: the Declaration of Helsinki versus the international consensus opinion. *J Med Ethics* (2004) **30**: 190–3.

18

Infectious Diseases

INGELA KRANTZ
Skaraborg Institute, Skövde, Sweden

INTRODUCTION

The field of infectious diseases has brought some useful innovations to the study of human illness and health care. As communicable diseases they are caused by biological agents (virus, bacteria, parasites). Infectious diseases were among the first human ailments that could be attributed to a demonstrable causal agent, identified by laboratory tests when the patients were still alive, prevented by vaccination of individuals at risk and avoided by sanitary interventions. Sanitation helped to prevent diseases in the general population of the community; vaccinations provided protection to individual persons.

The modern era of randomised trials in health care is said to begin with the evaluation of streptomycin for treatment of severe tuberculosis in the United Kingdom in 1948. The trial in which treatments first began to be allocated randomly to individuals was one designed to test the efficacy of immunisation against whooping cough from 1946, not the trial of streptomycin for treating pulmonary tuberculosis. The latter also started 1946, but was reported in 1948, three years earlier than the report of the whooping cough trial.[1] In the streptomycin trial the investigators assigned patients to an experimental group getting bed-rest and streptomycin and a control group just getting bed-rest (the conventional therapy) by random numbers. Streptomycin was in short supply, a factor contributing to the study being confined to a group of patients with rapidly advancing bilateral pulmonary tuberculosis. The investigators thus worked with a relatively homogeneous group of people; by so doing they avoided some of the problems that would later arise, when randomised controlled trials (RCTs) came into more general use.

At that time streptomycin was not part of routine treatment. The only specific treatment available for pulmonary tuberculosis before the introduction of streptomycin was induced collapse of the lung. This was performed on every patient in the study in need of it, irrespective of which group the patient belonged to.

Placebo was not used since one assumed that psychological factors would have little impact on such a serious disease. The main consideration was the welfare of the patient which was also behind the idea of not seeking formal informed consent; an acceptable ethical reasoning in its historical context. Medical care at that time belonged to a paternalistic worldview and doctors

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green

ruled in the best interests of their patients. It was felt there should not be separate ethical standards for therapeutic trials and routine medical care.[2]

Streptomycin showed substantial but temporary benefits, and two major disadvantages: toxicity and the early emergence of drug resistance. The investigators could present the results after a comparably short time.[1]

The establishment of RCTs as a tool for supplying evidence has meant many obvious benefits in the field of infectious diseases. The controlled trial is not the only way to study therapeutic efficacy,[3] but it certainly has meant many therapeutic advances and also eliminations of ineffectual treatments and diagnostic manoeuvres.

An RCT could be designed to obtain results of direct use for decisions in clinical practice but it could also be an explicatory experiment. Schwartz and Lellouch[4] coined the terms 'pragmatic' and 'explanatory' for these two approaches. We may seek to choose between two or more treatments or we may seek to verify a biological hypothesis. A controlled trial will require a different comparison in a different set of patients depending on which outlook is favoured by the investigator. The streptomycin trial was a straightforward pragmatic one intended to answer direct questions about the usefulness of streptomycin in clinical practice.

Since the 1940s countless reports of trials have been published in the field of infectious diseases – both pragmatic and explanatory ones – but many of these studies are biased and based on insufficient numbers of participants to yield reliable and relevant estimates of treatment effects. Therapy effects and outcomes as evident as the ones seen during the beginning of the antibiotic era are no longer to be expected, which means there is a quest for really large-scale randomised evidence.[5] Researchers using an RCT approach should put more effort into recruitment of larger numbers of eligible study subjects and collect less information per subject to safeguard the internal and external validity.

THE ARCHITECTURE OF A STUDY

Figure 18.1 shows a simple way of laying bare the basic structures in the 'architecture' of a controlled trial, where people under observation are followed from a baseline state through intervention by various manoeuvres to the occurrence of outcomes.[6]

When manoeuvres are imposed with an experimental plan for comparisons the design is called a controlled trial, otherwise an observational cohort. The intervention manoeuvres could be, for example, single- or multi-component antibiotics, vaccines, impregnated bed-nets, condoms and health education programmes.

In the streptomycin trial the manoeuvre consisted of a single antibiotic. The disease and the main aetiologic agent could be verified by laboratory tests, direct-smear examination, culture and X-ray, which made precision possible in the choice of study groups and outcome variables.

Trials in which randomisation is used to allocate subjects enable the investigator to isolate the effect of the study factor or the manoeuvre. Streptomycin gave an effect on radiographic appearances and survival (Figure 18.2). The architecture is not overly intricate and the outcome makes clinical sense. Unfortunately, things are not so straightforward today. RTCs are considered the gold standard for biomedical evidence but they are often controversial due to recruitment and ethical problems, and also due to differences in outlook – pragmatic or explanatory – between researchers and their audience.

R (randomisation)

Baseline	Intervention		Outcome	
Stratification	Manoeuvre	Side manoeuvres	Effects	Side effects
	Non-manoeuvre		Effects	Side effects

time →

Figure 18.1. The basic architecture of a clinical trial.

R

Baseline	No.	Intervention	Outcome		
"Acute progressive bilateral pulmonary tuberculosis of presumably recent origin, bacteriologically proved, unsuitable for collapse therapy", age group 15 to 30	55	streptomycin bed-rest	4 died	38 radiological improvement	36 had giddiness from toxic effects on the vestibular apparatus
Stratification: sex, treatment centre	52	no streptomycin bed-rest	14 died	17 radiological improvement	0 giddiness
time	**0**	**4**	**6 months**		

Figure 18.2. The MRC streptomycin trial.[1]

Infectious diseases have been the basic source of concepts, methods and technology in epidemiological research. Infectious diseases are good at illustrating the advantages and limitations of RCTs for all areas of medicine because the causal process is easily conceptualized in a sort of standard model (Figure 18.3). A microbial agent invades a host's body or part of that body and 'takes over' at the cost of that host. A measles patient coughs next to you and you get the disease, if you never had it before. A female *Anopheles* mosquito feeds on your blood and a plasmodium parasite enters your body, resulting in malaria. The virus, bacteria and parasites are necessary factors for you to get sick, but they are not sufficient for this to happen. Studying infectious diseases forces you to think contextually; no microbes act in a vacuum. Enabling and disabling factors have to be accounted for, a process necessary for the unavoidable discussion of confounding and relevance.

Antibiotics act against microbes and their effects can often be traced outside the specific patients being treated. Resistance among microbes is a phenomenon among the population of microbes with a dynamic of its own. In this aspect antibiotics differ from, say, cardiovascular drugs. The development of microbial resistance is a serious side effect with far-reaching consequences not just for the treated individual but also for the community at large.

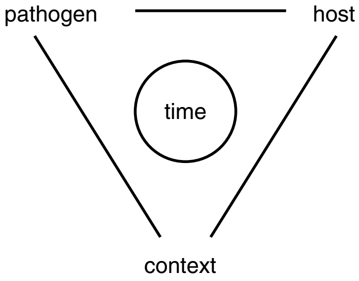


Figure 18.3. A model of the interactions between a susceptible host, a virulent pathogen and an environment/context favourable for disease development.

The RCT approach was adopted in medicine through clever instigations by statisticians and epidemiologists. Bradford Hill and his contemporaries had a concern for simplicity of design and clarity of presentation.[2] With computerised technical assistance the sophistication has reached unprecedented levels, very often at the cost of clarity. It is easy to forget about the basic rule that no statistical method can remove bias unwittingly incorporated in the design, when as a reader you are confronted with a barrage of statistical fireworks. Statistical methods are there to summarise findings and deal with variation, but they can never function as a remedy for missing information. Imputation methods and the like are only of use in models denuded of everyday messiness, as in experiments on defined strains of laboratory rodents.

Nowadays, when RCTs have moved outside the boundaries of hospitals and often concern broader topics than just the efficacy of a single drug, there is a need for inputs from social science to highlight many important questions about practicable therapeutic improvements in controlling infectious diseases and to help implement RCTs in real-life settings.

Sackett identified 24 different types of bias relating to the conduct of a study.[7] Among the most important of these are: selection bias, dealt with by random allocation; performance bias, dealt with by blinding; and detection bias, leading to differences in how outcomes are assessed. The streptomycin trial used random allocation, had some problem with performance bias (treatment with lung collapse measures upset the balance of the streptomycin and control groups) and reported outcomes with due regard for detection and attrition bias. To cite Richard Doll, "Modern authors please note."[2]

The somewhat fantastic development in molecular biology has certainly not made things easier in how to test and interpret the outcomes of therapies and preventive actions for infectious diseases. With more detailed knowledge of immunological processes disease manifestations are seen as products of complicated interactions between the microbial pathogen and the host. This interaction is the underlying basis of infectious disease. By understanding the molecular details of the interaction, we can identify host-defence strategies and virulence-associated microbial genes. Categorisation of bacteria on the basis of their genotypic characteristics, for instance, might supplant classification based on phenotypic markers in the future.

There are thus many more levels than the schematic one rendered in Figure 18.3 in a causal chain of an infection. Different levels of variation have to be considered: variations in the specific microbe population, the vector (if part of the transmission), the individual and the individual's internal milieu.

AETIOLOGY

Clinical studies to examine specific aetiologic relationships without a therapeutic or preventive component are seldom ethically feasible. The Tuskegee syphilis observational cohort is an appalling example of a long, inefficient and irrelevant non-therapeutic experiment on human beings with an approach that would not even be conceivable nowadays.[8] For 40 years between 1932 and 1972, the US Public Health Service conducted an experiment meant to discover how syphilis progressed without treatment and how it affected black people as opposed to white. Treatments for syphilis existed even in the 1930s and penicillin became part of the clinical armament in the 1950s.[9] Diseases have to be seen as part of the context and the context contains various therapies. A controlled trial to address the question of immediate effects of the then available treatments for the different stages of syphilis on random samples of patients might perhaps have been ethically defensible, but not if tested just against placebo, even in the 1930s. Nowadays there should be exciting possibilities for studies since the genome of *Treponema pallidum* has been sequenced.[10] One could envisage researchers taking this new knowledge as a basis for RCTs disentangling the not so simple pathology of syphilis and how to diagnose, treat and vaccinate against the disease.

In today's context treatment in the field of infectious diseases most often means an intervention with antibiotics. Specific aetiologic relationships have therefore often come to be explored by eradication and challenge experiments. The former approach is partly how *Helicobacter pylori* has been incriminated as cause for peptic ulcer disease.

The bacterium *H. pylori* occurs worldwide and infects humane gastric-type mucosal tissue. The infection is usually acquired in childhood and early adulthood. Around 50% of the world's population is said to carry the bacterium in the gastrointestinal tract, many without symptoms. There is an established association between the risk of developing duodenal and gastric ulcers and infection. Peptic ulcer recurrences are diminished if *H. pylori* is eradicated with antibiotics.[11]

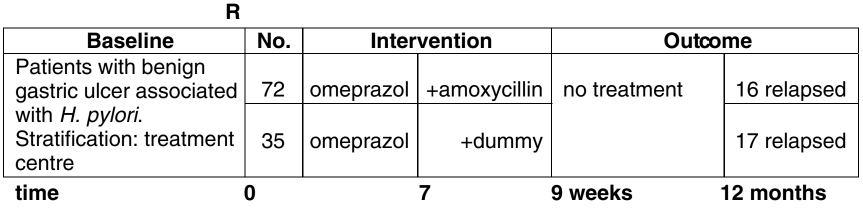
A study with a multi-centre, randomised, double-blind, controlled approach examined whether eradication of *H. pylori* infection reduces recurrence of benign gastric ulceration (Figure 18.4).[12] Patients were randomised to acid suppression (omeprazol) for eight weeks or the same treatment plus an antibiotic for weeks seven and eight. The patients were then untreated and followed for 12 months.

The recurrence of ulcers over the next 12 months in the eradication and non-eradication groups was 22% and 49%, respectively. The assessment showed that there was also a substantially better prognosis in those in whom *H. pylori* eradication had been achieved compared with those in whom the organism persisted, irrespective of drug regimen. Data thus suggest that the eradication of *H. pylori* from patients with gastric ulcer is associated with a lower rate of relapse. The antibiotic employed here was not as effective as the ones in the current recommendations.[13] Thus the outcome was not a success from a pragmatic point of view. The results, however, confirm that the use of a regimen eradicating *H. pylori* is more effective in the prevention of ulcer relapse than one designed merely to heal the ulcer by acid suppression. The eradication of *H. pylori* seems to change the natural history of gastric ulcer and one can assume that *H. pylori* is a sufficient but not necessary factor for this disease.

From an explanatory perspective this RCT has provided evidence for infection as part of the causal chain of gastric cancer. The gastric precancerous process is characterised by sequential lesions of the gastric mucosa, from chronic gastritis, atrophic gastritis, to intestinal metaplasia and dysplasia. The World Health Organization and its International Agency for Research on Cancer classified *H. pylori* as carcinogenic to humans in 1994.[14] The mechanisms by which *H. pylori* increases the risk for gastric cancer are unknown but virulence factors of *H. pylori* may have a role together with the type of T-cell response in the host.[15] The mere possibility of cancer development must mean that the natural history of this affliction in all its detail has to be puzzled out indirectly and preferably then by an RCT approach. A non-antibiotic, longitudinal, observational patient cohort with gastric cancer as outcome is now inconceivable and inhumane. In a 1990–3 recruited observational cohort, however, 1526 Japanese patients were followed with endoscopy for a mean of 7.8 years, and 82% had *H. pylori* infection. The endpoint was gastric cancer which developed in 36 of the infected and none of the uninfected patients. Eradication treatment was not part of the study.[16]

BASELINE

H. pylori prevalence varies with birth cohort and socio-economic factors and may be associated with crowding in childhood. Prevalence tends to be much higher (50–80%) in those born before 1950 in comparison with those born more recently in high-income countries. Its prevalence is highest in low-income countries and increases



R

Baseline	No.	Intervention		Outcome	
Patients with benign gastric ulcer associated with *H. pylori*. Stratification: treatment centre	72	omeprazol	+amoxycillin	no treatment	16 relapsed
	35	omeprazol	+dummy		17 relapsed
time	**0**		**7**	**9 weeks**	**12 months**

Source: Reproduced by permission.

Figure 18.4. The architecture of a randomised double-blind controlled study of recurrence of gastric ulcer after eradication of *Helicobacter pylori* infection.[12]

rapidly during the first two decades of life, such that 80–90% of the population may be infected by early adulthood irrespective of the period of birth.[17,18] The chronic infection rarely resolves spontaneously.

The possibility of secular trends and differing geographical contexts are factors that make 'horizontal' comparisons tricky and have an impact on recruitment for multi-centre studies and meta-analysis. Does *H. pylori* play the same role for peptic ulcer disease and gastric cancer in an individual living with the infection since childhood as in someone being infected as an adult? There are also different strains of the bacterium, some of them with virulence factors probably with an impact on disease outcome and with differing susceptibility to treatments.[17] In all epidemiological studies it is essential to have a clear definition of the symptoms, signs and other characteristics indicating that a person has a certain affliction. Baseline data collected on each patient before randomisation usually include demographics, medical history, current signs and symptoms, and infectious status by various microbiological tests like culture and serology. Effective randomisation is expected to ensure that there are no systematic differences between the comparison groups but important baseline differences may still arise by chance. Thus, the disease may have to be defined on a narrower level than species level to make groups comparable from randomisation in all features that are known or suspected to influence the outcome.

The natural history of tuberculosis could illustrate the complexity in selecting a baseline.[19] *Mycobacterium tuberculosis* is a relatively slow growing infectious agent with an incubation period said to range from a few weeks to a lifetime. Tuberculosis is classed as one of the granulomatous inflammatory conditions. The development of a granuloma is a characteristic of host response to mycobacterial infections. It functions as a barrier to bacterial dissemination and is a strategy to contain the infection. The bacteria are not eliminated with the granuloma, but they can alter their metabolism and replication, and become dormant, resulting in a latent infection. Latent infection can be diagnosed only by a tuberculin skin test.

Infection by *M. tuberculosis* begins when the bacteria reach the pulmonary alveoli, from where they potentially can be transported away via local lymph nodes. In people in whom the bacteria overcome the host defence and begin to multiply, there is progression from infection to disease. Disease could develop in the lungs, peripheral lymph nodes, kidneys, brain and bone. If the bacteria gain entry to the bloodstream they spread and develop tiny foci of infections throughout the body – miliary tuberculosis. In many patients the infection waxes and wanes and tissue destruction is balanced by healing and scarring. The disease most commonly affects the lungs and diagnosis is then usually made by direct microscopy or bacterial culture of sputum. Pulmonary tuberculosis may co-exist with extrapulmonary manifestations.

Disease may occur soon after infection – primary tuberculosis – or many years after infection – post-primary tuberculosis, secondary tuberculosis or reactivation tuberculosis of dormant bacteria. About 10% of infected persons with a normal immune system will develop disease in their lifetime. The risk of reactivation increases with an impaired immune system, such as in individuals infected with HIV, patients with diabetes mellitus, prolonged corticosteroid therapy and wasting diseases.

Helminth infections are long-lasting parasitic infections abundant in low-income countries. The parasites cause immune changes which may play an important part in the pathogenesis of various infections. Their similar geographic distribution to *M. tuberculosis* and HIV makes them of interest in the characterisations of baseline groups in most of the contexts where RCTs dealing with tuberculosis or HIV therapy take place.[20]

Drug-resistant tuberculosis is transmitted in the same way as drug-sensitive tuberculosis. Primary resistance develops in persons initially infected with resistant organisms. Acquired or secondary resistance may develop during therapy due to inadequate treatment or non-compliance.

Tuberculosis is just one example of how an infectious disease has a very varied way of presenting itself. Malaria is another example of how more knowledge of the processes of microbe – host interactions can mean more sophistication in the choice of baseline subjects for an RCT. Unknown genetic subpopulations within a larger population might cause spurious associations. Randomisation can of course be done within strata of the likely response to treatment, if only the clinicians can define the strata sufficiently clearly.

However, an excessive control over the study situation may be a weakness from the pragmatic point of view in that the study population differs from the target population on several characteristics with important implications for how the findings could be of use clinically and for public health activities. Broad inclusions are used with an aim to increase clinical usefulness as stated in an RCT from Swaziland (Figure 18.5).[21] Randomisation was made in four groups according to the clinical presentation of their tuberculosis. Direct observation of tuberculosis treatment (DOTS) seems to be an effective way of administering therapy but it requires considerable health care resources. The practical question was whether there was a difference in cure (as measured by smear-negativity) and compliance when the direct observation of tablet intake was made by community health workers or family members.

More baseline restrictions, e.g. for ethnicity, geographical area, genotype of microbe species, might be of importance from an explanatory perspective, though. Human immune responses to *Plasmodium falciparum* seem in part to be genetically restricted, which can give varying susceptibility to the severe form of malaria. Studies have shown that members of the Fulani ethnic group in Mali are less affected by malaria than those from the Dogon group living in the same area, in a similar socio-cultural context and having been similarly exposed to mosquito bites.[22] Genetic diversity in the parasite and the variation in host–parasite interactions could create an imbalance between groups in baseline variables that may influence outcome.

Large-scale randomised trials are needed to find evidence for worthwhile but moderate benefits. It is an arduous and often costly task, however, to recruit individuals and the need for precision in the baseline does not make it easier. It took four years for Axon *et al.* to get a satisfying number of patients together for their study on *H. pylori* and gastric ulcers, even though many centres were involved.[12] The streptomycin trial had to be extended to seven centres; the cases of the type defined were not easy to find, although the impression from the planning stage was that they should be.[1]

In most RCTs, individual patients are randomised to a treatment or control group, but sometimes groups of people may be randomised instead. A cluster RCT approach is often necessary for evaluating immunisation programmes. Vaccines, despite being limited to individuals,

R

Baseline	No.	Intervention	Outcome
Tuberculosis patients able to receive treatment as outpatients Stratification: 1. Smear-positive pulmonary disease 2. Smear-negative disease 3. Extrapulmonary disease 4. Previously treated with relapse	668	community health worker supervision of standard antibiotic treatment	453/664 were cured or completed treatment 91/664 died
	667	family member/carer supervision of standard antibiotic treatment	440/662 were cured or completed treatment 110/662 died

time **0** **6 months**

Figure 18.5. The architecture of an RCT of direct observation of treatment for tuberculosis in Swaziland.[21] ''Cured'' = smear-negative.

may affect others through a ‘herd effect’. People in a cluster that has been vaccinated not only have more resistance to the disease against which they have been vaccinated, but are also less likely to be exposed to the disease when incidence around them goes down. The herd effect will weaken any estimate of difference due to the intervention. Another example is health interventions addressing groups of people, such as insecticide spraying of whole villages or activities to raise the quality of maternal health care. The cluster approach adds to the problem with study size; the statistical power of a cluster RCT is greatly reduced in comparison with similarly sized individual RCTs.

INTERVENTION

After streptomycin other antibiotics and drug regimens proved to be effective in the treatment of tuberculosis. Sadly, the world is not relieved of the ravages of *M. tuberculosis*. The disease is nowadays responsible for more deaths in the world than any other infectious agent.[23] There have not been any really new specific medications for TB since the initial breakthrough during the 1950s and the decades thereafter. The resistance problem already noted in the MRC streptomycin trial is growing and concurrent poverty infections like HIV counteract the gains in better knowledge. New treatments and vaccines are urgently needed, which means that there should be a quest for fairly large studies with a randomised approach to give the reliable explanatory and pragmatic answers so greatly needed.

Treatment of TB requires six months of a multi-drug regimen. Relapses occur if patient adherence is inadequate, which contributes to the emerging resistance problem. Due to this and the variability in how TB disease manifests itself, trials necessarily become prolonged, large and costly. Unfortunately there is no way around this; too small studies will not give the much needed reliable answers and their smallness is also a dilemma in a meta-analysis approach. The very societies whose people are suffering from TB cannot afford “wrong answers”. They want RCTs telling them how, with a minimum of expenditure, they can accomplish a maximum of care.

In an RCT in India the efficacy of three-, four-, and five-month regimens using a fluoroquinolone in the intensive phase for the treatment of smear-positive pulmonary TB was studied.[24] Fluoroquinolones have been used increasingly for drug-resistant TB since 1985, since they are well tolerated and have a good sterilising effect, rendering patients sputum culture negative and by that lessening the contagiousness. The baseline consisted of patients with newly diagnosed drug-susceptible pulmonary TB. There was no standard control group but four arms with standard regimens of various lengths plus a fluoroquinolone (ofloxazin) in the intensive phase (Figure 18.6).

Figure 18.6 illustrates the intricate architecture of a trial where the overarching aim was to see whether a shortening of the standard approach was feasible. There were two main outcomes: the proportion of patients who became culture negative at the end of treatment and the proportion of patients who relapsed during follow-up.

Ofloxacin had shown promising results and was added in all the four regimens. The design is thus more of a pragmatic one, intended to give clinical useful answers on the length of treatment and if ofloxacin is feasible. The intended follow-up was five years, which is very long but totally apt considering the behaviour of *M. tuberculosis*. The report here is 24 months after treatment. Evidently a lot has happened during follow-up, restricting the numbers of patients available for examination and knowledge of whether the therapy works or not. Compared with the MRC RCT with its six month’s bed-rest,[1] researchers nowadays usually cannot have such a “controlled” situation. DOTS, the internationally recommended strategy, with its direct observation of the medication,[25] is of course a way of controlling compliance but only for the ones showing up. Evidently the Achilles heel of a TB treatment RCT approach is how to gather valid observations over lengthy periods of follow-up. Relapse is thought to arise from persisting foci of dormant infection contained within granulomas.

R

Baseline	No.	Intervention					Outcome
Adult patients with newly diagnosed sputum-positive pulmonary tuberculosis initially susceptible to all the drugs used Stratification: by the degree of bacteria in the sputum smear into two strata	All. 131 ITT 120	O IRP	O IRP	O IRP			Sputum culture negative at the end of treatment = 4/120 Bacteriological relapse requiring retreatment = 7/83
	All. 133 ITT 115	O IRP	O IRP	O IRP	IR		Sputum culture negative at the end of treatment = 6/115 Bacteriological relapse requiring retreatment = 3/81
	All. 134 ITT 118	O IRP	O IRP	O IRP	IR	IR	Sputum culture negative at the end of treatment = 5/118 Bacteriological relapse requiring retreatment = 2/86
	All. 131 ITT 116	O IRP	O IRP	IR	IR		Sputum culture negative at the end of treatment = 3/116 Bacteriological relapse requiring retreatment = 12/91
time	**0**	**1**	**2**	**3**	**4**	**5**	**. . . 29 months**

Figure 18.6. The architecture of an RCT comparing duration of therapy for treatment of pulmonary TB.[24] O = Ofloxacin, I = Isoniazid, R = Rifampicin, P = Pyrazinamide. All. = Allocated, ITT = intention to treat.

There is still uncertainty in relation to the mixture of antibiotics, their dose, the time factor and the importance of geographical variations when studying the efficacy of TB regimens. RCTs taking care of all these crucial aspects will by necessity have difficulties in recruiting enough patients for each treatment arm.

The Indian study consisted of four treatment arms and the intervention under scrutiny was the duration of a certain mixture of antibiotics.[24] Shortening regimens for treatment of pulmonary TB is urgently needed to facilitate compliance. Nowadays it is impossible to envisage a TB treatment trial without antibiotics. But other remedies could perhaps be added somewhat in analogy with what was done for *H. pylori* where antacids and antibiotics work together to eradicate the intruder. Cytokines like the tumour necrosis factor-alpha (TNF-α) play a key role in the pathogenesis of many chronic inflammatory diseases. TNF-α inhibitors reduce symptoms and signs in patients with rheumatoid diseases but at the same time they raise the risk of causing a latent TB infection to become active due to the importance of this cytokine in the granulomatous immune defence against TB. Granulomas are an essential part of host protection against mycobacterial infection. At the same time, however, they appear to protect *M. tuberculosis* bacteria during therapy.[26] RCTs examining targeted disruption of granulomas by substances like cytokine-inhibitors plus antibiotics could be of interest to better understand granuloma biology and to test new therapies, i.e. studies with both an explanatory and pragmatic approach. With an opening up of the bacteria's protecting harness one could perhaps hope for a faster eradication process and a reduced demand on the length of the observation time. The design of RCTs covering all these aspects and functional levels indeed poses formidable challenges.

Malaria is another infectious disease where complex intervention schedules are being tested by RCTs. In contrast to TB, malaria is a vector-borne disease, which adds another dimension to the interaction between agent and host. Both vectors and parasites can develop resistance to the various substances used for control and eradication. So far, public health programmes have to rely on effective case management and large-scale deployment of insecticide-treated nets while waiting for an effective and affordable vaccine.

Malaria, which still kills an estimated 1 million people per year,[27] brings forward the whole gamut of possible interventions, at both the individual and group level. We can thus easily find a range of various RCT designs, e.g. vaccines tested double-blinded against placebo,[28] intermittent treatments of malaria at time of childhood routine vaccinations compared with placebo[29] and wide-scale installations of insecticide-treated curtains in areas paired on the basis of ecology and demography and one area randomly selected within each pair.[30] Malaria shares the need with TB for observations over a fairly long time span. As for other vector-borne diseases, differing exposure patterns and seasonality will sometime hamper the possibility for generalisations of specific malaria interventions.

OUTCOME

Everything that happens after the imposition of an intervention can be regarded as an outcome event. Some of these events will have been anticipated; others will occur as surprises.[6]

In the MRC trial the clinical question was answered fairly quickly: yes, streptomycin could affect pulmonary tuberculosis (Figure 18.2). An unpleasant surprise was the emergence of resistance and the re-emergence of the disease.[1] For infectious diseases the desirable event for the sick individual is of course to get better; for the society the elimination of contagiousness is an important goal (Figure 18.6). Nothing, of course, hinders the exploration of both items.

The time factor is crucial, as usual. For acute infections treated in hospitals, such as septicaemia, meningococcal meningitis, etc., the situation could be as controlled as in a laboratory experiment. For malaria, TB and the like, the messiness of everyday life will always threaten to put an end to the controlled situation. The most easily handled outcome would be number of deaths, at least in societies like Sweden where a population register keeps track of the inhabitants from birth to demise. The dichotomy is absolute and the remaining task will be to single out the specific deaths from all-cause mortality.

Usually, however, the patient's true condition is equivocal and the results may not be easily classified in two sharply separated categories. Outcomes based on definitive diagnostic evidence are of course preferable. A blood thick film with abundant *M. falciparum* parasites does not leave much doubt as to the cause in an acutely ill and feverish Kenyan child. Finding the culprit is good, but all methods are not overly sensitive or specific. The technique of sampling three sputum smears examined by acid-fast stain (the fastest method for establishing contagiousness) usually carries a sensitivity between 50% and 80% for finding *M. tuberculosis*. Measurements of outcome thus often have to rely on contributory diagnostic evidence, in this case clinical signs and radiographic evidence. When a reference standard is available, the data can be checked against that standard. Cultures remain the gold standard for diagnosing TB, but the sensitivity still only lies at about 80% and even less among children.[19] In many medical circumstances a definitive standard is not available, or there may be problems with the standard itself as in the case of *M. tuberculosis*.

The definitive diagnostic evidence may be uncertain because a definitive result requires inappropriate invasion. In TB patients who cannot produce sputum, bronchoscopy should be considered. But what about the cited RCTs in Swaziland and India (Figures 18.5 and 18.6)? Appropriate and reliable tests are needed for contexts where laboratory facilities are not optimal – high incidence and prevalence of infectious diseases often do not correlate well with occurrence of sufficient infrastructure to serve routine as well as research.

The main outcome of the MRC trial was radiological improvement of chest films (Figure 18.2). Three members of a radiological panel worked separately after the close of the trial and did not know which treatment group the patient belonged to. Specimens or other material should be reviewed and interpreted by someone who is 'blind', i.e. unaware of the outcome of randomisation.

In TB, numerous abnormalities can be observed, including atelectasis, parenchymal consolidation, lymphadenopathy, pleural effusion, cavitation, miliary pattern, etc. The MRC trial had enough printing space to report how it handled the tricky observer variability.[1] This luxury is unfortunately not available today, unless the scientific periodicals offer the unlimited space existing on the web. Atypical radiographic findings in the lungs, for example, are extremely common in HIV-infected patients. These ambivalences make it important that detailed descriptions are given of how the outcome variables were measured and how decisions on categorisation were made. Detailed information like this is of course important if the study later becomes part of a meta-analysis.

A scientific study should be reported in such a way that it can be replicated even if this is not often the case. Everyone attempts to extend the state of knowledge in their own field of science and no studies are exactly alike. It becomes increasingly difficult to understand what all the published studies tell us. Meta-analysis could further knowledge as a sort of analysis of analyses. A collection of results from individual studies are used for statistical analysis for the purpose of integrating the findings.[31] Various factors influence publication decisions but for meta-analysis it is important to include all relevant studies in the synthesis, even the non-conclusive ones. Publication bias is a problem that may seriously distort attempts to estimate the effect under investigation in a meta-analysis or a clinical trial overview. However, prevention of publication bias by publishing all studies is an ideal that is hard to achieve.

In the beginning of the 1940s patulin (a metabolite produced by several species of *Penicillium* and *Aspergillus*; originally discovered as an antibiotic, it is toxic to both animals and plants)[32] was proposed as a remedy for the common cold. The big multi-centre MRC Patulin trial undertaken in 1943–44 in the United Kingdom paved the way for the somewhat later streptomycin trial.[33] The trial was controlled with alternative blinded assignments to patulin and placebo. The treatment had no detectable effect on the natural course of the disease. The trial is an early example of how a carefully conducted large trial can throw doubt on the validity of the results of less carefully designed trials. Patulin has long since been forgotten as a cure for common cold but the trial deserves to be remembered as a good example of the importance of negative results. The public was saved from having patulin put on the market in a non-adversarial collaboration between the manufacturers of the drug and the MRC.[34]

To rigorously evaluate novel laboratory tests by an RCT approach to determine whether their use would lead to improved clinical outcomes is another rather neglected field of enquiry. A good example, however, is a double-blinded study, which assessed whether use of multiple combination bactericidal antibiotic testing (MCBT) improved clinical outcomes in patients with acute pulmonary exacerbations of cystic fibrosis.[35]

Most new laboratory tests are assessed against an established gold standard (often another laboratory test) using a cross-sectional or a cohort design. We get information only indirectly via nosologic and diagnostic accuracy given as sensitivity, specificity and predictive values. Laboratory tests could, however, be further evaluated based on how they change clinical practice, or affect the process of health care delivery.

Patients who were chronically infected with multi-resistant bacteria had sputum tested at three-month intervals for conventional culture and sensitivity tests and for MCBT (Figure 18.7). The patients who developed an exacerbation of pulmonary disease were randomised to receive a 14-day course of any of two blinded intravenous antibiotics chosen on the basis of either results from the sputum culture with sensitivity testing or the result of MCBT. The primary outcome was time from randomisation until next pulmonary exacerbation. Antibiotic therapy directed by MCBT did not result in better clinical and bacteriological outcomes compared with therapy directed by the standard techniques in the pragmatic analysis.

The risk for unintended outcomes, such as adverse drug reactions, is often not known. Side effects might not make their appearance during the intended trial period and there is sometimes

R

Baseline	Intervention	Outcome	
Acute pulmonary exacerbations in 132/251 patients with cystic fibrosis infected with multi-resistant gram-negative bacteria	Antibiotics decided by multiple combination bactericidal antibiotic testing	Time to next pulmonary exacerbation	Lung function Dyspnoea Bacteriological density
	Antibiotics decided by conventional culture and sensitivity tests	Time to next pulmonary exacerbation	Lung function Dyspnoea Bacteriological density

time ⟶

Source: Reproduced by permission.

Figure 18.7. The architecture of a randomised double-blind controlled study to assess the use of combination antibiotic susceptibility testing.[35]

a need for extended observation time to detect undesirable side effects. An extended follow-up time could also be of importance for the refutation of alleged side effects, such as proposed sequelae a long time after vaccinations. Immunological adverse effects are often unpredictable, drug interactions likewise. If the risk for adverse effects is known, it can be handled by adequate exclusion criteria in the study design, e.g. removing individuals on salicylate therapy before studying effects of *H. pylori*. Randomisation becomes necessary if a putative adverse outcome does not differ from the studied disease or outcome. Adjuvant TB immunotherapy targeted at disruption of granulomas to expose dormant bacteria to antibiotics could also result in a flaring up of the disease itself. The design of RCTs to test the greatly needed new treatment approaches indeed poses several challenges!

The claim for making trials larger so as not to be seriously misled by chance is of course valid also for infectious diseases.[5] However, an infectious disease is usually the result of a complicated and multi-layered interaction. Latterly, the concept where many diseases can be aetiologically linked to more than one pathogen has gained increased attention. Trials will continue to be needed to understand this complexity, which makes place even for small but more intricate RCTs–albeit with participation in a meta-analysis in mind when designing and reporting the trial. Based on the conceptual approach with the baseline–exposure–outcome framework, RCTs could be assessed systematically for accuracy and suitability for taking part in generating unbiased answers to aetiological and therapeutic questions in the field of infectious diseases.

REFERENCES

1. MRC (Medical Research Council). Streptomycin treatment of pulmonary tuberculosis. *Br Med J* (1948) **ii**: 769–82.
2. Doll R. Controlled trials: the 1948 watershed. *Br Med J* (1998) **317**: 1217–20.
3. Hill AB. Reflections on the controlled trial. *Ann Rheum Dis* (1966) **25**: 107–113.
4. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Chron Dis* (1967) **20**: 637–48.
5. Peto R, Baigent C. Trials: the next 50 years. Large scale randomised evidence of moderate benefits. *Br Med J* (1998) **317**: 1170–1.
6. Feinstein AR. *Clinical epidemiology. The architecture of clinical research*. Philadelphia: W.B. Saunders (1985).
7. Sackett DL. Bias in analytic research. *J Chron Dis* (1979) **32**: 51–63.
8. Jones JH. *Bad blood: The Tuskegee syphilis experiment*. London: Collier Macmillan (1981).
9. Baker SM, Brawley OW, Marks LS. Effects of untreated syphilis in the negro male, 1932 to 1972: a closure comes to the Tuskegee study, 2004. *Urology* (2005) **65**(6): 1259–62.
10. Fraser CM, Norris SJ, Weinstock GM *et al*. Complete genome sequence of Treponema pallidum, the syphilis spirochete. *Science* (1998) **281**(5375): 375–88.
11. Ford A, Delaney B, Moayyedi P. Eradication therapy for peptic ulcer disease in *Helicobacter pylori*

positive patients. Cochrane Database Syst Rev (2003) Issue 4.

12. Axon ATR, Ó'Moráin CA, Bardhan KD *et al*. Randomised double blind controlled study of recurrence of gastric ulcer after treatment for eradication of Helicobacter pylori infection. *Br Med J* (1997) **314**: 565–68.
13. Kearney DJ. *Helicobacter pylori* infection. *Curr Treat Options Infect Dis* (2003) **5**: 197–206.
14. IARC. Infection with *Helicobacter pylori*. In: *IARC Monographs on the Evaluation of the Carcinogenic Risks to Humans*, Vol. 61, *Schistosomes, liver flukes and Helicobacter pylori*. Lyons: International Agency for Research on Cancer (1994) 177.
15. Achtman M, Suerbaum S, eds. *Helicobacter pylori: Molecular and Cellular Biology*. Wymondham, United Kingdom: Horizon Scientific Press (2001).
16. Uemura N, Okamoto S, Yamamoto S *et al*. *Helicobacter pylori* infection and the development of gastric cancer. *New Engl J Med* (2001) **345**: 784–9.
17. O'Mahony R, Vaira D, Holton J, Basset C. *Helicobacter pylori*: current status and future prospects. *Sci Prog* (2004) **87**(Pt 4): 269–96.
18. Marshall BJ. Helicobacter pylori. *Am J Gastroenterol* (1994) **89**(8): 116–28.
19. Frieden TR, Sterling TR, Munsiff SS, Watt CJ, Dye C. Tuberculosis. *Lancet* (2003) **362**: 887–99.
20. Borkow G, Weisman Z, Qibin, L *et al*. Helminths, human immunodeficiency virus and tuberculosis. *Scand J Infect Dis* (2001) **33**: 568–71.
21. Wright J, Walley J, Aby P *et al*. Direct observation of treatment for tuberculosis: a randomized controlled trial of community health workers versus family members. *Trop Med Int Health* (2004) **9**: 559–65.
22. Dolo A, Modiano D, Boubacar M *et al*. Differences in susceptibility to malaria between two sympatric ethnic groups in Mali. *Am J Trop Med Hyg* (2005) **72**(3): 243–48.
23. WHO (2005) Tuberculosis. http://www.who.int/topics/tuberculosis (accessed October 2005).
24. Tuberculosis Research Centre (Indian Council of Medical Research), Chennai. Shortening short course chemotherapy: a randomised clinical trial for treatment of smear positive pulmonary tuberculosis with regimens using ofloxacin in the intensive phase. *Indian J Tuberc* (2002) **49**: 27–38.
25. WHO. *Treatment of Tuberculosis: guidelines for national programmes*, 3rd edn. Geneva: WHO/CDS/TB (2003) **313**.
26. Wallis RS. Reconsidering adjuvant immunotherapy for tuberculosis. *Clin Infect Dis* (2005) **41**: 201–8.
27. Snow RW, Craig M, Deichmann U, Marsh K. Estimating mortality, mobility and disability due to malaria among Africa's non-pregnant population. *Bull WHO* (1999) **77**(8): 624–40.
28. Nosten F, Luxemburger C, Kyle DE *et al*. Randomised double-blind placebo-controlled trial of SPf66 malaria vaccine in children in northwestern Thailand. *Lancet* (1996) **348**: 701–7.
29. Schellenberg D, Menendez C, Kahigwa E *et al*. Intermittent treatment for malaria and anaemia control at time of routine vaccinations in Tanzanian infants: a randomised, placebo-controlled trial. *Lancet* (2001) **357**: 1471–7.
30. Cuzin-Ouattara N, van den Broek AHA, Habluetzel A *et al*. Wide-scale installation of insecticide-treated curtains confer high levels of protection against malaria transmission in a hyperendemic area of Burkina Faso. *Trans R Soc Trop Med Hyg* (1999) **93**: 473–9.
31. Glass GV. Integrating findings: the meta-analysis of research. *Rev Res Educ* (1978) **5**: 351–79.
32. Katzman PA, Hays EE, Cain JJ *et al*. Clavacin, an antibiotic substance from *Aspergillus clavatus*. *J Biol Chem* (1944) **154**: 475–86.
33. MRC Patulin Trials Committee. Clinical trials of Patulin in the common cold. *Lancet* (1944) **ii**: 373–5.
34. Chalmers I, Clarke M. Commentary: The 1944 patulin trial: the first properly controlled multicentre trial conducted under the aegis of the British Medical Research Council. *Int J Epidemiol* (2004) **32**: 253–60.
35. Aaron SD, Vandemheen KL, Ferris W *et al*. Combination antibiotic susceptibility testing to treat exacerbations of cystic fibrosis associated with multiresistant bacteria: a randomised, double-blind, controlled clinical trial. *Lancet* (2005) **366**: 463–71.

OPHTHALMOLOGY

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

19

Ophthalmology

JONATHAN C.K. CLARKE AND PENG-TEE KHAW
Moorfields Eye Hospital and Institute of Ophthalmology, University College London, London, UK

INTRODUCTION

Ophthalmologists treat diseases of the eye, the surrounding tissue and in association with neurologists the visual pathways. Modern ophthalmic practice has led to the organisation of subspecialist interests such as medical retina, surgical vitreoretina, adnexal (lids and associated structures), anterior segment (cornea and cataract), glaucoma, strabismus and paediatrics and neuro-ophthalmology. There is a certain amount of crossover between the subspecialities and the basic training of an ophthalmologist includes experience in all of these areas. Systemic diseases with eye involvement such as diabetes and hypertension are common and a close association with the hospital physician is often formed with the ophthalmologist. In addition, ophthalmologists work closely in multidisciplinary teams with dedicated ophthalmic nurses, orthoptists (who measure children's eyesight and eye movements), optometrists and associated technicians.

Most eye referrals are generated by a visit to a local optometrist whose letter is forwarded to the Eye Department by the General Practitioner who adds relevant medical details. Eye disease is common. A study at a general practice in a London community health centre identified that patients with eye symptoms represented 2.7% of all medical consultations.[1] Diseases that lead to blindness especially in the young and working population have grave implications for the patients' prospects both socially and economically. There is a network available to help such individuals and much of this is accessed once the patient is registered either blind or partially sighted. It is possible with the appropriate support for blind patients to function very well in society.

Globally the challenge of treatable blindness is tremendous. The World Health Organization developed the "Vision 2020: Fight for Sight" programme in response to this. Estimates of the number of people worldwide with either preventable or treatable blindness stood at 37.9 million in 1994 when the Vision 2020 programme was developed and subsequently launch in 1999. This number was expected to double by the year 2020. Five conditions – cataract, refractive errors, trachoma, onchocerciasis and vitamin A deficiency – were determined to be responsible for 75% of blindness in the developing world and for each of these conditions an effective, cost-efficient intervention is available.[2] In the UK

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

the number of people registered as blind is over 150 000 with an additional equal number registered as partially sighted.[3]

For the purposes of this chapter a limited number of specific diseases have been selected to represent a variety of issues that are pertinent to ophthalmology. These diseases account for a large proportion of blindness and visual morbidity in the UK. General issues concerning visual and non-visual outcomes are discussed and expanded further in the sections concerning surgical devices and treatment of chronic diseases. Issues concerning trial methodology are also described.

GENERAL OUTCOME ISSUES PERTAINING TO OPHTHALMOLOGY

Outcome measurements continue to cause debate. For chronic diseases such as glaucoma and diabetic retinopathy successful treatment may be considered the prevention of progression rather than reversal of the disease. Surgical treatment of cataracts with monofocal or multifocal intraocular lenses reverses the pathological state. Examples of trials which have addressed these issues are described.

MEASURING VISUAL ACUITY OUTCOMES

Nearly every trial in ophthalmology attempts to describe observed changes in visual acuity. Outcome can be measured in changes in visual acuity and this is performed with acuity charts. The commonest chart in use has been the *Snellen chart*. However, statistical analysis of results is difficult and unsatisfactory as the numbers of letters per line increases as the size of the letters reduces. The LogMAR chart is now considered the gold standard for statistical analysis of visual acuity as the gradation in size of letters is continuous and the number of letters per line is equal (Figure 19.1).

OTHER MEASURES OF VISUAL FUNCTION

There are other measures of visual functions; the case can be made that the ability to read letters at a set distance is not a good guide to

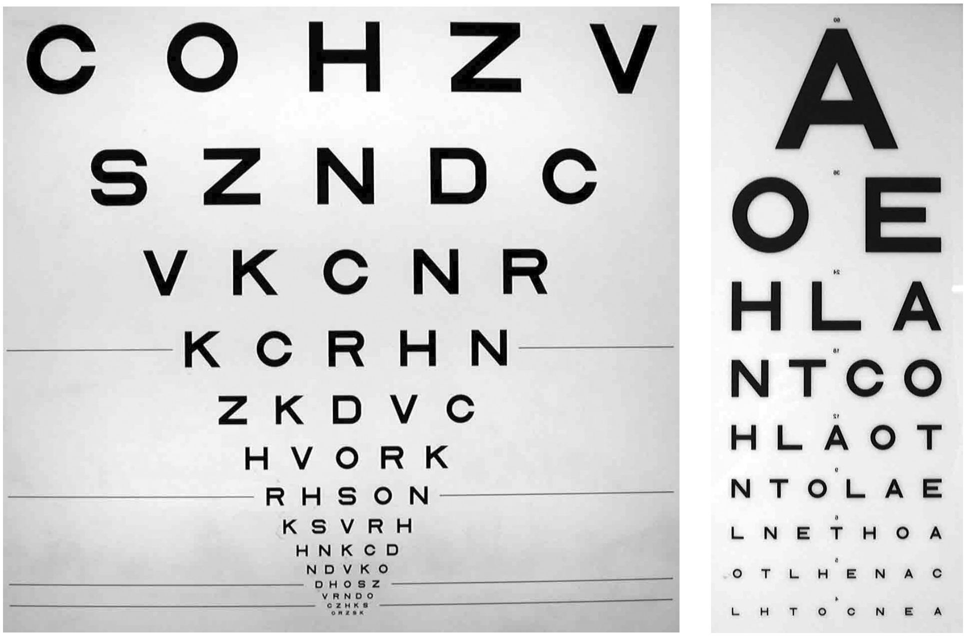


Figure 19.1. The LogMAR and Snellen charts.

visual function in real life. The Pelli–Robson contrast sensitivity chart may show a severe limitation of visual function in a patient who has good LogMAR visual acuity. There are other recognised measures of visual function such as kinetic visual fields, automated non-kinetic visual fields, colour vision assessments and methods of assessing Snellen equivalent visual acuity in children using distraction cards or picture cards as examples. The final decision as to which is the appropriate choice will depend on which disease is being studied. For example, diseases of the optic nerve are commonly assessed with colour vision, glaucoma with non-kinetic peripheral field analysis and cataract surgery outcomes with contrast sensitivity and visual acuity for both distance and near.

NON-VISUAL OUTCOMES

Visual function is nearly always recorded as an outcome in trials in ophthalmology. However, the primary outcome may not assess vision. The restoration of anatomy such as reattachment of the retina following detachment surgery is a valid outcome. Expectation of improved visual function would follow but can be unpredictable and the benefits of one technique over another may be better assessed with measuring reattachment rates. Some procedures are not expected to have a beneficial outcome on vision. The correction of strabismus (squint) to align both eyes and improve the cosmetic appearance is measured with prism dioptres of misalignment for both near and distance fixation. Many chronic eye diseases result in pain for the patient and measurements of pain with approved pain scales have been adopted. In an example of this a trial of excimer laser for bullous keratopathy, where corneal endothelial function is poor either due to inherited dystrophy or following trauma, reported changes in symptoms following various applications of the excimer laser.[4]

Intraocular pressure measured with a tonometer is the primary outcome in most glaucoma trials and new treatments are assessed by their ability to lower intraocular pressure. Further discussion of outcomes in glaucoma is addressed in the chronic diseases section.

Many clinical trials now attempt to assess patient satisfaction in addition to visual acuity which may give the truest representation of the patient's visual function.

OPHTHALMOLOGY AND CLINICAL TRIALS

The specialty lends itself well to clinical trials and over 1000 have been published in journals in the past 10 years. Cochrane has a dedicated section: The Cochrane Eyes and Vision Group with over 40 published reviews and many more submitted protocols. Randomised controlled trials are described as masked rather than blinded for obvious reasons in ophthalmology.

IMPORTANT METHODOLOGICAL ISSUES

CROSSOVER TRIAL

Crossover trials characteristically allow the participant to receive all of the study interventions in successive periods. The participant is given the intervention in a random sequence. Each participant acts as his or her control and can allow for statistically and clinically significant results in a smaller number of subjects. It might be considered that even more rapid results could be achieved if the two eyes of a patient were considered separately and the results of specific interventions on each eye compared. For example, two drops used to reduce intraocular pressure could be compared by placing drop A in the right eye and drop B in the left eye. However, these trials are not usually carried out as there is a certain amount of systemic absorption and a topical antihypertensive instilled into one eye will usually have a small effect on the other eye.

Carryover actions when the first intervention still has some effect even after it has been stopped is an issue e.g. glaucoma drops. However, these are often overcome with a short (two-week) period of wash-out.

Representative examples of published trials include a trial of the efficacy of topical anaesthetics to reduce pain in premature infants during eye examinations for retinopathy of prematurity.[5] Eye examinations for retinopathy of prematurity (ROP) are stressful and probably painful, but many ophthalmologists do not apply topical anaesthetics because their efficacy in reducing pain has not been established. In this trial the potential benefits of topical anaesthetic eye drops in reducing pain during neonatal eye examination for ROP were assessed. Neonates born at 30 or less weeks gestation and expected to have at least two examinations for ROP were included in the trial. Patients were randomly assigned to receive either proparacaine HCl ophthalmic solution 0.5% or NaCl 0.9% (saline) eye drops prior to an eye examination. In a subsequent examination, each patient received the alternate treatment. Eye drops were prepared in the pharmacy in identical tuberculin syringes, and physicians, nurses and pharmacists were blinded to the treatment given. Pain was measured using a scoring system with both physical and physiologic measures of pain, which had been validated in preterm infants. The results showed that patients experienced significantly less pain at speculum insertion with proparacaine than with saline. The investigators concluded that topical anaesthetic pre-treatment reduces the pain response to eye examination for ROP and should become routine practice.

A second example was a comparison of autologous serum eye drops with conventional therapy in a randomised controlled crossover trial for ocular surface disease[6] aimed at evaluating the efficacy of 50% autologous serum drops against conventional treatment in ocular surface disorders refractory to normal treatments in a prospective randomised crossover trial. Patients fulfilling ophthalmological and haematological entry criteria were randomised to either three months of autologous serum 50% followed by three months of their conventional treatment, or three months of conventional treatment, followed by three months of autologous serum. Clinical assessments of tear film quality (Schirmer's test, rose Bengal, and fluorescein staining) were carried out on entry and at monthly intervals. Impression cytology was performed at entry, three and six months. Grading was carried out on degrees of squamous metaplasia and goblet cell density. Subjective comfort was recorded daily using the 'faces' scale. These categorical scores were converted to linear measurement using Rasch analysis. Impression cytology available in 25 of 31 eyes showed significant improvement on serum treatment. Rasch-weighted faces scores were statistically significantly better with serum and it was concluded that the results of this randomised study provided evidence of the beneficial effects of autologous serum in severe ocular surface disorders. For most of these patients, autologous serum was superior to conventional treatment for improving ocular surface health and subjective comfort.

These examples illustrate the benefits of crossover trials in assessing treatments for rare diseases where large randomised control trials are not feasible.

READING CENTRES

A number of clinical trials, typically large multicentre type, have utilised Reading Centres to analyse their findings.[7–24] Reading Centres exist in many institutions in both the United States and Europe and are gaining a reputation for providing the gold standard in outcome assessment. The outcome being assessed is imaged, typically photographed, although any imaging technique is appropriate, and then provided to the Centre. The trained graders who are not usually medically qualified use protocols developed by the investigating team or published for use in previous trials. A series of categorical outcomes can then be presented and the benefit of treatment over no treatment or one treatment option over another can be identified. Where examined the inter- and intra-observer agreement is very high and agreement between the trained specialist and the graders is high.[25,26] Applications have been successful in many different ophthalmic subspecialties. For example, the European Glaucoma Prevention Study[21] assessed the benefit of treatment with the topical agent Dorzolamide against topical placebo in a large cohort of randomised ocular hypertensives. Part of the assessment, performed at a Reading Centre, was the analysis of optic nerve head cup to disc ratio which enlarges with progression of glaucoma.

Many of the earlier published uses of Reading Centres are in the field of medical retina and in particular age-related macular degeneration studies. The use of colour fundus photographs and fluorescein angiograms in the identification and treatment of choroidal neovascularisation in the Macular Photocoagulation Study[15] utilised a newly developed standard set of methods for assessing colour photographs and fluorescein angiograms on study patients. For pretreatment angiograms, these methods are used to determine the location and extent of the choroidal neovascularisation. For post-treatment colour fundus photographs, these methods are used to assess the extent and intensity of treatment. Although these methods were developed to judge eligibility and treatment of patients enrolled in the study, they provide an excellent way for practicing ophthalmologists to evaluate their patients' angiograms and to assess immediately the intensity and extent of laser photocoagulation. Since persistent neovascularisation is highly correlated with incomplete and/or inadequate photocoagulation treatment, clinicians were able to adopt these Reading Centre techniques to minimise the frequency of persistent neovascularisation and, possibly, to reduce the frequency of visual loss in treated eyes. The grading systems developed can have direct clinical implications and correlations with outcome and have helped unify the description of outcomes previously open to subjective interpretation.

CLINICAL TRIALS OF CHRONIC DISEASES

GLAUCOMA

Glaucomatous optic neuropathy is the worldwide leading cause of irreversible blindness; affecting around 70 million individuals with at least 6.8 million bilaterally blind.[27] The glaucomas as a group have in common a slow progressive degeneration of retinal ganglion cells and their axons, resulting in a distinct appearance of the optic disc and a concomitant pattern of visual loss.[28] More than 7 million outpatient visits per year are required to monitor glaucoma in the United States. The initial treatment typically is the application of topical antihypertensives. There have been recent additions to the classes of antihypertensives available and control of intraocular pressure (IOP) is often achieved with their use. However, a minority of patients require surgical procedures to control IOP and these are often augmented with topical antiscarring agents. With the increase in the number of both medical and surgical treatments available, identifying the most appropriate regime is not always straightforward.

The stimulus to investigate the outcomes of treatment in glaucoma came with the publication of a meta-analysis in 1993 which questioned the validity of previously published RCTs in glaucoma treatment.[29] The study found that serious methodologic problems with the trials that were reviewed existed. Areas of major concern were: use of unsatisfactory or unspecified methods of randomisation (89% of the trials reported no information), exclusion of some patients from the analysis (53% of the studies), failure to provide evidence of having estimated the number of patients needed to detect a prespecified treatment difference (96% failed to provide such an estimate), and incomplete description of patient characteristics (in 39% of the RCTs information on this item was insufficient). In conclusion for clinicians to make use of the results of clinical trials, future studies must be adequately designed and conducted. In particular, a proper method of randomisation, masking of the observers and inclusion of all randomised patients in the analysis must be used. Of perhaps even greater importance was the need for trials to measure clinically relevant outcomes. These conclusions even allowed some to question whether IOP lowering provided any benefit to patients with glaucoma and the search for better designed and performed RCTs commenced. It was already accepted that IOP was a strong risk factor for the development of glaucoma and not the underlying aetiology. Currently accepted treatments of glaucoma reduce the IOP and the benefit of one treatment over another has, until recently, depended on the treatment's ability to reduce IOP.

Table 19.1 contains a summary of the major trials that have been published in the past decade and a summary of their principal results which have confirmed the value of reducing IOP in patients with ocular hypertension (statistically raised IOP without glaucomatous field or optic disc damage), or primary open angle glaucoma (POAG) with pressure levels in both the statistically raised or normal range of IOP. This is to prevent the onset of glaucoma in the case of ocular hypertension and the progression of disease in the latter group of individuals.

VISUAL FIELD ANALYSIS IN GLAUCOMA

Glaucoma

The reduction of IOP has demonstrated that loss of peripheral vision, the primary type of visual loss in glaucoma, can be reduced. Characteristically visual field loss is assessed with automated, static field analysers (such as the Humphrey Visual Field Analyser). The process requires considerable concentration by the patient and except in the most dedicated patient fluctuations in outcome can be expected. This issue is usually addressed by asking the patient to perform at least two baseline field tests before treatment is initiated and the average of these tests would be used to assess further deterioration.

Identification of progression of field loss is controversial. Reliable identification of progressive visual field loss requires two consecutive tests to confirm the acquired defect and allow false positives to be excluded. The definition of the progression i.e. whether a patient's field test has worsened due to the disease, differs between some of the landmark trials in Table 19.1. Arguably, as long as one system is used throughout the course of a trial then there should be less cause for concern and this has always been the case. However, when the criteria for failure have been applied to one group of patients and agreement assessed the results were disappointing.[41] It is hoped that a uniform definition for progression of visual field loss can be agreed and this may come in the form of pointwise linear regression software such as Progressor.[42]

Worldwide variations

Racial differences in the type and age of presentation of glaucoma have been well documented and reveal dramatic variations. In surveys of European derived populations the rates of blindness in patients with a glaucoma diagnosis were estimated at 4.4% whilst in African Americans the estimate was 7.9%.[43] In addition blindness presents at a younger age in the African Americans and the burden of disease is far greater in this population. The prevalence of POAG is common in Europeans compared with the other forms of glaucoma (90%) but angle closure glaucoma accounts for 50–75% of glaucoma cases in East Asian populations.[44] This is due to the anatomical differences of the iridocorneal angle in this population. Interpretations of trial outcomes in one population and extrapolation to a second can raise validity doubts.

DIABETIC RETINOPATHY

Diabetic retinopathy is one of the commonest causes of new blindness in adults, with loss of vision most likely to be associated with proliferative diabetic retinopathy (PDR) in type I diabetes and with maculopathy in type II diabetes. Maculopathy causes 90% of blindness due to diabetes. There are conflicting reports regarding the incidence and prevalence of visual impairment,[45] but the Wisconsin Epidemiologic Study of Diabetic Retinopathy reported a frequency of any visual impairment in people with diabetes of 7.8% and an estimated annual incidence of blindness due to diabetes of 3.3 per 100 000 total population. As the prevalence of diabetes increases, the demand for ophthalmic health care is likely to rise.

Table 19.1. Some recent, major, multicentre clinical trials across the spectrum of primary open angle glaucoma

Study	Aim	Result
Ocular Hypertension Treatment Study (OHTS)[30,31]	Efficacy and safety of topical ocular medications in preventing or delaying the development of primary open angle glaucoma in individuals with elevated IOP (1636 patients)	The probability of developing glaucomatous change (optic disc or field change) was 4.4% in the medication group and 9.5% in the observation group at 60 months. Little evidence of systemic or ocular risk of treatment except an increased percentage of treated patients having cataract surgery (6.4% vs. 4.3%). Average reduction in IOP with treatment was 22.5%. Baseline age, vertical cup disc ration, visual field damage and intraocular pressure were good predictors of progression, corneal thickness was a powerful predictor of progression
Glaucoma Laser Trial (GLT)[32]	Efficacy and safety of argon laser trabeculoplasty or medical as initial treatment in primary open angle glaucoma (271 patients)	Eyes treated with LTP had a slightly lower IOP of 1.2 mmHg and (0.6 dB) better improvement in visual field with a median follow up of seven years
Collaborative Initial Glaucoma Treatment Study (CIGTS)[33]	Effects of randomising patients to either initial medical or surgical treatment (607 patients)	Surgery lowered the IOP more (average 14 to 15 mmHg vs. 17 to 18 mmHg) but with no statistical difference in visual field progression over five years. More initial visual acuity change in surgery group, more cataracts formation and more local symptoms in quality of life analysis
Early Manifest Glaucoma Treatment Trial (EMGTT)[34,35]	Effects of treatment with a beta blocker and laser trabeculoplasty versus observation in patients with newly detected primary open angle glaucoma (255 patients)	Progression (visual field or optic disc) was less frequent in the treatment group (45% vs. 62%) with a median follow-up of six years, Treatment was associated with a greater increase in lens opacity gradings. Other important predictors of glaucoma progression included lens exfoliation, bilateral glaucoma, IOP greater than 21 mmHg, more advanced visual field loss, disc haemorrhages and age ≥68 years
Fluorouracil Filtering Surgery Study (FFSS)[36]	Effects and safety of subconjunctival injections of 5-fluorouracil after glaucoma surgery in patients with a poor prognosis (213 patients)	After one year only 27% of the treated group had failed injections versus 50% of the control group. Corneal epithelial toxicity and transient visual acuity loss were more common in the 5-FU treated group, but were not significantly different at one year
Collaborative Normal Tension Glaucoma Study (CNTGS)[37,38]	Effect of pressure lowering (30%) on optic nerve damage and field loss in normal tension glaucoma (glaucoma with intraocular pressure in "normal" range) (140 patients)	In the treated patients only 12% progressed (optic disc and visual field progression) compared with 35% in the untreated group. There was a higher incidence of cataract progression in the treated group (38% vs. 14%), particularly in those that had glaucoma surgery.
Advanced Glaucoma Intervention Study (AGIS)[39,40]	Effect of treatment sequences of laser trabeculoplasty and trabeculectomy (surgery) in advanced glaucoma (776 eyes of 581 patients)	In this study the outcome (reduction in visual acuity or visual fields) depended on race. In patients who had laser trabeculoplasty first black patients were at a lower risk of failure than white patients of failure of first intervention, In patients who received surgery first black patients were at a higher risk of first failure than white patients

Moderate to severe visual loss from diabetes is preventable,[46] and a screening protocol currently being set up throughout the UK hopes to identify most patients with vision-threatening retinopathy.[47] Proliferative retinopathy is where new vessels grow in a poorly controlled manner either from the optic disc (DNV) or from other sites of the retina (NVE). These new vessels have the potential to bleed into the vitreous with a resulting vitreous haemorrhage or to fibrose and induce a tractional retinal detachment. In diabetic maculopathy exudative lipoprotein from leaking retinal capillaries and microaneurisms aggregates in the retina, causing thickening of the retinal tissue and loss of function. If this process is found within a specific distance from the fovea then it is classified as "clinically significant macular oedema" (CSMO) and treatment protocols have been developed if this finding is present.

Landmark studies have governed the treatment of diabetic retinopathy and were amongst the first to be published in ophthalmology. The Diabetic Retinopathy Study posed the question: does photocoagulation surgery reduce the risk of severe visual loss in diabetic retinopathy?[48] The objectives were twofold: to better establish the natural history of diabetic retinopathy without photocoagulation and to compare the effects of treatment techniques involving extensive scatter photocoagulation and focal treatment of new vessels with either a xenon or argon laser. Inclusion and exclusion criteria were defined from the outset and the trial was run from 15 centres. Fundus photographs were collected and analysed by the Wisconsin Reading Centre. A total of 1727 patients were enrolled. Patients were randomised to either xenon or argon laser treatment to one of their eyes. Their other eye acted as a control. Detailed treatment protocols were laid out that could be translated into clinical practice. The patients were followed up six weeks and four months after treatment and at four-month periods thereafter. Certified technicians recorded visual acuity, visual fields and other outcome indices in a masked fashion.

The results identified the risks of developing blindness from neovascularisation and in addition the considerable benefit for long-term prognosis that pan retinal argon laser treatment is able to provide. The results of this trial published in 1981[49] remain the foundation for the treatment of proliferative diabetic retinopathy. Subsequent trials investigating the role of laser photocoagulation are described in Table 19.2.

TRIALS OF SURGICAL DEVICES

TYPES OF DEVICES

There are a number of surgical devices in use and under trial in ophthalmology such as intravitreal implants for slow release of drugs, and silicon drainage tubes creating a fistula between the anterior chamber and subconjunctival space for controlling IOP in complex glaucoma. The most widely used device is the intraocular lens which is placed in the eye following cataract surgery. A comparison of outcomes between multifocal intraocular lenses and monofocal lenses is described below. This area has been the subject of a Cochrane Review and illustrates the many areas where poor trial design impedes outcome comparison.

Cataract surgery is accompanied by the implantation of an intraocular lens (IOL). Standard monofocal IOLs focus at one fixed distance – usually in the far distance. This means that most people will require spectacles in addition to monofocal IOLs. Multifocal intraocular lens implants reduce spectacle dependence after cataract surgery but at the expense of quality of vision, particularly contrast sensitivity. Multifocal IOLs allow the patient to focus at more than one distance.

Optical evaluation of multifocal IOLs suggests a two- to threefold increase in the depth of field is achieved at the expense of a 50% reduction in the contrast of the retinal image.[54,55] Clinical evaluation of a multifocal IOL is less clear cut. Several large studies, including non-randomised comparisons with monofocal IOLs, have indicated that the quality of vision with bifocal and multifocal IOLs is good.[56,57] Whether the optical trade-off inherent in a multifocal

Table 19.2. Trials of laser photocoagulation in proliferative diabetic retinopathy (PRP)

Study	Key points
The Diabetic Retinopathy Study[49]	Photocoagulation reduced the risk of severe vision loss compared with no treatment and identified a stage of retinopathy, termed high-risk PDR, in which the benefits of photocoagulation outweighed the risks
Diabetic Control and Complications Trial[50]	Patients who monitored their glucose closely (four measurements per day = tight control) do far better than patients treated with conventional therapy (one measurement per day). The former had a 76% reduction in the rate of development of any retinopathy and an 80% reduction in progression of established retinopathy versus those with conventional control. For advanced retinopathy, however, even the most rigorous control of blood glucose may not prevent progression
The Early Treatment Diabetic Retinopathy Study proliferative diabetic retinopathy outcomes[51]	PRP significantly retards the development of neovascularisation in high-risk characteristics (HRC) in eyes, very severe non-proliferative diabetic retinopathy and macular oedema. After seven years of follow-up 25% of eyes which received PRP developed HRC as compared with 75% of eyes in which PRP was deferred until HRC developed. However, visual loss can be prevented if patients are closely monitored and PRP performed once HRC develop. This reduces the number of patients that require PRP
The Early Treatment Diabetic Retinopathy Study macular oedema outcomes[52]	Clinically significant macular oedema (CSMO) defined as: Retinal thickening involving the centre of the macula Hard exudates within 500 μm of the centre of the macula An area of macular oedema > 1 DD but within 1 DD of the centre of macula Treatment strategy: to place a grid of light laser treatment burns over the affected area up to 500 μm away from the fovea Outcome: After three years of follow-up 15% of eyes with CSMO had a doubling of their visual angle as opposed to 32% of untreated eyes
The Diabetic Retinopathy Vitrectomy Study[53]	Eyes with recent severe diabetic vitreous haemorrhage reducing visual acuity to 5/200 or less for at least one month were randomly assigned either to early vitrectomy or deferral of vitrectomy for one year The proportion of eyes with visual acuity of 10/20 or better was higher in the early vitrectomy group than in the deferral group throughout the four-year follow-up period. Up to the 18-month visit, the early group had a higher proportion of eyes with visual acuity of no light perception An increased chance of obtaining good vision with early vitrectomy was clearly present in the type I diabetes group, particularly in patients who developed severe vitreous haemorrhage after less than 20 years of diabetes, a patient group tending to have more severe proliferative retinopathy. This advantage was not found in the type II diabetes group, in which patients were older and tended to have less severe retinopathy The findings support early vitrectomy in eyes known or suspected to have very severe proliferative diabetic retinopathy as a means of increasing the chance of restoring or maintaining good vision

IOL results in better or worse visual function compared with a monofocal IOL was the question posed by a Cochrane Review and a meta-analysis of the appropriate RCTs was published.[58]

Outcomes

The primary outcomes for the review were:

1. Distance visual acuity (unaided and corrected)
2. Near visual acuity (unaided and corrected)
3. Spectacle dependence.

The secondary outcomes included

depth of field, contrast sensitivity, glare and validated instruments assessing quality of life or visual function.

Study Designs

Only eight studies were included in this review of the 239 abstracts identified. This was due to the majority of the studies not meeting acceptable criteria listed below:

- Described as randomised.
- Double-masked procedure.
- Description of withdrawals.
- Acceptable method of randomisation.
- Method of masking–masking was considered to relate to masking of participants as to their IOL status, and of those performing post-operative assessments. Masking of the surgeon is in this case not realistically feasible, as the IOLs are demonstrably different at the time of surgical implantation.

Loss of information resulting from the non-acceptance of so many trials weakens the conclusion that can be made from the review.

Outcomes

Outcomes are recorded in Table 19.3.

Unaided near vision is critical to the assessment of multifocal efficacy, but was reported in a manner that makes comparison between studies difficult. Reading distances differ, and it is not made clear in most studies whether the reported print size read has been corrected for reading distance so as to allow a near visual acuity to be calculated. Only two studies explicitly reported near visual acuity but it was concluded from these outcomes that near vision is improved with the use of multifocals.

Table 19.3. Table of Outcomes

Outcome	Finding
Distance visual acuity	No difference
Depth of field	Improved with the multifocal IOL
Contrast sensitivity	Lower with the multifocal IOL
Objective glare	No significant differences

Subjective outcomes are fundamental to the evaluation of multifocal IOLs, but, like near vision, measurements were flawed in most of the studies. There was no consistent effect on visual satisfaction apparent from examination of the non-validated assessments. The two studies using a validated measure differed slightly with one finding a small but statistically significant increase in overall visual satisfaction using the multifocal IOL, and a larger effect with respect to near vision. The second found no difference in overall satisfaction (both groups had a median 8/10), with some increase in unaided near vision satisfaction (median 5/10 monofocal versus 7/10 for multifocal).

Adverse subjective visual phenomena, particularly haloes, or rings around lights, were more prevalent and more troublesome in participants with the multifocal IOL. The lack of a consistent drop in patient satisfaction despite the prevalence of these phenomena could be interpreted as evidence that patients do not perceive them as severe. Spectacle independence is more likely to be achieved with use of the multifocal IOL than monofocal IOL. However, in no study did more than half of the participants achieve spectacle independence.

Conclusions

The authors concluded that multifocal IOLs are effective at improving near vision relative to monofocal IOLs. Whether that improvement outweighs the adverse effects of multifocal IOLs will vary between patients, with motivation to achieve spectacle independence likely to be the deciding factor. Future research on these and similar IOLs should use validated subjective outcome criteria and strive for clarity in reporting of objective outcomes, particularly near vision.

The lack of well-designed trials has made evidence-based conclusions difficult to make and its is clear that despite many of the trials concerning multifocal lenses being designed in the past 10 years, many of them are of a poor quality.

SUMMARY OF TRIALS IN OPHTHALMOLOGY

There are many features, particularly concerning measuring outcomes, that are unique to ophthalmology. The embracing of Reading Centres is particularly prominent. High-quality trials have provided evidence-based medicine in this field for many years; however, many trials published even within the past decade are considered inadequately designed. There is a history of scrutinising trial design and outcomes within the specialty and many of the current trials underway are of the highest standard.

In the future improved technology may provide new and exciting surrogate outcomes such as *in vivo* retina cellular imaging[59].

ACKNOWLEDGEMENTS

Our research has been supported by the Medical Research Council G9330070, the Wellcome Trust, the Guide Dogs for the Blind Association, Moorfields Eye Hospital Trustees, Singapore National Medical Research Council, the Haymans Trust, the Michael and Ilse Katz Foundation and the Helen Hamlyn Trust in memory of Paul Hamlyn.

REFERENCES

1. Dart JK. Eye disease at a community health centre. *Br Med J (Clin Res Ed)* (1986) **293**: 1477–80.
2. World Health Organization. *State of the World's Sight: Vision 2020: The right to sight 1999–2005*. Geneva, WHO (2005).
3. Department of Health. *Registered Blind and Partially Sighted People Year ending 31 March 2003*. London: Department of Health (2003).
4. Maini R, Sullivan L, Snibson GR, Taylor HR, Loughnan MS. A comparison of different depth ablations in the treatment of painful bullous keratopathy with phototherapeutic keratectomy. *Br J Ophthalmol* (2001) **85**: 912–15.
5. Marsh VA, Young WO, Dunaway KK, Kissling GE, Carlos RQ, Jones SM *et al*. Efficacy of topical anesthetics to reduce pain in premature infants during eye examinations for retinopathy of prematurity. *Ann Pharmacother* (2005) **39**: 829–33.
6. Noble BA, Loh RS, MacLennan S, Pesudovs K, Reynolds A, Bridges LR *et al*. Comparison of autologous serum eye drops with conventional therapy in a randomised controlled crossover trial for ocular surface disease. *Br J Ophthalmol* (2004) **88**: 647–52.
7. The Age-Related Eye Disease Study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the Age-Related Eye Disease Study Report Number 6. *Am J Ophthalmol* (2001) **132**: 668–81.
8. Adamis AP, Altaweel M, Bressler NM, Cunningham ET Jr, Davis MD, Goldbaum M *et al*. Changes in retinal neovascularization after pegaptanib (Macugen) therapy in diabetic individuals. *Ophthalmology* (2006) **113**: 23–8.
9. Azab M, Boyer DS, Bressler NM, Bressler SB, Cihelkova I, Hao Y *et al*. Verteporfin therapy of subfoveal minimally classic choroidal neovascularization in age-related macular degeneration: 2-year results of a randomized clinical trial. *Arch Ophthalmol* (2005) **123**: 448–57.
10. Azar RG, Holdbrook MJ, Lemp M, Edelhauser HF. Two-year corneal endothelial cell assessment following INTACS implantation. *J Refract Surg* (2001) **17**: 542–8.
11. Barr JT, Schechtman KB, Fink BA, Pierce GE, Pensyl CD, Zadnik K *et al*. Corneal scarring in the Collaborative Longitudinal Evaluation of Keratoconus (CLEK) Study: baseline prevalence and repeatability of detection. *Cornea* (1999) **18**: 34–46.
12. Bressler NM, Bressler SB, Hawkins BS, Marsh MJ, Sternberg P Jr, Thomas MA. Submacular surgery trials randomized pilot trial of laser photocoagulation versus surgery for recurrent choroidal neovascularization secondary to age-related macular degeneration: I. Ophthalmic outcomes submacular surgery trials pilot study report number 1. *Am J Ophthalmol* (2000) **130**: 387–407.

13. Bressler NM, Arnold J, Benchaboune M, Blumenkranz MS, Fish GE, Gragoudas ES *et al*. Verteporfin therapy of subfoveal choroidal neovascularization in patients with age-related macular degeneration: additional information regarding baseline lesion composition's impact on vision outcomes-TAP report No. 3. *Arch Ophthalmol* (2002) **120**: 1443–54.
14. Bressler SB, Pieramici DJ, Koester JM, Bressler NM. Natural history of minimally classic subfoveal choroidal neovascular lesions in the treatment of age-related macular degeneration with photodynamic therapy (TAP) investigation: outcomes potentially relevant to management – TAP report No. 6. *Arch Ophthalmol* (2004) **122**: 325–9.
15. Chamberlin JA, Bressler NM, Bressler SB, Elman MJ, Murphy RP, Flood TP *et al*. The use of fundus photographs and fluorescein angiograms in the identification and treatment of choroidal neovascularization in the Macular Photocoagulation Study. The Macular Photocoagulation Study Group. *Ophthalmology* (1989) **96**: 1526–34.
16. Clemons TE, Milton RC, Klein R, Seddon JM, Ferris FL III. Risk factors for the incidence of advanced age-related macular degeneration in the Age-Related Eye Disease Study (AREDS) AREDS report no. 19. *Ophthalmology* (2005) **112**: 533–9.
17. Feuer WJ, Parrish RK, Schiffman JC, Anderson DR, Budenz DL, Wells MC *et al*. The Ocular Hypertension Treatment Study: reproducibility of cup/disk ratio measurements over time at an optic disc reading center. *Am J Ophthalmol* (2002) **133**: 19–28.
18. Jabs DA, Martin BK, Forman MS, Hubbard L, Dunn JP, Kempen JH *et al*. Cytomegalovirus resistance to ganciclovir and clinical outcomes of patients with cytomegalovirus retinitis. *Am J Ophthalmol* (2003) **135**: 26–34.
19. Johnson CA, Keltner JL, Cello KE, Edwards M, Kass MA, Gordon MO *et al*. Baseline visual field characteristics in the ocular hypertension treatment study. *Ophthalmology* (2002) **109**: 432–7.
20. Keltner JL, Johnson CA, Cello KE, Edwards MA, Bandermann SE, Kass MA *et al*. Classification of visual field abnormalities in the ocular hypertension treatment study. *Arch Ophthalmol* (2003) **121**: 643–50.
21. Miglior S, Zeyen T, Pfeiffer N, Cunha-Vaz J, Torri V, Adamsons I. The European Glaucoma Prevention Study design and baseline description of the participants. *Ophthalmology* (2002) **109**: 1612–21.
22. Miglior S, Zeyen T, Pfeiffer N, Cunha-Vaz J, Torri V, Adamsons I. Results of the European Glaucoma Prevention Study. *Ophthalmology* (2005) **112**: 366–75.
23. Spector SA, Weingeist T, Pollard RB, Dieterich DT, Samo T, Benson CA *et al*. A randomized, controlled study of intravenous ganciclovir therapy for cytomegalovirus peripheral retinitis in patients with AIDS. AIDS Clinical Trials Group and Cytomegalovirus Cooperative Study Group. *J Infect Dis* (1993) **168**: 557–63.
24. Tsilou ET, Thompson D, Lindblad AS, Reed GF, Rubin B, Gahl W *et al*. A multicentre randomised double masked clinical trial of a new formulation of topical cysteamine for the treatment of corneal cystine crystals in cystinosis. *Br J Ophthalmol* (2003) **87**: 28–31.
25. The age-related eye disease study (AREDS) system for classifying cataracts from photographs: AREDS report no. 4. *Am J Ophthalmol* (2001) **131**: 167–75.
26. Weinberg DV, Holbrook JT, Hubbard LD, Davis MD, Jabs DA, Holland GN. Clinician versus reading center assessment of cytomegalovirus retinitis lesion size. *Ophthalmology* (2005) **112**: 559–66.
27. Quigley HA. Number of people with glaucoma worldwide. *Br J Ophthalmol* (1996) **80**: 389–93.
28. Weinreb RN, Khaw PT. Primary open-angle glaucoma. *Lancet* (2004) **363**: 1711–20.
29. Rossetti L, Marchetti I, Orzalesi N, Scorpiglione N, Torri V, Liberati A. Randomized clinical trials on medical treatment of glaucoma. Are they appropriate to guide clinical practice? *Arch Ophthalmol* (1993) **111**: 96–103.
30. Kass MA, Heuer DK, Higginbotham EJ, Johnson CA, Keltner JL, Miller JP *et al*. The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol* (2002) **120**: 701–13.
31. Gordon MO, Beiser JA, Brandt JD, Heuer DK, Higginbotham EJ, Johnson CA *et al*. The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol* (2002) **120**: 714–20.
32. The Glaucoma Laser Trial (GLT) and glaucoma laser trial follow-up study: 7. Results. Glaucoma Laser Trial Research Group. *Am J Ophthalmol* (1995) **120**: 718–31.
33. Lichter PR, Musch DC, Gillespie BW, Guire KE, Janz NK, Wren PA *et al*. Interim clinical outcomes in the Collaborative Initial Glaucoma Treatment Study comparing initial treatment randomized to medications or surgery. *Ophthalmology* (2001) **108**: 1943–53.
34. Heijl A, Leske MC, Bengtsson B, Hyman L, Bengtsson B, Hussein M. Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial. *Arch Ophthalmol* (2002) **120**: 1268–79.

35. Leske MC, Heijl A, Hussein M, Bengtsson B, Hyman L, Komaroff E. Factors for glaucoma progression and the effect of treatment: the early manifest glaucoma trial. *Arch Ophthalmol* (2003) **121**: 48–56.
36. Fluorouracil Filtering Surgery Study one-year follow-up. The Fluorouracil Filtering Surgery Study Group. *Am J Ophthalmol* (1989) **108**: 625–35.
37. Collaborative normal-tension glaucoma study group. Comparison of glaucomatous progression between untreated patients with normal-tension glaucoma and patients with therapeutically reduced intraocular pressures. *Am J Ophthalmol* (1998) **126**: 487–97.
38. Collaborative normal-tension glaucoma study group. The effectiveness of intraocular pressure reduction in the treatment of normal-tension glaucoma. *Am J Ophthalmol* (1998) **126**: 498–505.
39. The Advanced Glaucoma Intervention Study (AGIS): 9. Comparison of glaucoma outcomes in black and white patients within treatment groups. *Am J Ophthalmol* (2001) **132**: 311–20.
40. The Advanced Glaucoma Intervention Study (AGIS): 7. The relationship between control of intraocular pressure and visual field deterioration. The AGIS Investigators. *Am J Ophthalmol* (2000) **130**: 429–40.
41. Katz J. Scoring systems for measuring progression of visual field loss in clinical trials of glaucoma treatment. *Ophthalmology* (1999) **106**: 391–5.
42. Viswanathan AC, Fitzke FW, Hitchings RA. Early detection of visual field progression in glaucoma: a comparison of PROGRESSOR and STATPAC 2. *Br J Ophthalmol* (1997) **81**: 1037–42.
43. Tielsch JM, Sommer A, Katz J, Royall RM, Quigley HA, Javitt J. Racial variations in the prevalence of primary open-angle glaucoma. The Baltimore Eye Survey. *JAMA* (1991) **266**: 369–74.
44. Husain R, Clarke JC, Seah SK, Khaw PT. A review of trabeculectomy in East Asian people – the influence of race. *Eye* (2005) **19**: 243–52.
45. Williams R, Airey M, Baxter H, Forrester J, Kennedy-Martin T, Girach A. Epidemiology of diabetic retinopathy and macular oedema: a systematic review. *Eye* (2004) **18**: 963–83.
46. Aiello LM. Perspectives on diabetic retinopathy. *Am J Ophthalmol* (2003) **136**: 122–35.
47. Bresnick GH, Mukamel DB, Dickinson JC, Cole DR. A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy. *Ophthalmology* (2000) **107**: 19–24.
48. Preliminary report on effects of photocoagulation therapy. The Diabetic Retinopathy Study Research Group. *Am J Ophthalmol* (1976) **81**: 383–96.
49. Diabetic retinopathy study. Report Number 6. Design, methods, and baseline results. Report Number 7. A modification of the Airlie House classification of diabetic retinopathy. Prepared by the Diabetic Retinopathy. *Invest Ophthalmol Vis Sci* (1981) **21**: 1–226.
50. The effect of intensive diabetes treatment on the progression of diabetic retinopathy in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial. *Arch Ophthalmol* (1995) **113**: 36–51.
51. Early photocoagulation for diabetic retinopathy. ETDRS report number 9. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology* (1991) **98**: 766–85.
52. Focal photocoagulation treatment of diabetic macular edema. Relationship of treatment effect to fluorescein angiographic and other retinal characteristics at baseline: ETDRS report no. 19. Early Treatment Diabetic Retinopathy Study Research Group. *Arch Ophthalmol* (1995) **113**: 1144–55.
53. Early vitrectomy for severe vitreous hemorrhage in diabetic retinopathy. Four-year results of a randomized trial: Diabetic Retinopathy Vitrectomy Study Report 5. *Arch Ophthalmol* (1990) **108**: 958–64.
54. Lang A, Portney V. Interpreting multifocal intraocular lens modulation transfer functions. *J Cataract Refract Surg* (1993) **19**: 505–12.
55. Holladay JT, Van Dijk H, Lang A, Portney V, Willis TR, Sun R *et al.* Optical performance of multifocal intraocular lenses. *J Cataract Refract Surg* (1990) **16**: 413–22.
56. Knorz MC, Claessens D, Schaefer RC, Seiberth V, Liesenhoff H. Evaluation of contrast acuity and defocus curve in bifocal and monofocal intraocular lenses. *J Cataract Refract Surg* (1993) **19**: 513–23.
57. Gimbel HV, Sanders DR, Raanan MG. Visual and refractive results of multifocal intraocular lenses. *Ophthalmology* (1991) **98**: 881–7.
58. Leyland M, Zinicola E. Multifocal versus monofocal intraocular lenses after cataract extraction. *Cochrane Database Syst Rev* (2001) CD003169.
59. Wirtitsch MG, Ergun E, Hermann B, Unterhuber A, Stur M, Scholda C *et al.* Ultrahigh resolution optical coherence tomography in macular dystrophy. *Am J Ophthalmol* (2005) **140**: 976–83.

PSYCHIATRY

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

20

Anxiety Disorders

M. KATHERINE SHEAR[1] AND PHILIP W. LAVORI[2]
[1]Western Psychiatric Institute and Clinic, Pittsburgh, PA 15213, USA
[2]VA Medical Center, Menlo Park, CA 94025 2539, USA

INTRODUCTION

Anxiety disorders are the most prevalent psychiatric conditions in the community with a lifetime community prevalence of 20–30%.[1] These disorders can be seriously impairing, reducing quality of life and causing disability. Recent studies suggest some forms of anxiety are associated with early mortality. Many who suffer from anxiety disorders have other serious medical problems, such as depression, pulmonary disease, cardiovascular illness and neurological conditions. Prevalent and debilitating, anxiety disorders are serious, persistent illnesses that warrant treatment. Clinical trials are needed to establish efficacy of promising interventions and to determine the best ways to deliver efficacious treatments in different contexts.

Methods for conducting efficacy trials in anxiety disorders have evolved over the past few decades. Reliable diagnostic instruments and symptom severity scales have been developed and tested. Strategies for medication administration have been identified and manuals written to standardise these procedures. Cognitive behavioural treatment methods have been specified and explained in manualised format. Treatment training and adherence measures are available. These methodological advances mean that studies of the efficacy of new interventions can be conducted efficiently and with confidence.

Given the availability of efficacious treatments, researchers are now turning their attention to studies that test these interventions in the community settings where they will be used, and in clinical contexts (such as maintenance of response) that go beyond the phase of acute illness that is the focus of most efficacy studies. With this shift in focus, new methodological problems appear. Generic problems that need to be addressed in designing such studies, often known as effectiveness studies, have been described in the literature.[2] In this chapter we discuss methodological issues pertaining to effectiveness studies of anxiety disorders. We identify some key features of these disorders and consider the problems they create for study questions and study design. Solutions to methodological problems in clinical trials often require trade-offs, and the problems we discuss are posed in this way. We provide our view of the best way to manage these

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

problems, and in some cases make suggestions for methodological innovations.

Clinical researchers regularly make methodological choices regarding subject recruitment, selection and characterisation of subjects, procedures for enrolment, assignment to experimental group, experimental manipulation, outcome assessment and follow-up process. Methods chosen will place specific limits upon what can be learned from a study. Thus, it is fundamental that study methodology be driven by the question the researcher seeks to answer. However, unlike efficacy studies, in effectiveness studies, the question is not always clear. Defining the study question is the first problem for the effectiveness researcher. Most experienced clinical researchers are expert in conducting efficacy studies to answer the question 'Does a new treatment produce better results than a control condition for a well-defined condition, under tightly controlled circumstances of use?' Both psychosocial and pharmacological treatment researchers have successfully undertaken such studies, and thus are poised to test efficacy hypotheses for new interventions.

The field of effectiveness research is far less developed. Investigators move forward in unmarked terrain as they decide upon the most important next questions. For example, a study of Long Term Strategies in the Treatment of Panic Disorder (MH045963-6) currently in progress under the direction of the authors is designed to answer the question 'Should non-responders to an initial trial of CBT receive medication or an additional dose of CBT?' This important question is not addressed by efficacy studies of either medication or psychotherapy. Having articulated such a question, decisions must be made about what methodological approach should be used, and what problems to anticipate. For example, Principal Investigators of the Long Term Strategies Panic Study had to confront the issue of what the right duration of the initial CBT trial might be, and what level of non-response to initial trial should be chosen to define intake into the randomised maintenance trial. Decisions such as these are not trivial, since neither the most important questions, nor the best way to approach a given question, is obvious. Given this ambiguity, we suggest anxiety disorder researchers might be guided by some of the key features of these disorders (Table 20.1). We discuss the methodologic relevance of five such features: (1) anxiety disorders are characterised by high community prevalence; (2) diagnostic boundaries are ambiguous, both between pathological and normal anxiety and among the different anxiety disorders; (3) phobic fear and avoidance is prevalent in these disorders; (4) anxiety disorders are treatable using either medication or cognitive–behavioural interventions; and (5) anxiety disorders frequently co-occur with other disorders. Each of these features will affect decisions about the research question and the choice of methods.

FEATURES OF ANXIETY DISORDERS THAT IMPACT STUDY METHODOLOGY

HIGH COMMUNITY PREVALENCE

Anxiety disorders are highly prevalent in the community. The high prevalence means there are many patients in need of treatment. Epidemiological studies document that most of these patients do not present for care in a specialty mental setting.[3] Instead, they can be found in a range of community service settings. Even among those who do seek specialty mental health treatment, only a subset will be enticed to an academic medical clinic, regardless of the incentives provided. For those who seek treatment at a community mental health setting, usual practice diagnostic procedures cannot be relied upon to identify anxiety disorders.[4] It is clear that we need to know how to recognise and treat the people with anxiety disorders who most researchers never see. Put another way, we need to study those who do not participate in studies. This obvious paradox underscores the principle that effectiveness studies will not be straightforward.

The job is not simply a matter of running an efficacy study in one or more community settings. Doing so would be important only if there are serious questions about whether patients in such settings respond to proven treatments.

Table 20.1. Implications of features of anxiety disorders for research design

Feature	Research issues
1. High community prevalence	Settings: which and how many Recruitment: reaching the unstudied patient Assessment: measuring outside the research clinic Awareness: bridging the patient's knowledge gap Human subjects protection: make or buy Technology: the right machine for the setting Comorbidity: adjusting to increased variability
2. Poorly defined nosologic boundaries	Normal vs disordered: a question of excess Differential diagnosis: core symptoms overlap Pluripotency: treatments with broad efficacy Double counting: symptoms Endpoints: ranking outcomes Aiming low: focus on preventing relapse Stability: the time frame for outcome
3. Phobic fear and avoidance	Evasion: measuring the avoidant subject Fear: recruiting the anxious subject Identification: personal choice vs avoidance
4. Discordant models of the disorders	Acknowledgement: both models have treatment successes Control: paying attention to the other intervention Comparing: accommodating preference for modality Targets: agreeing on the goals Dissemination: thinking ahead about the audience

If this is the case, it is important to frame the specific questions the study should answer, based on the reasons for predicting response differences. For example, if researchers suspect severity is an important treatment moderator, it might be important to conduct a standard randomised efficacy trial in settings with patients of varying severity. Likewise, some patients have co-occurring symptoms or syndromes, such as serious medical illness, along with an anxiety disorder such as panic disorder. It might make sense to recruit patients from medical clinics into an efficacy trial in order to study the influence of the medical illness on the treatment of the target condition. Other examples of parameters that might be predicted to render uncertain outcome with a proven efficacious treatment include organisational features of the setting or socio-economic status of the patient. Specific considerations like these ought to drive the important design decisions such as where the research will be conducted and in how many different kinds of settings.

A different kind of research question might be driven by the subject paradox (how to study patients who do not participate in studies): for example, ‘What is the most successful way of recruiting and engaging individuals who do not seek treatment in a research clinic?’ The investigator might compare a public education programme to a professional educational intervention. Or, the research aim might focus on evaluating alternative screening strategies in different settings. Another example might be ‘How much diversity of setting should be incorporated into a study?’ In addressing this question the researcher might investigate the variation of clinical presentation, treatment acceptance, or outcome across different ethnic or socio-economic groups. Alternatively, the investigators might examine the effect of different organisational structures or the impact of the organisational climate[5] on outcomes or study the implementation of organisational interventions to optimise the likelihood of dissemination of a treatment.[6]

Whatever studies are done, it is clear that for anxiety disorders, researchers need to extend their reach if they seek to make an impact on the great majority of individuals who suffer from these conditions. Methods need to be devised to study patients in primary care and specialty medical settings, dental offices, churches, schools, community centres, and a range of other community service or support settings (e.g. domestic violence or homeless shelters, or even highly utilised commercial operations such as supermarkets[7] or department stores). The use of such settings to deliver care may be particularly relevant for patients with anxiety disorders who have phobic restrictions, and are unable to travel outside a restricted area.

Designing a new clinical trial for an anxiety disorder outside of an established research centre raises other problems. Investigators make deliberate decisions about where to recruit, assess and treat patients, as well as whether to carry out any of these activities in more than one kind of setting. Existing clinical research methods for recognising and recruiting affected individuals may be too cumbersome to work in a setting where research activities are not customary. For example, a busy primary care practice or even a mental health facility may not be oriented towards identifying and tracking individuals who meet criteria for anxiety disorders. Frequently staff in such places are very busy, very dedicated and sometimes opinionated about what is best for their patients. The researcher who comes to study usual practice may be seen as challenging the skills, competence or even integrity of the staff. Still, recent studies in primary care[8] have succeeded in overcoming these barriers and have done much to provide information to inform processes to optimise care.

Protocol-driven treatments face additional barriers to acceptance in settings other than the research clinic. Assessment of outcomes is hard enough in a research clinic; assuring good follow-up and reliability of measurement in non-traditional research settings will tax the ingenuity of the next generation of effectiveness researchers. Recent work using adaptive testing methods holds promise as a technique. Given these challenges, it is tempting to suggest that researchers concentrate on one research setting, and hope or assume that results will generalise. But the decision to limit the setting has uncertain implications for interpreting and generalising results. There is a trade-off between generalisability and the cost of dealing with heterogeneity of setting. These costs must be borne, and the methodological challenges met, in order to produce research-grade answers to the question of effectiveness.

In working in almost any non-mental-health community setting, the investigator must address stigma and self-criticism that can be associated with the idea of having a mental disorder. Researchers need to take steps to minimise the difficulties that may be caused by identifying a person as ill, especially when the person in question has not already identified their symptoms as problematic. In such a situation, the news may come as an unwelcome surprise, or may be perceived as insulting or embarrassing. The newly diagnosed individual may feel suddenly stigmatised and this may lead to a rejection of the diagnosis and/or the researcher bearing the news. There may be anger or discouragement towards the community setting in which the person sought help. The researcher needs to be sensitive to these possibilities and proactive in dealing with untoward reactions associated with identification of an anxiety disorder. For example, if there is a decision to recruit subjects in a non-psychiatric setting such as a church or supermarket, the researcher would need to include an introductory phase of the work that addresses fears and stigma associated with a diagnosis. This can be done in a variety of ways. A community educational phase might be undertaken prior to initiation of recruitment. Individual or group consciousness raising might be offered. Focus groups are a very useful strategy being increasingly used by researchers. In this situation, small groups of individuals with different types of anxiety might be invited to participate in a focus group. Participants are paid and the group leader might guide the group in discussion of topics such as the meaning of having an anxiety disorder, the perceived response of others, including family, friends and/or the community at large. A focus group might also be asked to discuss how researchers might best approach undiagnosed people in the community who suffer from these disorders, or the group might be asked how to best present treatment options, or how to explain and encourage participation in research. Thus armed, the researcher will be more successful in recruiting and retaining subjects for a clinical trial. An example of a very innovative approach to the problem of community recruitment[9] utilised an intensive telephone engagement strategy in which mothers of inner city minority children were invited to identify and problem-solve an important difficulty they were experiencing. Only after the intake recruiter had successfully helped with this practical problem did they explain the availability of services for other kinds of problems. This approach was shown to significantly increase attendance at the first clinic appointment. Whatever the approach, it is clear that the prospective patient research volunteer must be given opportunities to understand their anxiety symptomatology as a treatable condition underlying what may be just an awareness of limitation or fear. These individuals further need to decide for themselves which treatment programme they wish to access. The researcher needs to present a clinical trial in this context.

Sometimes stigma can be best addressed at the level of the service provider – such as a primary care physician, or administrators and service providers in different kinds of agencies. In order to access patients in a given facility it may be very important to first understand the headaches of the facility administrators. A researcher who takes the time to both identify and respond to the problems faced by those attempting to deliver care will be rewarded with a much higher level of enthusiasm and support for the research project. Researchers in the field of services research have understood and successfully accomplished this kind of work.[10] Partnering with administrators in different service agencies to find ways to improve their efforts is likely to provide easier access to subjects and better support for implementation of study procedures. Careful attention to such issues can determine the feasibility of the study.

Assessment Strategies in Community Settings

The standard research diagnostic interview and follow-up batteries were designed to achieve careful, reliable descriptions of different well-specified phenotypes of psychiatric illness. While highly successful in meeting this goal, such instruments have not been designed to maximise efficiency and minimise patient and staff burden. These extensive and time-consuming inventories will not survive transplantation into a primary care setting, a hospital emergency room or a dental office, let alone a church or school. Instead, radically simplified tools must be developed that utilise innovative statistical and psychometric methods (e.g. adaptive testing) and/or study sample sizes must be increased to compensate for extra variance.

The assessment strategy used in a community setting may need to be altered in other ways as well. No matter how prevalent an anxiety disorder may be in the community, it will be lower in the community setting, compared to the prevalence in the enriched intake stream of a specialty anxiety clinic. The odds on a disease may easily vary fivefold or more from clinic to community. Given that the specificity and sensitivity of the diagnostic instruments will be no better in the community setting, and may well be worse, the 'Bayes factor' of the test (sensitivity divided by $1 - \text{specificity}$) will be smaller in the community setting. For example, if the sensitivity and specificity both decline from 90% to 80% then the Bayes factor declines from $0.9/(1 - 0.9) = 9$ to $0.8/(1 - 0.8) = 4$. If both the prior odds of disease and the Bayes factor are lower, the positive predictive value will also be lower (the odds of disease given a positive test are just the prior odds of disease times the Bayes factor). In the numerical example above, given only a fivefold difference in prior odds, the posterior odds on disease given a positive test may vary by an order of magnitude from clinic to community, making interpretation of intake diagnosis problematic. Multi-step diagnostic procedures may be necessary to avoid over-diagnosis.[11]

Research Recruitment

Study subjects volunteer to participate in research. In a clinical trial, the manner of presentation of treatment options may make or break a study. The highly selected population of patients who present to an academic centre often come preconditioned to the value of research protocols, and may have specifically sought out the clinic because of its reputation as a research centre. The potential research volunteer in a community setting has not voted for research 'with his feet', and may need a gradual, informative and upbeat approach, to accept the idea of protocol-driven treatment and randomisation. Institutional review boards may well regard placebo control as especially unattractive in such a context, and may also be concerned about 'overselling' the potential benefits of research to patients. Yet, there is reason to believe that the patient in the community context may be the one with the most to gain from participation in research, because of the likelihood that her illness would otherwise go unrecognised, or the equally disadvantageous likelihood of inadequate treatment.

Context-Relevant Treatment Protocols

To optimise study results, strategies must be developed for providing protocol treatments in a context-relevant manner. This may include adjusting to the absence of third-party payers, or making use of setting-specific para-professional personnel for some of the interventions. Or, it may mean incorporating 'escalation' strategies into the treatment protocol, so that subjects identified with substantial needs are transferred to a more traditional setting.

Practical and Administrative Issues

Human subjects review may need to be coordinated among several kinds of organisations. Some may be willing to enter into agreements to accept the investigator's home institutional review, others may need to develop their own review processes and obtain Federal-Wide Accreditation. Template agreements that have been shown to work would be a valuable resource.

The investigator must choose methods of data capture and processes for the data edit cycle that work in diverse settings at sites that are distant from the coordinating institution. Data monitoring, correction of errors and tracking of follow-up are all affected. Technological limitations need to be respected. For example, fax-based methods may be more easily deployed than internet-based methods, especially in settings where a fax machine is already in use. On the other hand, as personal digital assistants become ubiquitous, patient follow-up may be individualised, remote and remotely cued. We can imagine technology that rings a telephone number or sends an instant message, asking for self-report follow-up, and that can schedule and connect the subject with a live interviewer, all implemented on the same small wireless device that might be cheap enough to give away as a free incentive to participation.

Despite the difficulties associated with exporting clinical trials to the community, it is clear that the next generation of clinical research in anxiety disorders needs to be rolled out into the settings where individuals with these disorders live and work. In addition to the many issues related to the setting of a study in the community, there are many design considerations related to which patients should be included in a given study. Patients in different community settings are likely to be heterogeneous in different ways, and to differ from patients who seek treatment at traditional research clinics. Existing studies document a high rate of comorbidity among anxiety disorders, between anxiety disorders and depression, and between anxiety disorders and medical illnesses. There is also comorbidity of anxiety disorders with for example psychotic disorders[12] and substance abuse.[13,14] Such comorbidity may increase the likelihood that a patient seeks treatment at research clinics, and therefore it is possible that studies in the community will have to deal with less comorbidity than studies in the research clinic. Nevertheless, an effectiveness researcher must decide how to manage comorbidity. There are many consequences of decisions to include or exclude comorbidities from study eligibility criteria. There are a variety of assessment considerations that are different in comorbid versus non-comorbid subjects. Symptoms of co-occurring depression or substance abuse may be difficult to disentangle from anxiety symptoms. Many medical disorders produce symptoms of autonomic nervous system activation, as do anxiety disorders. Such medical comorbidity may be especially likely in primary care or medical clinic settings. The trade-off between heterogeneity and its attendant increase in measurement variance, and homogeneity and its attendant restrictions on generalisability, must be carefully considered. In addition, rigid exclusion criteria may be less acceptable in community settings than in the specialty research clinic; patients who are surprised by a diagnosis may be disappointed if they are ruled out from studies by being 'too complicated'. An alternative for the researcher is to simply accept comorbidity and heterogeneity of the population and evaluate a treatment that targets a specific symptom, behavioural pattern or symptom cluster, without regard to the context in which it occurs. To make this decision the researcher accepts the 'noise' this will cause in the system and powers the trial accordingly.

Other considerations related to patient heterogeneity include the fact that illness severity and typical background treatment history may vary across settings. Patients in some settings may have already received multiple treatments, while in other settings they may be treatment naïve. Given the findings from multiple studies that have documented that affective and anxiety disorders are under-recognised and under-treated in the community, it is likely that patients recruited from non-mental-health settings will have had little exposure to proven efficacious treatment. Often such patients have sought help from clergy or other informal sources. In the case of anxiety disorders, the awareness of the 'irrationality' of symptoms often means these individuals suffer in silence, embarrassed to reveal their self-perceived defects. Such patients are often enormously relieved when they learn that their disorder is understood. Even when treatment has apparently been offered, it may be less vigorous than the versions that have been proven efficacious in clinical trials. Inadequate doses and durations of pharmacotherapy may be the rule, and specific psychotherapies may be offered in name only. It may be particularly important not to assume (for example) that a patient has demonstrated a lack of response to treatment, and therefore be ruled out as ineligible.

If patients are identified in settings other than the specialty clinic, they may not view their anxiety disorder (which may be news to them) as the main problem they should be concerned with (along with their hypertension, macular degeneration, current spousal abuse or arthritis). They may be unwilling to make accommodations in schedules and may have needs for unusual availability of research staff in time and space. Some patients may not understand the usual standard procedures for treatment in a mental health clinic. They may need to be approached in an accommodating way.

POORLY DEFINED NOSOLOGIC BOUNDARIES

A second feature of anxiety disorders is that the boundaries between normal and pathological anxiety and among the pathological disorders are ill defined. Unlike most psychiatric disorders, the symptoms that comprise the diagnostic criteria for anxiety disorders are recognisable in normal people every day. The pathological state is defined by excess. However, the definition of excess is not precise. Because anxiety is a normal emotion, it is not always clear where the boundary between normal and pathological lies. This is particularly true in the context of stressful life events and ongoing difficulties. The boundary with normal may arise in defining a clinical population in need of treatment. Boundary issues are also relevant to treatment targets and definition of remission. In general, there is no consensus on what comprises remission of an anxiety disorder. We discuss this problem and suggest some ways it might be addressed. The problem of the boundary between normal and pathological is not a question raised only in the area of anxiety disorders, but rather is a continued question in the ongoing discussion related to definitions of psychopathology. A relatively recent paper[15] provides a good summary of current issues. As these authors point out, it is also relevant to consider the relationship between mental and physical disorders. These considerations are important for clinical researchers to keep in mind but a detailed discussion of the various issues is beyond the scope of this chapter.

However, as noted above, anxiety is a normal emotion, and so its pathological state must be distinct from normal variation. It is best to experience anxiety in moderation. While anxiety can be disabling in excess, a deficiency of anxiety can also be impairing. The question of how much anxiety is optimal is not a philosophical one. Rather, it is one of the conundrums that currently face the clinical trials investigator. Namely, the investigator must decide how much symptom relief is optimal and how much is sufficient to declare a meaningful response to treatment. Given that anxiety is normal, is there some expected floor for the intensity of anxiety symptoms, or is symptomatic anxiety qualitatively different?

Another design question relates to the level of anxiety that results in optimal long-term outcome. Still another relates to the definition of remission of a given disorder. The field has not reached consensus on how to define remission for any of the anxiety disorders. This is a critical methodological problem that needs to be addressed. Investigators need to consider whether there is a way to overshoot the mark or is less always more? This is a serious question, as researchers are not agreed upon whether it is useful to have some anxiety symptoms in order to keep coping functions operative and/or provide 'toughening up' experiences. Perhaps some continued symptomatology is a good idea to encourage continued exposure. The continued presence of low-level symptoms may increase the chances that the patient does not become complacent[16] and/or provide opportunities to confirm the absence of more severe symptoms.

On the other hand, since anxiety disorders are clearly debilitating, perhaps it is best to eliminate symptoms as fully as possible. Perhaps if we leave residual symptoms, this indicates that we have not eliminated the underlying vulnerability and relapse will be more likely. Ideally, we would like to eliminate pathological anxiety while leaving 'normal' anxiety intact. Yet this distinction may be difficult to define. If we have a pharmaceutical compound that reduces anxiety, might we overshoot the mark? If so, would that be as problematic as undertreatment? Common sense, and the results of a famous study,[17] suggest that a moderate level of anxiety is associated with optimal performance in situations like test-taking. Laurence Olivier is known to have suffered, as many actors do, from tremendous stage fright. His view of this was that this fear was an essential motivator that ensured his performance would be undertaken with the highest possible focus and concentration. Every researcher knows that the approach of the deadline for grant submission generates substantial anxiety which again motivates the highest possible level of energy and productivity.

Threshold issues relate to the decision to begin as well as the decision to end treatment. At what point do we declare anxiety to be at a clinically significant level that warrants intervention? If Laurence Olivier experienced intense anxiety at each performance, should he be treated? The goal of treatment of an unhappy but successful person should be first and foremost to prevent failure (inability to perform, because of paralysing fear or shoddy performance, because of cavalier attitude) while, if possible, reducing the discomfort of unhappiness. In this context we echo a famous quote of Freud, concerning the goals of psychoanalysis vis-à-vis unhappiness. Anxiety clearly exists on a continuum yet a treatment decision is a binary one. We do not attempt to administer a partial treatment. The decision of who to treat, of the minimal level of symptomatology an eligible subject may have, will have implications for interpreting and acting on study results. It is likely that there is a distance from the boundary with normality associated with optimal effect of treatment. The closer to the boundary, the more likely the study will show non-specific or placebo effects. The farther from the boundary (i.e. the more severe and complicated the symptoms are) the less likely that the treatment will be fully effective. One consideration in deciding who to treat in a research study of anxiety disorders is the life context and the personal context in which the anxiety disorder symptoms arise.

Considerations Related to Life Context and Individual Psychology

Because of the salience of environmental stimuli as a trigger for normal anxiety, and the importance of coping mechanisms and social supports as responses to anxiety, it might be argued that these context measures are of particular importance in anxiety studies. Little is known about the relationship between onset, course and treatment of anxiety disorders and these external factors. There is a need to examine what the nature of these relationships may be. For example, it is not known whether faulty coping mechanisms play a role in the vulnerability to one or another of the anxiety disorders. If so, perhaps this should be a target of a treatment intervention. If not, perhaps coping skills are variable across individuals and/or across stressors and may act as a moderator of treatment response. In this case, improving coping may be a strategy for treatment of non-responders.

Strong social support is well known to be an important contributor to a sense of safety. Anxiety disorder patients experience the world as more dangerous. Safety is not necessarily the opposite of danger, but a sense of safety can mitigate the perception of likelihood of danger and/or the perception of consequences of the danger. Anecdotally, some anxiety disorder patients are thought to have unusually good interpersonal skills. Turning to others may be one way a patient with panic disorder copes with a world perceived as persistently and unpredictably frightening. For other individuals with anxiety disorders, anxiety may be exacerbated by relationships with others. A patient with social phobia fears scrutiny by others and this may motivate them to avoid relationships or to concentrate on developing a small group of 'safe' people. Someone with OCD may fear contamination from others, or an OCD patient may fear harming other people. Either may lead them to have reduced social contacts and less overall sense of safety. It is also possible that deficits in internal representations of other people can lead to problems in regulation of emotions. This can contribute to anxiety symptoms and perhaps even to the onset of anxiety symptomatology. There is some indication that relationships with others help regulate neuroendocrine and autonomic nervous system functioning. Whether to include measures of social support and attachment into clinical trials in anxiety disorders is a decision researchers must begin to consider. Such information is an additional patient burden. However, it may be difficult to determine optimal interventions for patients in the community if researchers do not begin to address some of these issues.

Also ill defined are boundaries between different diagnostic categories with similar symptomatology, and frequent comorbidity. Given the fact that fears, worries, somatic symptoms and behavioural manifestations are shared across disorders, distinctions can sometimes be blurred. There have been changes in diagnostic criteria sets since DSM III, especially for panic disorder and generalised anxiety disorder (see below). There has also been a change in the relationship between panic and agoraphobia, and between these disorders and other DSM IV anxiety disorders. These changes reflect growing recognition of the occurrence of panic and phobic symptoms across disorders. In addition, the core generalised anxiety disorder (GAD) symptom, worry, is also frequently found in other anxiety disorders and in depression. The content of worry apparently differs in depression and GAD.[18] However, this is a nuance for an assessment strategy, and again time is required to tease this apart.

The diagnostic groups that comprise the anxiety disorders share cognitive manifestations of fear or worry, behavioural manifestations of efforts to cope with the anxious thoughts (phobia or compulsions), and somatic symptoms that often accompany the anxiety. Yet each disorder has a different 'flavour' of symptoms, and each may reflect different aetiological underpinnings. The similarities have implications for clinicians and researchers alike and will have a bearing on the design of new treatments and dissemination of efficacious interventions. As study intake moves into the community, we can expect the diagnostic overlap to increase, if only because mild versions of disorders are harder to separate than severe ones. Efficacy studies have focused on specific diagnostic groups, rather than on anxiety as a loose complex of symptoms. This means that the ostensible usefulness of study results depends on clear, reliable identification of the specific disorders in patients, at best a dubious proposition in the community setting. However, it should be recognised that while the efficacy studies have been carried out in carefully constructed 'pure cultures', the results of those studies show a startling uniformity of options for treatment across the spectrum of anxiety disorders. It may be that the careful and expensive nosologic dissection characteristic of the first generation of clinical trials is added to the precision and power of those studies, but may be relaxed in the next generation of effectiveness studies, making a virtue of necessity.

We now know how to reliably identify anxiety disorders and we know how to provide efficacious acute interventions, but these demonstrations have occurred only in the research clinic. The traditional decision point for clinical intervention, i.e. clinical (DSM IV) diagnosis, is fairly clear, though there are remaining controversies about psychiatric diagnosis. For the most part, these are beyond the scope of this chapter. Researchers have developed tools to use for screening, diagnosis and severity ratings of anxiety symptoms. Ensuring that clinicians are aware of these and that the instruments are user-friendly is a focus of current work. There are many existing publications related to different assessment instruments, so we will not provide this information here. Instead, we suggest that even with a good instrument, there are some difficulties in establishing clearly a single target condition, and that the attempt may be a useless diversion of effort if discriminatory precision is less important than inclusiveness of intake.

The high rate of co-occurrence of anxiety disorders is an area of confusion that concerns the diagnostic nomenclature. For this and other more theoretical reasons, there is controversy in the field with regard to whether different DSM IV diagnoses describe truly distinct illness categories. Moreover, even if they are different, their co-occurrence creates measurement problems. For example, if a patient in treatment for a panic disorder has a co-occurring specific phobia of heights, should avoidance of bridges be rated a symptom of panic disorder, of height phobia, or both?

As noted above, it may be possible to aim current established treatments on the generic symptoms that occur across the disorders: fear or worry, somatic symptoms and behavioural changes such as avoidance or compulsive ritualising. The broad spectrum effects of serotonin-active medications lend themselves to such an approach, as do the psychosocial treatments which may reduce generic cognitive, somatic and behavioural symptoms across disorders. An investigator planning a community study needs to decide whether to test treatments in the classical disorder-specific trial or in a more broadly-based group of patients defined by the common symptoms and behaviours of the anxiety disorders. We think that the latter choice deserves serious consideration.

Defining Outcomes and Measuring Results

Katschnig and Amering[19] point to the considerable complexity of symptoms in panic disorder, suggesting that spontaneous and situational panic attacks, anticipatory anxiety, phobic avoidance, disabilities, comorbid depression and substance abuse must be considered. One might add to this the presence of other anxiety disorders, personality disorders and physical illnesses. These authors list methodological difficulties that emerge from this complexity. They raise questions such as which of the phenomena are most important in assessing course and outcome, what are the relevant time intervals for symptom assessment, at what point should the clinician consider that the illness has transitioned to a remitted state? Such considerations are relevant for each anxiety disorder. All are composed of multiple domains of symptoms, including panic, anticipatory anxiety, worry, phobia, obsessions or compulsions. Since the diagnostic criteria do not require the presence of each of these, it is possible to meet criteria for one or another anxiety disorder with prominent symptoms in one domain and none in another. Over time this may change. In some situations different domains within a disorder may be negatively correlated. For example, an individual may experience a reduction in panic attacks while becoming increasingly phobic. Is this improvement or worsening of the overall condition? A patient with OCD may experience a decrease in obsessions as the compulsive behaviours grow and become instantiated. Should this be considered a change in severity? A person with a phobia may experience lower overall impairment and/or fear as their phobic behaviour become more fixed, and they begin to accommodate the phobia in their lives. Does this mean the phobia is in partial remission? What if the intensity of symptoms is actually worse than when the disorder was first diagnosed, and yet there is less impairment? Similarly, if an individual with OCD has prominent obsessions and intermittent compulsions are they better off, worse off, or the same as if the opposite is true? What is the role of impairment and/or quality of life in determining outcome? What criteria should we use for illness severity? What about treatment response or remission? It is clear that response entails a clinically significant, noticeable change in symptom levels while remission entails a return to functioning with symptoms at a level that they cause no noticeable distress. Studies are underway to identify precise markers of these important clinical transitions.

The fact that the symptoms of a single disorder do not necessarily travel together creates difficulties in defining the endpoint for a treatment. Such a ‘carousel course’ (Figure 20.1) of symptoms leads to assessment quandaries that can be daunting. Again, taking the case of panic disorder with agoraphobia, if a patient starts treatment with several full panic attacks per week, and then has a marked reduction in panic attacks, but continues to have frequent limited symptom episodes, and remains moderately agoraphobic, how much improved is the patient compared to baseline? How should life context be factored into assessment of illness severity? If a social phobic gets a new job which requires less public performance than previously, but the job is below his or her level of competence, social anxiety symptoms may diminish noticeably but is the patient really better?

Several authors have drawn attention to these problems and the general recommendations have been to use composite measures of severity over an extended period of time. Such composite measures are available for most of the disorders, and most are quite user-friendly: The Yale–Brown Obsessive Compulsive Scale has been widely used and is available in a self-report format. The same is true for the more recently developed Panic Disorder Severity Scale. The Social Phobia Inventory (SPIN) is also brief and comprehensive. There is little agreement in the field about the one or two best measures for each disorder. The same measures can sometimes be used for screening diagnosis and outcome though it makes sense to pick the instrument most relevant to the goal of the assessment. The use of a composite measure presupposes that it is possible in principle to rank order the outcomes of the patients, although there may be many outcomes that are distinct but not ordered. From a statistical point of view, the ability to reliably order patient outcomes into as few as four or five categories provides considerable increase in the power to

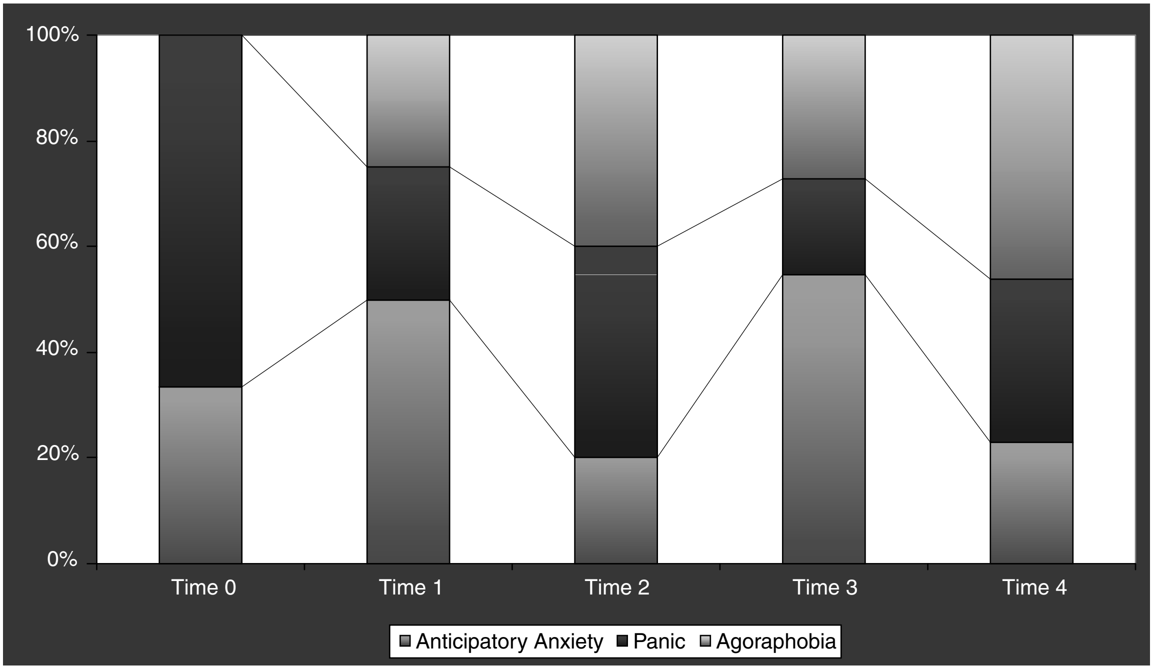


Figure 20.1. Panic disorder as an example of a 'carousel' symptom pattern.

detect treatment differences, compared to a simple dichotomy of response/no response. There are diminishing returns even to perfectly reliable orderings with more than five levels. Given even modest unreliability, it may not pay to push composite measures beyond a few levels of discrimination. The challenge posed by the 'carousel course', and the pleiotropic outcomes, of anxiety disorder is fundamental: the ability to rank patient outcomes is the most basic feature of scientific measurement and study.

As studies extend into the community, they will explore the less severe forms of the disorder, and may be even more vulnerable to the problem discussed above. This raises the possibility that the target of measurement should not be improvement (alone) but prevention of significant worsening. The advantage of this approach is that it may move the measurement into a more reliable regime, in which there is less controversy about the meaning of the outcome. The disadvantage is that it may also require large sample sizes, in order to detect modest effects on low probability outcomes.

Choosing a Time Frame for Outcome Assessment

The specifics of time frame are also controversial. While a group of senior panic disorder researchers achieved consensus on a recommendation of an optimal period of four weeks for assessment of symptoms, and a minimum of two weeks, these recommendations are not always followed. In fact, frequently symptom status is reported without specification of the time frame of the assessment. The issues around time course are further complicated by variability between domains and within a domain, depending on life circumstances. Some domains of symptoms, such as phobic symptoms, are very stable, and a change in them, even over a fairly brief period, e.g. a few weeks, can be an indicator of change in illness severity or course. (A caveat here is the change in life context.) However, other symptoms are very unstable. For example, it is typical for panic attacks to occur in clusters and then to subside. The problem is further compounded by difficulties inherent in rating panic frequency. Anticipatory anxiety can be far worse if there is a specific environmental demand to confront an anxiety-provoking situation. For example, if a social phobic must go to a wedding. This raises the question of the time frame over which different types of symptoms should be assessed, and the situations in which the symptoms should be evaluated. Again, a focus on long-term prevention of serious worsening may help.

We do not have definitive answers to these myriad questions, but suggest that we should be paying attention to these assessment challenges. It may be possible to undertake secondary data analyses that target these questions. In the meantime, we suggest that outcome assessment must take into consideration multiple domains to make a meaningful judgement of response or remission. Further, it makes sense to establish and publish conventions for raters so that others can understand clearly results of studies. Reports of study results rarely describe conventions for rating phobia, including changes in life context and/or situational demands. Many published panic disorder studies use panic attack frequency as the only outcome. Reporting conventions should be broadened to address these issues.

Phobic Fear and Avoidance

A third issue specific to anxiety disorders is the occurrence of phobic symptomatology. One of the trickiest problems in anxiety disorders treatment is the assessment and management of avoidance. Avoidance is a natural reaction to fear and is usually successful in reducing anxiety in the short run. However, the longer-term effect is virtually always to increase anxiety. Avoidance also causes substantial functional impairment. Avoidance may lead a social phobic to seriously curtail his or her education, or resist career development because of fear of speaking in a group. The net result can be highly significant to income and productivity. Thus, avoidance is both a coping mechanism and a symptom. By its nature, it can be difficult to measure, since many anxiety disorder patients try to avoid thinking about anxiety-provoking situations, in addition to avoiding confronting these situations. This means that asking a person if there is anything they are avoiding often results in under-reporting. It is necessary to enquire about avoidance by asking specific questions, and this can be time-consuming. Some behaviour therapists argue that phobias can only be assessed using a behavioural challenge protocol. However they are measured, it is clear that phobic symptoms are important as they are among the strongest and most consistent predictors of long-term outcome.

Avoidance can also play a role in silencing anxiety symptoms and reducing the impetus to seek help. This may be one way that phobic symptoms act to worsen the course of illness. Silencing of symptoms is also reminiscent of the hypertension analogy where serious consequences result from lack of awareness of symptoms and difficulty adhering to treatment regimens. In fact, phobic avoidance has now been found, like hypertension, to be a predictor of cardiovascular mortality, at least for men. A further issue related to the silencing of distress is that it can be difficult to distinguish pathological from normal avoidance behaviours. Phobic symptomatology may become so integrated into the patient's life that it seems normal. Avoidance of some situations may be treated as though they are simply life choices. The patient may say that they simply do not enjoy shopping in a mall when the fact is that they are afraid to go to a mall because they may have a panic attack. The problem of distinguishing normal from pathological anxiety is broader than the issues related to phobia.

This realm of symptoms causes methodological problems that involve both assessment and treatment. Phobic symptoms entail avoidance of cues that evoke fear, anxiety or other dysphoric affects. An individual with phobic avoidance will make every effort to evade exposure to the feared situation. Evasion often extends to thinking or talking about the situation. This means that the phobic individual cannot be counted on to talk about their symptoms spontaneously. In fact, to obtain a clear picture of the extent of behavioural avoidance often requires a detailed enquiry. Such an assessment takes time and is not desirable in many community settings. An investigator must decide whether this time is worth the trade-off of information. The answer to this question will be influenced by the type of study and the population being studied. However, it is important for researchers to be aware that more co-occurring phobia has been consistently associated with poorer response to treatment and, greater likelihood of relapse.[20,21] The clinical significance of phobic symptoms underscores the need to attend to this component of anxiety symptomatology.

THE TWO CULTURES: DISCORDANT MODELS OF THE ANXIETY DISORDERS

A fourth characteristic of anxiety disorders is based upon the fact that there are two powerful models of these disorders that are not yet fully integrated. Specifically, neurobiological (generally biomedical) and learning theory (generally academic psychological) researchers use different paradigms to explain symptom onset and to guide treatment. When studying treatment in community settings, it is important that neither group ignore the other. In anxiety disorders, perhaps more than any other conditions, there is a need to build on information obtained from both of these academic disciplines, given that each field can claim clinical results.

Incorporating Information from Biomedicine and Academic Psychology in Study Designs

Anxiety disorders are unusual in that they have been the focus of intensive and more or less independent study by both biomedical/psychiatric and behavioural/psychological researchers. Efficacious treatments have been devised by each group. Yet, most treatment studies test interventions in one, but not both areas. The existence of two very different types of efficacious intervention for each of the anxiety disorders presents some especially challenging methodological issues for which there is no simple solution. The practice of ignoring the findings of the other modality when conducting studies is increasingly problematic. If not specifically instructed, pharmacotherapists may vary widely in their knowledge and use of efficacious behavioural interventions. This variation can be highly problematic for a treatment study. On the other hand, much effort must go into controlling the interaction of the pharmacotherapists with the patient. Researchers must decide how much behavioural intervention the medication therapist should administer. Complicated and time-consuming procedures are often required to ensure that such interventions are provided uniformly.

On the other hand, patients in the community often receive medication that can be efficacious for treating anxiety disorders, even before presenting to the cognitive behavioural therapist for treatment. Investigators must decide how to manage this situation. Should such patients be eliminated from a CBT trial? Should they be eligible and left on medication that is not fully effective? Or should all medication be discontinued? Each of these decisions is problematic since a partially effective medication can affect outcome whether it is continued or discontinued. Omission of this increasingly large group of patients, on the other hand, can also be an important threat to generalisability of the study.

Another problem for researchers is how to decide whether to compare medication and psychosocial treatments, and, if so, to decide how best to do so. There is clearly a need in the field to address this problem, but the solution to the problem is not trivial. Among the problems are that patients often have treatment preferences. Many will simply refuse to participate in a study in which they must agree to be assigned to a treatment modality at random. Others will agree and drop out when they receive an unwanted treatment assignment. There are a number of possible designs for a comparison treatment study. These include a full factorial design (Figure 20.2) in which each active treatment is compared to an

	Active medication	Placebo medication	No pill
Active psychotherapy	X	X	X
‘Placebo’ psychotherapy	X	X	X
No psychotherapy	X	X	X

Figure 20.2. Full factorial design.

inactive (placebo) control treatment and no treatment, and the two treatments are combined in all possible combinations. While this is the most complete design it is often impractical because of the treatment combinations (or lack of treatment) or because of the number required. It is difficult to undertake such a study at a single site and then there are problems with multiple sites in equivalence of providing all treatments and in minimising patient heterogeneity. An alternative class of designs has been described fairly recently[22] that allows patients and investigators to describe preferences in advance of randomisation and then be randomised within their preference set (‘equipoise stratum’).

Other design issues are related to the different putative underpinnings of symptoms as conceptualised from different points of view. These different viewpoints sometimes translate to different treatment targets. For example, a CBT approach to panic disorder focuses on underlying fear of bodily sensations, while the pharmacotherapist targets bursts of autonomic arousal. Pharmacotherapists and psychosocial researchers typically use different assessment strategies, and may or may not accept those of the other camp. In recent years, a series of multisite studies undertaken as a collaboration between neurobiologic and cognitive behavioural scientists has produced a more comfortable meeting ground for both groups. Still, there are disagreements.

In addition to the scientific differences, there are social and political differences between the two groups of researchers that can complicate methodological decisions in treatment trials. The investigators need to be clear about who the audience for their results will be. Design decisions may influence who will listen to their results. Ideally, a study can be designed so that it will be convincing to any treatment researcher. However, there are turf issues that may influence the mutual acceptance. Clinicians and researchers from one camp may feel the other is poaching on their turf. This is more likely to occur if there is insufficient attention to the issues of efficacy of the alternative treatments.

Guild issues are prominent in this field, and few pharmacotherapists understand the principles and techniques of administering cognitive behavioural treatment. Similarly many psychosocial researchers are not well informed about pharmacotherapy. Investigators from each group tend to have strong allegiance to the unique validity of their own methods. At times, there has been rancour and contentiousness between them, though this has improved in recent years. In the few instances that there has been a head-to-head comparison of medication and cognitive behavioural treatment, they have been similar in efficacy. It is not yet clear when or how combination treatments might be best administered. There is a need to take into consideration both biomedical and behavioural–psychological perspectives in designing treatment studies.

CONCLUSIONS

Although proven efficacious treatments have been identified for each of the anxiety disorders, the work of clinical trials remains unfinished. There are many unanswered questions, and much left to study in order to inform clinicians about how to optimise treatment decisions for patients with these debilitating conditions. In spite of achievements in documenting treatment efficacy over the past decades, treatment research has just scratched the surface. Innovations are needed in treatment development and in dissemination of proven interventions. To accomplish this there is a need for innovations in methodology. Efficacy studies, designed to meet US FDA regulatory needs,[23] will continue to have a role in the clinical research pipeline. But, there is a need for new clinical methods to support studies before and after efficacy. It is not our purpose to provide a comprehensive review of such methods. Rather, we have focused on a few key issues in anxiety disorders that require special consideration.

Existing clinical trials in anxiety disorders, like those in other areas of psychiatry, have provided information telling us which treatments are active in reducing target symptoms. Unanswered are a myriad of critical questions that relate to daily life decisions in the clinic. For example, do impairments as well as symptoms respond to efficacious treatments? If so, what is the time course of response? What is the optimum dose and duration of treatment to achieve maximal results? How often can we produce remission with existing treatments? What is the best way to define remission? Is maintenance treatment needed after remission is achieved? If so, how long? What if a patient does not achieve symptom remission? How should such a patient be managed over the long run? Do patients with complex comorbid conditions respond to treatment in a way that is similar to or different than those with less comorbidity? Can a clinician be confident that proven efficacious treatments are appropriately utilised in a patient whose symptoms meet criteria for the target disorder, but who differs in demographic characteristics, social supports, or other ways from those seen in efficacy studies? How closely must procedures in the community follow those used in research studies in order to achieve the same results?

These and other questions like them are often broadly grouped under the rubric of 'effectiveness' studies. We focus especially on characteristics of anxiety disorders that make these decisions complicated and that comprise methodological challenges for researchers. We confess that we may raise more questions than we can answer. However, where possible, we will at least provide suggestions about possible ways to address the problems.

Decisions about primary, secondary and tertiary prevention interventions are not so well specified. There is accumulating evidence for psychological and neurobiological precursor states for anxiety disorders and for psychosocial risk factors. This information raises hopes for the possibility of primary prevention. Clearly it would be advantageous to be able to intervene early, before the development of the disorder and, ideally, even before the onset of a precursor state. Once established, anxiety disorders are chronic relapsing conditions. With or without treatment, patients are likely to experience symptoms that wax and wane, to meander in and out of full-fledged symptom states in different configurations, to experience temporary, partial or incomplete states of remission. We need more information about how to manage anxiety disorders, once established, in order to best prevent complications and recurrence of full symptomatic episodes.

REFERENCES

1. Kessler RC, McGonagle KA, Zhao S, Nelson CB, *et al*. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States:

Results from the National Comorbidity Survey. *Arch Gen Psychiat* (1994) **51**(1): 8–19.

2. Hohmann A, Shear MK. Community based intervention research: coping with the noise of real life in study design. *Am J Psychiat* (2002) **159**: 201–7.
3. Kessler RC, Soukup J, Davis RB, Foster DF, Wilkey SA, Van Rompay MI, Eisenberg DM. The use of complementary and alternative therapies to treat anxiety and depression in the United States. *Am J Psychiat* (2001) **158**(2): 289–94.
4. Shear MK, Greeno C, Kang J, Ludewig D, Frank E, Swartz HA, Hanekamp M. Diagnosis of non-psychotic patients in community settings. *Am J Psychiat* (2000) **157**(4): 581–7.
5. Glisson C, Hemmelgarn A. The effects of organizational climate and interorganizational coordination on the quality and outcomes of children's service systems. *Child Abuse Neglect* (1998) **22**(5): 401–21.
6. Rosenheck R. Stages in the implementation of innovative clinical programs in complex organizations. *J Nerv Ment Dis* (2001) **189**(12): 812–21.
7. Swartz H, Shear MK, Frank E, *et al.* A pilot study of community mental health case for depression in a supermarket setting. *Psychiatr Serv* (2002) **53**: 1132–7.
8. Roy-Byrne PP, Katon W, Cowley DS, Russo JE, Cohen E, Michelson E, Parrot T. Panic disorder in primary care biopsychosocial differences between recognized and unrecognized patients. *Gen Hosp Psychiat* (2000) **22**(6): 405–11.
9. McKay M, McCadam K, Gonzales J. Addressing the barriers to mental health services for inner city children and their caretakers. *Commun Mental Health J* (1996) **32**(4): 353–61.
10. Essock S, Goldman H. Outcomes and evaluation: system, program and clinician level measures. In: Minkoff K, Pollack D, eds, *Managed Mental Health Care in the Public Sector: A Survival Manual. Chronic Mental Illness*, Vol. 4. Amsterdam: Harwood Academic (1997) 295–307.
11. Dohrenwend BP, Shrout PE. Toward the development of a two-stage procedure for case identification and classification in psychiatric epidemiology. *Res Commun Mental Health* (1981) **2**: 295–323.
12. Dixon L, Green-Paden L, Delahanty J, Lucksted A, Postrado L, Hall Jo. Variables associated with disparities in treatment of patients with schizophrenia and comorbid mood and anxiety disorders. *Psychiat Serv* (2001) **52**(9): 1216–22.
13. Kandel DB, Huang F-Y, Davies M. Comorbidity between patterns of substance use dependence and psychiatric syndromes. *Drug Alcohol Depend* (2001) **64**(2): 233–41.
14. Brady KT. Comorbid posttraumatic stress disorder and substance use disorders. *Psychiat Ann* (2001) **31**(5): 313–19.
15. Widiger TA, Sankis LM. Adult psychopathology: issues and controversies. *Ann Rev Psychol* (2000) **51**: 377–404.
16. Fava GA, Rafanelli C, Ottolini F, Ruini C, Cazzaro M, Grandi S. Psychological well-being and residual symptoms in remitted patients with panic disorder and agoraphobia. *J Affect Disord* (2001) **65**(2): 185–90.
17. Yerkes RM, Dodson JD. The relation of strength of stimulus to rapidity of habit formation. *J Comp Neurol Psychol* (1908) **18**: 459–82.
18. Diefenbach GJ, McCarthy-Larzelere A, Williamson DA, Mathews A, Manguno-Mire GM, Bentz BG. Anxiety, depression, and the content of worries. *Depress Anxiety* (2001) **14**: 247–50.
19. Katschnig H, Amering M. The long-term course of panic disorder and its predictors. *J Clin Psychopharm* (1998) **18**(6, Suppl 2) 6S–11S.
20. Scheibe G, Albus M. Predictors and outcome in panic disorder: a 2-year prospective follow-up study. *Psychopathology* (1997) **30**: 177–84.
21. Cowley DS, Flick SN, Roy-Byrne PP. Long-term course and outcome in panic disorder: a naturalistic follow-up study. *Anxiety* (1996) **2**: 13–21.
22. Lavori PW, Rush AJ, Wisniewski SR, Alpert J, Fava M, Kupfer DJ, Nierenberg A, Quitkin FM, Sackeim HA, Thase ME, Trivedi M. Strengthening clinical effectiveness trials: equipoise-stratified randomization. *Biol Psychiat* (2001) **50**(10): 792–801.
23. Rubinow DR. Practical and ethical aspects of pharmacotherapeutic evaluation. In: Ginsburg BE, Carter BF, eds, *Premenstrual Syndrome: Ethical and Legal Implications in a Biomedical Perspective*. New York: Plenum Press (1987) 47–55.

21

Cognitive Behaviour Therapy

NICHOLAS TARRIER[1] AND TIL WYKES[2]
[1]School of Psychiatry and Behavioural Sciences, University of Manchester, Manchester M23 9LT, UK
[2]Institute of Psychiatry, De Crespigny Park, London SE5 8AF, UK

BACKGROUND

We have chosen in this chapter to provide an overview of the difficulties for the investigation of psychological therapies using the methodology of randomised control trials. In order to do so we have selected studies of a new treatment for psychosis, cognitive behaviour therapy (CBT). This is a new therapy that, following a period of development, has now resulted in four large randomised control trials.

Cognitive behaviour therapy is a therapy that targets the symptoms of one disorder, schizophrenia. The lifetime risk is 1%. This disorder is characterised by a cluster of specific symptoms that are typically divided into two categories, positive and negative. Positive symptoms include auditory hallucinations and delusions, both of which produce much distress. Negative symptoms include lack of drive, emotional apathy as well as poverty of speech and social withdrawal.

In many, if not most, cases the disorder follows a relapsing course.[1] A significant proportion of, but not all, people suffering from the disorder have poor outcomes, i.e. with high levels of dependence on continuing psychiatric care, low levels of financial independence and little social fulfilment. There is some underlying variation in the disorder,[2] which is probably affected by interactions with other clinical, social and environmental demands and supports such as life events (death of parent), absence of a supportive family (or presence of a critical one) and economic conditions (high unemployment).

Several different sorts of psychological therapies have been developed to address the following outcomes:

- Total number of symptoms
- Distress caused by symptoms
- Relapse
- Social functioning
- Family engagement
- Quality of life
- Skills/thinking style, e.g. problem solving, coping skills.

The currently accepted treatment for the positive symptoms of schizophrenia is medication. This has been shown to reduce them significantly and does reduce relapse. However, it also has costs as well as benefits in that there is the risk of

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

developing side effects such as tremour, restlessness and uncontrollable mouth movements. Most side effects disappear on stopping the medication but there is the chance that the mouth movements will develop into a condition known as tardive dyskinesia that is irreversible. Some patients, about one-third, also continue to experience positive symptoms despite adequate doses of medication. It is these symptoms that were the targets for change in this further development of a psychological therapy, CBT.

Because of the potential risks of long-term medication and the unpleasant side effects also experienced on short-term treatment, consumers of mental health services have been extremely positive about the development of psychological treatment. This has led to further pressure on funders of health service research to provide more data on acceptable alternatives or adjuncts to medication treatment. Hence the recent trials of CBT in the UK sponsored by either the UK Department of Health directly, government research agencies or large UK research charities.

WHAT IS COGNITIVE BEHAVIOUR THERAPY?

The main developmental roots for CBT have been in depression and anxiety. This began over 20 years ago but more recently the approach has been applied to people with schizophrenia. This later development produced changes in the way the intervention is presented, although the underlying model of change may be similar to that adopted for the other disorders. The main aim of the intervention is to reduce distress, disability and emotional disturbance as well as the relapse of the acute symptoms.[3] Cognitive behaviour therapies are active and structured therapeutic methods and should be distinguished from psychoeducation which tends to be simple, didactic and educational. Brief educational packages have been shown to be ineffective either with families[4] or with individual patients.[5]

Although there are specific components of CBT that would be accepted by all its proponents, these ingredients may be given in different proportions by different groups of professionals and for different individuals within a single service. Below is a list of the ones that we have identified as being used by most groups:

- Engagement with the client.
- Problem identification.
- Agreeing on a collaborative formulation of the problems to be assessed.
- Use of alternative explanations to challenge delusional and dysfunctional thoughts.
- Establishing the link between thoughts and emotions.
- Encouraging the patient to examine alternative views of events.
- Encouraging the patient to examine the link between thoughts and behaviour.
- Use of behavioural experiments to reality test.
- The setting of behavioural goals and targets.
- Developing coping strategies to reduce psychotic symptoms.
- Development and acquisition of relapse prevention strategies.

Some groups have also included:

- Improvement in self-esteem.
- Increasing social support and social networks.
- Schema focused therapy.

TREATMENT DEVELOPMENT

New treatments usually evolve through a number of stages. Initially the problem is identified and suggestions, involving theoretical and pragmatic elements in varying degrees, are advanced for its solution. Innovative case studies are carried out. Replications and developments in other case studies and case series follow. The next stage consists of uncontrolled and small exploratory controlled trials. These are often innovative but methodologically weak. Finally the 'gold standard' of evaluation, the large randomised controlled trial (RCT), is carried out if the new treatment is showing sufficient promise. RCTs are increasingly large and methodologically rigorous and therefore more expensive, often now involving numerous sites and large numbers of patients. A further theme is that of identifying what is responsible for the improved outcome following the treatment; that is, the trial includes an explanatory element. Trials that identify the key components responsible for the changes are essential to the further development of treatment and to the dissemination of the treatment package into the wider health service. However, this lack of an accepted theoretical base does not (and should not) prevent a number of different and successful treatment innovations from being introduced into health services. Pragmatic trials, which address the issue of whether or not a new treatment works within a routine service setting, are usually large, simple and multi-centred and evaluate a small number of outcomes. These trials do not address the question of why a treatment works. But this understanding is essential because the costs of care might be reduced considerably if it is discovered that a rather simple and cheap component of treatment is responsible for the majority of the variance in treatment outcome. These treatment extensions, although important, rarely get adequate funding following the initial innovative RCTs.

The development and evaluation of CBT for psychosis is no different from the development of other treatments in mental health and has followed a characteristic path. Numerous case studies were published, some as far back as the 1950s. For example Dr A.T. Beck initially worked with psychotic patients and published a case study of the cognitive treatment of a patient suffering from delusional disorder[6] before moving to start his seminal work on depression. Other case studies were published in the 1970s and 1980s (see Tarrier[7,8] and Haddock *et al.*[9] for reviews), but it was not until the recent decade that randomised controlled trials were carried out. Small trials with methodological weaknesses were initially published. For example, Tarrier *et al.*[10] compared coping training with problem solving but assessments were not blind and drop-outs were not included in the analysis. Garety *et al.*[11] compared cognitive behaviour therapy to treatment-as-usual but again assessments were not blind and group allocation was not random. Drury *et al.*[12,13] evaluated cognitive therapy with acutely ill patients, but the treatment included individual and group treatment of patients and families while assessments were neither independent nor blind. However, three medium size methodologically robust trials of CBT variants have been carried out with chronic schizophrenic patients,[14–16] and one large multisite trial with recent onset acute patients (the SoCRATES Trial[17]). It is therefore appropriate to review not only these trials but also the changes in clinical trial methods in this field in order to begin to define the most optimal strategy for the future evaluation of this and other psychological therapies.

WHY CARRY OUT CLINICAL TRIALS OF PSYCHOLOGICAL TREATMENTS?

PURPOSES AND OUTCOMES

There are a number of different beneficiaries from clinical trials. From the health services perspective there is an increase in knowledge about what treatments are likely to provide the most benefits (see section on Evidence-Based Practice below). In addition, for clinical academics there may be elements of the design of a trial that will allow certain models of aetiology or treatment efficacy to be tested which can inform theories of the disorder as well as leading to improvements in treatment. For therapists the trial may produce clinical improvements that mean that the participants can make health gains and for the patients the treatments may provide them with changes that are valued, such as increased social inclusion. So it cannot be assumed that there is a single purpose for carrying out a clinical trial. These different purposes change the type of trial performed, particularly in relation to how outcomes are defined. We have set out a number of different outcomes below which may be variously valued by different groups (in the list respectively health service, clinical academic, therapist and participant) and which could be targets in CBT trials. It cannot be assumed that all groups will value all outcomes to the same extent, or that the same outcome would be measured in the same way from the different viewpoints. For instance, symptoms can be measured as a simple change over treatment, by a threshold amount or by the effects on the emotional life of the patient, for example the distress caused by the symptom.

Possible outcomes of treatment:

- The occurrence or frequency of a particular event: e.g. number of relapses, time to relapse (survival functions).
- The use of services or other resources: e.g. days spent in hospital, use of community mental health care.
- Improvement in symptoms at a level assumed to be of clinical significance: e.g. at least 20% or 50% improvement, return to within normal range.
- Change in a single symptom or other continuous outcome that is considered central or primary to the disorder: e.g. severity of delusions or hallucinations.
- Change in psychopathology that is general or secondary to the disorder: e.g. scores on a standard measure of psychopathology, severity of distress or anxiety.
- Changes to other important aspects of the person's life: e.g. social functioning, number of friends, quality of life.

Trials are also expensive and so the chances of funding are dependent on the types of trials wanted by the funding agencies. The main beneficiary (and funding) of clinical trials is the health service that would prefer pragmatic rather than model testing trials. But in the UK there has also been a new trend that may also affect the type of trial – the inclusion of mental health service users (consumers) and, where appropriate, carers on the trial management committees. There are examples of this; users and carers were represented in this way on a trial of CBT in dual diagnosis patients[18] and of effectiveness of family interventions,[19] and also on the research steering group which generates the research designs for the Centre for Recovery in Severe Psychosis at the Institute of Psychiatry in London. The involvement of service users in clinical trials in the UK is now defined in guidelines provided by the Consumers in Research Unit within the Department of Health. This new undertaking does not seem to be prevalent in other countries.

The difficulty for research into psychological treatments is that studies are usually funded from public resources even at the early stages. This is in contrast to trials of medications where particular companies not only are required to carry out specific research for licensing but are also likely to benefit financially from the results of trials. Unlike drugs, psychological treatments do not have a specific product champion and therefore have to compete with other health care trials for scarce resources.

THERAPEUTIC RELEVANCE

In addition to the list of possible outcomes above there are other measures that may be essential in the assessment of outcome in a trial. For instance, if one of the hypotheses is that the therapy works through a specific mechanism then a sole outcome measure without recourse to either qualitative and/or process measures would not provide a test of this hypothesis. This is an extension of the sorts of questions stipulated by a clinical academic but is also essential to the health services. It may be that the treatment provides its effects through a simple mechanism which could be provided in a less sophisticated way; that is, not requiring high levels of training and supervision.

It has been suggested that psychological therapies may all work through a common pathway: that the non-specific effects of psychotherapy may account for much of the effect of treatment outcome.[20,21] This is hardly surprising as psychological therapies have much in common with each other – they involve, for example, an interaction, negotiation of goals, an agenda for each session. The improvement could be produced by these commonalities and not through the specific model of therapy adopted. For example, treatments that were designed as non-specific placebo controls (e.g. befriending in Sensky *et al.*[16] and supportive counselling in Tarrier *et al.*[15]) performed much better than expected, although never better than CBT. Therefore, the choice of a comparison group is extremely important. If psychological therapy is compared to treatment-as-usual (TAU), which includes less individual attention than the psychological therapy, its effectiveness may be due to shared common themes of psychological therapy not to ingredients of a particular therapy. Tarrier *et al.*[10] investigated the effects of expectation of therapeutic benefit by the use of a demand and counter-demand manipulation. Half of the participants were told that they should expect therapeutic benefit to accrue with their progress through treatment (demand condition) and the other half were told that they should expect benefit but that it would not be apparent until after the end of treatment and post-treatment assessment (counter-demand condition). This manipulation of expectancies had no effect on clinical measures, suggesting that at least the anticipation of treatment benefit was not influential in this patient group.

Alternatively, as psychological therapies include specific attributes in common it may be wrong to conclude, in a comparison of two types of psychological therapy, that CBT is not the best form of therapy when the two treatments do not differ significantly from each other. There is always the danger that the study will be underpowered to demonstrate an advantage of CBT when the non-specific control group does better than expected. However, CBT may be significantly better than TAU, whereas the alternative may not give such an advantage. When the health services have to decide which of several forms of psychological therapy to choose to add to their therapeutic armantarium, selecting CBT would be their best choice.

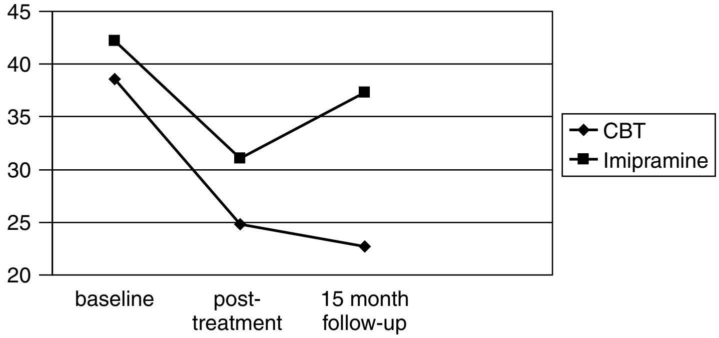
ACUTE CARE, MAINTENANCE THERAPY AND DURABLE EFFECTS

Schizophrenia is most often a chronic relapsing condition. If we take the metaphors from treatments with medication then psychological therapy could be provided in a number of different ways.

- Acute antibiotic treatment which kills off the bacteria causing the disease (intensive psychological treatment which changes a key factor in the psychological make-up of the individual, e.g. cognitive behaviour therapy for panic disorder[22]).
- Acute treatment of symptom exacerbations and maintenance treatment, e.g. asthma treatment with steroids followed by maintenance with Salbutamol and/or sodium chromoglycate (CBT for chronic depression[23]).
- Prophylactic treatment for malaria (e.g. debriefing treatments for possible post-traumatic stress disorder[24]).

Psychological therapy often sets itself the same target as treatment using antibiotics, with an acute phase followed by a follow-up during which there is no active treatment. This protocol mainly resulted from the lack of specialist input in the health services, making it imperative to ration services. It also follows a set of expectations that come from the behavioural tradition in the treatment of psychological problems and recent CBT interventions for anxiety disorders where interventions are brief and the effects durable. For example, Figure 21.1 shows the effects of imipramine and CBT for panic disorder from a trial by Clark *et al.*[22] In their trial the drug and the psychological treatment had similar effects at the end of treatment, but psychological treatment had a more permanent effect and the differences between the two treatments were significant at follow-up. The improvement was predicted by the change in cognitions following treatment. In other words, the psychological treatment changed a maintenance factor for the disorder.

This expectation of intensive treatment producing durable gains may not be appropriate for schizophrenia as it is a relapsing condition that, in some cases, may have a deteriorating course. Furthermore, residual symptoms may be present between episodes of exacerbation. Residual positive symptoms at discharge are a risk factor for relapse.[5] Many maintenance and causal factors have been proposed and are in multiple domains, such as social, biological as well as



Source: Reproduced from Clark *et al.*,[22] with permission.

Figure 21.1. Psychological treatment for panic disorder.

psychological. It is possible that CBT could have a successful effect on one of these factors but fail later when other factors become crucial in the progress of the disorder. This would be shown as a successful outcome at post-treatment but a lack of durability of gains at follow-up. However, the usual interpretation of this pattern of results is that the effect on the disorder was only temporary. If gains were 'only temporary' in a group of patients who were chosen because their symptoms were 'residual' then even this 'temporary' gain would be welcome. This set of results, rather than dismissing the treatment effect, actually generates a further question – how do we maintain the gains made during treatment when treatment is withdrawn?

The therapeutic protocol adopted for schizophrenia with medication is to provide medication intensively at the acute stage that is followed by maintenance treatment at lower dose of similar drugs. It may be that psychological treatment needs to be provided in an equivalent way.

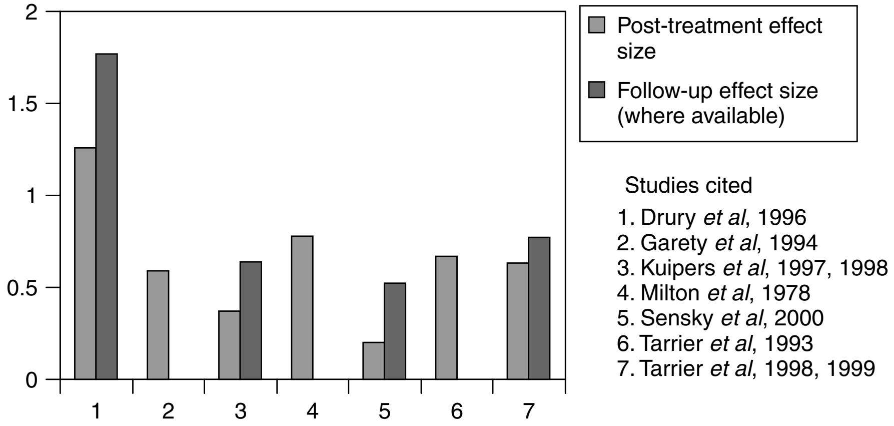
An alternative mechanism and pattern of results for CBT could be improvement in one factor, such as self-esteem, which then allows further improvements in other factors to occur, such as increased social support through the extension of a support network by increased social contact. This would produce an improvement at post-treatment and even greater gains at follow-up. However, it would appear that CBT was not only durable but conferred greater benefits as time passes, although it would not be clear to the research team how this latter improvement came about. This poses the question of how do we explain increases in effect size post-therapy which cannot be explained merely by the loss to follow-up of those people for whom the therapy conferred hardly any benefit at all?

Trials of acute CBT have shown significant effects of therapy mostly at the cessation of treatment but always after a follow-up period. Figure 21.2 provides data from Gould *et al.*[25] of the effect sizes of seven trials calculated from the following equation:

$$\text{Effect size} = (M_\text{t} - M_\text{c})/\text{SD}_\text{c}$$

where M_t is the mean of the treatment group, M_c is the mean of the control group and this is divided by the standard deviation of the control group of participants.

The mean effect size for the trials studied by Gould *et al.* is 0.65 (95% CI 0.56–0.71). This average is called a medium to large effect size according to Cohen[26] (p. 40). Patients continued to improve over the follow-up period with the combined effect size cited by Gould *et al.*[25] of 0.93 for the four studies reporting a follow-up period. These results are encouraging given that schizophrenia is a relapsing condition where life events and other stressors may trigger new episodes of illness. However, the



Source: Reproduced from Gould *et al.*,[25] with permission from Elsevier.

Figure 21.2. Effect sizes of CBT trials.

interpretation of the results of individual trials has since changed, mainly because the accepted standards for trials have changed. Several trials that make up this figure are methodologically weak with difficulties in random assignment, blindness of ratings, adequate outcome assessment and problematic or unsophisticated analyses (see below).

FAIRNESS AND CHANGING STANDARDS IN TRIAL DESIGN AND REPORTING

Standards have changed and what was reported in papers a number of years ago would have been adequate and satisfactory for the times. However, there are now clear guidelines on how trials should be reported, formalised in the CONSORT Statement.[27,28] CONSORT is a checklist and flow diagram that were designed to improve the quality of reports of randomised controlled trials. The checklist gives detailed instruction on describing the study's method and design, assignment and randomisation, masking (blinding), participant flow and follow-up, and analysis. The flow diagram provides readers with a clear picture of the progress of all participants in the trial, from the time they are referred to the trial until the end of their involvement. It should include the number assessed for eligibility for the trial, reasons for exclusion, who was randomised and what happened to them prior to final assessment and analysis of the trial results.

These standards on reporting, by implication, provide strong recommendations to researchers about what they need to consider and action when designing and managing a trial. Table 21.1 contains a list of those points of the design or analysis that can seriously bias the interpretation of the results.

The majority of the current CBT trials do not conform to the reporting guidelines as set out in here. For some trials the significant discrepancies between Table 21.1 and the trial may lead to a biased interpretation of the results. For example, in Drury *et al.*[12,13] it is not clear what specific therapy is provided as a variety of different components were being tested at the same time. In Garety *et al.*[11] the participants were not randomly allocated to treatment groups. Kuipers *et al.*[14] had no blind assessment of treatment outcomes. Sensky *et al.*[16] recruited by repeatedly canvassing local services for referrals. However, the current meta-analyses do show that despite these methodological difficulties there seem to be significant changes in overall symptoms following treatment with CBT.

Table 21.1. Items that should be included in reports of randomised trials

Heading	Subheading	Descriptor
Title		Identify the study as a randomised trial
Abstract		Use a structured format
Introduction		State prospectively defined hypothesis, clinical objectives, and planned subgroup or covariate analyses
Methods	**Protocol**	**Describe**
		Planned study population with inclusion or exclusion criteria
		Planned interventions: their nature, content and timing
		Primary and secondary outcome measure(s) and the minimum important difference(s), and indicate how the target sample size was estimated
		Reasons for statistical analyses chosen, and whether these were completed on an intention-to-treat basis
		Mechanisms for maintaining intervention quality, adherence to protocol and assessment of fidelity
		Prospectively defined stopping rules (if warranted)
	Assignment	**Describe**
		Randomisation (e.g. individual, cluster, geographic)
		Allocation schedule method
		Method of allocation concealment
	Masking (blinding)	**Describe**
		Mechanism for maintaining blind and allocation schedule control
		Evidence for successful blinding
Results	**Participant flow and follow-up**	Provide a trial profile summarising participant flow, numbers and timing of randomisation assignment, interventions, and measurements for each randomised group
	Analysis	State estimated effect of intervention on primary and secondary outcome measures, including a point estimate and measure of precision (confidence interval)
		State results in absolute numbers when feasible (for example, 10/20, not 50%)
		Present summary data and appropriate descriptive and interferential statistics in sufficient detail to permit alternative analyses and replication
		Describe prognostic variables by treatment group and any attempt to adjust
		Describe protocol deviations
	Discussion	State specific interpretations of study findings, including sources of bias and imprecision (internal validity) and discussion of external validity, including appropriate quantitative measures when possible
		State general interpretation of the data in light of the available evidence

Source: Modified from Begg *et al.*[27] and Moher *et al.*[28]

EVIDENCED-BASED PRACTICE

There is considerable current enthusiasm for evidence-based health care in general and mental health care, but there is a current debate on how evidence-based practice can be consistently implemented in routine settings both in the UK[29] and abroad.[30,31]

Evidence-based practice is the delivery of interventions for which there is strong scientific evidence that they improve relevant patient outcomes. Although the type of scientific evidence does vary, the gold standard for treatment outcome is the RCT. Where several trials exist they can be considered together through meta-analysis. Knowledge concerning evidence-based practice accrues through the accumulating results of efficacy and effectiveness studies.

Thus the purpose of evidence-based practice is (a) to ensure that the wealth of research evidence informs clinical practice so that those who are in receipt of treatment will receive the treatment that is the best available and represents the current knowledge base, and (b) to ensure that planning and policy is determined by empirical evidence, for those purchasing services to be able to make informed choices and for those receiving services to be empowered by such knowledge. Furthermore, the establishment of an evidence-based practice knowledge base of what works allows the practice of mental health services and individual clinicians to be compared to the evidence base. This increases accountability and establishes guidelines for good practice and improves the quality of mental health services. Limitations in evidence also set the research agenda for the future.

There are, however, critics of the collation of data for evidence-based practice. This mainly focuses around the use of specific meta-analytic techniques that have very limited entry criteria. The Cochrane database, for example, provides valuable searches and evaluations of randomised control trials with strict criteria for entry. Although the evidence may be strong for a particular practice it may be based on a very small number of studies. The main criteria for exclusion are the lack of randomisation of the participants within the trial and the lack of data on all those participants who entered the trial. Although clearly the results of such trials should be less weighted in the final evaluation, such information may be valuable when few other data are available.

TRIALS METHODOLOGY AND AIMS

Efficacy trials are devised to test whether the therapy has an effect overall on the outcomes of interest. They are carried out in relatively controlled environments, usually by sophisticated university-based research teams, and often involve highly expert therapists. For CBT trials the outcomes of interest are a reduction in overall psychotic symptoms, reductions in relapse or reduced rates on admission to hospital, reduced psychopathology and improvements in functioning. These trials may also include various control groups and process measures to help understand why the treatments work. An effectiveness trial attempts to more closely resemble the real world of routine services, inclusion criteria are wider so the sample treated is more heterogeneous and includes the atypical patients, and the therapists are recruited from the routine services. The measured outcomes are reduced to the minimum and tend to be gross measures that are clinically significant such as relapse or hospital admission; health economic measures to assess cost are also desirable. In special cases an equivalence trial may be designed, in which a new treatment is expected to match the clinical efficacy of an established treatment but may have other benefits, for example in terms of acceptability or cost. These trials have special methodological features that distinguish them from simple comparative trials.[32]

PARTICIPANT SAMPLES

Recruitment Bias

Figure 21.3 shows how the patient flow in a study should be described. The box of particular interest is the one at the very top that describes those who have been assessed for eligibility. In order to prevent bias in recruitment the best method for ascertaining samples of patients for a trial is to recruit them from a cohort of patients in contact with a service that covers a geographic area (as in Tarrier *et al.*[15] and Lewis *et al.*[17]). This ensures that the people who are in the trial do represent those who have the disorder. In the UK it is largely assumed that those patients with schizophrenia in contact with the services will represent those with the disorder requiring clinical intervention. For example, Tarrier *et al.*[15]

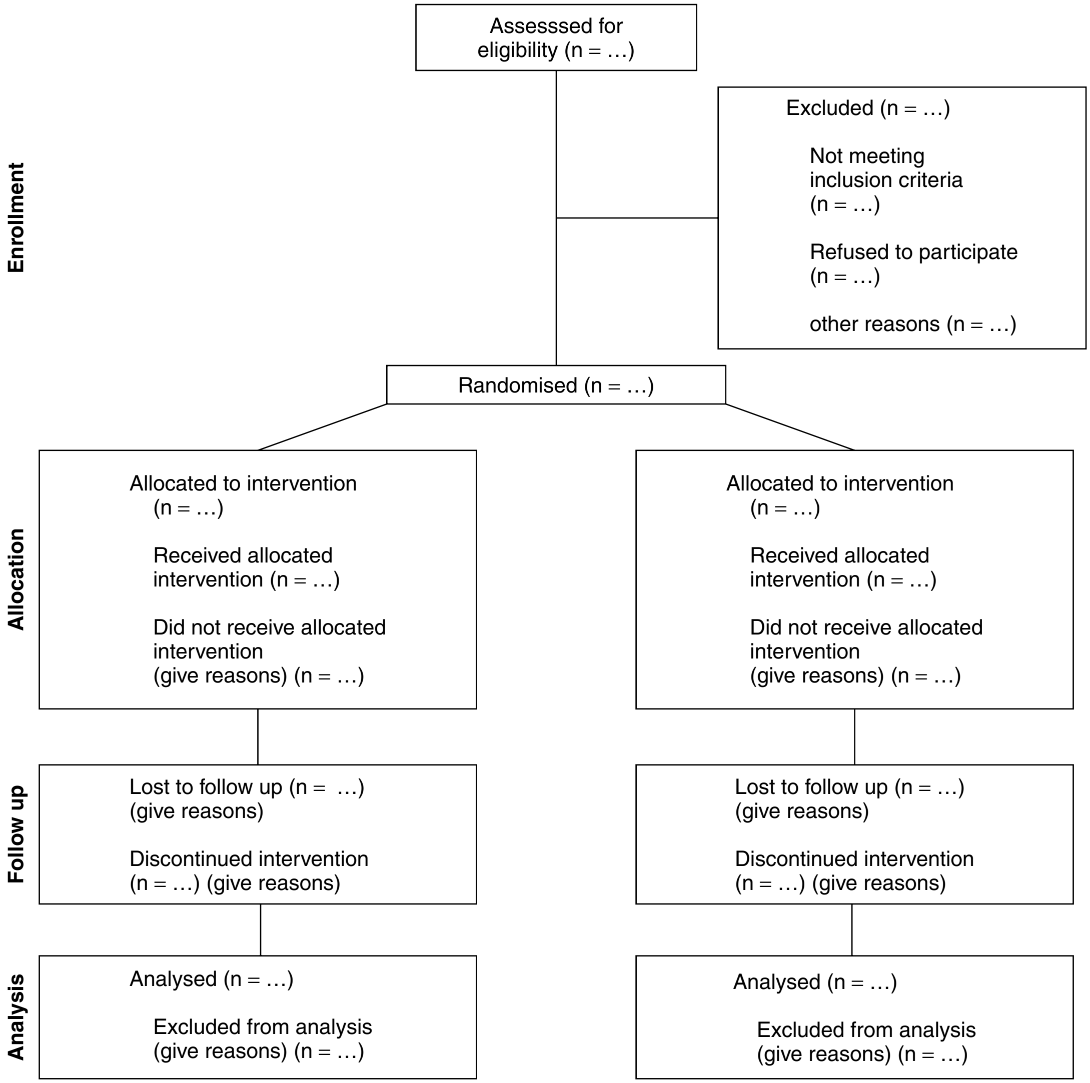

Assesssed for eligibility (n = …)
Excluded (n = …)
Not meeting inclusion criteria (n = …)
Refused to participate (n = …)
other reasons (n = …)
Enrollment
Randomised (n = …)
Allocation
Allocated to intervention (n = …)
Received allocated intervention (n = …)
Did not receive allocated intervention (give reasons) (n = …)
Allocated to intervention (n = …)
Received allocated intervention (n = …)
Did not receive allocated intervention (give reasons) (n = …)
Follow up
Lost to follow up (n = …) (give reasons)
Discontinued intervention (n = …) (give reasons)
Lost to follow up (n = …) (give reasons)
Discontinued intervention (n = …) (give reasons)
Analysis
Analysed (n = …)
Excluded from analysis (give reasons) (n = …)
Analysed (n = …)
Excluded from analysis (give reasons) (n = …)

Figure 21.3. Participant recruitment and flow.

screened all patients who might have a diagnosis of schizophrenia in a number of NHS trusts, selected those who achieved predetermined criteria and examined their notes further. All putative candidates following this procedure were interviewed to ascertain whether they satisfied the entry criteria. This method has been used as a gold standard and other trials of CBT have used the data from Tarrier *et al.*[15] to compare with their study sample in order to conclude that their sample was representative (see Sensky *et al.*[16]). What a comparison of samples allows is just that – if the samples are similar then the results of the trials can be usefully compared, but this information cannot be used as evidence of sample representativeness.

Convenience samples which recruit from clinic attenders or, even more problematically, patients referred to the project by their clinicians are at risk of selection bias. The referrer may only select those possible participants who they view as good candidates for the treatment or conversely patients who are difficult or treatment refractory. Recruitment of referred patients is unfortunately the norm.[14,16] Even though it may be possible to compare the recruited sample to the whole population of patients who may be eligible in terms of socio-demographic and clinical service contact, this will not be enough data to rule out a systematic bias. In the treatment of panic disorder, Klein[33,34] has argued vigorously that in comparisons of psychotherapy verses drug, a pill placebo – drug comparison is necessary to ensure that the sample is not atypical since the efficacy of the drug (in this case imipramine) is well established. This is largely an argument about how representative or typical any sample is, given a reliance on convenience samples.

Selection

There are a number of different factors that need to be considered as part of the recruitment process such as service delivery system, academic support, socio-economic status of the area and geography (urban, suburban and rural area). It is unlikely that these will have a specific interaction with the outcome from therapy, but as these factors will affect the generalisation of the trial results it is probably important for the sample to represent a variety.

But ethnicity and cultural mix may potentially affect therapy outcomes. As we know very little about how to target psychological therapy to different cultural groups, it seems reasonable to start investigating a new treatment with a culturally homogeneous group and in later trials modify to accommodate cultural diversity, if such modification would be a requirement of effectiveness in cultural subgroups.

Diagnosis

In psychological therapies, especially in the field of psychosis, there has been a dilemma about whether to adopt medical diagnoses as entry criteria to studies. Some clinical psychologists (e.g. Bentall *et al.*[35]) would prefer the adoption of symptomatic entry criteria as schizophrenia is a term covering a group of people with a wide variety of abnormal experiences. So some trials have as their entry criteria a specific symptom experienced as distressing rather than membership of a single diagnostic category.[11,36] However, even in these studies some patients were excluded on the basis of diagnosis because of not fulfilling other criteria (see below), and it was certainly the view of one of the authors (TW) that in feasibility studies of group CBT some patients with diagnoses other than schizophrenia, e.g. personality disorder or bipolar affective disorder, did not respond in similar ways to the patients with diagnoses of schizophrenia or schizoaffective disorder. Current CBT studies have generally included patients from the schizophrenia spectrum and it is certainly the view of some CBT therapists that the type of therapy offered to people with bipolar affective disorder is different from that designed for schizophrenia.[37]

Even when diagnosis is used there are too many different systems to choose from (e.g. clinical case note diagnosis, research diagnoses (RDC), DSMIV, ICD10, etc.). The choice of a different system will change the characteristics of the sample. For instance, if people are drawn on RDC criteria they will not necessarily be as chronic as those fulfilling the DSMIV criteria.

Exclusion Criteria

As well as criteria for inclusion into trials most studies also exclude people on the basis of specific issues. In trials of psychological therapy for psychosis one usual criterion is that the people who enter the trial are those whose symptoms have remained despite adequate doses of medication. The group chosen on this basis is extremely chronic and refractory and provides an extremely stringent test of the efficacy of psychological treatment.

A further thorny issue is that of co-morbid substance abuse. Most studies will exclude individuals when the abuse is severe, but the criteria for severity are rarely set out clearly so that it is impossible to compare between trials. Patients who are recruited from inner city areas are unlikely to be free of recreational drug use. A small consumption of cannabis may not affect the therapy efficacy, but it is not clear whether any use of class A drugs affects the therapeutic effect of psychological treatment. A more recent trial has been designed to test the efficacy of CBT and family intervention to treat dual diagnosis patients (those diagnosed as suffering from schizophrenia and substance abuse) in which the substance abuse is thought to increase the risk of poor outcomes in the primary disorder.[18] In this case severe substance abuse was an entry criterion.

Again some people may have a co-morbid organic condition such as epilepsy that may warrant exclusion, although most trials again would evaluate whether the organic condition is primarily responsible for the symptoms of the disorder which they are trying to alleviate. Deteriorating brain disorders such as Alzheimer's disease may be a reasonable exclusion criterion as CBT relies on the carry-over of changes in one session to subsequent sessions. Similarly, people who have learning disabilities may also have some difficulties with CBT as it is currently devised, although therapists have extended treatment for depression to the learning disabilities field. Current trials also do not support the idea that lower IQ prevents therapeutic changes.[38] But all current trials do have a lower cut-off for IQ, usually around 65.

Drop-out or Lost to Follow-up

Two main issues affect the inferences about the trial results. The first is the effect of those people who drop out of the therapy and the second is those people who are lost to contact at any stage of the trial. Different systems of dealing with drop-outs can be adopted. Some systems assume that the person would not have changed at all since leaving the trial (LOCF), but this approach has its problems.[39] But assuming that the group who drop out would have performed in the same way as those who remained also produces difficulties. Drop-outs may be those people who might never have achieved any change following therapy. Clearly if a treatment produces high levels of drop-outs this might imply something about the acceptability of treatment. A precise description of drop-outs is required but, from the trials submitted so far, this is missing in all but a few cases.

More research on drop-outs is clearly required. But in the area of mental health in particular, the research is difficult, if not impossible, to carry out. The guidelines on research governance[40] do not allow for the harassing of people who have dropped out of trials, for their reasons for dropping out or for data on their current health status. However, it is not only of interest academically, as it provides some information on the veracity of the theory underlying the disorder, but also essential to inform the health services. For example, Tarrier *et al.*[41] reported that patients who dropped out of treatment tended to be male, unemployed and unskilled, single, with a low level of educational attainment and a low premorbid IQ. They had a lengthy duration of illness although at the time of discontinuation they were not severely ill and functioned at a reasonable level. They were likely to be paranoid but not suspicious of the therapist. They were unlikely to be grandiose. They did not understand the rationale for therapy or the potential for benefit but feared it could make them worse.

It is not clear whether it is appropriate or ethical to collect personal information that is kept for routine monitoring purposes for a person who has dropped out of a trial. This information may consist of health service contacts kept on health care databases such as case notes as well as information from third parties. For trials involving people with severe disorders, third-party information from key workers is nearly always included as part of the measurement of outcome. The lack of data on drop-out may affect the relevance and benefit of the trial results to the wider community. It may therefore be unethical not to collect as much data as possible. The interpretation of legal rights such as the UK Human Rights Act should make the position of researchers clearer, but it is also possible that the idiosyncratic interpretations made by local

Research Ethics Committees will lead to further confusion in this already complicated area.

PLATFORM AND ORDNANCE

A naval military analogy between the vehicle of delivery, the platform (e.g. battleship, frigate, etc.) and what is delivered, the ordnance (e.g. shell, missile, etc.), is helpful in understanding the difference between service organisation and therapy.[42] In terms of this analogy the platform would be aspects of the mental health service, such as assertive outreach, case management and so on; whereas ordnance would consist of different types of therapeutic intervention, such as CBT and family interventions. This distinction is useful in clarifying what is being tested. For example, a trial of different service organisations (platforms) would be the UK 700 trial,[43] in which 708 psychotic patients in four centres were randomly assigned to standard or intensive case management. In this trial the only specific difference between the two trial limbs was the number of patients the case managers had in their case loads. No investigation was made about the therapeutic input that the case managers implemented. The results indicated that there was no advantage in clinical or social outcomes of intensive case management. In contrast there are examples of therapy trials in which a comparison was made between CBT plus routine care, supportive counselling plus routine care, and routine care alone for chronic patients[15] and acute patients.[17] In these trials patients are recruited across a number of sites so that variations in routine care and service delivery are accommodated. It may be questioned whether trials of services are of much value if they do not include effective therapies. A battleship is unlikely to perform well in a naval engagement firing a bow and arrow!

BACKGROUND SERVICES

The background mental health services and their accessibility may affect trials in a number of ways. Recruitment may be affected by what services are already available and who has access to them. For example, recruitment is likely to be different if there is free universally available health care provided by a service committed to research and development. A large proportion of the population will use this service and potentially be available for recruitment and eligible for the trial. This is essentially what happens in the UK National Health Service. This case is very different when health care is provided, funded by reimbursements in a fragmented manner to certain groups of the population by private services who are unlikely to be committed to research. In this case the proportion of the population available for recruitment will be much reduced and biased towards certain subgroups who may, for whatever reasons, have no access to private care. Here recruitment is likely to be highly selective and potentially biased. The provision of different services to different income groups or other population subgroups mitigates against representativeness of trial populations in the USA, Australia and some European countries[44,45] and compromises the value of such trials.

RANDOMISATION

Purpose

To give an equivalent chance of a recruit being in any of the groups in the trial design, some researchers think that one of the purposes is to balance the groups on every factor that may be relevant to the treatment response, but purely random allocation will not provide such matching. If there is strong evidence that a particular factor may affect the outcome then this should be included as a factor in the analysis. In the past researchers have said they have provided evidence of the equivalence of their samples in analyses of pre-treatment group comparisons and on the basis of finding no statistically significant differences on factors pertinent to the treatment response have then not included these factors in their analyses of therapeutic outcome. The current advice is not to carry out such pre-treatment comparisons but to include pertinent factors in the outcome assessment. However, there is a need for a clear description of the people who dropped out of treatment in relation to those who remained as this may bias the interpretation of the results.

For studies of psychological treatments in psychosis the most relevant factors are listed below.

- The *chronicity of the illness*, measured in months or years since first diagnosis, is likely to affect treatment outcome because it is well known in the field of psychiatry that those with longer illnesses may be less likely to change. Tarrier *et al.*[15] found that this modestly predicted treatment outcome in an intensive CBT trial.
- The *duration of untreated psychosis* (DUP) is the time spent experiencing symptoms prior to the diagnosis and treatment of the disorder. Several studies suggest that DUP affects the success of other treatments, particularly medication, and it may be that this is also a factor in the efficacy of psychological disorders.
- The *severity of the symptoms* has been shown to affect treatment outcome.[15,38] In the London–East Anglia trial the best outcomes following CBT were found for those people who said they were not absolutely certain that their delusions were true. The effect of this same factor was also alluded to several years ago in a small trial by Watt *et al.*[46] But, although the outcome at post-treatment was affected by this factor there was no measurable influence at follow-up nine months after the end of therapy.[47]
- *Gender* was investigated by Gould *et al.*[25] in their meta-analysis of CBT trials. They found no relationship between effect size and the proportion of men in the trial but as they themselves point out no data were available for the specific outcomes for men and women that preclude a definitive evaluation. However, young men are usually thought to have a poorer outcome and are more likely to drop out.[41]
- *Intellectual status* has been suggested by critics of psychological therapy to be a bar to significant treatment effects. Although one study has not found this to be true,[38] trialists should consider this factor in their analysis if only to counter such criticisms.
- *Interactions with other treatments* need to be considered where these are variables within the group of patients entered into the trial. The most pertinent for CBT studies is the issue of the use of medication. Medication is now often divided into two main types, typical antipsychotics which have been available for a number of years and atypical antipsychotic medications which have become available recently. Most published CBT studies were carried out before the wide availability of these newer medications. However, medication was not a predictor of outcome in Garety *et al.*[38] or Tarrier *et al.*[15] Kuipers *et al.*[47] comment that in their CBT group, because symptomatic improvement would be achieved, these patients would be less likely to be prescribed clozapine (an atypical antipsychotic) and would generally be prescribed lower doses of medication. These predictions, although only a trend towards significant, were shown in their data. Pinto *et al.*[48] chose their sample on the basis of a failure of at least two medications to reduce positive symptoms. In their study, which was not methodologically strong, the effect size was extremely large with the combined effect of CBT and the new medication (clozapine) producing an effect size of 2.18.[49] This suggests that medication should be taken into account in randomisation or at least in subgroup analyses.

Entry to the study may be stratified if the variable has a known interaction with treatment or the variable can be used as a covariate in the analyses. Being clear about which variables may interact with psychological treatment is essential at the outset of the trial because the trial must be defined and have a sufficiently large sample to be adequately powered to test for these effects.

Details

Details of the process of randomisation must be supplied in the paper. For instance in Kuipers *et al.*[14] a randomised permuted blocks allocation was adopted in each centre which contributed participants to the trial. Other studies with multiple centres (e.g. Sensky *et al.*[16]) randomised participants at each centre; this they then argued allows them to control for within-centre effects and allows them to test between-centre effects of treatment efficacy.

Blindness

In clinical trials blindness usually refers to two aspects of the trial. The first relates to the allocation of participants to the different treatment limbs so that the allocation process is independent and concealed from those involved in the assessment or treatment. This prevents people from choosing who to put into the trial on the basis of the patient's own preference, resulting in more enthusiastic people being in the treatment arm, or the research worker's preference which may result in those with more favourable prognoses being allocated to the experimental treatment.

The second use of the word blindness relates to the concealment of which treatment the participant received from those involved with assessment, especially of outcome. This is an extremely important issue and one that is difficult to ensure. The aim is to prevent any bias, conscious or otherwise, entering the assessment process through knowledge of which treatment the participant received. For example, knowledge that the hypothesis to be tested was that CBT would be better that treatment-as-usual because previous studies had demonstrated this may bias an assessor to rating the patient as more improved if they knew the patient had received CBT.

The importance of adequate concealment was demonstrated in a study by Moher *et al.*[50] who examined the quality of concealment in treatment trials in circulatory and digestive disease, mental health, obstetrics and childbirth. The examined trials had already passed a number of quality assessments and been included in a number of meta-analyses. They found that trials with poorer quality blinding were associated with an increased estimate of benefit of 34%. This replicated a similar earlier finding of Schulz *et al.*[51] who also reported exaggerated treatment efficacy of 30–40% in trials with inadequate concealment. There is good evidence that the poorer the trial methodology, the better, and more inflated, the treatment results obtained.

The assessment and treatment procedures must be separate and independent, in other words the person who carries out the assessment should be different from the person who delivers the treatment. This is not always the case in published trials: for example, Brooker *et al.*[52,53] trained mental health nurses in family intervention and assessed the effectiveness of the intervention by having the nurses perform the assessments. It could be argued that any problem of bias could be avoided in cases such as these by performing the assessments through patient self-report. However, this does not address the problem of social approval that may introduce bias where patients give results they think their therapist would want to receive.

Independent assessors and therapists will not ensure that assessors remain naive to treatment allocation. Accidental knowledge of allocation can be minimised by using separate administrative procedures and geographically separating therapists from assessors in terms of office location and administrative procedure. This should prevent assessors bumping into patients about to receive therapy and such similar accidents. Patient allocation should be multiply coded so that learning of one patient's allocation does not break the whole trial code. Patients should be instructed not to reveal any detail of their treatment or who has treated them to the assessors at the start of the trial and before each assessment. It is unlikely that this will be fool-proof but it will minimise revelations. See Tarrier *et al.*[15] for further details of efforts to maintain blindness in a clinical trial of CBT.

Opinions differ as to whether verification of maintenance of blindness is desirable. It is possible to ask assessors to guess the allocation of trial participants. This can be used as evidence of successful blindness.[15,54] Assessors should not be informed that they will be asked to guess as this would prime them to the task. Guessing is less likely to be successful when there are more than two treatment groups. With two groups, or even more than two, an assessor could adopt a strategy that patients who improved should be in the experimental treatment group because this would be in line with the study hypothesis. If the trial had been successful this strategy would have been correct and the assessor would most likely have guessed right in many cases although for the wrong reasons. This would not be an indication that the assessor knew of the treatment allocation and was hence biased in their assessment but that they knew who improved, which aided them in guessing group allocation. The problem for the trial investigators here would be that their assessors appear not to have been blind. If the assessors were not able to guess correctly using this strategy it would probably mean that the experimental treatment had not been effective and the trial was a failure anyway. Having assessors guess allocation holds the investigators hostage to fortune, although with multiple treatment groups it can be effective in demonstrating blindness.

Even if assessors do maintain blindness to treatment allocation they will still be aware of the timing of the assessment, pre-, post-treatment or follow-up. Thus the only way to ensure blindness of both treatment allocation and assessment time is to separate the gathering of information from its rating. Thus all assessment interviews should be audio-taped independently of their rating and rating should be carried out by a different assessor who is unaware of the allocation or assessment. This would also allow the audio-tapes to be edited of any accidental revelation of identifiers. To be successful interviews need to follow a protocol as to the procedure of the interview so that adequate and sufficient information is available to make ratings. In most studies of CBT in general and for psychosis in particular, the process for blind allocation is rarely described, for example Kuipers *et al.*[14] In contrast, Sensky *et al.*[16] and Tarrier *et al.*[15] both describe the method for ensuring blindness and the maintenance of allocation of subjects to groups.

PROTOCOL

Design Protocol

There are various ways of testing whether a particular treatment is efficacious but the accepted method is to compare the treatment with a placebo control that allows for a comparison of client expectations of improvement during therapy with the active ingredient itself. We have discussed above the importance of these non-specific factors in psychological treatments. Social contact, social support and the modelling of interpersonal behaviour are all an integral part of psychological therapy. Tests of the effectiveness of individual CBT have used a variety of designs. Table 21.2 gives the outline of the main recent trials.

There are a variety of designs that will allow the examination of both the effectiveness and specificity of the effect of CBT above the effects found for psychological interventions in general in this group. The results show that there are significant effects over treatment-as-usual. There is also one study[16] that shows a

Table 21.2. Designs for randomised control trials of CBT for psychosis

Comparison with treatment-as-usual	Comparison with alternative 'placebo' therapy	Comparison with 'placebo' therapy and treatment-as-usual
Garety *et al.*[11] Kuipers *et al.*[14,47] Barrowclough *et al.*[18]	Sensky *et al.*[16] Drury *et al.*[12,13]	Tarrier *et al.*[10] Tarrier *et al.*[15,73] SOCRATES[17]

difference between CBT and a 'placebo' therapy (befriending), but only at follow-up. However, Tarrier *et al.*[15] found a significant difference between CBT and TAU but no overall difference between the two therapies at any stage of the study.[55] Analysis of specific symptoms found that there was a significant advantage of CBT over supportive counselling in the treatment of hallucinations.[56]

Treatment Protocol

It is essential to have a clear and unambiguous treatment protocol for psychological treatments. However, even when a manual is available it is much harder to evaluate exactly whether the protocol has been adhered to. In treatments with medication this process is relatively easy as the dose and timing of the treatment can be verified using simple procedures. For psychological treatment the verification process relies on taped interviews of treatment sessions that are then rated later for fidelity with the treatment protocol. However, there are several problems that may interfere with this process. Firstly the patient must agree to the recording of the session and in some studies, e.g. Chadwick *et al.*,[57] the patients refused to have any sessions taped. Once taped sessions have been collected the independent rating must answer a number of questions:

- Does the session represent the treatment to be provided? In other words is it possible to differentiate the experimental treatment from the placebo treatment? Sensky *et al.*[16] and Tarrier *et al.*[15] were able to show that their independent assessors was able to assign 100% and 97% of the tapes rated to the appropriate treatment arm.
- Is the experimental treatment manual being adhered to? This requires that the researchers have a specific rating scale that will allow the rating of key areas of their treatment. Haddock *et al.*[58] have developed a rating scale (the Cognitive Therapy Scale for Psychosis–CTS-Psy) to assess quality of therapy. This allows assessment of general (e.g. interpersonal effectiveness) and specific (e.g. guided discovery) aspects of therapy to be assessed. However, the scale allocates equal weighting to all items. There is, as yet, no empirical evidence to support such equal weightings, and it may well be, for example, that *'agenda setting'* is less important than the *'choice of intervention'*.
- Does the progression of therapy cover all the key topics of the manual? This requires that several sessions of therapy are recorded at different times and that the content of these is scored for the timing of the interventions in the treatment programme. This is probably the most important part of rating CBT trials because although it can be clear how many sessions are provided to a patient it is not clear whether the content given is the same. CBT researchers (personal communications, including one of the authors, NT) observe that some patients are able to travel through the whole manual whereas others cover much less. So although the therapy duration may be equivalent, exposure to the complete protocol can be different. This dosage of treatment may be an important factor in defining treatment outcome as some patients are clearly getting more treatment than others. However, the number of treatment sessions is not related to effect size as presented in Gould *et al.*[25]

Progression through therapy may be affected by a number of factors such as the level of disturbance or cognitive impairment of the patient. Therapists will also differ in their ability to progress therapy and their skills in different aspects – determined by skills, training, profession, trial provided training (see Tarrier *et al.*[59]). So far in CBT trials these factors have not been investigated in any detail.

INDIVIDUALISED TREATMENTS

Turkington and Siddle[60] claim that a case formulation is essential to treating psychotic patients successfully. We do not disagree that a case formulation is desirable but with psychotic patients it is not always possible and a purely symptomatic approach has to be adopted on occasions.[61] In spite of our support for case formulation the evidence from general adult mental health that a case formulation is necessary is poor and the results equivocal. Schulte *et al.*[62] treated a mixed group of 120 phobic patients with a standardised treatment and individualised behaviour therapy based on functional behaviour analysis. They also included a yoked control group in which treatment was based not on the individual's assessment but that of the yoked patient. The standardised treatment group showed the most improvement and patients who acted as yoked controls improved as well as the other patients. Similarly Emmelkamp *et al.*[63] allocated 22 obsessive–compulsive patients to either tailor-made cognitive behavioural therapy or standardised exposure *in vivo* therapy. There were no significant differences between groups but the group sizes were small ($n = 11$ in each group). Jacobson *et al.*[64] treated 30 distressed marital couples with either a manual-based version of marital therapy or a clinically flexible version of the same treatment in which treatment plans were individually based and the number of treatment sessions were not specified. Both treatments resulted in significant improvements at post-treatment but at six-month follow-up the couples treated with the structured format were more likely to have deteriorated and flexibly treated couples were more likely to have maintained their treatment gains. There appears little advantage of case formulation-based treatment over a standard package. This result is not surprising given the sample sizes of these studies. Standard treatment programmes are effective for a wide range of psychological disorders and even if an individualised treatment was superior the difference in effect sizes will most probably be small and the sample size to significantly demonstrate such a difference would necessarily be large. Therefore, the studies that have been done are massively underpowered. To substantiate this Tarrier and Calam[65] have estimated the sample sizes required to show significant differences with 80% power and 0.05 significance level based on the data provided in the published report of Emmelkamp *et al.*[63] On the basis of their data the numbers in each group required to show a significant difference for the five outcome measures would be between 25 and in excess of 15 000, with a median of 800 patients in each treatment group. The issue of case formulation-based treatment versus protocol-based treatment is unlikely to be resolved by a direct head-to-head comparison, which would be too large and costly.

TREATMENT COMPONENTS

CBT treatments also differ in other ways from each other. Although they have a basic set of ingredients the emphasis may be placed differently. For example, the different emphases on behavioural activation and cognitive schema in the changes in thinking thought to be the cause of the treatment effect. Changing behaviour can have an effect on thinking as studies of CBT for panic disorder have discovered.[22] Patients in the Clark study were treated with behavioural activation programmes that are embedded in CBT. They showed that the prediction of outcome was dependent on one main factor, cognitions about their bodily sensations. The behavioural experiments seemed to have an effect on cognition. But, other groups in the field of psychosis emphasise more distal stimuli such as the developmental path of the delusion. The particular component of CBT that accounts for most of the variance in outcome has not yet been differentiated and these more subtle differences are not used in meta-analyses of the treatment studies. Figure 21.4 shows the effect sizes taken from Gould *et al.*[25] on a scale devised by ourselves on the amount of behavioural activation that the treatment emphasises. As can be seen from the graph the effect size is increased when more behavioural activation is included. It may be that a simple change in behaviour via a behavioural experiment may provide enough evidence to reduce delusional conviction. For instance Birchwood and Chadwick[66] suggest that the perceived powerfulness of an auditory hallucination directly predicts the distress experienced. Adopting one successful coping strategy may provide enough evidence to reduce the perceived power of an auditory hallucination and increase the amount of perceived control the patient has over their symptoms. Wykes *et al.*[67] provide some evidence of this relationship in a waiting list control trial of group CBT for auditory hallucinations. If this is true then successful behavioural experiments should always be included and should predict successful treatment outcome. As yet no study has attempted to measure this process.

Turkington and Siddle[60] maintain that cognitive therapy with psychotic populations will result in long-term improvement because it involves schema change whereas cognitive behaviour therapy (as carried out by Tarrier *et al.*[15]) will result in short-term change only. They go on to say that 'schema change seems vital in terms of the durability of any achieved benefits' (p. 302). However, there is little evidence that schema are causative in psychosis or that change is important for treatment effects (see Figure 21.4). The direct transport of Beck's model of depression to psychosis in the absence of evidence for its explanatory value in this population has been criticised.[68] The evidence for the effect of schema work on outcome is anyway sparse even in the area of depression for which it was designed. Jacobson *et al.*[75] randomised 150 people to three treatment arms: (i) behavioural activation, (ii) behavioural activation and work on dysfunctional thoughts, and (iii) total CBT with work on cognitive schema. The results of this trial showed no differences in outcome either at post-treatment or follow-up between the three groups. The effect of the other components of CBT was no different to the effects of behavioural activation alone (see Figure 21.5).

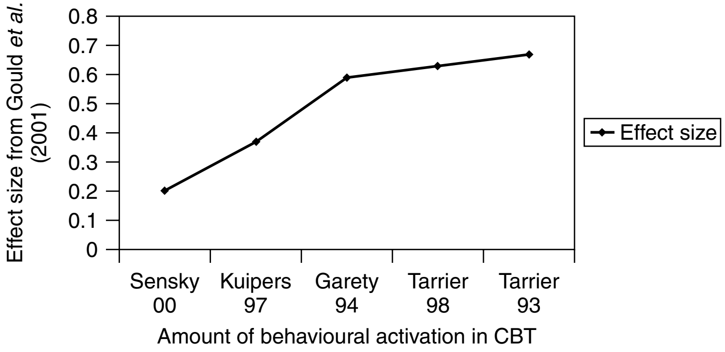
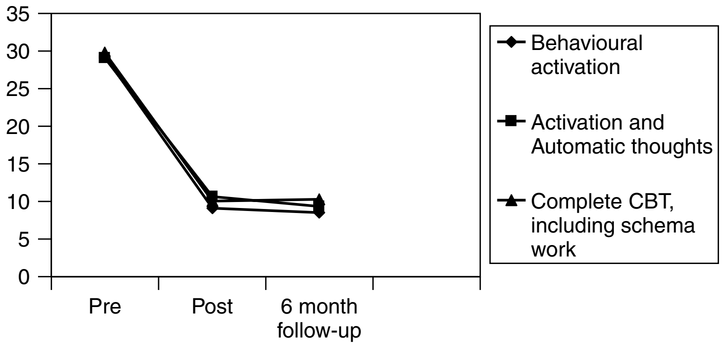


Figure 21.4. How much 'B' in CBT for psychosis.



Source: Data from Jacobsen *et al.*[75]

Figure 21.5. Evidence of the effect of schema work in depression.

OUTCOMES

Current thinking from methodologists on trials in psychiatry is that designs should be simplified and that outcome measures should be kept to a minimum. In particular the use of rating scales should be restricted to 'one or two which are best understood' (Johnson,[69] p. 229). Follow-up should be carried out on 'few occasions rather than many' and entry criteria should be as broad as possible.[69] This advice is rather in conflict with that given previously which suggests that, in the treatment of psychosis, multiple outcomes which reflect the complex nature of the disorder and its effects should be used[70] and that data on the process of therapy are essential. The multiple effects of therapy may, but not necessarily, have a common outcome in relapse prevention or total symptoms. For instance cognitive therapy should affect cognition, CBT should affect cognitions and behaviour, and family interventions should affect families' interactions. All these effects could in some way change the outcome of the disorder but via multiple pathways. Because of these multiple effects, multiple outcomes may be recommended in order to differentiate the route to effectiveness in explanatory trials.

All current trials measure overall symptoms. However, they all do so in different ways and even when the measure appears to be a standard measure (e.g. BPRS[71]) it is often adapted. For instance Kuipers *et al.*[14] added items to the BPRS making it difficult to compare their results with others adopting the conventional version of the same instrument. The use of non-standard instruments to measure awareness of stigma, coping skills, etc., prevents comparisons being made and does need replication with standardised rating scales with no known psychometric properties.

In all medical trials a statistically significant difference in outcome may provide little benefit to the patients. What needs to be defined for trials of CBT for psychosis is the clinical significance of outcomes. This was alluded to earlier in this chapter. Clinically significant outcomes may be reductions in the distress associated with the disorder. Currently clinical significance is defined as the sorts of improvements that are achieved in drug trials – 20% change in symptoms. This may be a low threshold for what could be achieved through psychological therapy. Many trials of psychological therapy adopt only a statistically significant test of effectiveness, but Tarrier *et al.*[15] and Sensky *et al.*[16] adopt a 50% change criterion for their measures, although Kuipers *et al.*[14] use the lower threshold of 20%. Where trials use such a threshold of achieving clinical significance or not, comparisons can be made by comparing the Number Needed to Treat (NNT), which represents the number of patients that need to be treated to achieve one clinically significant outcome.[72]

CONCLUSIONS

Because psychological therapy has no product champion as found in the drug industry, pragmatic trials are needed initially to convince people that therapy is worthwhile. However, these need to be followed by explanatory trials that can establish the specificity of the treatment. Currently the trials in the field of psychosis have mainly been pragmatic and these have shown that the therapy is worthwhile with improvements in positive and negative symptoms at post-treatment and follow-up in some studies. However, the trials that have been designed to test the specificity of treatment have not been so successful. Very few differences emerge between CBT for psychosis and alternative therapies are shown at post-treatment. Of the two studies with long follow-ups one showed an advantage for CBT over the alternative therapy[16] but the other showed equivalent benefits of both therapies (CBT and supportive counselling) over routine care.[55,73] One resolution of this conflict may be the nature of the therapy chosen. In the Sensky *et al.*[16] study the comparison condition was befriending which is described as a therapy with equivalent amounts of therapist contact but where the content was a discussion focusing on neutral topics such as hobbies, sports and current affairs where the therapist is instructed to be empathic but non-directive. Even this therapy was associated with reductions in symptoms at the end of therapy that may be a testament to the paucity of the social contact and lack of warm relationships of people with continuing active psychosis. However, the effects of this therapy were not sustained to a follow-up where the patients in the comparison therapy condition actually got worse. In the Tarrier *et al.*[15] study the comparison therapy chosen was supportive counselling in which the therapist tried to achieve a supportive relationship which fostered rapport and provided emotional support. This therapy proved to be successful in reducing symptoms over the course of the therapy and follow-up. This result suggests that some of the essential ingredients of CBT encompassed within the counselling framework may be either shared with supportive counselling or are as effective as these other ingredients. It is of course possible that within the model of schizophrenia which encompasses a vulnerability stress model the two forms of therapy may work via different pathways. Supportive counselling may work by emphasising self-esteem through rapport within the therapeutic relationship. Unlike befriending this support produces more stable changes that are durable to follow-up. However, it should be noted that supportive counselling did significantly worse at treating hallucinations when compared to CBT on this symptom alone.[56]

Currently there is no evidence that cognitive behaviour therapy works via a cognitive system, although training in coping skills has been shown to improve coping.[74] None of the studies have yet produced analyses showing that the cognitive change established during therapy is the key to later improvement. In fact few have even provided analyses which test this possibility. Because of the complex nature of the aetiology of psychosis with its multiple causal processes it may be impossible to identify a single route to change. There may be many idiosyncratic routes that will only be established in large trials with several hundred patients included. One such trial is taking place in Manchester, the Socrates trial.[17] In this trial hundreds of acute patients who are at the beginning of their illness have been provided with therapy and followed up over a long period. It is only through these sorts of studies that it will be possible to establish routes to change that can then inform the development of therapy. It is only by these later developments that it will be possible to develop training and therefore provide larger numbers of people with psychosis with effective psychological therapy.

REFERENCES

1. Wiersma D, Nienhuis FJ, Sloof CJ, Geil R. Natural course of schizophrenic disorders: a 15 year follow-up of a Dutch incidence cohort. *Schizophr Bull* (1998) **24**: 75–85.
2. British Psychological Society. *Recent Advances in Understanding Mental Illness and Psychotic Experiences*. Leicester: British Psychological Society (2000).
3. Fowler D, Garety P, Kuipers E. Cognitive therapy for psychosis: formulation, treatment, effects and service implications. *J Mental Health* (1998) **7**: 123–33.
4. Tarrier N, Barrowclough C, Vaughn C, Bamrah JS, Porceddu K, Watts S, Freeman HL. The community management of schizophrenia: a controlled trial of a behavioural intervention with families. *Br J Psychiat* (1988) **153**: 532–42.
5. Cunningham-Owens DG, Carroll A, Fattah S, Clyde Z, Coffey I, Johnstone EC. A randomised, controlled trial of a brief educational package for schizophrenic outpatients. *Acta Psychiat Scand* (2001) **103**: 362–9.
6. Beck AT. Successful outpatient psychotherapy with a schizophrenic with delusion based on borrowed guilt. *Psychiatry* (1952) **15**: 305–12.
7. Tarrier N. Psychological treatment of schizophrenic symptoms. In: Kavanagh D, ed., *Schizophrenia: An Overview and Practical Handbook*. London: Chapman & Hall (1992).
8. Tarrier N. Modification and management of residual psychotic symptoms. In: Birchwood M, Tarrier N, eds, *Psychological Approaches to Schizophrenia: Innovations in Assessment, Treatment and Services*. Chichester: John Wiley & Sons, Ltd (1992).
9. Haddock G, Tarrier N, Spaulding W, Yusupoff L, Kinney C, McCarthy E. Individual cognitive-behaviour therapy in the treatment of schizophrenia: a review. *Clin Psychol Rev* (1998) **18**: 821–38.
10. Tarrier N, Beckett R, Harwood S, Baker A, Yusupoff L, Ugarteburu I. A trial of two cognitive behavioural methods of treating drug-resistant

residual psychotic symptoms in schizophrenic patients. I: Outcome. *Br J Psychiat* (1993) **162**: 524–32.

11. Garety P, Kuipers E, Fowler D, Chamberlain F, Dunn G. Cognitive-behaviour therapy for drug resistant psychosis. *Br J Med Psychol* (1994) **67**: 259–71.
12. Drury V, Birchwood M, Cochrane R, Macmillan F. Cognitive therapy and recovery from acute psychosis. I. Impact on psychotic symptoms. *Br J Psychiat* (1996) **169**: 593–601.
13. Drury V, Birchwood M, Cochrane R, Macmillan F. Cognitive therapy and recovery from acute psychosis. II. Impact on recovery time. *Br J Psychiat* (1996) **169**: 602–7.
14. Kuipers E, Garety P, Fowler D, Dunn G, Bebbington P, Freeman D, Hadley C. London–East Anglia randomised controlled trial of cognitive-behavioural therapy for psychosis: 1. Effects of the treatment phase. *Br J Psychiat* (1997) **171**: 319–27.
15. Tarrier N, Yusupoff L, Kinney C, McCarthy E, Gledhill A, Haddock G, Morris J. A randomised controlled trial of intensive cognitive behaviour therapy for patients with chronic schizophrenia. *Br Med J* (1998) **317**: 303–7.
16. Sensky T, Turkington D, Kingdon D, Scott JL, Scott J, Siddle R, O'Carroll M, Barnes TRE. A randomised controlled trial of cognitive-behavioral therapy for persistent symptoms in schizophrenia resistant to medication. *Arch Gen Psychiat* (2000) **57**: 165–72.
17. Lewis SW, Tarrier N, Haddock G, Bentall R, Kinderman P, Kingdon D, Siddle R, Drake R, Everitt J, Leadley K, Benn A, Glazebrook K, Haley C, Akhtar S, Davies L, Palmer S, Faragher B, Dunn G. Randomised controlled trial of cognitive-behaviour therapy in early schizophrenia: acute phase outcomes. *Br J Psychiat* (2002) **181**(Suppl 43): 91–7.
18. Barrowclough C, Haddock G, Tarrier N, Lewis S, Moring J, O'Brian R, Schofield N, McGovern J. Randomised controlled trial of motivational interviewing and cognitive behavioural intervention for schizophrenia patients with associated drug or alcohol misuse. *Am J Psychiat* (2001) **158**: 1706–13.
19. Barrowclough C, Tarrier N, Sellwood W, Quinn J, Mainwaring J, Lewis S. A randomised controlled effectiveness trial of a needs based psychosocial intervention service for carers of schizophrenic patients. *Br J Psychiat* (1999) **174**: 506–11.
20. Horvath A, Symonds D. Relationship between working alliance and outcomes in psychotherapy: a meta-analysis. *J Counsel Psychol* (1991) **38**: 139–49.
21. Wampold BE, Mondin GW, Moody M, Stich F, Benson K, Ahn H. A meta-analysis of outcome studies comparing bona fide psychotherapies: empirically, "all must have prizes". *Psychol Bull* (1997) **122**: 203–15.
22. Clark DM, Salkovskis PM, Hackmann A, Middleton H, Anastasiades P, Gelder M. A comparison of cognitive therapy, applied relaxation and imipramine in the treatment of panic disorder. *Br J Psychiat* (1994) **164**: 759–69.
23. Paykel ES, Scott J, Teasdale JD, Johnson AL, Garland A, Moore R, Jenaway A, Cornwall PL, Hayhurst H, Abbott R, Pope M. Prevention of relapse in residual depression by cognitive therapy. *Arch Gen Psychiat* (1999) **56**: 829–35.
24. Yammey G. Psychologists question debriefing for traumatised employees. *Br Med J* (2000) **320**: 140.
25. Gould RA, Mueser KT, Bolton E, Mays V, Goff D. Cognitive therapy for psychosis in schizophrenia: an effect size analysis. *Schizophr Res* (2001) **48**: 335–42.
26. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum (1988).
27. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simelk D, Stroup DF. Improving the quality of reporting of randomized controlled trials: The CONSORT Statement. *J Am Med Assoc* (1996) **276**: 637–9.
28. Moher D, Schulz K, Altman D. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel group randomisation. *J Am Med Assoc* (2001) **285**: 1987–91.
29. Geddes J, Reynolds S, Streiner D, Szatmari P, Haynes B. Evidence-based practice in mental health. *Evidence-Based Mental Health* (1998) **1**: 4–5.
30. Barlow DH. Health care policy, psychotherapy research and the future of psychotherapy. *Am Psychol* (1996) **51**: 1050–8.
31. Drake RE, Goldman HH, Leff HS, Lehman AF, Dixon L, Mueser KT, Torrey WC. Implementing evidence-based practices in routine mental service settings. *Psychiat Serv* (2001) **52**: 179–82.
32. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *Br Med J* (1996) **313**: 36–9.
33. Klein DF. Preventing hung juries about therapy studies. *J Consult Clin Psychol* (1996) **64**: 81–7.
34. Klein DF. Discussion of 'methodological controversies in the treatment of panic disorders'. *Behav Res Ther* (1996) **34**: 849–53.
35. Bentall RP, Jackson HF, Pilgrim D. Abandoning the concept of schizophrenia: some implications of validity arguments for psychological research into

psychotic phenomena. *Br J Clin Psychol* (1989) **27**: 303–24.

36. Wykes T, Parr A-M, Landau S. Group treatment for auditory hallucinations – a waiting list controlled study. *Br J Psychiat* (1999) **175**: 180–5.
37. Lam D, Bright J, Jones S, Hayward P, Schuck N, Chisholm D, Sham P. Cognitive therapy for bipolar illness – a pilot study of relapse prevention. *Cognit Ther Res* (2000) **24**: 503–20.
38. Garety P, Fowler D, Kuipers E, Freeman D, Dunn G, Bebbington P, Hadley C, Jones S. London–East Anglia randomised controlled trial of cognitive-behavioural therapy for psychosis. II Predictors of outcome. *Br J Psychiat* (1997) **171**: 319–27.
39. Everitt BS. The analysis of longitudinal data: beyond MANOVA. *Br J Psychiat* (1998) **172**: 7–10.
40. Department of Health. *Research Governance Framework for Health and Social Care*. Department of Health, UK (2001).
41. Tarrier N, Yusupoff L, McCarthy E, Kinney C, Wittkowski A. Some reason why patients suffering from chronic schizophrenia fail to continue in psychological treatment. *Behav Cognit Psychother* (1998) **26**: 177–81.
42. Marshall M. A systematic review of the effectiveness of acute day hospitals versus admission. Paper presented at The Making Mental Health Services Effective: Now and Tomorrow Conference, 7–8 September 2000, Manchester, UK (2000).
43. Burns T, Creed F, Fahy T, Thompson S, Tyrer P, White I, & the UK 700 Group. Intensive versus standard case management for severe psychotic illness: a randomised trial. *Lancet* (1999) **353**: 2185–9.
44. Wykes T, Tarrier N, Lewis S. *Outcome and Innovation in the Psychological Treatment of Schizophrenia*. Chichester: John Wiley & Sons, Ltd (1998).
45. Tarrier N. What can be learned from clinical trials? Reply to Devilly and Foa (2001). *J Consult Clin Psychol* (2001) **69**: 117–18.
46. Watt F, Powell G, Austin S. Modification of delusional beliefs. *Br J Med Psychol* (1983) **46**: 359–63.
47. Kuipers E, Fowler D, Garety P, Chisholm D, Freeman D, Dunn G, Bebbington P, Hadley C. London–East Anglia randomised controlled trial of cognitive-behavioural therapy for psychosis. III Follow-up and economic evaluation at 18 months. *Br J Psychiat* (1998) **171**: 319–27.
48. Pinto A, La Pia S, Mennella R, Giorgio D, DeSimone L. Cognitive-behavioral therapy and clozapine for clients with treatment-refractory schizophrenia. *Psychiat Serv* (1999) **50**: 901–4.
49. Rector NA, Beck AT. Cognitive-behavioral therapy for schizophrenia: an empirical review. *J Nerv Ment Dis* (2001) **189**: 278–87.
50. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* (1998) **352**: 609–13.
51. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effect in controlled trials. *J Am Med Assoc* (1995) **280**: 178–80.
52. Brooker C, Tarrier N, Barrowclough C, Butterworth A, Goldberg D. Training community psychiatric nurses for psychosocial intervention: report of a pilot study. *Br J Psychiat* (1992) **160**: 836–44.
53. Brooker C, Falloon I, Butterworth A, Goldberg D, Graham-Hole V, Hillier V. The outcome of training community psychiatric nurses to deliver psychosocial intervention. *Br J Psychiat* (1994) **165**: 222–30.
54. Basoglu M, Marks I, Livanou M, Swinson R. Double-blindness procedure, rater blindness, and ratings of outcome. *Arch Gen Psychiat* (1997) **54**: 744–8.
55. Tarrier N, Kinney C, McCarthy E, Humphreys L, Wittkowski A, Morris J. Two year follow-up of cognitive-behaviour therapy and supportive counselling in the treatment of persistent positive symptoms in chronic schizophrenia. *J Consult Clin Psychol* (2000) **68**: 917–22.
56. Tarrier N, Kinney C, McCarthy E, Wittkowski A, Yusupoff L, Gledhill A, Morris J. Are some types of psychotic symptoms more responsive to CBT? *Behav Cognit Psychother* (2001) **29**: 45–55.
57. Chadwick P, Sambrooke S, Rasch S, Davies E. Challenging the omnipotence of voices: group cognitive behavior therapy for voices. *Behav Res Ther* (2000) **38**: 993–1003.
58. Haddock G, Devane S, Bradshaw T, McGovern J, Tarrier N, Kinderman P, Baguley I, Lancashire S, Harris N. An investigation into the psychometric properties of the Cognitive Therapy scale for psychosis (CTS-Psy). *Behav Cognit Psychother* (2001) **29**: 93–106.
59. Tarrier N, Barrowclough C, Haddock G, McGovern J. The dissemination of innovative cognitive-behavioural treatments for schizophrenia. *J Mental Health* (1999) **8**: 569–82.
60. Turkington D, Siddle R. Improving understanding and coping in people with schizophrenia by changing attitudes. *Psychiat Rehab Skills* (2000) **4**: 300–20.
61. Tarrier N, Haddock G. Cognitive-behaviour therapy for the treatment of schizophrenia: a case

formulation approach. In: Hofman SG, Thompson MC, eds, *Handbook of Psychological Treatments for Severe Mental Disorders*. New York: Guilford Press (2001).

62. Schulte D, Kunzel R, Pepping G, Schulte-Bahrenberg T. Tailor-made versus standardised therapy of phobic patients. *Adv Behav Res Ther* (1992) **14**: 67–92.
63. Emmelkamp PMG, Bouman TK, Blaauw E. Individualised versus standardised therapy: a comparative evaluation with obsessive–compulsive patients. *Clin Psychol Psychother* (1994) **1**: 95–100.
64. Jacobson NS, Schmaling KB, Holtzworth-Munroe A, Katt JL, Wood LF, Follette VM. Research-structured vs clinically flexible versions of social learning-based marital therapy. *Behav Res Ther* (1989) **27**: 173–80.
65. Tarrier N, Calam R. New developments in cognitive-behavioural case formulation. Epidemiological, systemic and social context: an integrative approach. *Cognit Behav Psychother* (2002) **30**: 311–28.
66. Birchwood M, Chadwick P. The omnipotence of voices: testing the validity of cognitive model. *Psychol Med* (1997) **21**: 1345–53.
67. Wykes T, Reeder C, Corner J, Williams C, Everitt B. The effects of neurocognitive remediation on executive processing in patients with schizophrenia. *Schizophr Bull* (1999) **25**: 291–308.
68. Patience DA. Cognitive-behaviour therapy for schizophrenia. *Br J Psychiat* (1994) **165**: 266–7.
69. Johnson T. Clinical trials in psychiatry: background and statistical perspective. *Stat Methods Med Res* (1998) **7**: 209–34.
70. Barrowclough C, Tarrier N. Psychosocial interventions with families and their effects on the course of schizophrenia: a review. *Psychol Med* (1984) **14**: 629–42.
71. Ventura J, Green M, Shaner A, Liberman R. Training and quality assurance with the Brief Psychiatric Rating Scale: the drift busters. *Int J Methods Psychiat Res* (1993) **3**: 221–44.
72. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine*. London: Churchill Livingstone (1998).
73. Tarrier N, Wittkowski A, Kinney C, McCarthy E, Morris J, Humphreys L. The durability of the effects of cognitive behaviour therapy in the treatment of chronic schizophrenia: twelve months follow-up. *Br J Psychiat* (1999) **174**: 500–4.
74. Tarrier N, Sharpe L, Beckett R, Harwood S, Baker A, Yusupoff L. A trial of two cognitive behavioural methods of treating drug-resistant residual psychotic symptoms in schizophrenic patients. II: Treatment-specific changes in coping and problem-solving skills. *Social Psychiat Psychiat Epidemiol* (1993) **28**: 5–10.
75. Jacobson NS, Dobson KS, Truax PA, Addis ME, Koerner K, Gollan JK, Gortner E, Prince SE. A component analysis of cognitive – behaviour treatment for depression. *J Consult Clin Psychol* (1996) **64**: 295–304.

22

Psychotherapy for Depression

GRAHAM DUNN
Biostatistics Group, Division of Epidemiology & Health Sciences, University of Manchester, Manchester, UK

INTRODUCTION

The purpose of the present chapter is to explore the pitfalls in and challenges to the valid estimation of the effects of psychotherapies. Although 'depression' appears in the title, the discussion will be relevant to psychological treatments for any mental illness (including psychotic disorders such as bipolar depression and schizophrenia). Most of the illustrative examples, however, will refer to the treatment of depression. Right from the start, despite pointing out all of its potential problems, I will assume that by the far the best way of trying to estimate treatment effects is via the use of a *randomised controlled trial* (RCT). I have little sympathy with the increasingly popular view that we can learn much of real value about treatment effects from systematically collected outcome data in routine clinical practice (see Dunn[1] for a critique of this view). Nor do I have any sympathy with the often-heard view that the RCT and the use of statistical methods are inappropriate vehicles for the evaluation of something as complex as psychotherapy. As we have written elsewhere: "Clinicians who claim that statistical methods are inappropriate for the evaluation of psychotherapies because they are limited to analysing means, or do not account for individual differences, are simply revealing their ignorance of statistics and of recent developments in statistical methodology".[2]

A belief in the fundamental role of randomisation, however, does *not* imply that the naive implementation of the RCT in outcome research cannot lead to some invalid or unsafe conclusions. The design of the trial and statistical analysis of the results have to be appropriate to the setting. Psychotherapy involves complex interactions between patient and therapist and sometimes (as in group therapy) involves the interaction of a group of patients with each other as well as with their therapist. It is not as simple as taking a tablet! A psychotherapy trial is likely to be far more complex than most drug trials, both in its implementation and in the analysis and interpretation of the subsequent results. There are also far more opportunities for invalid inferences concerning treatment effects.

First, and often primarily, we are concerned with *internal validity*: the valid estimation of a causal effect of treatment from the data actual collected (given set of patients, therapists, treatment centres, and psychotherapy actually delivered). Are the group differences we see the causal

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

effects of treatment? Or can they be explained by other factors? Later we may be concerned with *external validity*: the generalisation of the inferred causal effects to other patients, therapists, treatment centres and, perhaps, other forms of psychotherapy. To help clarify the discussion I have made much use of Rubin's counterfactual model of causality and its use in the estimation of treatment effects.[3,4] The kernel of the present discussion, applied to the problems of patient choice and non-adherence to treatment, can be found in Dunn.[5]

THE CAUSAL EFFECT OF TREATMENT

Mr Smith has suffered from severe depression, on and off, for several years. Six months ago his family doctor advised him to undergo a course of psychotherapy. He accepted this advice, has had several what he thinks were very helpful sessions with the psychotherapist, and now is feeling considerably better. Let's assume, for the sake of argument, that he has a total score of 10 points on the Beck Depression Inventory (BDI),[6] having started with a score of 20 six months ago. What is the effect of psychotherapy? How do we measure this effect? Putting it another way, what proportion, if any, of the drop from 20 to 10 points might be attributed to the receipt of therapy? Or perhaps I should have written "What proportion might *validly* be attributed to the receipt of therapy?" But, before we attempt to answer this question, let us consider another patient, Mrs Jones, who also suffers from chronic depression. Like Mr Smith, she was advised to have psychotherapy by her family doctor six months ago but, for various reasons, she never managed to keep any of her appointments and has not received any help from the therapist. A third patient, Mr Adams, refused outright to have anything to do with the therapist. Mrs Jones' present BDI score is 12 and that for Mr Adams is 15. For Mrs Jones and Mr Adams one might ask what would have been the effect of therapy offered if they had actually received it? What might their BDI scores have been?

In the above paragraph we are trying to find a way to estimate what may be called the *causal effect* of a treatment. The essence of the solution to the problem is a comparison. For each of the three patients, Mr Smith, Mrs Jones and Mr Adams, there are two possible outcomes of the referral to see a psychotherapist. The first is to receive therapy and have the severity of depression measured after a given interval after the onset of the course of treatment. The second is to fail to get the offered help, but again have the severity of depression measured after the allotted time. Let the variable T represent treatment received. It has two possible values, $T = \mathrm{t}$ (therapy) and $T = \mathrm{c}$ (no therapy). I use 'c' for no therapy to indicate that it can be regarded as a control condition. Let i indicate the identity of the patient ($i = 1$ for Mr Smith, 2 for Mrs Jones and 3 for Mr Adams). Finally, let $Y_{\mathrm{T}}(i)$ indicate the final BDI score for patient i after receiving treatment option T. There are two *potential* outcomes for each of the three patients as indicated in the following table:

Patient	BDI with therapy	BDI without therapy
Mr Smith	$Y_{\mathrm{t}}(1)$	$Y_{\mathrm{c}}(1)$
Mrs Jones	$Y_{\mathrm{t}}(2)$	$Y_{\mathrm{c}}(2)$
Mr Adams	$Y_{\mathrm{t}}(3)$	$Y_{\mathrm{c}}(3)$

We define the causal effect of the receipt of therapy as the difference between the BDI score for the ith patient after therapy and the corresponding BDI score after receiving no therapy. That is, by the difference $Y_{\mathrm{t}}(i) - Y_{\mathrm{c}}(i)$. It is a random variable that varies from one patient to another. Unfortunately, it can never be observed. The obvious problem is that each patient receives one of the treatment conditions, or the other, but not both. Either the patient receives psychotherapy or he or she does not. That is, the ith patient provides a value for either $Y_{\mathrm{t}}(i)$ or $Y_{\mathrm{c}}(i)$, but not both. Mr Smith provides $Y_{\mathrm{t}}(1)$ but not $Y_{\mathrm{c}}(1)$, Mrs Jones provides $Y_{\mathrm{c}}(2)$ but not $Y_{\mathrm{t}}(2)$, and so on. We provide a statistical solution to this problem in the following section but we will re-emphasise the point that the causal effect of psychotherapy is the comparison of the outcome actually observed with that which would have been observed if, *contrary to fact*, the other treatment option had been taken.

Similar arguments apply to the comparison of the effects of different types of psychotherapy, or to the comparison of a specific type of psychotherapy with, for example, a psychopharmacological intervention such as a tri-cyclic antidepressant. The essence is always to try to get an estimate of the difference between the patient's observed response with that which would have been observed if the patient had received the alternative treatment.

COMPARISON OF GROUP AVERAGES AND THE ROLE OF RANDOMISATION

Now let's assume that we have access to a large population of eligible patients – the target population about which we wish to draw causal inferences about the value of psychotherapy or counselling. And let us concentrate on the *average causal effect* (ACE)[3,4] of the therapy for this target population. The average for the population is called an expected value in statistics and the ACE can therefore be written as

$$\mathrm{ACE} = \mathrm{E}[Y_\mathrm{t}(i) - Y_\mathrm{c}(i)] \qquad (22.1)$$

where the expectation E[] is over all values of i. From the mathematical properties of expectations (averages) it follows that

$$\mathrm{ACE} = \mathrm{E}[Y_\mathrm{t}(i)] - E[Y_\mathrm{c}(i)] \qquad (22.2)$$

This simple formula shows us that information on *different* patients can be used to estimate $\mathrm{E}[Y_\mathrm{t}(i)]$ and $\mathrm{E}[Y_\mathrm{c}(i)]$ separately and the difference between these two expectations (averages) can be used to estimate the average of the differences (i.e. the ACE). We *can* observe the $Y_\mathrm{t}(i)$ in patients receiving therapy and we can also observe the $Y_\mathrm{c}(i)$ for those in the control condition. All that we need is to be sure that the observed averages for the treated (therapy) and untreated (control) patients are unbiased estimates of $\mathrm{E}[Y_\mathrm{t}(i)]$ and $\mathrm{E}[Y_\mathrm{c}(i)]$, respectively. In general, however, the average of the $Y_\mathrm{t}(i)$ for the *whole* of the population (i.e. all possible i) is *not* the same as the average of the $Y_\mathrm{t}(i)$ for those patients who have happened to receive the treatment (psychotherapy).

Expressed mathematically,

$$\mathrm{E}[Y_\mathrm{t}(i)] \neq \mathrm{E}[Y_\mathrm{t}(i)|T = \mathrm{t}] \neq \mathrm{E}[Y_\mathrm{t}(i)|T = \mathrm{c}] \qquad (22.3)$$

and

$$E[Y_\mathrm{c}(i)] \neq \mathrm{E}[Y_\mathrm{c}(i)|T = t] \neq \mathrm{E}[Y_\mathrm{c}(i)|T = \mathrm{c}] \qquad (22.4)$$

where '|' means 'given that'. To summarise, the ACE, is defined by the difference between $\mathrm{E}[Y_\mathrm{t}(i)]$ and $\mathrm{E}[Y_\mathrm{c}(i)]$ but what we actually observe are the estimators of $\mathrm{E}[Y_\mathrm{t}(i)|T = \mathrm{t}]$ and $\mathrm{E}[Y_\mathrm{c}(i)|T = \mathrm{c}]$. How do we ensure that our observed averages are also valid estimators of $\mathrm{E}[Y_\mathrm{t}(i)]$ and $\mathrm{E}[Y_\mathrm{c}(i)]$? If we are able to do this then we have replaced an impossible-to-observe causal effect on an individual patient with a possible-to-estimate average of the causal effects for our target population.[4]

If both $\mathrm{E}[Y_\mathrm{t}(i)] = \mathrm{E}[Y_\mathrm{t}(i)|T = \mathrm{t}]$ and $\mathrm{E}[Y_\mathrm{c}(i)] = \mathrm{E}[Y_\mathrm{c}(i)|T = \mathrm{c}]$ then the potential outcomes (both the $Y_\mathrm{t}(i)$ and the $Y_\mathrm{c}(i)$) are statistically independent of the mechanism of assigning (or choosing) treatment options. Otherwise, they are not. If either the patient's family doctor or the patient himself, or both, were to decide which treatment option to choose then it is almost certain that this choice will not be statistically independent of the of the *potential* outcome. This is the familiar problem of *confounding*. The difference in *observed* outcomes may arise from the fact that the patients with the best (or worst) prognosis, on average, might be the ones that opt for therapy. The observed outcomes in this situation might tell us something about the selection mechanism (treatment choice) but are not very informative about the causal effect of therapy. Knowing the values of all of the prognostic variables, together with a little knowledge of experimental design, might lead us to match or stratify the patients prior to estimation of the treatment effects. But we cannot guarantee that we are aware of all possible confounders. There is always the possibility that we have not thought of, or forgotten, something that is vitally important. Although we may be able to convince ourselves that we have not missed an important confounder the only way we can ensure that we can convince a sceptical reviewer is to allocate treatment options *randomly*. Random allocation ensures that both $\mathrm{E}[Y_\mathrm{t}(i)] = \mathrm{E}[Y_\mathrm{t}(i)|T = \mathrm{t}]$ and $\mathrm{E}[Y_\mathrm{c}(i)] = \mathrm{E}[Y_\mathrm{c}(i)|T = \mathrm{c}]$ providing that t and c are the *allocated* treatments (not, necessarily, those actually received). Randomisation is the only sure way of coping with all confounders, and it copes with them irrespective of whether we are aware of them or not. Randomisation does not guarantee that treatment groups will be exactly comparable in any given comparison, but it does ensure that on average there will be comparability. Our conclusion is that if we wish to be sure that we are estimating the desired ACE we need an RCT.

An essential corollary of randomisation is that we obtain outcome data on all of the randomised patients and that we calculate our group averages from the patients as they were randomised and *not* according to whether they actually received or adhered to the treatment option that they were allocated to. This is the *intention-to-treat* (ITT) principle (see, for example, Sheiner and Rubin).[7] If we do not use ITT then the fundamental assumptions concerning our estimates of the causal effect of treatment (ACE) no longer hold. Loss to follow-up (i.e. a failure of the patient to provide outcome data) is a major threat to all RCTs, but in this part of the discussion we will simplify matters by assuming that outcome has, indeed, been obtained for all patients entering the trial. But what if some of the patients choose a treatment option other than the one they were randomly assigned to? Or perhaps some patients adhere to the allocated treatment much less than others – they turn up to the occasional session of therapy, for example, but not all of those which had been planned. This will clearly dilute (attenuate) the effect we wish to estimate. In fact, our ACE estimator (the difference between the observed mean outcomes for the two randomly allocated groups) provides us with an estimate of the causal effect of *offering* treatment (i.e. randomisation) rather than the effect of actually receiving it. It is a valid estimator of a causal effect but many investigators (particularly psychotherapists!) might claim that it is an estimator of the wrong effect. As an estimator of the causal effect of receiving therapy the ITT estimate is likely to be biased. However, many other investigators might be convinced that this is the estimator of real interest – it measures the effect of a decision to treat in a given way and is therefore is vitally important for people involved in making these decisions (or, at the very least, those paying for them!). It is the standard approach to the analysis of drug trials and that usually expected by the regulatory authorities and other bodies such as the US Food and Drug Administration (FDA) and the UK National Institute for Clinical Excellence (NICE).

CHOICE OF, AND ADHERENCE TO, AN APPROPRIATE FORM OF PSYCHOTHERAPY

What constitutes the active treatment for our required comparison? There are several common forms of psychotherapy that are regularly used for patients with depression, including behaviour therapy, cognitive-behaviour therapy (CBT),[8] interpersonal psychotherapy (IPT),[9] brief dynamic psychotherapy.[10] Usually the therapy involves the treatment of individual patients, but there is also the possibility of working with groups of patients with similar problems, or with the patient and his or her family. For a general review, see for example, Scott[11] or Roth and Fonagy.[12] If our aim is to evaluate the efficacy of one of these forms or models of treatment, or to compare its efficacy with another model of psychotherapy or even pharmacotherapy, then it must be self-evident that we need to be able to describe explicitly and precisely what treatment using any of the specified models actually involves; that is, they must be standardised. Crits-Christoph and Gladis[13] give two main reasons for standardisation. First, from a clinical viewpoint, there is a need to be able to describe what actually seems to work (or does not) so that clear treatment recommendations can be made to other potential therapists. Second, from a research viewpoint, therapies need to be replicable. Standardisation of psychotherapies, however, is not easy, and it is a topic beyond both the scope of the present chapter and the competence of the present writer. Briefly, it involves the creation of a detailed treatment manual, the selection and subsequent training of appropriate therapists in the use of the manual, certification of therapists based on adherence to the treatment model, and continued assessment of therapist adherence and competence during a clinical trial.[13] Clearly, when critically appraising the results of a particular RCT, we need to be able to convince ourselves that the therapy has been undertaken as intended, and that the therapy as given was exactly what it is said to be. For this we need a published treatment manual and a well-validated method of measuring adherence to the therapy as described in the manual.

CHOICE OF AN APPROPRIATE CONTROL GROUP

Standardisation of psychotherapy might be thought to be a difficult problem but it is often far more difficult to come up with a valid and convincing control condition. Crits-Christoph and Gladis[13] consider this as perhaps the single most vexing problem for research into the outcome of psychotherapy. Too often, we see that researchers have used 'no treatment' or 'waiting-list' controls. Too often we see the phrase 'routine care' used for the control condition when, in many circumstances, it implies routine neglect. It is important that when patients are invited to take part in an RCT they are convinced that they will receive adequate levels of advice, support and care if they are allocated to the nominal control condition. Otherwise, why should they consent to randomisation? Otherwise, why should an ethics committee grant its approval for the trial? It is also important that the test psychotherapy is being compared with a care package that might be regarded as potentially as good as the therapy on offer. The test therapy, for example, might involve supportive counselling in addition to the specific elements implied by the psychotherapeutic model, and the natural control condition would be the receipt of the same level of support in the absence of the psychotherapeutic elements under test. If, however, we wish to evaluate supportive counselling itself, then we still have a problem. Trialists often refer to 'equipoise' in justification of randomisation in an RCT. To maintain equipoise we need to be convinced that the control group patients are at least provided with the best available routine care and that they are not allocated to a condition that might cause harm.

I will illustrate the choice of control groups by referring to a particularly well-known and influential psychotherapy trial. The National Institute of Mental Health (NIMH) Treatment of Depression Collaborative Research Program (TDCRP)[14] involved the use of four treatment arms. Groups 1 and 2 received CBT and IPT, respectively. Group 3 received pharmacotherapy with imipramine (administered double-blind) together with a care package called 'clinical management (CM)'. Finally, Group 4 received a pill placebo (again administered double-blind) and CM. Elkin *et al.*[14] state that:

> The CM component of both pharmacotherapy conditions was introduced into the study to ensure standard clinical care, to maximize compliance, and to address ethical concerns regarding the use of a placebo on depressed patients. The CM component provided guidelines, not only for the management of medication and side effects and review of the patient's clinical status, but also for providing the patient with support and encouragement and direct advice if necessary. Although specific psychotherapeutic interventions were proscribed (especially those that might overlap with the two psychotherapies), the CM component approximated a 'minimal supportive therapy' condition.

The essence of the imipramine–CM and placebo–CM conditions was to provide a fully standardised package of clinical care, either of which could be used as a control group for the evaluation of the efficacy of the psychotherapies. So, the two major questions addressed by the NIMH TDCRP study were: '(1) Is there evidence of the effectiveness of each of the psychotherapies, as compared both with the standard reference treatment of imipramine–CM and with the placebo plus CM (PLA–CM) control condition? (2) Are there any differences in the effectiveness of the two psychotherapies?' These questions emphasise the comparative nature of this and any other well-designed RCT. When one asks questions about the effectiveness of psychotherapy one should always add the rider 'relative to what?' A valid and well-standardised control condition is as vital to the comparison as is the standardised package of therapy. Crits-Christoph and Gladis,[13] referring on the TDCRP trial, comment that whilst the placebo–CM is perhaps not the ideal control condition for psychotherapy, it serves a practical function. That is, if a specific psychotherapy can do no better than the placebo–CM control, should the psychotherapy be pursued as a treatment option? Beware of authors who make claims about the improvement of patients in a particular treatment group without reference to that in other comparison groups. Roth and Fonagy's[12] (p. 64) comment that the small differences between the four TDCRP trial groups, in terms of their outcome, is due to the unexpectedly good outcome under placebo–CM (explained by the fact that it contains non-specific elements of psychotherapy) seems to be missing the point.

CHOICE OF ASSESSMENT METHOD AND OUTCOME MEASURES

It is very difficult to see how one could possibly design a so-called double-blind RCT in the field of psychotherapy evaluation. The patients are likely to know what is going on, unless they have been deceived by their therapists, and it would be rather bizarre if the therapists were unaware of what treatment was being offered! Blind assessment by a third party (a clinician or research worker not involved in the provision of therapy or clinical support) is often the preferred option, but even here it is frequently difficult to maintain blindness. The therapists, themselves, should not undertake the assessment of outcome. One should always bear in mind, however, that irrespective of who carries out the assessment, there is always the possibility of subjective biases in the assessments. The Hamilton Rating Scale for Depression (HRSD)[15] and the Beck Depression Inventory (BDI)[6] are the two most commonly used measures of depressive symptoms in RCTs for the treatment of depression. In fact, they are frequently both used within the same RCT to assess different aspects of symptomatology. The HRSD is a clinician-rated measure, based on an interview with the patient, which gives more weight to the 'biological' or somatic symptoms of depression, whilst the BDI is a patient-completed questionnaire which concentrates more on the cognitive aspects. There have been suggestions that different forms of therapy (drugs as opposed to psychological treatments, for example) might have a differential effect on these outcome measures (drugs doing better according to the HRSD and the BDI favouring CBT, for example). The expected treatment group by outcome measure interaction needs to be specified (and preferably published) as part of the trial protocol and, if it is regarded as being important, the trial needs to be powered accordingly. In reality, it is hard to imagine a convincing justification for a trial of the size and expense needed for such a test.

What about missing outcome assessments? Drop-outs and other sources of missing data lead to real problems for the valid estimation of the effects of treatment. A detailed discussion of this topic is beyond the scope of the present chapter, but it must be stressed that the only effective way of dealing with missing data is to ensure that there are none. Investigators should make every effort to ensure that outcome data (however brief) are collected on *all* of the patients randomised, irrespective of their subsequent treatment history.

In particular, data collection should not be abandoned simply because the patient has not taken up the offer of therapy or has not adhered to the prescribed course of treatment.[16,17] There is no logical reason why patients should refuse to be assessed even though they have decided that therapy is not for them, although, in practice, the two types of protocol violation are likely to go together.

But what if you have drop-outs and haphazardly missing data? What is the best way of dealing with them in the statistical analysis? If we use a naive complete-case analysis (i.e. base the inferences concerning causal effects on those patients with complete data) then we are likely to have two problems: lack of statistical efficiency (low statistical power) and bias. Bias may be caused, for example, by the drop-outs not occurring completely at random. Ideally, analyses should be available on all available data (and should include where possible all patients randomised to the competing treatment arms) and should compensate for the observed patterns of drop-out. Possibilities for dealing with the missing values include imputation (ranging from rather crude and unsatisfactory methods – at least from the point of view of estimation – such as last-observation-carried-forward to the much more realistic stochastic imputation methods including hot-decking and multiple imputation), the use of inverse probability weighting and, finally, a full likelihood analysis based on statistical models for both the missing data process and for the outcome given that it has been assessed. For further details, readers are referred to a series of reviews in *Statistical Methods in Medical Research*, **8**(1), particularly the primer on multiple imputation by Schafer.[18] The use of inverse probability weights is widespread in survey statistics but has only occasionally been used to allow for drop-outs in RCTs. The use of this technique in longitudinal clinical trials is explained and illustrated in Everitt and Pickles.[19] Its use is also illustrated in a recent depression trial by Dowrick *et al.*[20] Finally, a discussion of the analysis of longitudinal data with drop-outs, paying particular attention to the NIMH TDCRP trial, is provided by Gibbons *et al.*[21]

CENTRE, GROUP AND THERAPIST EFFECTS

It is clear that the outcome of psychotherapy is dependent upon characteristics of the therapist. These include training and experience of the therapist, degree of adherence to the therapeutic model (use of a manual, for example) and the capacity to develop a therapeutic alliance with the patient (see, for example, Crits-Christoph *et al.*[22] and Roth and Fonaghy[12]). In a multicentre RCT there are also likely to be differences in the effectiveness of the collaborating clinical centres. Therapists in some centres may have considerably more experience in the use of a given treatment approach than in others, reflected, for example, in their degree of adherence to a given therapeutic model. In addition, if patients are treated as groups rather than individually there are also likely to be differences between groups arising not only from the characteristics of the patients and of the therapist but also from interactions between the patients.[23] If they get on well together the group might thrive. If, on the other hand, there is a particularly disruptive or difficult patient within a particular group then the group as a whole may not do as well as it might otherwise have done.

Consider a hypothetical single-centre RCT with two treatment arms. In Group A all the patients receive CBT individually from an internationally respected pioneer of CBT, Dr Garner. In Group B all of the patients, again individually, receive supportive counselling from a recently trained community psychiatric nurse, Mr Martin. Let's assume that Dr Garner's patients do considerably better than those of Mr Martin. What can we infer about the causal effect of CBT from such a trial? What are the threats to the validity of the trial? Dr Garner is likely to be a very experienced, highly skilled and highly motivated 'Brand champion'. Mr Martin, on the other hand, lacks experience. Is the observed difference due to the difference in abilities and experience of the two therapists, or is it an effect of CBT? We cannot tell. The two effects are completely confounded in this simple design. This is a severe threat to the internal validity of the trial. If, however, we believe that the observed differences are an effect of CBT, then what? We still cannot be sure that there is not some attribute of Dr Garner that enables him to be particularly successful in delivering this particular variant of CBT. Could other clinicians use the same model and achieve the same or, at least, comparable results? We do not know. This is a threat to the external validity or generalisability of the trial's findings.

In practice, many RCTs, which otherwise have admirable design characteristics and quality control procedures, involve the use of only two or three highly skilled and experienced therapists. They are often the academic clinicians who have been involved in the development or modification of the therapy under evaluation. Does this invalidate the findings of these trials? No, but it does limit their generalisability. They should, perhaps, be regarded as the equivalent of the pharmacotherapist's 'Phase II' drug trials, being a necessary preliminary to the design and conduct of a full 'Phase III' evaluation using a large and representative sample of therapists. It would be inappropriate and certainly difficult to justify a large multi-centre trial involving large numbers of therapists without first being able to establish that the 'experts' or 'Brand champions' are able to achieve promising results. If the latter cannot demonstrate worthwhile effects then it would be pointless to move on to the larger trial. If they can, however, we then (but only then) need to ask how well the therapy might work in routine clinical practice.

In a large multi-centre trial we need to involve as many therapists as possible. Each therapist is likely to be based in only one of the centres and to be delivering treatment in only one arm of the trial (i.e. therapists are *nested* within both centres and treatments) but it is possible for a therapist to deliver more than one of the forms of therapy in a comparative trial (i.e. therapistics, like centres, are *crossed* with treatments). These designs have implications for the statistical analysis and for the validity of statistical inferences based on these analyses.[24–26] Both centre and therapist effects should be incorporated into an appropriate statistical model. Using the notation of Roberts,[26] such a model, for a quantitative outcome measure, for example, will have the form

$$y_{ijk} = \alpha + \lambda_j + \sum_p \beta_p x_{ijkp} + u_{jk} + \varepsilon_{ijk} \quad (22.5)$$

Here y_{ijk} is the outcome of the ith patient of the kth therapist within the jth treatment arm of the trial. Assuming that λ_j is zero in the control arm, λ_j is the effect of the treatment effect for the jth arm of the trial. Each x_{ijkp} is the baseline measurement of the pth patient characteristic (such as a demographic or other prognostic variable, including treatment centre) and β_p is the corresponding regression coefficient. The term u_{jk} is the average effect of the kth therapist within the jth treatment arm of the trial. It is a random variable (i.e. randomly varying from one therapist to another) with an assumed mean of zero and variance of σ_j^2. This variance may vary from one arm of the trial to another (there is no a priori reason why the variation between therapists within different arms of the trial should be the same) and, in particular, if the control arm does not involve the use of therapists at all, then, for the controls $\sigma_j^2 = 0$ (i.e. in this situation there are no therapist effects in the control group). In the case of the possibility of one or more arms of the trial involving group therapies the statistical model would be even more complex.

Models such as that described in equation (22.5) are called *random effects*, *random regression* or *multilevel* models.[27] Technical details of their use are beyond the scope of this chapter and interested readers are referred to Roberts[26] for an illustration of their use in the context of RCTs involving therapist effects. What readers should note, however, is that failure to allow for appropriate therapist effects in the statistical analysis (assuming that they are present in the data) is likely lead to spurious statistical significance (i.e. the stated P-values will be too low) and estimated confidence intervals or standard errors for treatment effects that are too optimistic (i.e. smaller than they should be). A corollary of this is that, even when the analysis is correct, a trial whose sample size has not been determined after allowance for the possibility of therapist effects is very likely to be underpowered (too small!). This is the same problem as those faced by the designers of cluster randomised trials.[28] Again, Roberts[26] provides details of the required adjustments to sample size calculations, on the assumption that therapist variation is the same for each of the arms of the trial.

But there is more to the problem of therapist effects than can be solved by the technical device of allowing for them in an appropriate statistical model. Nor is the main problem one of generalising from the impact of therapists in a given trial to the wider community of therapists. We started the discussion in the present section by comparing the outcome of CBT as delivered by Dr Garner with that of IPT as delivered by Mr Martin. We pointed out that the required treatment effect is fully confounded with the difference between the two therapists. Now let us move on to a larger trial in which each of the patients in Group A receive CBT from a randomly selected therapist from a team of, say 50, experienced and highly competent cognitive therapists. Each of the patients in Group B, on the other hand, receive IPT from a randomly selected therapist from a team of experienced and highly competent interpersonal therapists. We still have a problem. Again, the required treatment effect (the difference between outcomes for CBT and IPT) cannot be disentangled from the difference between the average effects of the two groups of therapists. In general, the λ_j in equation (22.5) can be interpreted either as an average of the within-Group j therapist effects or as an effect of Therapy j – that is, the two interpretations are equivalent.

> At the present time, researchers and consumers of psychotherapy research findings are left with a basic dilemma when interpreting the findings of studies focusing on the efficacy of specific treatments: how to disentangle the effects dues to the therapeutic approach from those due to the particular therapists who have carried out the approach. It is particularly pressing when different therapists carry out each of the treatments in a comparative outcome study'.[29]

The cognitive therapists and interpersonal therapists in the above hypothetical trial (or even in a real one such as the NIMH TDPRC study) might differ in lots of ways and these therapist differences may be the causal effects of the treatment difference, not the difference in psychotherapeutic approach. Consider therapeutic competence, for example. The CBT therapists might be either more or less competent than their IPT counterparts. But how could we assess this? How could we possibly compare the competence of Dr Garner as a cognitive therapist, for example, with that of Mr Martin as an interpersonal psychotherapist? It is akin to asking whether I am more competent as a statistician than my scientific colleague is as a laboratory worker. And moving to a crossed design (both types of therapy being offered by every therapist) does not solve our problem. If Dr Garner, for example, were to be experienced and highly competent as both a cognitive therapist and an interpersonal therapist we still would not be able to compare the competences in the two approaches. Elkin[29] concluded that: 'We may never be able to truly "disentangle" the effects due to the therapist from those due to the therapy, because they may often be inherently intertwined and also very interactive with particular patient attributes.'

Further, as stated by Elkin *et al.*: 'The treatment conditions being compared ... are, in actuality, "packages" of particular therapeutic approaches and the therapists who chose and are chosen to administer them'.[30] The interpretation of the results of RCTs should explicitly acknowledge this fact. As well as very carefully defining both the treatment and the control conditions, authors should provide critical information about the therapists carrying out the treatments, and the information should be included in the dissemination concerning supposedly empirically validated results.[29] The latter is particularly important when we come to systematic review and/or meta-analysis of the results from a disparate collection of individual trials, and in the formulation of any subsequent clinical guidelines based upon the results of these trials.

WHAT WORKS FOR WHOM?

The question implies a belief that there is no constant treatment effect. That is, it implies that a given form of treatment has a greater effect on some patients than it does on others; that the receipt of Psychotherapy A will be more beneficial for Mr Smith than receipt of Psychotherapy B, for example, but that B might be better than A, however, for Mrs Jones. Mr Smith has a particular attribute (presenting symptoms, clinical or family history, for example) that indicates therapy A. Mrs Jones, on the other hand, has characteristics that indicate therapy B. In the epidemiological literature this is called 'effect modification' – a particularly useful term as it should remind us that 'causal effect' implies comparison of observed outcome with that which would have been observed under different circumstances. In terms of statistical modelling (analysis of variance, or covariance, for example) it will provide an example of a treatment group by patient attribute interaction, where the attribute could be one of a potentially vast range of measures made on the patients at or prior to randomisation. Supposed examples of such interactions are rarely convincing. Even if based on a valid statistical analysis (i.e. a test of an appropriate two-way interaction) they are usually 'discovered' as part of a *post hoc* 'fishing trip'. More frequently their existence has been based on an invalid analysis. All too often the investigators are looking for a so-called 'predictor of outcome' by searching in the relevant treatment group for patient attributes that are associated with good outcome. This tells us nothing about effect modification – the same attributes might lead to the better outcomes within the control group(s). One should always remember that valid inferences from an RCT involve comparison of the randomised groups. Here we are concerned with the question 'Does the treatment effect (e.g. comparison of outcomes in Groups A and B) depend on, say, patient attribute C?' The identity of attribute C should be clearly specified in the trial protocol, together with a prior estimate of the size of the proposed interaction. The sample size for the trial should then be determined such that there is sufficient power to detect this interaction through the use of an appropriate statistical significance test. One good candidate for attribute C might be patient preference[31] (see below), but there is little, if any, methodologically sound work in this area.

The quality of the therapeutic alliance is also a good candidate effect modifier but, again, there is little sound work in this area. There are several technical challenges to the valid exploration of therapeutic alliance effects. Therapeutic alliance is always subject to substantial amounts of measurement error and, by definition, is only measured in the group(s) receiving therapy (i.e. we have a rather dramatic missing data problem). Remembering that we are aiming to look at comparisons between those patients receiving therapy and those who are not, a naive regression analysis in the treated group is likely to produce estimates that are a mixture of the required treatment effect and selection biases (the effects of hidden confounders). Patients who can form a strong therapeutic alliance are likely to be those who would have a relatively good outcome even in the absence of treatment. Estimated regression coefficients obtained from the analysis of data from the treated group are not valid estimates of treatment effects. Valid ways of attempting to cope with these problems are only in their infancy and the methodology is well beyond the scope of the present chapter. We know of no published examples.

Returning to more straightforward attempts to detect effect modification, Crits-Christoph and Gladis[13] comment that two of the largest randomised clinical trials ever undertaken to evaluate psychotherapies (although not specifically for depression) failed to provide much support for specified patient–treatment interactions.[22,32] However, readers interested in pursuing this topic should consult the recent text by Aguinis.[33]

ESTIMATION OF CAUSAL EFFECTS IN AN RCT WITH NON-ADHERENCE TO ALLOCATED TREATMENT

Consider a hypothetical RCT in which 200 eligible depressed patients have been randomly allocated to receive either counselling plus routine care ($T = t$) or routine care alone ($T = c$). For simplicity, assume that all of the 100 patients allocated to routine care receive exactly that (they do not have access to counselling unless they have been allocated to that treatment arm of the trial). Of the 100 patients offered counselling, however, only 70 accept the offer. After a fixed time interval after randomisation (six months, say) the patients' clinical status (improved versus not improved) is assessed and used as the primary outcome of the trial. The effects of either treatment allocation or treatment actually received are to be estimated from the differences between average outcomes as before, the only difference being that we are averaging binary outcomes ($1 =$ Improved; $0 =$ not improved, say) to obtain observed proportions. The results of this hypothetical trial are summarised in Table 22.1 (note that we have simplified the issue by assuming that there is no loss to follow-up).

The estimate of the ITT effect is both simple and familiar. The proportion of those receiving counselling who improve is 0.70 (i.e. 70/100) and the corresponding proportion for the control group is 0.50 (i.e. 50/100). The difference (the ACE for being offered counselling) is 0.20. For readers who prefer a number needed to treat (NTT – the reciprocal of the difference between the two proportions), this is 5 (i.e. 1/0.20). But what about estimating the causal effect of receiving treatment? There are two commonly used, but potentially invalid,[34] methods of analysis – analysis *per protocol* or analysis *as treated*. There is also the correct (correctness, of course, being vitally dependent on the validity of a few key assumptions) but much less familiar estimator – the *complier average causal effect* (CACE).[35–37] The *per protocol* analysis compares the outcome in those people in the counselling group who actually receive counselling with that in the control group (i.e. it excludes patients who have violated the treatment protocol from the analysis). Here the difference is $60/70 - 50/100 (= 0.36)$. The *as treated* analysis compares outcome in those patients who receive counselling with that in those who do not receive it (all patients are included in this analysis). Here it is $60/70 - 60/130 (= 0.40)$. The problem with both of these estimators is that it is impossible to interpret them as a causal effect in the sense of comparing potential outcomes *on the same patient*. The patient groups are not comparable. The estimated effects are merely associations, subject to confounding. And association, as we all know, does not imply causality!

Table 1. Results of a hypothetical trial of counselling

	$T = $ t		$T = $ c	
	Improved	Total	Improved	Total
Comply	60	70		
Do not comply	10	30		
Overall	70	100	50	100

What about the CACE? This is an estimate of the difference between the outcome in the compliers (i.e. those who accepted and received the offered counselling) and that which would have been expected *in the same patients* if they had not been offered counselling. This is where we need two key assumptions.[38,39] The first one is easy to defend for a randomised trial. The second needs a bit more careful thought:

> Assumption 1: the proportion of patients who are *potential* compliers is the same in the two randomly allocated groups. This follows directly from the random allocation mechanism.
>
> Assumption 2: the proportion of potential non-compliers who improve is independent of treatment allocation. In other words, it makes no difference to the outcome of a patient who would refuse the offer of counselling whether or not the patient is in the group actually offered counselling. The offer, in itself, is not beneficial. This is called an exclusion restriction.

Assumption 1 allows us to estimate the proportion of potential compliers in the control group. In our example it is 70/100. The estimated number of non-compliers in the control group is 30 and the number of compliers is 70.

Assumption 2 allows us to estimate the proportion (number) of patients who improve amongst the non-compliers in the control group. In our example the number of patients who improve in this group is estimated to be 10 (the proportion is 10/30). Now, there were a total of 50 patients who were observed to improve in the control group and therefore the estimated number of potential compliers who improve in the control group must be 40 (that is $50 - 10$). Otherwise the numbers do not add up! So, the proportion of patients improving in the counselling group amongst those who actually receive counselling is 60/70. The proportion in the corresponding control group (i.e. those who would have accepted the offer) is estimated to be 40/70. The CACE estimator is the difference between these two proportions, $60/70 - 40/70 (= 0.29)$. The corresponding NNT is 3.5.

Note that in the above example the potential compliers did better than the non-compliers, irrespective of which treatment arm they were allocated to. This is not unexpected and not too difficult to rationalise. But now consider a second, more 'difficult' example. The results of a second hypothetical trial are shown in Table 22.2. The ITT effect (ACE) is estimated by $50/100 - 30/100 (= 0.20)$. The corresponding NTT is 5. The CACE estimate is $35/70 - 15/70 (= 0.29)$. The corresponding NNT is again 3.5, but note that this time the potential compliers in the control group are doing a lot worse than the non-compliers (15/70 vs. 15/30). Again, this is reasonably straightforward to rationalise. The patients who accept the offer of counselling are those with the worst prognosis or, equivalently, those that turn it down (or would turn it down if offered it) are those who are getting better anyway. The latter do not need treatment. But what this should do is to prompt the data analyst to ask whether Assumption 2 is really justified. Might the offer of help on its own be of benefit? And if so, of how much benefit? Or perhaps those patients in the control group who would have accepted the offer feel let down (resentful demoralisation) and do worse than they would otherwise have done if they had known nothing about the possible offer of help.

Table 2. Results of a second hypothetical trial of counselling

	$T = \mathrm{t}$		$T = \mathrm{c}$	
	Improved	Total	Improved	Total
Comply	35	70		
Do not comply	15	30		
Overall	50	100	30	100

In practice (at least when there are no complications arising from missing outcome data), we do not have to work through the estimation from first principles in the above way. It can be shown that the required estimates can be obtained from the following simple formula:[35,40]

$$\text{CACE estimate} = \frac{\text{ITT estimate for outcome}}{\text{ITT estimate for receipt of treatment}} \qquad (22.6)$$

So, for the first example above, $\text{CACE} = (70/100 - 50/100)/(70/100 - 0) = 0.29$, as before.

So far, we have ignored missing outcome data. But all trials have some missing outcomes and it is highly likely that those patients who fail to comply with the treatment protocol have an increased probability of failing to provide outcome data. The analysis of data from psychological treatment trials with both non-adherence to treatment and missing outcome data is beyond the scope of this chapter, but interested readers are referred to papers by Dunn *et al.*[41–43] The latter two[42,43] contain detailed illustrations of CACE estimation using real data from the Outcomes of Depression International Network (ODIN) trial.[20]

HOW MUCH THERAPY? 'DOSE'–RESPONSE EFFECTS

Is the treatment effect for patients who receive 10 sessions of psychotherapy different to that in patients who receive a course of 12 sessions? The obvious way to answer this is through the use of an RCT in which, for example, patients are randomly allocated to receive 12, 6 or no sessions. Assuming that there is full adherence to the treatment protocol (perhaps unlikely) and no missing outcomes (again unlikely) a very straightforward analysis will tell us whether there is an apparent 'dose' effect. Examples of such trials are provided by the work of Shapiro *et al.*,[44] Barkham *et al.*[45] and Dekker *et al.*[46] We must be a little wary of interpreting the results, however. Although we are estimating the differences in treatment effects between specified courses of therapy of different lengths we must not necessarily equate the number of planned sessions as a measure of the amount of therapy received. The shorter course might be more intensive and easily make up for the more leisurely approach provided by 12 sessions.

A second, complementary way of looking at the 'dose'–response problem is to allocate patients to a fixed number of sessions (12 or none, say) and observe how many sessions they actually attend. We then investigate the relationship between the number of sessions attended and the outcome of the therapy. Very often investigators do this via a naive regression analysis of the results from one arm of a trial (those offered the course of therapy), ignoring the controls. Estimates of apparent dose–response effects obtained from such an analysis are likely to be biased. They will be subject to selection effects (hidden confounding).[47] A preferable approach is instrumental variable regression (using two-stage least squares, for example). The instrumental variable here is randomisation. As an instrumental variable (or instrument), randomisation is assumed to have the following two properties: it has a strong effect on the number of sessions attended (particularly if none of the patients in the control group have access to therapy) and its only influence on outcome is through the number of sessions received (i.e. the effect of randomisation is mediated completely by the treatment received). The second characteristic is analogous to the exclusion restriction (Assumption 2) as described under CACE estimation. In fact, CACE estimation is just a specific example of the use of instrumental variables in the estimation of an average treatment effect. In the absence of missing outcome data the required estimate of the slope for the 'dose'–response is, again, simply given by equation (22.6). In this case the ITT effect for receipt of treatment is the difference in mean number of sessions attended in the two randomised arms of the trial.

Of course, we could randomise patients to receive programmes of therapy of different length (including a group given no access to therapy) and still observe how many sessions were, in fact, attended within each treatment group. In this situation we could then combine the data analytic strategies to look at both the effect of allocation (an ITT effect) and 'dose'–response within each allocated group (using the no-therapy group as a control in each case). We know of a few examples of the use of instrumental variables to look at 'dose'–response effects in observational studies;[47,48] the only example from an RCT of psychological treatments is the analysis of ODIN data briefly summarised in Dunn *et al.*[43] The latter was complicated by a significant amount of missing outcome data. Readers are referred to Permutt and Hebel,[49] Imbens and Angrist[50] and Fischer-Lapp and Goetghebeur[51] for more information on the estimation of causal effects with varying intensity of treatment.

RANDOMISED CONSENT AND PATIENT PREFERENCE DESIGNS

A serious issue in the design of RCTs concerns the amount of information given to the patient about the aims of the trial. So-called informed consent is a prerequisite for most trials but it is not always obvious what 'informed consent' actually means or whether, strictly speaking, it is ever possible. In the context of our example illustrating the effect of patient compliance to a treatment offer, is it ever ethically justified to randomise and then only seek consent to treat in the group allocated to receive therapy? This is an example of Zelen's original form of the randomised consent design.[52] All patients in the trial are asked to provide outcome data, of course, but those in the control group may never know that they had taken part in a trial. Is this really a serious ethical problem? This design would circumvent the potential problem of resentful demoralisation amongst the controls. I will not attempt to answer the question raised.

In an attempt to solve some of the serious problems surrounding the issue of patient preference, Brewin and Bradley[53] (see also Bradley[54] and Silverman and Altman[55]) have proposed what they describe as *patient preference design*. In this design, eligible patients are told about the reasons for the trial and the treatments on offer. Patients who do not have a strong preference (i.e. they are prepared to be randomised) are entered into a conventional RCT. Those patients with a strong preference are offered the treatment of their choice. So, for the comparison of two treatments, A and B, for example, the patient preference trial finishes up with four groups: those who prefer A; those without preference who are randomly allocated to A; those who prefer B; those without preference who are randomly allocated to B. In the context of the present discussion, the comparison of the randomly allocated groups can lead to an ACE or CACE estimate as described above. But what can the two patient preference groups provide? Merely an estimate of association. Like *per protocol* or *as treated* estimators, they do not appear to be able to provide estimates of causal effects. And for this reason they cannot be used to check the (external) validity of the estimates of causal effects provided by the randomised groups. Whether the difference between the two preference groups is the same as or completely different from that provided within the core RCT, so what? What does it tell us? That there are selection effects? The treatment effect may, indeed, be different in those patients without a strong preference (i.e. those prepared to be randomised) when compared with the rest, but the rest cannot provide the valid information from which we can test whether it is true. Perhaps readers should see the results of such a trial and decide for themselves. An example of the use of a patient preference design is provided by a published trial of counselling for depression[56,57] (also see the recent systematic review by King *et al.*).[58]

Another design possibility, which, in my view, has much more promise, is simply to ask the participants of a conventional RCT what their preferences are prior to randomisation. The aim here is not to allow patient preference to influence treatment received (but in the presence of non-compliance this will be inevitable) but to incorporate patient preferences into the analysis of the resulting data. This has been tried by Torgerson *et al.*[31] Although these authors do no pursue all of the possibilities in terms of estimating treatment effects, the design offers ways, at least partially, of testing the validity of the assumptions necessary for the above CACE estimator, or, equivalently, looking for a poor prognosis/demoralising effect in the potential compliers of the control group. Getting preference information prior to randomisation would also improve the precision of the estimates of the CACE, but this is well beyond the scope of the present chapter.

CONCLUSIONS

The design and analysis of a convincing RCT for the estimation of the effects of psychotherapy are difficult. It is not safe simply to assume that the theoretical and logistical problems are similar to those of the average drug trial. Life here is much more complex. Psychotherapy (at least in its individual form) involves the interaction of two people (the patient and the therapist) and it is the involvement of these two people that is the essence of the complexity. Added to this are the problems of the choice of adequate control groups (in particular, the absence of a convincing placebo) and the impossibility of conducting a trial that is double-blind. In the critical appraisal of such trials we should not, perhaps, be searching for methodological perfection but, instead, be aware of the pitfalls to valid inferences concerning treatment effects and temper our judgements accordingly (and this applies just as much to the trials that we have been involved as it does to the trials of other investigators).

This chapter has not considered systematic review and meta-analysis of trials of psychotherapies but the authors (and appraisers) of such studies should be fully aware of all of the methodological pitfalls of the individual trials. A meta-analysis of a series of trials that have naively ignored random therapist effects, for example, or ignored the structure of a group therapy trial, simply summarises the faulty analyses of the originals. Unfortunately, the consumers of meta-analyses (particularly if they are produced under the auspices of such august bodies as the Cochrane Collaboration) seem to place far too much faith in their findings. Consumers need to be aware that the authors of systematic reviews are capable of missing subtle (or not so subtle) methodological flaws in the original trials. Consumers should resist taking the conclusions of the authors of these meta-analyses, and the clinical guidelines that result from them, on trust. In order to appraise critically a published systematic view one needs not only to know about the mechanism (and quality) of the review itself, but also to have a detailed knowledge of what the reviewers should have been looking for in the way of methodological problems in the original trials. Reporting guidelines such as CONSORT[59,60] are having a substantial impact on the quality of clinical trials and on the appraisal methodologies of systematic reviewers. In the case of psychotherapy trials, however, the CONSORT recommendations only cover a small part of the key components of the trial. Sticking to CONSORT guidelines is necessary for a good trial report, *but is not sufficient*. I hope the present chapter succeeds in stimulating readers to think of other aspects of such trials that need to be equally well reported.

REFERENCES

1. Dunn G. Statistical methods for measuring outcomes. In Tansella M, Thornicroft G, eds, *Mental Health Outcome Measures*. London: Gaskell (2001) 5–18.
2. Garety P, Dunn G, Fowler D, Kuipers E. The evaluation of cognitive behaviour therapy for psychosis. In: Wykes T, Tarrier N, Lewis S, eds, *Outcome and Innovation in Psychological Treatment of Schizophrenia*. Chichester: John Wiley & Sons, Ltd (1998), 101–18.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* (1974) **66**: 688–701.
4. Holland PW. Statistics and causal inference (with discussion). *J Am Stat Assoc* (1986) **81**: 945–70.
5. Dunn G. Patient choice and non-adherence to treatment in randomised controlled trials of counselling or psychotherapy. *Understand Stat* (2002) **1**: 19–29.
6. Beck AT, Ward CH, Mendelson M *et al.* An inventory for measuring depression. *Arch Gen Psychiat* (1961) **4**: 561–71.
7. Sheiner LB, Rubin DB. Intention-to-treat and the goals of clinical trials. *Clin Pharmacol Ther* (1995) **57**: 6–15.
8. Fennell MJV. Cognitive-behaviour therapy for depressive disorders. In Gelder MG, Lopez-Ibor J-L Jr, Andreasen NC, eds, *New Oxford Textbook of Psychiatry*. Oxford: Oxford University Press (2000), 1394–1405.
9. Markovitz JC, Weissman MM. Interpersonal psychotherapy for depression and other disorders. In Gelder MG, Lopez-Ibor J-L Jr, Andreasen NC, eds, *New Oxford Textbook of Psychiatry*. Oxford: Oxford University Press (2000), 1411–20.
10. Ursano RJ, Ursano AM. Brief individual psychodynamic psychotherapy. In Gelder MG, Lopez-Ibor J-L Jr, Andreasen NC, eds, *New Oxford Textbook of Psychiatry*. Oxford: Oxford University Press (2000), 1421–32.
11. Scott J. Psychological treatments for depression: an update. *Br J Psychiat* (1995) **167**: 289–92.
12. Roth A, Fonagy P. *What Works for Whom? A Critical Review of Psychotherapy Research*. New York and London: Guilford Press (1996).
13. Crits-Cristoph P, Gladis M. The evaluation of psychological treatment. In Gelder MG, Lopez-Ibor J-L Jr, Andreasen NC, eds, *New Oxford Textbook of Psychiatry*. Oxford: Oxford University Press (2000), 1259–69.
14. Elkin I, Shea T, Watkins JT *et al.* National Institute of Mental Health Treatment of Depression Collaborative Research Program (1989). *Arch Gen Psychiat* (1989) **46**, 971–81.
15. Hamilton, M. A rating scale for depression. *J Neurol Neurosurg Psychiat* (1960) **23**: 56–61.
16. Lavori PW. Clinical trials in psychiatry: should protocol deviation censor patient data? *Neuropsychopharmacology* (1992) **6**: 39–48.

17. Siqueland L, Chittams J, Frank A *et al*. The protocol deviation patient: characterization and implications for clinical trials research. *Psychothe Res* (1998) **8**: 287–306.
18. Schafer D. Multiple imputation: a primer. *Stat Methods Med Res* (1999) **8**: 3–15.
19. Everitt BS, Pickles A. *Statistical Aspects of the Design and Analysis of Clinical Trials*. London: Imperial College Press (1999).
20. Dowrick C, Dunn G, Ayuso-Mateos J-L *et al*. Problem solving treatment and group psychoeducation for depression: a multicentre randomised controlled trial. *Br Med J* (2000) **321**: 1450–5.
21. Gibbons RD, Hedecker D, Elkin I *et al*. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Arch Gen Psychiat* (1993) **50**: 739–50.
22. Crits-Cristoph P, Siqueland L, Blaine J *et al*. Psychological treatments for cocaine dependence: National Institute for Drug Abuse Collaborative Cocaine Treatment Study. *Arch Gen Psychiat* (1999) **56**: 493–502.
23. Burlinghame G, Kircher J, Taylor S. Methodological considerations in group psychotherapy research: past present and future practices. In Fuhriman A, Burlinghame G, eds, *Handbook of Group Psychotherapy and Counseling*. New York: John Wiley & Sons, Inc. (1994) 41–80.
24. Martindale C (1978). The therapist-as-fixed-effect fallacy in psychotherapy research. *J Consult Clin Psychol* (1978) **46**: 1526–30.
25. Crits-Cristoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *J Consult Clin Psychol* (1991) **59**: 20–6.
26. Roberts C. The implications of variation in outcome between health professionals for the design and analysis of randomized clinical trials. *Stat Med* (1999) **18**: 2605–15.
27. Goldstein H. *Multilevel Statistical Models*, 3rd edn. London: Arnold (2003).
28. Donner A. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold (2000).
29. Elkin I (1999) A major dilemma in psychotherapy outcome research: disentangling therapists from therapies. *Clin Psychol: Sci Pract* (1999) **6**: 10–32.
30. Elkin I, Parloff MB, Hadlet SW, Autry JH. National Institute of Mental Health Treatment of Depression Collaborative Research Program: background and research plan. *Arch Gen Psychiat* (1985) **42**: 305–16.
31. Torgerson DJ, Klaber-Moffett J, Russell IT. Patient preferences in randomised trials: threat or opportunity? *J Health Serv Res Pol* (1996) **1**: 194–7.
32. Project MATCH Research Group. Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes. *J Stud Alcohol* (1997) **58**: 7–29.
33. Aguinis, H. *Regression Analysis for Categorical Moderators*. New York: Guilford Press (2004).
34. White I. Uses and limitations of randomisation-based efficacy estimators. *Stat Methods Med Res* (2005) **14**: 327–47.
35. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *J Am Stat Assoc* (1996) **91**: 444–55.
36. Little R, Yau LHY. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychol Methods* (1998) **3**: 147–59.
37. Little R, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Ann Rev Public Health* (2000) **21**: 121–45.
38. Bloom HS. Accounting for no-shows in experimental evaluation designs. *Eval Rev* (1984), **8**: 225–46.
39. Sommer A, Zeger SL. On estimating efficacy from clinical trials. *Stat Med* (1991) **10**, 45–52.
40. Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. *Stat Med* (1997) **16**: 1017–29.
41. Dunn G. Estimating the causal effects of treatment. *Epidemiol Psichiat Soc* (2002) **11**: 206–15.
42. Dunn G, Maracy M, Dowrick C *et al*. Estimating psychological treatment effects from an RCT with both non-compliance and loss to follow-up. *Br J Psychiat* (2003) **183**: 323–31.
43. Dunn G, Maracy M, Torgerson B. Estimating treatment effects from randomized clinical trials with non-compliance and loss to follow-up: the role of instrumental variable methods. *Stat Methods Med Res* (2005) **14**: 369–95.
44. Shapiro D, Barkham M, Rees A *et al*. Effects of treatment duration and severity of depression on the effectiveness of cognitive-behavioral and psychodynamic- interpersonal psychotherapy. *J Consult Clin Psychol* (1994) **62**: 522–34.
45. Barkham M, Rees A, Shapiro D *et al*. Outcomes of time-limited psychotherapy in applied settings: replicating the second Sheffield psychotherapy Project. *J Consult Clin Psychol* (1996) **64**: 1079–85.
46. Dekker J, Molenaar S, Kool G *et al*. Dose-effect relations in time-limited combined psychopharmacological treatment for depression. *Psychol Med* (2005) **35**: 47–58.
47. Foster M. Is more better than less? An analysis of children's mental health services. *Health Serv Res* (2000) **35**: 1135–58.

48. Bickman L, Andrade AR, Lambert EW. Dose response in child and adolescent mental health services. *Mental Health Serv Res* (2002) **4**: 57–70.
49. Permutt T, Hebel JR. Simultaneous-equation estimation in a clinical trial of the effect of smoking and birth weight. *Biometrics* (1989) **45**: 619–22.
50. Imbens G, Angrist JD. Identification and estimation of local average treatment effects. *Econometrika* (1994) **62**: 467–76.
51. Fisher-Lapp K, Goetghebeur E. Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clin Trials* (1999) **20**, 531–46.
52. Zelen M. A new design for randomized clinical trials. *New Engl J Med* (1979) **300**: 1242–5.
53. Brewin C, Bradley C. Patient preferences and randomised clinical trials. *Br Med J* (1989) **299**: 313–5.
54. Bradley C. Designing medical and educational intervention studies. *Diabet Care* (1993) **16**: 509–18.
55. Silverman WA, Altman DG. Patient's preferences and randomised trials. *Lancet* (1996) **347**, 171–4.
56. Ward E, King M, Lloyde M *et al*. Randomized controlled trial of non-directive counselling, cognitive-behaviour therapy, and usual general practitioner care for patients with depression. I: Clinical effectiveness. *Br Med J* (2000) **321**: 1383–8.
57. Bower P, Byford S, Sibbald B *et al*. Randomized controlled trial of non-directive counselling, cognitive-behaviour therapy, and usual general practitioner care for patients with depression. II: Cost effectiveness. *Br Med J* (2000) **321**: 1389–92.
58. King M, Nazareth I, Lampe, F *et al*. Impact of participant and physician intervention preferences on randomized trials. *J Am Med Assoc* (2005) **293**: 1089–99.
59. Moher D, Schultz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* (2001) **134**: 657–62.
60. Altman DG, Schultz KF, Moher D *et al*. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* (2001) **134**: 663–94.

REPRODUCTIVE HEALTH

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

23

Contraception

GILDA PIAGGIO AND TIMOTHY M.M. FARLEY
Department of Reproductive Health and Research, World Health Organization, Geneva, Switzerland

INTRODUCTION

Contraception deals with the prevention of pregnancy. The basic pillar of family planning programmes is a wide spectrum of contraceptive methods that enables men and women to make informed choices about timing and size of their family. Effective and safe methods should be available such that they fit the needs of women and men in very diverse social and cultural settings worldwide. Ideally, there should be reversible and permanent methods that also protect against sexually transmitted infections, can be controlled by women without requiring their partner's consent or cooperation, can be used by adolescents and by breast-feeding women. The choice of a contraceptive method involves personal decisions and depends on the stage of life, family situation or civil status, age, preferences and health profile of individuals and couples.

Contraceptives are typically used by healthy individuals to prevent pregnancy – they are used not only for personal convenience and lifestyle, but also to improve the health of women and children by allowing optimal timing and spacing of births, and freeing women from the burden of repeated unintended pregnancy. Contraceptive methods need to be very safe so that the benefits from their use are not offset by excessive or unacceptable risks. Contraceptive efficacy, effectiveness and risks must be well defined to enable users and providers to make the best choice of a contraceptive method.[1]

The development of effective and safe methods of contraception poses special challenges. First, the complex physiology of human reproduction must be understood to identify targets where the normal processes can be safely interrupted. Second, the effectiveness of some methods depends not only on the degree of protection provided by the method itself (called the *efficacy* of the method), but also on how consistently and correctly it is used, which depend on complex behavioural and social factors. Third, some methods are used by the man but failure (pregnancy) is always observed in the woman.

Progress in contraceptive research and development since the first oral contraceptives were discovered in the 1960s has resulted in a wide variety of safe and effective methods.[2] New methods have been developed, the safety and effectiveness of existing methods have been improved and there is now a much better understanding of the characteristics of users who could

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

safely and effectively use different methods. There has been particular emphasis on the development of new or improved reversible hormonal methods for women, with a transition from high-dose to low-dose combined oral contraceptives, from inert to copper- and levonorgestrel-releasing intrauterine devices (IUDs), the development of combined injectable methods, a combined hormonal patch and vaginal ring, and progestogen-only implants have been introduced. Hormonal methods for men are still not widely available, despite considerable research efforts to date. There is a need to continuously update guidelines and recommendations based on the most recent evidence, and make these accessible to all providers and programme managers. The WHO guidelines follow an evidence-based approach to balancing the risks and benefits of use of different methods and recommend 'who' can use which contraceptive methods safely.[2] These are complemented by information on 'how' to use the different methods safely and effectively.[3] These international guidelines need to be adapted by countries and programme managers according to the availability of different methods, the service delivery requirements, as well as the risks associated with unplanned pregnancy.

CONTRACEPTIVE METHODS: AN OVERVIEW

Contraceptive methods can in general be classified into hormonal and non-hormonal methods. Hormonal methods used by women include oral contraceptive pills (OCs), injectable preparations, implants, hormone-releasing devices (vaginal rings and progestogen-releasing IUDs) and post-coital oral pills (visiting pills and emergency contraceptive (EC) pills). Non-hormonal methods used by women include IUDs, barrier methods (diaphragm and female condom), spermicides, natural methods (calendar and lactational amenorrhoea) and sterilisation, as well as immunocontraceptives, under development. Hormonal methods for men consist of injectable preparations and implants, still under development. Non-hormonal methods include condoms, withdrawal and sterilisation (vasectomy and vas occlusion), while immunocontraceptives or vaccines for men are under development. Extensive and detailed descriptions of old and new contraceptive methods are available.[4,5]

These broad classes of contraceptive methods differ in their length of action, in the mechanism of action, in the interval and mode of administration or insertion, in the possibility of control by the woman, in their effectiveness[6] and in their possible effects on health and indications for their use.[2] They also differ in the way they meet the interests of men and women in different social and cultural settings. Table 23.1 presents a list of selected currently used contraceptive methods with their timing of use or duration of action, typical and perfect failure rates and continuation rates

In what follows we briefly define each of these methods, refer to its effectiveness, its mechanisms of action, a brief history of its development, any safety concerns applied to medical eligibility criteria, and the extent of its use if known.

HORMONAL CONTRACEPTIVES FOR WOMEN

Hormonal methods prevent conception by inhibiting ovulation, preventing implantation or changing the quality of cervical mucus and thus preventing sperm access to the cervix. Oral methods exert their action within the cycle of administration, while the duration of action of injectable preparations, implants and hormone-releasing intrauterine devices are not limited to single cycles.

Oral Contraceptives

OCs comprise combined oral contraceptives (COCs) and progesterone-only pills (POPs). Modern low-dose COCs contain a combination of oestrogen and a progestin (typically 20 to 35 mcg of oestrogen and 150 mcg or less of levonorgestrel, or 200 to 300 mcg of norgestrel or

Table 23.1. Selected contraceptive methods, their timing of use or duration of action, typical and perfect failure rates and continuation rates

Type	Method	Timing of use or duration of action	Percentage of women experiencing an unintended pregnancy within the first year of use		Percentage of women continuing use after one year
			Typical use	Perfect use	
Hormonal for women					
Oral contraceptives	Combined oral contraceptive	Daily	8	0.3	68
	Progestogen-only contraceptive	Daily	1 (breast-feeding)	0.5	–
Injectables	Depotmedroxy-progesterone acetate	3 months	3	0.3	56
	Combined injectable	Once a month	3	0.05	56
	Norethisterone acetate	2 months	3	0.3	56
Implants	Norplant	5 years	0.05	0.05	84
Vaginal rings	Combined	3–12 months	8	0.3	68
	Low-dose levonorgestrel	3–12 months	4.5	3.2	–
Intrauterine device	Levonorgestrel-releasing IUD	5–7 years	0.1	0.1	81
Contraceptive patch	Combined	1 day per week for 3 weeks followed by 1 week free	N/A	N/A	N/A
Non-hormonal for women					
Intrauterine device	Copper	8–10 years	0.8	0.6	78
Barrier	Female condom	Coitus-related	21	5	49
	Diaphragm w/spermicide	Coitus-related	16	6	57
	Cervical cap	Coitus-related			
	in nulliparous women		16	9	57
	in parous women		32	26	46
Spermicides	Spermicides	Coitus-related	29	18	42
Natural	Periodic abstinence	Daily	25	1–9	–
	Lactational amenorrhoea	Duration of breast-feeding	2	0.5	–
Sterilisation	Female	Permanent	0.5	0.5	100
Non-hormonal for men					
Barrier	Male condom	Coitus-related	15	2	53
Natural	Withdrawal	Coitus-related	27	4	43
Sterilisation	Male (vasectomy)	Permanent	0.15	0.1	100

Note: This table has been adapted from Table 1 in WHO Medical Eligibility Criteria for Contraceptive Use 2004.[2]

400 to 1000 mcg of norethindrone or the equivalent of another progestin). There are monophasic formulations, with constant daily doses of oestrogen and progestin, biphasic ones, in which the dose of progestin changes in each of two periods, and triphasic ones, in which the dosages change in each of the three seven-day periods of pill intake during the 21 days of pill cycle.

OCs require daily attention by the woman and they have a high discontinuation rate,[7,8] although this seems to have improved.[2] Both COCs and POPs are very effective under perfect use, and under typical use they are still effective (Table 23.1).

COCs prevent conception through the suppression of ovulation via hypothalamic and pituitary effects and progestin-mediated alterations in the consistency and properties of cervical mucus. It is still unconfirmed if the mechanism of action also includes alterations in the endometrial lining and of tubal transport mechanisms. POPs have a lower dose of progestin than do COCs (typically 75 mcg of norgestrel or 350 mcg of norethindrone). They prevent conception through a combination of mechanisms including suppression of ovulation, alteration of cervical mucus, of the endometrium and of the fallopian tubes.

Synthetic oestrogens were first developed in the early 1930s and the more potent ethinyl oestradiol in 1938, while synthetic orally active progestins were first produced in the early 1950s. In this decade the first generation progestins, like ethynodiol and lynesterol, were developed and OCs became available in the United States in 1959. A major breakthrough in the development of OCs was the finding that the oestrogen and progestin acted synergistically to inhibit the pituitary. This allowed the transition from high dose to low dose of both the oestrogen and the progestin. Information on efficacy and common side effects was obtained from randomised clinical trials (RCTs), with the findings that low-dose oestrogen COCs have less frequent complaints about breast tenderness, nausea and leg cramps[9] and that COCs have advantages over POPs.[10,11] The progestins that have been most widely studied are norethindrone (or norethisterone) and levonorgestrel (often called second-generation). Around the 1990s, three new progestins have been introduced (often called third-generation): norgestimate, desogestrel and gestodene.

OCs have been shown to affect lipid and carbohydrate metabolism as well as the coagulation system, and products which have minimal effects are preferred. The metabolic changes caused by modern hormonal methods are small.[12–15] However, it is not known exactly how predictive these changes are of cardiovascular problems. The association between OCs and cardiovascular diseases, in particular venous thrombosis (VTE), ischaemic heart disease and cerebrovascular disease, had been noted since the first oral contraceptives were introduced, and have been the subject of numerous epidemiological studies. Since the first products were introduced, there has been a progressive lowering in the doses of the oestrogen and progestogen components, a shift towards new synthetic hormones with fewer adverse pharmacological effects, as well as a more careful selection of the types of women who can or cannot safely use these methods. There is a two- to threefold increase in the risk of VTE among COC users, though the absolute risk in young women is very low. Considerable controversy surrounds the larger increase for the OCs containing third-generation progestins compared with those containing second-generation progestins. However, 'modern, low oestrogen dose OCs are extremely safe if used appropriately in young women'.[16] The most recent evidence suggests that myocardial infarction and stroke are rare and limited to users who smoke cigarettes or have hypertension or other cardiovascular risk factors.

Regarding the effect on cancer risk, studies have shown that the use of hormonal contraceptives is protective of cancer of the ovary and the endometrium.[17] However, OCs are not recommended for women with a history of breast cancer because of a possible increased risk in such women. Side effects of COCs are nausea, headaches, dizziness, spotting, weight gain, breast tenderness and chloasma. For POPs, the main side effect is menstrual irregularities.

OCs constitute the most common form of steroidal hormonal contraception and are also the most common method of reversible contraception in countries other than China. It is estimated that 60 to 80 million women are OC users worldwide.[18] COCs are a safe method of contraception, only not to be recommended for women with cardiovascular diseases, diabetes mellitus, breast cancer or smoking.[2] POPs can be taken by lactating women, but are not recommended in cases of thromboembolism or deep vein thrombosis.

Injectable Preparations

The most common injectable contraceptive is the progestin-only preparation depot-medroxyprogesterone acetate (DMPA) that provides contraceptive protection for three months. Norethisterone enantate (NET-EN) is also a progestin-only preparation that provides protection for two months. Combined oestrogen–progestin once-a-month injectable contraceptives are Cyclofem, which combines DMPA with an oestrogen, and Mesigyna, which combines NET-EN with an oestrogen. Injectable preparations are very effective contraceptives[2] (Table 23.1) and require the intervention of health care professionals to administer the injection at the required interval.

The mechanism of action of DMPA is suppression of ovulation and changes in the cervical mucus and the endometrium. Combined injectable preparations seem to have a mechanism of action similar to COCs.

DMPA was first used as a contraceptive in the 1960s. Subsequently other alternative injectable contraceptives were developed among which NET-EN gained widespread use. Once-a-month injectables were developed with the purpose of overcoming the inconvenience of the disruption of the menstrual bleeding pattern of progesterone-only preparations. WHO undertook the evaluation and optimisation of the dose and oestrogen/progesterone ratio of Cyclofem and Mesigyna.[19] Also, the Chinese Injectable No. 1, with a complicated administration schedule, was developed in China. A multicentre trial was important to decide between the 100 mg or 150 mg dose for DMPA.[20] A number of clinical trials showed that NET-EN was highly effective.[21] Other trials determined that NET-EN needs to be administered every two months and also compared DMPA and NET-EN.[22]

A large multicentre study provided reassurance that the use of DMPA was not associated with cancer and thus DMPA was registered in the United States as a long-acting contraceptive.[23,24] There is still a concern that long-term use of DMPA might be associated with low bone mass density. Headache is a common complaint; side effects are weight gain and delay in the return of fertility. Menstrual irregularities are frequent, including prolonged and heavy bleeding, mostly during the first months of use, and long periods of amenorrhoea. Progesterone-only contraceptives can be taken by breast-feeding women when they do not have access to other methods. They are not recommended before six weeks postpartum, since there might be a risk of exposure for the neonate, neither for women with multiple risk factors for arterial cardiovascular disease or with unexplained vaginal bleeding or with a history of or current breast cancer.[2]

It is estimated that about 16 million women worldwide are users of injectable contraceptives: 13 million DMPA users in 90 different countries, 1 million NET-EN users and 2 million once-a-month injectables users in Latin America and China, and introductory studies have been conducted in several other countries.[25]

Implants

Implants used for contraception in women consist of a silicone rubber tube or capsule inserted subdermally in the arm, containing a steroid or progestin released through it at a constant rate for several years. Implants are very effective[2] (Table 23.1) long-acting contraceptives and require the intervention of health care professionals for insertion and removal. Norplant is the most widely used implant, consisting of six levonorgestrel-releasing rods with contraceptive action during five years. Jadelle is a two-rod levonorgestrel implant with a five-year duration.

Implanon is a single-rod implant releasing 3-keto-desogestrel during three years. Another single implant releases ST 1435, a progestin rapidly metabolised, making the implant suitable for lactating women.

The mechanism of action, similarly to that of POPs, includes a combination of effects, there being indications that ovulation suppression is not the only one,[26] since only about 50% of women show suppression of progesterone levels.[27]

Norplant became available in the United States in 1991, after regulatory approval based on large clinical trials, which provided information on discontinuation rates and side effects.[28] Norplant II is a two-rod levonorgestrel implant easier to insert and remove and less conspicuous than Norplant, with a modified manufacturing design. The pregnancy rates were found to depend on the type of tubing used to manufacture the implant, the soft tubing being an improvement over the hard tubing.

The safety of Norplant has been confirmed in a post-marketing surveillance study.[29] Implants are considered safe, other than for occasional occurrences of infections at the implant site. Medical eligibility criteria are similar to those of other progestogen-only contraceptives mentioned above. The main side effect observed among Norplant users is disturbances in the menstrual bleeding pattern, with episodes of prolonged and heavy bleeding, mostly during the first months of use. Common complaints are headache, weight gain, mood change and depression. It is estimated that 1 million women are Norplant users.

Vaginal Rings

Vaginal rings are very effective contraceptive devices[2,30] (Table 23.1) releasing either a combination of a progestin and an oestrogen or only a progestin, the most common progestins being levonorgestrel and progesterone. They are attractive because they can be discontinued easily by the woman herself and are thus under her control, they do not require daily attention like the OCs, and they are not coitus-related like condoms.

The mechanism of action of levonorgestrel-releasing rings is similar to that of POPs, but with an increased effect on cervical mucus.

The first ring was progesterone-only, then the progesterone dose was reduced and combined rings were developed. Several designs were studied before an active core ring surrounded by an active silastic membrane was developed, leading to multi-compartment rings. A low-dose levonorgestrel contraceptive ring (20 mcg/day) was studied in WHO multicentre trials.[31,32]

Safety concerns related to the levonorgestrel ring are menstrual disturbances, vaginal symptoms, lesions and repeated expulsion. Medical eligibility criteria are similar to those of combined injectable contraceptives mentioned above.[2]

Emergency Contraception

Emergency contraception (EC) based on pills is a post-coital method that is recommended up to three to five days after an act of unprotected intercourse. The standard EC method was the Yuzpe regimen of COCs (ethinylestradiol 100 mcg and levonorgestrel 0.5 mg or *dl*-norgestrel 1.0 mg repeated 12 hours later). A superior regimen consists of two 0.75 mg doses of levonorgestrel 12 hours apart[33] or a single 1.5 mg dose.[34] Single doses of the anti-progestin mifepristone, ranging from 10 mg to 600 mg, is another method. EC is a backup method and cannot be used regularly.

In women receiving EC up to 72 hours after unprotected intercourse, 85% of pregnancies can be prevented with the recommended levonorgestrel treatment with typical use.[33] With the coitus-to-treatment interval extended to 120 hours, pregnancies prevented were in the range of 81% to 85% after 10 mg of mifepristone.[34,35] With perfect use, 89% pregnancies were prevented after levonorgestrel.

Regarding the mechanism of action, if unprotected intercourse occurs within a few days of ovulation, the only time when fertilisation can occur, ECs will exert their effect prior to implantation being completed (day 6–7 after fertilisation) and thus an established pregnancy would not be disrupted.[36,37] If EC is administered when a woman is already pregnant there is evidence from a study with pregnant women that 'ethinyl oestradiol is not a reliable abortifacient ... and that its efficiency as a postcoital contraceptive may be limited to a relatively short period following ovulation prior to implantation'.[38]

Although EC started in the 1980s in Europe with the Yuzpe regimen, it was only in 1997 that the FDA in the United States declared the regimen safe and effective.[39] The standard EC method until the late 1990s was the Yuzpe regimen, when levonorgestrel was shown in a trial to be more effective and have a better side-effect profile.[33] It was registered in 1999 in the United States and is now approved in over 100 countries around the world. Mifepristone as an EC method was initially used at the dose of 600 mg, until it was shown that doses as low as 10 mg can be used instead of higher doses, with the advantage that menses is not delayed as much as with higher doses.[35]

EC pills are relatively benign and they pose no serious safety concerns. Nausea and vomiting are common with high-oestrogen regimens. Levonorgestrel and mifepristone regimens have a better side-effect profile than the Yuzpe regimen.[33,35] A concern with mifepristone, mainly with high doses, is the delay of menses,[35] which is undesirable because it gives the opportunity for more acts of unprotected intercourse and is a source of anxiety for the woman.

NON-HORMONAL CONTRACEPTIVES FOR WOMEN

IUDs

IUDs are inert intrauterine rings or plastic devices with or without drug loading (copper or levonorgestrel). They are long-acting and require the intervention of health care professionals for insertion and removal. IUDs inserted after an unprotected coitus are also effective as EC. Levonorgestrel-releasing IUDs actually belong to hormone-releasing methods, but they are naturally included in this section.

Once an IUD is correctly inserted, women do not have to worry about compliance, but there could be discontinuation for several reasons. High efficacy, low risk and low discontinuation rates have been observed for new copper and for levonorgestrel-releasing devices in large trials.[40–43] IUDs are not recommended for women with current sexually transmitted infections or at risk of acquiring them.[2] If inserted in pregnant women they might cause abortion.

The mechanism of action of IUDs is to inhibit sperm and ovum transport and fertilisation.

IUDs were first introduced for contraceptive purposes at the beginning of the 1900s. The first IUD consisted of a loop of silk thread. Then a metal copper-releasing ring was developed. In the 1960s plastic coils became popular, such as the Lippes Loop. An IUD used in the 1970s, the plastic Dalkon Shield, was associated with high pregnancy rates and high infection rates. Safety problems with old devices included the risk of contracting pelvic inflammatory disease, and that of septic abortion and infertility, with consequent high discontinuation rates. There has been a progressive increase in effectiveness with continued research.[27] The superiority of collared Copper-T was established in RCTs published in 1975 comparing the Dalkon Shield with the Lippes Loop D and the new Copper-7 and Copper-T 200.[44–46] In the early 1980s trials were conducted including NOVA T, MLCu250, Copper-T 220C and MLCu 375.[47] In other trials the Copper-T 380 showed superiority over the MLCu 375.[48,49] Many IUD trials were conducted in China to try to design copper IUDs adapted to Chinese women. IUDs releasing 20 mcg/day of levonorgestrel constituted a major development and have been shown to be very effective.[40]

IUDs are the most commonly used contraceptive methods after sterilisation, and the most commonly used reversible method in China. It is estimated that about 120 million women are IUD users worldwide.[50]

Barrier Methods

Barrier methods used by women are the diaphragm, the female condom and spermicides. The importance of developing effective dual-protection barrier methods that provide protection against sexually transmitted diseases (STDs) has increased in the last few years with the spread of the HIV/AIDS epidemic. Condoms are barrier methods providing this feature of dual protection.

Barrier methods prevent conception by avoiding contact between sperm and the ovum. They act as a mechanical barrier (condom, diaphragm) or by inactivating the sperm (spermicides) or both (diaphragm with spermicide and cervical cap).

The female condom has become an important alternative because it is under the woman's control and can provide women with the ability of protecting themselves against STDs in situations where men refuse to use condoms. It is coitus-related and thus pregnancy can be the result of either method failure or inconsistency of use. It is effective under perfect use and only somewhat effective under typical use[2] (Table 23.1).

The diaphragm is an elastic membrane with cavity rim, which may be attractive to potential users but it lacks the advantage of protection against STDs. A new microbicide-releasing diaphragm is being developed to address this concern. It is not recommended for women with a history of toxic shock syndrome or allergic to latex.[2] It is effective under perfect use (Table 23.1).

Spermicides are in the form of creams, jellies, foams in pressurised containers, foaming tablets or suppositories. They are not very effective when used by themselves, but can be used in combination with other methods.[2] (Table 23.1). Self-administered topical preparations with spermicidal and microbicidal activities, such as cellulose sulphate gel, that provide both contraceptive and anti-infection protection and are under the control of the woman, have been developed,[51] and others are being studied.

The cervical cap or sponge is a mushroom-cap-shaped device releasing a spermicide and whose concave side is applied over the cervix. Its maximal insertion time is 24 hours. It is effective under perfect use (Table 23.1).

Natural Methods

Periodic abstinence restricts intercourse to the infertile phase of the woman's cycle, which depends on the ability of the woman to identify the fertile period.[52] It acts through prevention of fertilisation. It is effective under perfect use (Table 23.1).

The lactational amenorrhoea method is an accepted method of contraception when the interest of the woman is birth-spacing. It can be effectively used in women fully or nearly fully breast-feeding who are within six months of delivery and are amenorrhoeic.[2,53]

Sterilisation

Sterilisation in women is a very effective[2] (Table 23.1) surgical procedure involving the blockage of the fallopian tubes, which transport mature ova from the ovaries to the uterus. The most widely practised techniques are minilaparotomy and laparoscopy. Sterilisation is the only permanent contraceptive method and the most prevalent, since 180 million couples have been reported to be sterilised (male or female). A non-surgical procedure is under investigation.

Immunocontraceptives

Research is in progress for the development of a female vaccine based on the human chorionic gonadotrophin molecule (hCG), for action after fertilisation and before implantation.

HORMONAL CONTRACEPTIVES FOR MEN

Currently, the only widespread methods of fertility regulation for men are non-hormonal. Efforts are ongoing to develop additional options, in order to better meet couples' needs. Research on hormonal injectable methods for men that reduce the production of spermatozoa to infertile levels is based on results obtained from proof-of-concept studies investigating weekly injections of testosterone. Studies are in progress for the development of a longer-acting injectable androgen preparation, in combination with a progestin, e.g. DMPA or NET-EN, for male fertility control.

Implants for men are also under investigation. A depot progestin and androgen combination has recently demonstrated high contraceptive efficacy with satisfactory short-term safety and recovery of spermatogenesis.[54] In this relatively small trial, a hormonal implant was given every four months to replace testosterone and the progestin DMPA was injected every three months. A trial is being conducted to investigate the efficacy of various regimens of an etonogestrel implant, in combination with the long-acting injectable testosterone undecanoate, in suppressing sperm production to levels compatible with contraception.

NON-HORMONAL CONTRACEPTIVES FOR MEN

Available non-hormonal methods of fertility regulation for men are condoms (a barrier method), vasectomy (sterilisation) and withdrawal (a natural method).

Barrier Methods

Condoms used by men are tubes closed spherically on one side, normally made of a latex membrane 0.06 to 0.07 mm thick.[55] Most are lubricated, and some contain spermicides. They are effective if used correctly and consistently and only somewhat effective under typical use[2] (Table 23.1). The feature of dual protection and the mechanism of action are the same as those described for the female condom.

Research on the male condom has dealt with efficacy and acceptability issues. Old condoms were made of hard material, acceptability was low and they were not very resistant to adverse storage conditions. Improvement has been made with the latex condoms. A new polyurethane condom was compared with latex condoms in RCTs.[56]

The male condom is the most widely used barrier method but its use is not more widespread because it is often not accepted, mainly by the male partner. Condoms have practically no risk of side effects. The only concern has been allergy to latex in latex condoms.

Sterilisation

Surgical sterilisation, or vasectomy in the male, is an effective means of fertility control and is the only permanent contraceptive method for men. Vasectomy in the male is a simple procedure, very effective[2] (Table 23.1), but it is not well accepted in some cultural settings due to its non-reversibility and the requirement for incision. No-scalpel vasectomy has proved to be more acceptable and results in fewer and less severe side effects. Research is in progress for the development of a reversible procedure.

A possible association between vasectomy and prostate cancer was a safety concern, but observational studies have shown that if there is such an association, the increased risk in vasectomised men compared with non-vasectomised men is small.[57] Other studies have concluded that there is no increased risk of cancer or cardiovascular disease associated with vasectomy.[58,59]

Natural Methods

Withdrawal, or *coitus interruptus*, is a low-effectiveness method (Table 23.1) which depends on the man successfully withdrawing the penis from the vagina before ejaculation starts, and thus preventing fertilisation.

Immunocontraceptives

Research on vaccines for men is in progress based on antibodies that neutralise the biological effect of the gonadotrophin hormone-releasing hormone (GnRH) and follicle-stimulating hormone (FSH), with a resulting oligospermia or azoospermia

CLINICAL TRIAL METHODS IN CONTRACEPTION

Observational studies constitute the source of information for comparisons of efficacy, discontinuation rates or safety across broad classes of contraceptive methods, e.g. implants and IUDs. Women cannot usually be randomised to different broad classes of methods because the woman's choice of contraceptive is determined by social, cultural and personal reasons. An exception to this was one large RCT which allocated women at random to OCs or to vaginal methods (consisting of diaphragms, jellies, creams or foams).[60] However, results were difficult to interpret because there were many women switching methods and lost to follow-up.[61,62]

RCTs, on the other hand, have been an important tool to find new, safe and effective regimens or devices within each broad class of contraceptive methods and improve existing ones by answering questions about the best compound, the best dose, the best interval or route of administration (compounds) or the best physical properties (devices).

Sometimes partially randomised trials are used to compare two types of hormonal contraceptives within the same broad class with a non-hormonal one, used as a placebo control group. For example, in a WHO trial (data not published) on the effect of two injectable contraceptives (DMPA and NET-EN) on lipid and lipoprotein metabolism, women requesting an injectable contraceptive were allocated at random to one of two preparations, and a group of non-hormone-releasing IUD users was the control. In another WHO ongoing trial, two types of implants allocated at random to women choosing implants are compared with regard to efficacy in preventing pregnancies, and women in the implant groups are compared with a control group of IUD users to assess the effect of hormones on the bleeding pattern.

Clinical trials generally include an insufficient number of women to provide conclusive information on rare events, like cancer and cardiovascular diseases.[27] However, a careful documentation of serious adverse events and predisposing risk factors in the conduct of clinical trials should always be provided.[1]

General principles applying to the conduct of clinical trials and post-registration assessment of steroidal contraceptive drugs have been established.[1,63] The development of a new contraceptive method involves a long process until it is registered and reaches the market. The methodology used depends on the stage of development and will be treated separately for Phase I/II and Phase III trials.

PHASE I/II TRIALS

Objectives

Phase I trials deal with drug safety and aim to determine an acceptable drug dosage, and also study drug metabolism and bioavailability. In contraceptive research, Phase I trials are conducted to investigate the pharmacology of steroidal contraceptive or other contraceptive drugs in healthy volunteers. Phase I trials must be preceded by initial toxicity studies in rodents and pharmacokinetic studies in primates, which give an indication for the dose used in the first clinical study.[63]

When a contraceptive has been assessed to be safe in Phase I trials, research progresses to Phase II trials, using the optimal dose and administration schedule. Contraceptive Phase II trials are small-scale investigations into the effectiveness and safety of a contraceptive method, carried out on closely monitored patients. They have the goal to establish its mechanisms of action, metabolic effects and provide preliminary estimates of the frequency of common side effects, the effectiveness and the acceptability. It is recommended to previously conduct repeated-dose toxicity and reproduction studies in animals. Phase II studies are conventionally subdivided into Phase IIA, studies on the pharmacology of the drug in patients, and Phase IIB, definitive dose-finding studies.[63]

Since steroidal contraceptives are used by healthy people, it is desirable to assess the minimum effective dose at the initial stages of clinical testing.[63] This can be done in Phase I trials instead of in later stages. The direct assessment of efficacy of a steroid drug for pregnancy prevention in small trials is not possible because with reasonably effective contraceptives, pregnancy is a rare event. Phase I trials on contraceptives, therefore, are often also used to look at surrogates of efficacy in addition to safety issues, so that Phase I and Phase II trials are combined to evaluate both safety and endocrinological endpoints. Examples of surrogates of efficacy are hormone levels as indicators of inhibition of ovulation in contraceptives for women, sperm concentration in long-acting androgen–progestogen formulations for men as an indicator of inhibition of spermatogenesis, and amount of serum hCG antibodies in immunocontraceptive trials for women. Serological and clinical diagnoses of pregnancies are also conducted. In the case of hormonal contraceptives for women, clinical pharmacological parameters to assess the inhibition of ovulation have been described in detail.[1]

Recruitment, Design, Trial Size and Ethical Considerations

Recruitment into Phase I trials to study contraceptives for women is conducted among volunteers of reproductive age, not pregnant or lactating, regularly menstruating, identified in family planning clinics or selected community groups, who are IUD users or sterilised, and therefore not at risk of becoming pregnant. Users of hormonal contraceptives other than that being studied are not acceptable because their use might interfere with the assessment of clinical and laboratory parameters. Other selection criteria depend on the contraceptive being studied. For example, for contraceptive vaccines, acute hypersensitivity to the carrier should be excluded, and if reversibility cannot be assured, participants should be 25 years or older and have had children previously.

A series of sequential studies using different dose levels are conducted to assess the minimum effective dose. These studies involve doses that are two or three times the initial dose. For each dose level, a study is conducted in 10–20 healthy volunteers.[63]

Selection criteria for Phase II trials to study contraceptives for women are similar to those for Phase I trials, except that women should currently be exposed to the risk of pregnancy and have proven fertility. At this stage, if volunteers participating in Phase I studies are IUD users and they are willing to continue, then the IUD should be removed to assess efficacy.

Phase IIB trials require about 50–100 subjects to assess efficacy and side effects of the dosage determined in early trials (Phase I–IIA).

When conducting Phase I/II trials, the fact that contraceptive methods are used by healthy individuals implies a different risk/benefit assessment compared with therapeutic drugs for life-threatening conditions. This justifies the assessment of the minimum effective dose at early stages of development. When volunteers are advised on the risks and benefits of the study in order to seek their informed consent, the specific risks of receiving a steroidal contraceptive should be explained.

Examples of Phase I and Phase II clinical trials are the ones conducted with injectable preparations to evaluate well-known potent synthetic progestins in combination. A Phase I trial tested the use of progesterone as an alternative.[64]

Several examples for injectable contraceptives are summarised in a review by Newton *et al.*[25] An early pharmacological trial on Cyclofem with 11 women involved one pre-treatment cycle, a three-month treatment phase with an injection of Cyclofem every 28 days and then a three-month recovery phase. It confirmed that ovulation was inhibited and that inhibition of luteal activity persists after the last injection for several cycles.[65]

A comparative non-randomised study of Cyclofem and Mesigyna with 15 women, 8 receiving Cyclofem and 7 Mesigyna, involved one pre-treatment cycle, three treatment cycles of 28 days and a 90-day follow-up period. The results showed that the suppressive effect of Cyclofem was greater than Mesigyna.[66]

A four-arm trial of reduced dose of medroxyprogesterone acetate and oestradiol cypionate in Cyclofem recruited 88 women into the following groups: Cyclofem full dose, Cyclofem half dose, DMPA full dose, DMPA half dose. All four preparations were found to be effective in inhibiting ovulation for at least one month after the injection, and the combined preparations showed more regular bleeding patterns.[67]

Metabolic Studies

Specific Phase II studies on biochemical variables are conducted when required. These variables include lipid and lipoprotein metabolism, coagulation, fibrinolysis and platelet function as well as other physiological events such as vaginal blood loss.[63] The parallel group design has been the most common in this type of study, but factorial designs have also been used. The crossover design was applied in a metabolic study to compare three different progestogens (norgestrel, norethisterone and medroxyprogesterone) in treatment periods of three-week duration immediately preceded by three weeks of 'wash-out'.[66] Newton *et al.* describe examples of these studies for injectable once-a-month preparations.[25]

PHASE III TRIALS

Objectives

After a contraceptive is shown to be reasonably effective in Phase II trials, it is essential to compare it with any current standard contraceptive(s) within the same broad class in a large trial involving a substantial number of patients, with the goal of establishing its efficacy.[27] Phase III trials permit more refined estimates of safety, effectiveness and acceptability in comparison with a standard. In contraceptive research, this information provides the basis for introducing a hormonal contraceptive into family planning practice in field trials in various settings, as a prerequisite for registration with drug regulatory authorities (see introductory trials in the section on OTHER ISSUES below).

Design and Trial Size

The most common design to compare methods within each broad class of contraceptives has been the parallel group design, with simple randomisation in single-centre trials, and stratified by centre in multicentre trials. This was the case for the development of OCs,[9–11] injectables,[19,20,23,68] implants,[28] IUDs,[45–49] condoms[69] and EC regimens.[33–5,56]

The control used in RCTs to compare efficacy of methods is typically an active control, since a placebo control would not be ethical. Examples of comparisons of new versus standard, respectively, are the following: NET-EN versus DMPA (injectables), Norplant II versus Norplant (implants), steroid-releasing versus copper IUDs, polyurethane condom versus latex condom, Yuzpe versus levonorgestrel regimens (EC). Placebo controls have been used to assess efficacy of a treatment to improve the bleeding pattern disrupted by the use of progestin-only contraceptives.

In contraceptive trials, the main endpoint for efficacy is based on pregnancies, a rare event. The number of subjects required per group to detect as significant a difference between groups corresponding to a doubling of the rate, in a two-sided 5% level test, with 80% power, is usually large (1140 for a control rate of 2%, 4700 for a control rate of 0.5%).

When the effect of two factors is of interest and if an interaction between the two factors is likely to be present, the sample size needed for a four-arm trial is approximately double that for a two-arm single-factor trial. This might be a reason why factorial designs have not been commonly used in contraceptive efficacy trials. In the study of bleeding patterns among users of progestogen-only contraceptives, an example of a factorial design is provided by a trial comparing the effect of low-dose aspirin and vitamin E alone or in combination on Norplant-induced prolonged bleeding.[70]

For registration of a steroidal contraceptive, some regulatory agencies require clinical trials with 200 (FDA) or 400 subjects completing two years of observation, while some others require even fewer.[27] It is clear that this number does not provide sufficient power to detect a difference in rare events with the control. Nor does it provide sufficient precision for a confidence interval estimation of a rare event: five events observed in 200 subjects gives a rate of 2.5% with a 95% CI of 1% to 10%. On the other hand, the Committee for Proprietary Medicinal Products (CPMP) recommends that 20 000 cycles be observed, which at 13 28-day cycles per year is equivalent to 1540 women-years or 770 women followed completely for two years. This calculation is based on the criterion that the difference between the upper 95% confidence limit for the Pearl index (number of pregnancies per 100 women-years) and the point estimate does not exceed 1.[1]

Once effective contraceptives exist, a non-inferiority design is often needed to find a new treatment equally effective to the standard but presenting some other advantage, e.g. greater availability, reduced cost, fewer side effects or greater ease of administration, for instance one single dose rather than a split dose.[34] Because proof of exact equality is impossible, a pre-stated margin of non-inferiority (Δ) for the treatment effect in a primary patient outcome is defined. The new treatment will be recommended if it is similar to or better than an existing one, but not if it is worse (by more than Δ). Equivalence trials are very similar, except that equivalence is defined as the treatment effect being between $-\Delta$ and Δ. Most trials intended to show that a new contraceptive is equally effective to the standard with added advantages address the question of non-inferiority, and in the event the new treatment is found to be superior, this would be an added bonus. True (two-sided) equivalence trials, on the other hand, are relevant mainly to address bioequivalence hypotheses (e.g. to verify that two formulations of a drug are bioequivalent regarding metabolic parameters).

The rationale of testing for non-inferiority has been common in contraceptive trials, but very few trials have been both designed and reported as such. This sometimes resulted in underpowered trials to demonstrate non-inferiority or equivalence within a clinically relevant difference.[71]

An example of a trial with a non-inferiority rationale is given by the WHO Yuzpe–levonorgestrel trial.[33] The Yuzpe regimen using combined oral contraceptives had been used in EC as an effective method to prevent unwanted pregnancy. However, like other regimens containing oestrogen, it is associated with side effects like nausea and vomiting. The progestogen regimen levonorgestrel was shown to be better tolerated and equally or more effective, and it was recommended as a better alternative to the Yuzpe regimen. Another example is a trial establishing non-inferiority of a single dose compared with a split dose of 1.5 mg of levonorgestrel for EC.[34]

Recruitment

Participants in Phase III contraceptive trials are usually recruited in family planning clinics. A majority of clients requesting contraception in family planning clinics (other than STD clinics) are healthy. On arrival, subjects (women or men) or couples requesting or using the method under study are screened for eligibility. An eligibility criterion common to contraceptive efficacy trials is good general health, but others are specific to the contraceptive method, depending on the corresponding safety concerns and eligibility criteria.[2]

Trial participants are not therefore a random sample from women in reproductive age, and their particular characteristics affect external validity, making difficult the generalisation of results to wider populations.[6,72] First, women choosing a particular broad class of contraceptive are likely to be self-selected. For example, implants are often selected by older women. Second, clinicians are likely to select different types of women for different contraceptives. Third, women who are long-term users of a method and are happy with it do not come to the clinic and are less likely to be enrolled.

According to current ethical principles, all eligible subjects have to provide informed consent before being enrolled into the trial. In contraceptive trials, obtaining this consent from adolescents is problematic because some countries require a minimum legal age to provide consent. Consent from relatives is not always possible due to the need to maintain confidentiality in sensitive issues like contraception.

Randomisation, Allocation Concealment and Blinding

Randomisation and allocation concealment strategies for contraceptive RCTs are similar to those for RCTs in other areas.

Most clinical trials comparing implants, IUDs and other devices cannot be blinded because insertion or placement of the device usually implies that both the administrator and the user will see it, touch it or smell it. The situation is similar in sterilisation trials in which surgical procedures are compared. Some trials comparing IUDs or sterilisation techniques can be blinded to the woman but not to the device or procedure administrator. Depending on the treatments being compared, many clinical trials comparing injectable preparations, drugs for EC and possibly spermicides can be blinded to users, treatment administrators and outcome evaluators ('double-blind').

Blinding in contraceptive trials can avert bias after treatment allocation by preventing the following causes of bias. First, it is possible that the health care provider or the user will tend to discontinue one treatment more than the other. Second, ascertainment bias could be introduced in the evaluation of subjective outcomes, like lesions in contraceptive rings trials. The delay in the recognition of pregnancy, the imprecision in the estimate of the date of conception and the occurrence of chemical pregnancies noted above are sources of uncertainty which also pave the way for the introduction of ascertainment bias. Bias could still be present even in blinded trials due to unblinding caused by ancillary information, like differential side effects from the treatments being compared. For example, in EC trials, higher doses of a compound might cause nausea more frequently than lower doses.

Effectiveness and Efficacy of Contraceptive Methods: Theoretical Model

Effectiveness of a contraceptive method can be defined as 'the proportionate reduction in fecundability caused by the use of a method'.[6] As such, it is not measurable because one would have to compare the rate of conception under use of the method with that in the same population not practising contraception nor lactating. The common use of effectiveness is to denote how well a method works. Sometimes efficacy is used with this meaning.

Steiner *et al.*[73] proposed a theoretical model in which the couple's ability to conceive and the timing and frequency of intercourse determine the unobservable pregnancy rate in the absence of contraception. In the presence of (perfect or imperfect) contraceptive protection this pregnancy rate is reduced, determining the 'typical' pregnancy rate. This typical rate is composed of the perfect use pregnancy rate and the imperfect use pregnancy rate, weighted by the proportion of each type of user.

A measure of efficacy that implies a comparison with the same treated population under placebo is the proportion of pregnancies prevented out of the expected pregnancies, or preventable fraction, given by 1 − observed pregnancy rate/expected pregnancy rate.

The contraceptive method efficacy is the preventable fraction under conditions of perfect use, and the effectiveness is the preventable fraction under conditions of typical use. The difference between these two rates depends on both the pregnancy rates under each condition and the proportions of the two types of users.[52]

Estimation of Effectiveness and Efficacy of Emergency Contraceptives

The proportion of observed pregnancies (number of pregnancies divided by the number of women treated) is a crude (inverse) measure of how well a method works, but it is affected by the distribution of timing of intercourse with respect to women's menstrual cycle. In order to obtain this information, common eligibility criteria for women recruited into EC trials are to have had a single unprotected act of intercourse within the last three to five days and to report its date and the date of onset of the last menstrual period. The number of pregnancies occurring in the same population under no use of method, or expected pregnancies, is unobservable, but it can be estimated using external probabilities of conception. The number of expected pregnancies on a cycle day is estimated by multiplying the number of women having unprotected intercourse on that day by the probability of conception on that day, and then the expected pregnancies are summed over all days of the cycle.[74] The proportion of pregnancies prevented, or preventable fraction, is given by 1 − (observed pregnancies/expected pregnancies).

A technique for the construction of confidence intervals for the preventable fraction is available, using variance–covariance matrices from the external estimates of conception probabilities.[74] Since in large trials it is expensive to conduct daily monitoring of follicular growth or urinary metabolites, the day of ovulation, and thereof the day of the cycle in which intercourse took place, is estimated from the date of the last menstrual period as reported by women, and thus subject to imprecision.

The success of EC depends on not having further unprotected acts of intercourse during the same cycle, since the EC treatment does not prevent pregnancies resulting from these acts.[33–35] Therefore user compliance can affect the effectiveness of the method. If the EC treatment includes two doses, its success also depends on the treatment compliance, i.e. on the woman taking the second dose (at home) at the correct interval.

Thus, in trials comparing EC regimens, the preventable fraction under typical use, including all subjects, estimates the effectiveness, and that under perfect use, including only perfect users, estimates the efficacy.

Estimation of Effectiveness and Efficacy of Regular Use Contraceptives

In large trials comparing regular use contraceptives it is difficult to obtain data on pattern and timing of intercourse, therefore the common inverse measure of how well a contraceptive method works in preventing pregnancy is failure, or the occurrence of pregnancy in the period of time during which the contraceptive is used.[6] Thus, the pregnancy rate is used as an inverse measure of efficacy.

The estimation of the pregnancy rate using the Pearl index (number of pregnancies divided by woman-years of observation, typically expressed per 100 woman-years) has been shown to be not appropriate due to the decline in fertility of the cohort with duration of the contraceptive use. This decline in fertility has been illustrated by Sivin and Schmidt[47] from long-term IUD studies, where a progressive increase in the effectiveness of each device with age was observed, as well as a wider difference in failure rates among devices and a progressive increase in effectiveness with continued research.

Life table techniques have been used in the analysis of regular use contraceptive trials, using the single decrement method, in which women who exit for other reasons than pregnancy are censored at the time of exit.[52,72,75] The estimation of the pregnancy rate is given by the cumulative life table rate (net rate). The daily life table method, using the Kaplan–Meier product-limit estimate of the cumulative pregnancy probabilities, gives similar results and leads naturally to the log rank test to compare groups.[75]

A difficulty in the estimation of pregnancy rates is the presence of reasons for discontinuation other than pregnancy. For IUDs, the commonly analysed discontinuation reasons are expulsion, medical removal (due to pain, bleeding or pelvic inflammatory disease), non-medical removal (wish to become pregnant, no further need of contraception) and loss to follow-up. The use of net rates from life table techniques deals with competing causes by censoring, assuming independence of the different reasons for discontinuation, which might overestimate the rate for each reason.[76] Kaplan–Meier estimates are appropriate when comparing the effectiveness of contraceptive methods, but cumulative incidence estimates might be more appropriate when making programmatic decisions regarding contraceptive methods.[77]

For reversible methods (e.g. IUDs and long-acting hormonal methods), the assessment of the pregnancy status might be difficult due to the following sources of uncertainty:[52,72] (1) when the decision to stop using a method is made, the pregnancy might be recognised after the method is stopped; (2) imprecision in the estimate of the date of conception when the estimate is based on the date of start of the last menstrual period; (3) occurrence of early 'chemical' pregnancies, of which a considerable percentage are lost before reaching the stage of clinical pregnancy; and (4) early foetal losses, which might be unnoticed by the woman.

In clinical trials comparing regular use contraceptives, women are usually required to return to the clinic at specified intervals during a follow-up period. The timing of reporting pregnancies varies among women. It is important that the method of counting pregnancies does not depend on this timing. The 'active follow-up' prevents this problem by defining a cut-off date and contacting women three months later to learn their pregnancy status at the cut-off date.[72]

One of the main problems affecting data quality in trials comparing regular use contraceptives that require long periods of follow-up, is the loss to follow-up. Bardin and Sivin[27] discuss the bias introduced in comparative trials by the failure to observe all subjects through the completion of the study. The magnitude of the bias depends on the proportion of subjects lost to follow-up and the outcome mean or proportion in this group.

The Importance of Behavioural Patterns in the Estimation of Effectiveness and Efficacy

Sterilisation, which acts continuously and is permanent, and methods which act continuously but are reversible, like IUDs and long-acting hormonal methods, are non-coitus-related in the sense that they do not require any particular action by the user to be effective. For these methods, effectiveness (preventable fraction under conditions of typical use) and efficacy (preventable fraction under conditions of perfect use) are the same.

Methods that are used around the time of intercourse, like barrier methods and spermicides, are coitus-related and require a high degree of user compliance with the correct way of using the method in order to prevent pregnancy reliably.[52] For these methods, a pregnancy can be the result of a method failure or lack of use or incorrect use. OCs are not coitus-related but have to be taken daily by women, posing similar types of problems. Similarly, periodic abstinence relies for its effectiveness on rules of when to abstain from sexual intercourse in order to avoid pregnancy, and users may depart from these rules.

In order to separate a method failure from a lack of use or an incorrect use of a method which is coitus-related, investigators denoted pregnancies in which the method had not been used or had been incorrectly used as 'user failures'. Pregnancies in which the method had been correctly used were denoted as 'method failures'. Trussell[72] illustrated the inadequacy of computing pregnancy rates corresponding to these two sources using the same denominator that includes all exposure from both 'perfect' and 'imperfect' use. He proposed to collect information on 'imperfect' use for all months of exposure, or alternatively obtain information on correct and consistent use at the end of the trial, and calculate separate rates for 'perfect' users and for 'imperfect' users.

For comparative trials, this issue is addressed by conducting a stratified analysis by imperfect and perfect use. The comparison of the effectiveness between treatments for all cycles (whether perfect or imperfect use took place) provides the treatment effect under conditions of typical use. The comparison of the efficacy between treatments is given by a subgroup analysis with cycles of perfect use.

Caveats in Comparing Efficacy and Effectiveness Between Groups: Intention-to-Treat and Subgroup Analysis

In RCTs, the comparison of estimates of effectiveness obtained with two treatments corresponds to an intention-to-treat (ITT) analysis (analysing all patients within their randomised groups, regardless of whether they completed allocated treatment, which is only possible in the absence of lost to follow-up), while that of efficacy corresponds to a subgroup analysis of perfect users. In large RCTs, the proportions of perfect and imperfect users are likely to be similar in the treatments being compared, so that differences in effectiveness between two treatments will depend mainly on differences between the pregnancy rates under the two treatments. Thus, the comparisons of effectiveness between treatments within the RCT are not biased (internal validity).

On the other hand, the comparison of efficacy between treatments has the limitations of a subgroup analysis. In the first place, the advantages of randomisation are lost, since imperfect users are excluded from the analysis. When the subgroup analysis is based on subject characteristics that are not affected by treatment, like baseline variables, each smaller subgroup is like a smaller randomised trial.[78] But when the subgroup is defined by a variable observed after randomisation and potentially affected by the treatment, then the treatment effect may influence classification into the subgroup. The treatment effect observed in the subgroup would then be biased. This caveat is illustrated by an RCT to compare mifepristone and levonorgestrel for EC. The main variable to define perfect use is adhering to the protocol requirement of not having further acts of unprotected intercourse before the start of next menses. Mifepristone is known to delay ovulation and thus is associated with a delay in the start of menses, while levonorgestrel is not.[34,35] This provides women under both regimens with a differential opportunity to violate the requirement, and then the effects of treatment under perfect use and under typical use are likely to be of different magnitude. In the second place, unless the trial was designed to have sufficient power at the subgroup level, a relevant treatment effect in the subgroup will not be detected, or the confidence interval estimate of the effect will be imprecise. Stratification into perfect and imperfect users is another way of reporting results but subject to the same caveats.[33,35]

Assessment of Side Effects and Acceptability of Contraceptive Methods

Expected side effects and complaints such as nausea, vomiting, diarrhoea, fatigue, dizziness, headache, lower abdominal pain and breast tenderness, as well as adverse events, can be collected in follow-up visits. Differences between groups in events which have a rate of 5 or more per 100 can be detected with small trials. Rates of 1 to 5 per 100 require larger trials. Detection of differences for lower rates would require even larger trials.[27]

Acceptability of a contraceptive method depends not only on the characteristics of the method but on the service delivery setting and the socio-demographic and economic factors of a particular country.[25] It can be assessed by questions to the user regarding satisfaction, willingness to recommend the method to others and to pay to have access to the method. Many side effects of regular use contraceptives are reflected in discontinuation. Some of these discontinuation reasons are related to the acceptability of the method by the user. For long-acting hormonal methods, for example, the main discontinuation reason is disturbances in the menstrual bleeding pattern and is largely determined by cultural and social factors. An Egyptian study on the acceptability of once-a-month injectable contraceptives found differences between women discontinuing and those continuing in all measures of acceptability.[79] Factors important in determining acceptability were: age, contraceptive history, learning about injectables, the husband's attitude and knowing about another user's satisfaction.

OTHER ISSUES

Vaginal Bleeding Patterns

Hormonal contraception is often associated with disturbances in the vaginal bleeding pattern. These disorders are common with the use of progestogen-only hormonal methods and they do not imply a health risk per se, since it has been shown that the amount of blood loss is not a problem. They may be tolerated by the woman, and this depends on cultural and behavioural patterns. The measurement of bleeding patterns can be achieved by direct questions to women, by their completing menstrual diaries or by measuring blood loss.[27]

The most used method of analysis of menstrual diaries is the reference period method,[80] which was standardised by WHO using a 90-day reference period.[81] It consists of analysing bleeding/spotting records in women's menstrual diary cards by taking fixed-length segments of time (the reference period, for which a 90-day segment has been used as a convention) and deriving measures of bleeding pattern, or indices. The following 10 indices have been recommended:[82] number of bleeding/spotting days, number of spotting days, number of bleeding/spotting episodes, number of spotting-only episodes, mean, range and maximum value of lengths of bleeding/spotting episodes, mean, range and maximum value of lengths of bleeding-free intervals. These indices have been analysed using box–whisker plots and non-parametric analysis techniques. To summarise the information provided, Belsey and Carlson[83] conducted a principal component analysis with data from women using different contraceptives, and concluded that most of the essential information about a woman's bleeding pattern was contained in four indices: number of bleeding/spotting episodes, mean length of episodes, mean length of bleeding-free intervals and the range of bleeding-free interval lengths. Based on the indices, the following clinically important patterns are derived:[82] no bleeding (amenorrhoea), prolonged, frequent, infrequent and irregular bleeding.

The 90-day reference period method was applied to diary data collected from women treated with Cyclofem, Mesigyna, a low-dose levonorgestrel-releasing ring and DMPA taking part in Phase III WHO clinical trials. Among women using once-a-month injectable and the levonorgestrel-releasing ring the incidence of acceptable patterns was higher than among DMPA users, although the patterns were different from those of normally menstruating women.[84]

Several placebo-controlled RCTs have been conducted to investigate the therapeutic effectiveness of one or more treatments for bleeding irregularities. An example is given by a trial comparing the bleeding pattern of untreated DMPA users with groups treated with ethinyl oestradiol or oestrone sulphate.[85]

Introductory Trials and Phase IV Trials

Introductory trials are field studies to assess acceptability, effectiveness, continuation of use, side effects and service-related needs of a method in specific populations, in the context of family planning services.[63] They are expanded Phase III trials. Some countries may require to conduct these trials in a network of 5–10 centres, including an acceptability component. Such studies might involve 1000 to 5000 subjects.

Phase IV trials are those conducted after a drug has been approved for marketing, to further investigate adverse effects of the drug. Very rare events cannot be rigorously assessed before the contraceptive drug reaches the market because even Phase III trials do not have sufficient power. Strategies for post-registration surveillance of contraceptive drugs are reports of adverse reactions, large-scale experimental studies, formal epidemiological studies and indirect correlational studies.[63] The most commonly used strategy consists of epidemiological studies. Post-registration RCTs are costly, lack sufficient power to detect uncommon but important reactions, cannot last long enough to identify long-term effects and the experimental group cannot be compared with a placebo.[63,86] This last limitation implies that when comparing two active treatments through an RCT, the absence of effect does not mean that either has no risk compared with a placebo. Another limitation of RCTs as a strategy at this stage of development of the contraceptive is that RCTs are conducted on healthy women and the risk of adverse reactions might be relevant in women with risk conditions.[62]

Systematic Reviews

Systematic reviews on contraceptive methods are available in the Cochrane Library.[87] A search was done using the word 'contraception' in the title, abstract or keywords, obtaining 36 results out of 4200 records. Only complete reviews addressing comparisons of efficacy, side effects or acceptability were included. Reviews including contraceptives as treatment for complications or diseases, those comparing treatments for complications due to contraceptive use, those on subfertility and education were not included. The title and if necessary the abstract were examined to assess whether the review was eligible. The 25 reviews satisfying these criteria are listed in Table 23.2.

As an example, the systematic review 'Interventions for emergency contraception' included 33 trials, most of which were conducted in China. The authors conducted 46 meta-analyses with different comparisons and various outcomes comprising efficacy (pregnancies) and side effects, including delay of menses. For the mifepristone dose comparisons they grouped the doses used in different trials in low, mid and high doses.[88]

ACKNOWLEDGEMENTS

The authors are grateful to Dr Paul David Griffin for proof-reading the section on Phase I/II trials and to Dr Kirsten M. Vogelsong for her contribution to the sections on contraceptives for men.

EP is supported by the UNDP/UNFPA/WHO/World Bank Special Programme of Research, Development and Research Training in Human Reproduction, Department of Reproductive Health and Research, WHO, Geneva, Switzerland.

The views expressed are solely those of the authors and do not necessarily reflect those of the World Health Organization.

Table 23.2. Systematic reviews in The Cochrane Database of Systematic Reviews addressing efficacy or side effects of contraceptive methods

Method	Review
Oral contraceptives	Biphasic versus monophasic oral contraceptives for contraception
	Biphasic versus triphasic oral contraceptives for contraception
	Continuous or extended cycle vs cyclic use of combined oral contraceptives for contraception
	Progestogens in combined oral contraceptives for contraception
	20 mcg versus > 20 mcg estrogen combined oral contraceptives for contraception
	Skin patch and vaginal ring versus combined oral contraceptives for contraception
Injectables	Combination injectable contraceptives for contraception
More than one hormonal method	Combined hormonal versus nonhormonal versus progestin-only contraception in lactation
	Strategies to improve adherence and acceptability of hormonal methods for contraception
	Combination contraceptives: effect on weight
Hormonal methods for men	Steroid hormones for contraception in men
Emergency contraception	Interventions for emergency contraception
Intrauterine devices	Frameless versus classical intrauterine device for contraception
	Hormonally impregnated intrauterine systems (IUSs) versus other forms of reversible contraceptives as effective methods of preventing pregnancy
	Immediate post-partum insertion of intrauterine devices
	Immediate postabortal insertion of intrauterine devices
Barrier	Condom effectiveness in reducing heterosexual HIV transmission
	Diaphragm versus diaphragm with spermicides for contraception
	Sponge versus diaphragm for contraception
	Cervical cap versus diaphragm for contraception
	Non-latex versus latex condoms for contraception
	Spermicide used alone for contraception
Natural	Lactational amenorrhoea for family planning
	Fertility awareness-based methods for contraception
Sterilisation	Minilaparotomy and endoscopic techniques for tubal sterilisation

REFERENCES

1. Committee for Proprietary Medicinal Products (CPMP). Clinical investigation of steroid contraceptives in women. The European Agency for the Evaluation of Medicinal Products (2000).
2. World Health Organization, Dept. of Reproductive Health and Research. *Medical Eligibility Criteria for Contraceptive Use*, 3rd edn. Geneva: WHO (2004).
3. World Health Organization, Dept. of Reproductive Health and Research. *Selected Practices Recommendations for Contraceptive Use*, 2nd edn. Geneva: WHO (2005).
4. Van Look PFA, Pérez-Palacios G. *Contraceptive Research and Development 1984 to 1994*. Oxford: Oxford University Press (1994).
5. Rabe T, Runnebaum B. *Fertility Control Update and Trends*. Berlin: Springer (1999).
6. Trussell J, Kost K. Contraceptive failure in the United States: a critical review of the literature. *Stud Fam Plann* (1987) **18**(5): 237–83.
7. Hagenfeldt K. Contraceptive research and development today: an overview. In: Van Look PFA, Pérez-Palacios G, eds, *Contraceptive Research and Development 1984 to 1994*. Oxford: Oxford University Press (1994) 3–22.
8. Rivera R. Oral contraceptives: the last decade. In: Van Look PFA, Pérez-Palacios G, eds, *Contraceptive Research and Development 1984 to 1994*. Oxford: Oxford University Press (1994).
9. Dionne P, Vickerson F. A double-blind comparison of two oral contraceptives containing 50 μg. and 30μg. ethinyl estradiol. *Curr Ther Res Clin Exp* (1974) **16**(4): 281–8.
10. WHO Task Force on Oral Contraceptives. A randomized, double-blind study of six combined oral contraceptives. *Contraception* (1982) **25**: 231–41.
11. WHO Task Force on Oral Contraceptives. A randomized, double-blind study of two combined and two progesterone-only oral contraceptives. *Contraception* (1982) **25**: 243–52.
12. Fotherby K. Update on lipid metabolism and oral contraception. *Br J Fam Plann* (1990) **15**: 23–6.
13. Gaspard UJ. Metabolic effects of oral contraceptives. *Am J Obstet Gynecol* (1987) **157**: 1029–41.
14. Runnebaum B, Rabe T. New progestogens in oral contraceptives. *Am J Obstet Gynecol* (1987) **157**: 1059–63.
15. Sabra A, Bonnar J. Hemostatic system changes induced by 50 mcg and 30 mcg estrogen/progestogen oral contraceptives. *J Reprod Med* (1983) **28**: 85–91.
16. Farley TMM, Meirik O, Collins J. Cardiovascular disease and combined oral contraceptives: reviewing the evidence and balancing the risks. *Hum Reprod Update* (1999) **5**(6): 721–35.
17. WHO. Oral contraceptives and neoplasia. Report of a WHO Scientific Group, 817. WHO Technical Report Series (1992).
18. Rabe T, Runnebaum B. The future of oral hormonal contraception. In: Rabe T, Runnebaum B, eds, *Fertility Control Update and Trends*. Berlin: Springer (1999) 73–89.
19. WHO Task Force of Long-Acting Systemic Agents for Fertility Regulation. A multicentred phase III comparative study of two hormonal contraceptive preparations given once-a-month by intramuscular injection. II. The comparison of bleeding patterns. *Contraception* (1989) **40**(5): 531–51.
20. WHO Task Force of Long-Acting Systemic Agents for Fertility Regulation. A multicentred phase III comparative trial of 150 mg and 100 mg of depot-medroxyprogesterone acetate given every three months: Efficacy and side-effects. *Contraception* (1986) **34**: 223–35.
21. Hall P. Long-acting formulations. In: Diczfalusy E, Bygdeman M, eds, *Fertility Regulation Today and Tomorrow*. New York: Raven Press (1987) 119–41.
22. WHO Task Force of Long-Acting Systemic Agents for Fertility Regulation. A multinational comparative clinical trial of long-acting injectable contraceptives: norethisterone enantate given in two dosage regimens and depot-medroxyprogesterone acetate, final report. *Contraception* (1983) **28**: 1–20.
23. WHO Collaborative Study of Neoplasia and Steroid Contraceptives. Breast cancer and depot-medroxyprogesterone acetate: a multinational study. *Lancet* (1991) **338**: 833–8.
24. WHO Collaborative Study of Neoplasia and Steroid Contraceptives. Depot-medroxyprogesterone acetate (DMPA) and risk of epithelial ovarian cancer. *Int J Cancer* (1991) **49**: 191–5.
25. Newton JR, d'Arcangues C, Hall PE. Once-a-month combined injectable contraceptives. *J Obstet Gynaecol* (1994) **14**: S1–34.
26. Croxatto HB, Diaz S, Miranda P, Elamsson K, Johansson ED. Plasma levels of levonorgestrel in women during longterm use of Norplants. *Contraception* (1980) **22**(6): 583–96.
27. Bardin CW, Sivin I. Lessons learnt from clinical trials. In: Michal F, ed, *Safety Requirements for Contraceptive Steroids*. Geneva: WHO (1989) 400–13.
28. Sivin I. International experience with Norplant and Norplant-2 contraceptives. *Stud Fam Plann* (1988) **19**: 81–94.
29. International Collaborative Post-Marketing Surveillance of Norplant. Post-marketing surveillance

of Norplant® contraceptive implants: II. Non-reproductive health(1). *Contraception* (2001) **63** (4): 187–209.

30. Sivin I, Mishell DR Jr, Victor A, Diaz S, Alvarez-Sanchez F, Nielsen NC *et al*. A multicenter study of levonorgestrel–estradiol contraceptive vaginal rings. I-Use Effectiveness. An international comparative trial. *Contraception* (1981) **24**(4): 377–92.
31. Koetsawang S, Ji G, Krishna U, Cuadros A, Dhall GI, Wyss R *et al*. Microdose intravaginal levonorgestrel contraception: a multicentre clinical trial. I. Contraceptive efficacy and side effects. World Health Organization. Task Force on Long-Acting Systemic Agents for Fertility Regulation. *Contraception* (1990) **41**(2): 105–24.
32. Koetsawang S, Ji G, Krishna U, Cuadros A, Dhall GI, Wyss R *et al*. Microdose intravaginal levonorgestrel contraception: a multicentre clinical trial. II. Expulsions and removals. World Health Organization. Task Force on Long-Acting Systemic Agents for Fertility Regulation. *Contraception* (1990) **41**(2): 125–41.
33. WHO Task Force on Postovulatory Methods of Fertility Regulation. Randomised controlled trial of levonorgestrel versus the Yuzpe regimen of combined oral contraceptives for emergency contraception. Task Force on Postovulatory Methods of Fertility Regulation. *Lancet* (1998) **352**(9126): 428–33.
34. Von Hertzen H, Piaggio G, Ding J, Chen J, Sonj S, Bártfai G *et al*. Low dose mifepristone and two regimens of levonorgestrel for emergency contraception: a WHO multicentre randomised trial. *Lancet* (2002) **360**: 1803–10.
35. WHO Task Force on Postovulatory Methods of Fertility Regulation. Comparison of three single doses of mifepristone as emergency contraception: a randomised trial. Task Force on Postovulatory Methods of Fertility Regulation. *Lancet* (1999) **353**(9154): 697–702.
36. Wilcox AJ, Weinberg CR, Baird DD. Timing of sexual intercourse in relation to ovulation: effects on the probability of conception, survival of the pregnancy, and sex of the baby. *New Engl J Med* (1995) **333**(23): 1517–21.
37. Wilcox AJ, Weinberg CR, Baird DD. Post-ovulatory ageing of the human oocyte and embryo failure. *Hum Reprod* (1998) **13**(2): 394–7.
38. Bacic M, Wesselius dC, Diczfalusy E. Failure of large doses of ethinyl estradiol to interfere with early embryonic development in the human species. *Am J Obstet Gynecol* (1970) **107**(4): 531–4.
39. Grimes DA. Emergency contraception expanding opportunities for primary prevention. *New Engl J Med* (1997) **337**(15): 1078–9.
40. Sivin I, Stern J, Coutinho E, Mattos CE, el Mahgoub S Diaz S, *et al*. Prolonged intrauterine contraception: a seven-year randomized study of the levonorgestrel 20 mcg/day (LNg 20) and the Copper T380 Ag IUDS. *Contraception* (1991) **44**(5): 473–80.
41. Luukkainen T, Allonen H, Nielsen NC, Nygren KG, Pyorala T. Five years' experience of intrauterine contraception with the Nova-T and the Copper-T-200. *Am J Obstet Gynecol* (1983) **147**(8): 885–92.
42. WHO Task Force on the Safety and Efficacy of Fertility Regulation. The TCu 380A, the TCu 220C, Multiload 250 and Nova T IUDs at 3, 5 and 7 years of use. Results from three randomized trials. *Contraception* (1990) **42**: 141–58.
43. WHO. Long-term reversible contraception. Twelve years of experience with the TCu380A and TCu220C. *Contraception* (1997) **56**(6): 341–52.
44. Tatum HJ. Comparative experience with newer models of the Copper T in the United States. In: Hefnawi F, Segal SJ, eds, *Analysis of Intrauterine Contraception*. New York: American Elsevier Publications (1975) 155–63.
45. Jain AK. Safety and effectiveness of intrauterine devices. *Contraception* (1975) **11**(3): 243–59.
46. Sivin I, Stern J. Long-acting, more effective copper T IUDs: a summary of U.S. experience, 1970–75. *Stud Fam Plann* (1979) **10**(10): 263–81.
47. Sivin I, Schmidt F. Effectiveness of IUDs: a review. *Contraception* (1987) **36**(1): 55–84.
48. Champion CB, Behlilovic B, Arosemena JM, Randic L, Cole LP, Wilkens LR. A three-year evaluation of TCu 380 Ag and multiload Cu 375 intrauterine devices. *Contraception* (1988) **38**(6): 631–9.
49. Cole LP, Potts DM, Aranda C, Behlilovic B, Etman ES, Moreno J *et al*. An evaluation of the TCu 380Ag and the Multiload Cu375. *Fertil Steril* (1985) **43**(2): 214–17.
50. Wagner H. Intrauterine contraception: past, present and future. In: Rabe T, Runnebaum B, eds, *Fertility Control Update and Trends*. Berlin: Springer (1999) 151–71.
51. Malonza I, Mirembe F, Nakabiito C, Odusoga LO, Osinupebi OA, Hazari K, Chitlange S, Ali MM, Callahan M, Van Damme L. Expanded Phase I safety and acceptability study of 6% cellulose sulphate vaginal gel. *AIDS* (2005) **19**: 2157–63.
52. Farley TM. Reproduction. In: Armitage P, Colton T, eds, *Encyclopedia of Biostatistics*. Chichester: John Wiley & Sons, Ltd (1998).
53. Labbok M, Jennings VH. Advances in fertility regulation through ovulation prediction during lactation (lactational amenorrhoea method) and during the menstrual cycle. In: Van Look PFA,

Pérez-Palacios G, eds, *Contraceptive Research and Development 1984 to 1994*. Oxford: Oxford University Press (1994).

54. Turner L, Conway AJ, Jimenez M, Liu PY, Forbes E, McLachlan RI, Handelsman DJ. Contraceptive efficacy of a depot progestin and androgen combination in men. *J Clin Endocrinol Metab* (2003) **88**(10): 4659–67.
55. Rabe T, Vladescu E, Runnebaum B. Contraception: historical development, current status and future aspects. In: Rabe T, Runnebaum B, eds, *Fertility Control Update and Trends*. Berlin: Springer (1999) 29–72.
56. Population Council Center for Biomedical Research. *Annual Report 1991*. New York: The Population Council (1991).
57. Farley TMM, Meirik O, Metha S, Waites GMH. The safety of vasectomy: recent concerns: *Bull WHO* (1993) **71**(3–4): 413–19.
58. Cox B, Sneyd MJ, Paul C, Delahunt B, Skegg DCG. Vasectomy and risk of prostate cancer. *JAMA* (2002) **287**: 3110–15.
59. Goldacre MJ, Wotton CJ, Seagroatt V, Yeates D. Cancer and cardiovascular disease after vasectomy: an epidemiological database study. *Fertil Steril* (2005) **84**: 1438–43.
60. Fuertes-de la Haba A, Bangdiwala IS, Pelegrina I. Success of randomization in a controlled contraceptive experiment. *J Reprod Med* (1973) **11**(4): 142–8.
61. Potts M, Feldblum PJ, Chi I, Liao W, Fuertes-de La HA. The Puerto Rico oral contraceptive study: an evaluation of the methodology and results of a feasibility study. *Br J Fam Plann* (1982) **7**: 99–103.
62. Skegg DCG. The epidemiological assessment of safety of hormonal contraceptives: a methodological review. In: Michal F, ed, *Safety Requirements for Contraceptive Steroids*. Geneva: WHO (1989) 21–37.
63. Special Programme of Research, Development and Research Training in Human Reproduction. Guidelines for the toxicological and clinical assessment and post-registration surveillance of steroidal contraceptive drugs. In: Michal F, ed., *Safety Requirements for Contraceptive Steroids*. Geneva: WHO (1989) 415–54.
64. Garza-Flores J, Fatiikun T. Future directions in fertility regulation. In: Negro-Vilar A, Pérez-Palacios G, eds, *Reproduction, Growth and Development*. New York: Raven Press (1991) 383–97.
65. Fotherby K, Benagiano G, Toppozada HK *et al.* A preliminary pharmacological trial of the monthly injectable contraceptive Cycloprovera. *Contraception* (1982) **25**: 261–72.
66. Aedo AR, Landgren BM, Johannisson E, Diczfalusy E. Pharmacokinetic and pharmacodynamic investigations with monthly injectable contraceptive preparations. *Contraception* (1985) **31**: 453–69.
67. WHO Task Force of Long-Acting Systemic Agents for Fertility Regulation. A multicentred pharmacokinetic, pharmacodynamic study of once-a-month injectable contraceptives. I. Different doses of HRP112 and of DepoProvera. *Contraception* (1987) **36**(4): 441–57.
68. WHO Task Force of Long-Acting Systemic Agents for Fertility Regulation. A multicentred phase III comparative study of two hormonal contraceptive preparations given once-a-month by intramuscular injection: I. Contraceptive efficacy and side effects. World Health Organization. *Contraception* (1988) **37**(1): 1–20.
69. Family Health International. *Contraceptive technology and family planning research*. Family Health International (1992).
70. d'Arcangues C, Piaggio G, Brache V, Ben Haissa R, Hazelden C, Mossai R *et al*. Effectiveness of vitamin E and low-dose aspirin, alone or in combination, on Norplant induced prolonged bleeding. *Contraception* (2004) **70**: 451–62.
71. Piaggio G, Pinol AP. Use of the equivalence approach in reproductive health clinical trials. *Stat Med* (2001) **20**(23): 3571–7.
72. Trussell J. Methodological pitfalls in the analysis of contraceptive failure. *Stat Med* (1991) **10**(2): 201–20.
73. Steiner M, Dominik R, Trussell J, Hertz-Picciott I. Measuring contraceptive effectiveness: a conceptual framework. *Obstet Gynecol* (1996) **88**(3 Suppl): 24S–30S.
74. Trussell J, Rodriguez G, Ellertson C. Updated estimates of the effectiveness of the Yuzpe regimen of emergency contraception. *Contraception* (1999) **59**(3): 147–51.
75. Farley TM. Life-table methods for contraceptive research. *Stat Med* (1986) **5**(5): 475–89.
76. Tai BC, Peregoudov A, Machin D. A competing risk approach to the analysis of trials of alternative intra-uterine devices (IUDs) for fertility regulation. *Stat Med* (2001) **20**(23): 3589–600.
77. Farley TM, Ali MM, Slaymaker E. Competing approaches to analysis of failure times with competing risks. *Stat Med* (2001) **20**(23): 3601–10.
78. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* (1991) **266**(1): 93–8.
79. Hassan EO, el Nahal N, el Hussein M. Acceptability of the once-a-month injectable contraceptives Cyclofem and Mesigyna in Egypt. *Contraception* (1994) **49**(5): 469–88.
80. Rodriguez G, Faundes-Latham A, Atkinson LE. An approach to the analysis of menstrual patterns

in the critical evaluation of contraceptives. *Stud Fam Plann* (1976) **7**(2): 42–51.

81. Belsey EM, Farley TM. The analysis of menstrual bleeding patterns: a review. *Contraception* (1988) **38**(2): 129–56.
82. Belsey EM, Machin D, d'Arcangues C. The analysis of vaginal bleeding patterns induced by fertility regulating methods. World Health Organization Special Programme of Research, Development and Research Training in Human Reproduction. *Contraception* (1986) **34**(3): 253–60.
83. Belsey EM, Carlson N. The description of menstrual bleeding patterns: towards fewer measures. *Stat Med* (1991) **10**(2): 267–84.
84. Fraser IS. Vaginal bleeding patterns in women using once-a-month injectable contraceptives. *Contraception* (1994) **49**(4): 399–420.
85. Said S, Sadek W, Rocca M, Koetsawang S, Kirwat O, Piya-Anant M *et al*. Clinical evaluation of the therapeutic effectiveness of ethinyl oestradiol and oestrone sulphate on prolonged bleeding in women using depot medroxyprogesterone acetate for contraception. World Health Organization, Special Programme of Research, Development and Research Training in Human Reproduction, Task Force on Long-acting Systemic Agents for Fertility Regulation. *Hum Reprod* (1996) **11**(Suppl 2): 1–13.
86. Petitti DB, Shapiro S. Strategies for post-registration surveillance of contraceptive steroids. In: Michal F, ed, *Safety Requirements for Contraceptive Steroids*. Geneva: WHO (1989) 335–47.
87. *The Cochrane Library (Issue 1)*. Oxford: Update Software (2006).
88. Cheng L, Gulmezoglu AM, Ezcurra E, Van Look PFA. *Interventions for emergency contraception. The Cochrane Library*. Oxford: Update Software (2001).

24

Gynaecology and Infertility

SILADITYA BHATTACHARYA[1] AND JILL MOLLISON[2]
[1]Department of Obstetrics and Gynaecology, University of Aberdeen, Aberdeen AB25 2ZD, UK
[2]Centre for Statistics in Medicine, Wolfson College Annexe, Linton Road, Oxford OX2 6UD

INTRODUCTION

The randomised clinical trial is widely accepted as a gold standard for scientific evaluation of treatments. Data from clinical trials are considered to represent the highest level of evidence that can be used to inform effective treatment strategies. Yet, there are fewer trials in gynaecology in comparison with other disciplines. Those reported in the literature account for a minority of published papers in major journals.[1] Gynaecological trials incorporate a wide spectrum of clinical conditions and proposed interventions. Women can differ substantially in terms of age, physical and psychological disability, while treatments can range from drug therapy to surgical procedures, from information provision to physiotherapy and dietary advice. The aim of this chapter is to examine, test and explore the basic principles of clinical trials in gynaecology. An overview of different types of trials is provided and reference will be made to specific challenges, including identification of sample populations, choice of appropriate outcomes and tools, randomisation and arrangements for follow-up. Examples are drawn from general gynaecology, infertility and fertility control. Trials in obstetrics, contraception and gynaecological oncology, which are discussed elsewhere, will be excluded from our discussion.

TAXONOMY OF CLINICAL TRIALS

Clinical trials may be classified in a number of ways (see Table 24.1). Some of these are discussed below.

PHASE I CLINICAL TRIALS

These preliminary studies generally address drug safety rather than efficacy, and may be performed on healthy volunteers. Examples include studies of drug metabolism and bio-availability of recombinant gonadotrophins in infertile women.[2] Most Phase I trials are either directly or indirectly supported by the pharmaceutical industry and involve relatively small numbers of subjects. Women are required to adhere to a strict protocol and agree to fairly extensive evaluation often involving multiple investigations such as blood counts, biochemistry, endocrine profile and liver and kidney function tests. In this context, it may

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

Table 24.1. Taxonomy of clinical trials

Phased trials	Phase I
	Phase II
	Phase III
	Phase IV
Conduct	Pragmatic
	Explanatory
Design	Parallel group
	Crossover
	Factorial
	Patient preference
	Cluster
	Equivalence/non-inferiority
Randomisation	True
	Quasi randomisation

be useful to be aware of the fact that finding 'normal' subjects for such trials in reproductive medicine can sometimes be challenging as a large proportion of young, fit, healthy women may either be on oral contraception or actively trying for a pregnancy.

PHASE II CLINICAL TRIALS

These are also fairly small-scale investigations into the efficacy and safety of a drug and require close monitoring of each patient. Sometimes they can be employed as a screening process for drugs, which are either potentially inactive or toxic. They may also be used to determine the most appropriate dose and route of administration of a drug. Examples of these types of trials include those involving the use of misoprostol for medical termination of pregnancy.[3] Where patient acceptability of the route of administration i.e. vaginal, oral or sublingual, is an important outcome, these trials may need to break out of the traditional mould of strictly controlled explanatory trials and assume the pragmatic approach associated with Phase III trials.

PHASE III CLINICAL TRIALS

Following the Phase II trial, the next step is to evaluate the drug's effectiveness using a pragmatic approach and compare it with the current standard management for the same condition and/or placebo in a large trial involving a substantial number of patients. This is also the design used for non-pharmacological interventions, which are increasing in number. The majority of the trials referred to in this chapter are Phase III trials. This is the point of evaluation following which interventions are introduced into clinical practice.

PHASE IV CLINICAL TRIALS

Most Phase III trials are not powered to detect differences in adverse effects. Thus, even after a treatment finds general acceptance, unanswered questions about its safety and long-term effectiveness continue to be addressed in the context of Phase IV trials. The long-term implications of new methods of treatment of menorrhagia such as endometrial ablation are still under evaluation a decade after the results of the first Phase III trials were reported. Medium-term data have been presented in a number of publications.[4,5]

PRAGMATIC AND EXPLANATORY TRIALS

In terms of design, clinical trials are often described as either explanatory or pragmatic. Explanatory trials measure efficacy – the benefit a treatment produces under ideal conditions. Pragmatic trials measure effectiveness – the benefit the treatment produces in routine clinical practice.[6] Examples of the former include evaluation of drugs used to treat menorrhagia or those used to undertake medical termination of pregnancy. The aim of an explanatory trial is to assess the outcome of a new drug under controlled conditions using a homogeneous group of patients. Eligibility criteria are strict, and protocol violations are not allowed. In an explanatory trial comparing recombinant follicle-stimulating hormone with a urinary preparation, any woman who fails to receive the appropriate drug in the prescribed dose will be excluded from the study on grounds of protocol violation.

In contrast, a pragmatic trial aims to mirror the normal variations between patients that occur in real life. For example, a pragmatic trial of medical versus surgical treatment for menorrhagia will include all women with a subjective complaint of heavy menstrual loss. Women randomised to drug treatment, who find the intervention unacceptable and elect for surgery, do not face disqualification from the trial. This somewhat relaxed policy is justified on the grounds that women's decisions to reject their allocated treatment are likely to reflect real-life situations and can actually be interpreted as a measure of dissatisfaction. Furthermore, the treatment offered to patients in the surgery arm may not be identical, as operations may be performed by more than one surgeon, each with a slightly different technique. A similar attitude would apply to a pragmatic trial of physiotherapy for prolapse or counselling for premenstrual syndrome, where identical interventions cannot be guaranteed by different physiotherapists or counsellors.

There are other differences between the two types of trials. Blinding is more likely to be used in an explanatory trial such as one comparing oral metformin with placebo in women with polycystic ovarian syndrome. Pragmatic trials may also be blinded, but this is often not feasible (e.g. in surgical versus medical trials) nor always desirable. There is also less of a compulsion to use placebos, as the objective is to compare the new intervention, not with a placebo, but with the 'gold standard' or best of the existing treatments. Clinician and patient biases caused by the absence of blinding may not necessarily be detrimental to the trial, but could actually be seen to be part of the response to treatment. The outcome in a pragmatic trial such as one comparing oral clomifene citrate (a drug treatment) versus expectant management in unexplained infertility incorporates the total difference between the interventions that are being evaluated. This may include the effect of the treatment as well as the associated placebo effect as this best reflects the likely clinical response in practice.[9]

TRIAL DESIGN

TWO-ARM PARALLEL GROUP

This is the simplest and commonest trial design involving a comparison between two groups, i.e. an experimental versus a control group. As this type of trial is most easily understood by researchers as well as patients, examples abound.

MULTIPLE ARM PARALLEL GROUP

A trial may have more than two arms, e.g. intrauterine insemination (IUI) versus ovarian stimulation + intrauterine insemination (SO/IUI) versus *in vitro* fertilisation (IVF) for the treatment of unexplained infertility.[7] Multi-arm trials have specific statistical issues that require consideration at the design stage. It is important that a priori objectives and a clear principal comparison of interest are stated indicating how the groups are to be compared. Researchers should acknowledge the impact of multiple testing and, where appropriate, take steps to minimise the Type I error. Sample size for multi-arm trials will reflect these objectives. For example, in a trial to compare IVF with IUI with stimulated IUI (SO/IUI), the key comparisons are likely to be (1) IUI versus SO/IUI and (2) IUI versus IVF. Pre-planned comparison may include IVF versus the two IUI arms combined (if no difference is found between the two IUI arms), or IVF may be compared with IUI and stimulated IUI in two separate comparisons. The sample size required for such trials will be dictated by the minimum clinical difference in outcome between any two arms. Due to the nature of the interventions, this minimum clinical difference between any two of the three arms will vary. The minimum clinical difference in live birth rate to be detected would be greater between IVF and IUI than between IUI and SO/IUI. Thus, the number of women required to show a clinically important difference in live birth between IUI and IVF is smaller than that necessary to show a difference between IUI and stimulated IUI.

CROSSOVER

Crossover trials have the advantage of using participants as their own controls, thus reducing the sample size required. Women are randomly allocated to either the control or the intervention arm first, followed by the other. Often a 'wash-out period' is introduced between the two interventions to reduce the risk of contamination (i.e. the effect of the first intervention being carried over to the second intervention). Unfortunately this design is far more suited for medical treatments of chronic conditions as opposed to surgical trials or infertility trials. In surgical trials the practicalities of the situation render such a design inappropriate. In infertility trials, a definite outcome such as pregnancy has the natural effect of preventing women from completing later phases of the trial.[8] In such situations, exaggerated estimates of treatment effect can occur, leading to erroneous clinical decisions. From a practical point of view, only data from the first phase of the crossover trial may be valid. However, this design may well be suitable for exploring drug treatment of chronic conditions such as premenstrual syndrome or sexual function.

FACTORIAL

Factorial designs are often efficient as they can address two research questions within the context of a single trial. The simplest is a 2 × 2 factorial design, an example of which is described in Figure 24.1. Women with unexplained infertility are randomised into four groups to receive no treatment, clomifene alone, insemination (IUI) alone, or clomifene and insemination treatment as shown in Figure 24.1.

The advantage of this design in self-evident; it enables two clinical questions to be addressed within one trial. However, this is only possible if the two interventions are unlikely to interact with one another (e.g. the effect of clomifene is unchanged by the concomitant use of IUI, and vice versa). In this case, the effect of IUI can be assessed by comparing groups C and D with A and B, while the effect of clomifene can be evaluated by comparing B and D versus A and C. However, if clomifene and IUI interact with one another (e.g. the effect of IUI is influenced by the presence of clomifene), the analysis will lose power. For example, the effectiveness of clomiene, in the presence of an interaction with IUI, would involve a comparison of B versus A, using only 50% of couples in the trial. It is important that investigators consider whether there is a possibility of interaction between treatments when choosing the factorial design and appropriately analyse and report the trial. A recent systematic review concluded that transparent reporting is required to ensure accurate interpretation of factorial trials.[9]

	Intrauterine insemination (IUI)	
Clomifene	No	Yes
No	No treatment *Group A*	IUI *Group C*
Yes	Clomifene *Group B*	Clomifene and IUI *Group D*

Figure 24.1

On the other hand, when the aim is to assess whether interaction does exist between treatments, the factorial design is the only design that will facilitate this. However, the study must be adequately powered to detect a clinically relevant interaction.

CLUSTER RANDOMISED TRIAL

Some interventions in gynaecology and infertility are not targeted at individual patients, but at groups of patients (e.g. general practices, hospitals). This can happen where the intervention is an information package for the management of menorrhagia in primary care[10] or a clinical guideline for the management of infertility.[11] In these studies, randomising patients to receive management using the information package or clinical guidelines would have introduced contamination, since GPs would have been expected to manage both study (information leaflet, clinical guidelines) and control patients. Potentially this could underestimate the true effectiveness of the intervention. Therefore, clusters of patients (e.g. general practices) were randomly allocated to receive intervention (e.g. information leaflets, clinical guidelines) or control. Cluster randomisation should only be carried out when there is a strong justification for doing so.

The primary implication of cluster randomised trials is that the measurements on individuals are not statistically independent of one another; that is, measurements from individuals within the same cluster will be correlated to one another. This has implications in the design (e.g. sample size), conduct (e.g. informed consent), analysis and reporting. Cluster randomised trials should adjust for this clustering when determining the number of patients required. The sample size that would be required if patients were to be randomised must be inflated by a factor which takes into account the extent of the clustering and the size of the cluster.[12] The extent of the correlation is measured by the intracluster correlation coefficient (ICC)[13] and researchers are required to have some estimate of this, in order that the study can be adequately powered. Studies that fail to inflate the sample size will suffer from a Type II error (i.e. failure to detect a difference between the interventions, when in fact one exists).

Similarly, the correlated responses obtained from each cluster have an implication for the statistical analysis, since standard statistical tests (e.g. *t*-test) assume that observations are independent of one another. There are a number of approaches to analysing cluster randomised trials and these are detailed elsewhere.[14] Failure to account for the correlated responses in the analyses will result in an increased Type I error (i.e. a difference between the interventions is detected when in fact none exists).

Clustering of outcomes can also occur in infertility trials where alternative treatments are being compared. For example, in randomised controlled trials comparing IVF with ICSI the unit of allocation varies between patients (RCTs)[15], oocytes[16] and cycles.[17] Often, outcomes such as implantation rate and fertilisation rate are considered. These are both expressed as percentages out of the total number of oocytes retrieved. Hence, in trials that randomise patients (couples) or cycles and report implantation or fertilisation rates, there will be clustering of the outcome since oocytes are clustered within patients or cycles.[15,17]

Hence, in these studies adjustment should be made in the analysis for the correlated outcomes assessed (on each oocyte) within patients (or cycles). In trials that randomise by patients and report fertilisation of implantation rates, some adjustment is required. However, for outcomes such as live birth rate or pregnancy rate no adjustment is required since the percentages are expressed out of the total number of couples randomised. Bhattacharya *et al.*[17] randomised cycles and reported implantation rates per transferred embryo. However, they noted that the confidence interval for the difference in implantation rates was likely to be wider than that reported due to failure to adjust for the clustering of embryos/oocytes transferred to each woman. Studies where oocytes have been randomised have no clustering implications since oocytes retrieved from the same women are randomly allocated to receive ICSI or IVF.

When conducting RCTs in infertility, consideration should be given to the unit of randomisation and the outcome measures to be applied. When there is implicit clustering in the data, the statistical analysis should account for this using the methods described above.[14]

QUASI-RANDOMISED TRIALS

These are controlled experimental studies where treatment allocation is performed on the basis of odd or even patient unit numbers or days of the week when the patients are recruited. Although this design of treatment allocation affords an element of chance, it cannot be considered to be genuine randomisation. This type of design may still appeal to those involved in laboratory trials involving incubation or cryopreservation of human embryos. In these cases, it may be easier and cheaper to use a certain protocol for all embryos on alternate days or alternate weeks rather than change the protocol or a freezer setting for each embryo or each woman. The consequent loss of allocation concealment can lead to serious selection bias as some patients may be deliberately excluded. This, is turn, can exaggerate treatment effects.

PATIENT PREFERENCE TRIAL

A potential problem in some randomised trials arises when patients or their clinicians refuse to be randomised on grounds of strong treatment preferences. Exclusion of these patients may affect the generalisability of the results as participants may not be representative. Yet recruitment of these patients may introduce substantial bias, especially since it is impossible to blind them. In addition, compliance and satisfaction may be higher with the preferred intervention.[18] This is particularly so when the 'new' treatment is only available within the context of the trial or when as in trials in unexplained infertility one of the arms comprises a 'no-treatment' or 'expectant management' group. Dissatisfaction with the allocation may lead to differential compliance and follow-up resulting in groups which cannot then be assumed to be similar. The outcome measures could also be affected by how satisfied patients are with their allocated treatment. The effect of patient preference on outcome would depend to a great extent on the specific outcomes being assessed. If the principal outcome is death or live birth, then the effect of patient preferences is likely to be small. If the principal outcome is satisfaction with care, then the effect of patient preference is large.[19] Under such circumstances the conventional randomised trial will underestimate the relationship between the intervention and the outcome, i.e. show the minimum effect size. Conversely a comparison between two groups of patients who have chosen their treatment and thus optimised their treatment choice will be considered to represent the maximum effect size. An intervention in question will have an effect size between the minimum and maximum as derived from the randomised and the preference part of a partially randomised patient preference trial.[19]

To deal with patient preferences within a trial, the use of a partially randomised patient preference (PRPP) trial has been suggested.[20] Patients with strong preferences are allowed their desired treatment. Those without such views are subjected to randomisation. For example, in a trial of medical and surgical termination of pregnancy we end up with four groups: randomised to medical, randomised to surgical, prefer medical and prefer surgical.

One potential disadvantage with PRPP trials is the impact on the trial size. The size of a total PRPP cohort will need to be much larger than for a conventional RCT. As already mentioned, the size of the randomised cohort needs to be the same as in a conventional trial. In addition, the number of patients in the non-randomised preference arms needs to be of equivalent size. The numbers quickly add up to generate a total sample size double that for a conventional trial. This has the predictable effect of adding to the cost and duration of the trial. Entry of a disproportionate number of patients into either the randomised or the preference arms is also a problem, as the trial will not be completed unless the appropriate number have been recruited into the two components of the trial. The situation may be further complicated by patients favouring one treatment over another, making comparison of the two groups in the preference arm more difficult or at worst result in only one preference arm. Thus it is important that, prior to a PRPP trial design being adopted, pilot work is carried out to estimate the likely extent of preferences.

A further problem with this approach lies in the analysis. Any comparison using the non-randomised groups is unreliable because of unknown and uncontrolled confounders. Patient preference designs have been used in trials of termination of pregnancy[21,22] and menorrhagia.[23] The evidence to support the use of PRPP trials compared with conventional randomised trials is limited. A randomised comparison of the two strategies by Cooper *et al.*[23] suggested that the extra cost and complexity were not justified in the context of medical versus surgical treatment of menorrhagia.

A conventional randomised trial could address the effect that patient preference has on outcome by recording this information before allocation.[24] This would allow resources to be concentrated on recruiting as many patients as possible into the randomised comparison group but would allow stratification of the results by initial preference.

EQUIVALENCE TRIALS

Often in reproductive health care the aim is to show that one treatment is as effective (equivalent), or no less effective (non-inferior), as another. The methodology for equivalence trials differs to that of superiority trials in design, analysis and interpretation. In designing equivalence trials, attention must be given to defining an equivalence margin. That is, the difference in effect that would be deemed to be 'clinically insignificant'.[25] To clarify, we consider a trial to compare the efficacy of recombinant and urinary HCG with the primary outcome being the number of oocytes retrieved. In this trial the researchers defined the equivalence margin to be ±3 oocytes retrieved as this was deemed to be clinically acceptable.[26] By definition a clinically acceptable difference will be smaller than a clinically worthwhile difference, as defined in a superiority trial. Therefore, larger sample sizes are needed to demonstrate equivalence. In the analysis of equivalence trials, conventional statistical testing has little relevance and interpretation of results should be conducted though use of confidence intervals in relation to the predefined equivalence margin.[27] Statistical significance is demonstrated if the upper and lower limits of the 95% confidence interval do not cross the equivalence margin.[27] In the trial of recombinant and urinary human chorionic gonadotrophin (hCG), the two treatments were declared equivalent since both the upper and lower limits of the 90% confidence interval, for the difference between treatments in the number of oocytes recovered, did not exceed 3 oocytes.[26] The choice of a 90% confidence interval was not justified in the trial. It is important to note that a lower level of confidence (90% as opposed to 95%) will produce a narrower confidence interval and thus greater chance of declaring the result equivalent. In superiority trials, the most conservative analysis is by intention to treat (ITT). That is, participants are analysed in the groups to which they were randomised, irrespective of the treatment they actually received. However, in an equivalence trial a 'per protocol' (PP) analysis is usually considered statistically more conservative. This is because an ITT analysis may blur the comparison between groups and lead to an increased chance of declaring the two treatments as equivalent when they are not. The decision about which should be the primary form of analysis (ITT or PP) in an equivalence study is not straightforward.[28] It depends on the particular characteristics of the study, including the definitions adopted for the ITT and PP analyses and the risk of bias.[29] It is generally recommended that both ITT and PP analyses are conducted.

PERFORMING AN RCT

SYSTEMATIC REVIEW

A systematic review of the literature is an essential component of the pre-trial work-up. It enables the researcher to define the clinical question in the light of work that has gone before and assess the need for a trial. It also provides vital information about the limitations of previous trials, outcome measures used and nature of follow-up. This is useful in planning the design of the proposed study. A recent systematic review[30] has identified typical problems associated with previous trials in unexplained infertility including small sample sizes, inappropriate outcome measures (pregnancy rates per cycle) and lack of cost data. The Cochrane Collaboration conducts systematic reviews on many other topics in gynaecology.

DEFINING THE STUDY POPULATION

Definition of the study population is a vital part of any clinical trial. Unfortunately this aspect of trial design can be contentious. The diagnostic criteria of many gynaecological conditions continue to generate debate amongst clinicians. Disagreement about the definition of a particular condition can lead to dismissal of the conclusions of a trial as irrelevant. Certain clinical terms continue to pose particular problems. For example, infertility is defined as ‘the lack of pregnancy after one year of regular unprotected coital exposure’. Thus it refers to the lack of a clinical endpoint, i.e. pregnancy, rather than a particular disorder. Thus in a group of infertile couples there may be contributory factors from both sexes. To complicate the issue, definitions of subgroups such as unexplained infertility can vary depending on the rigour of the diagnostic tests involved. Variation in laboratory procedures (such as semen analyses) may affect the diagnosis of male infertility while the investigations used for tubal patency (laparoscopy versus hysterosalpingogram) may have an effect on the identification of tubal disease and endometriosis in infertile women.

With menorrhagia, the problem is different. The conventional textbook definition of menorrhagia (menstrual blood loss > 80 ml) is impractical as menstrual loss is not measured routinely in clinical practice. A more pragmatic approach is to include all women with a subjective complaint of heavy menstrual blood loss. This allows some critics to question the external validity (i.e. generalisability) of trials where women have been included both on the basis of objective measurement of menstrual blood loss as well as on the basis of subjective complaints of heavy menstrual bleeding. Purists will argue that efficacy of treatments for menorrhagia cannot be evaluated accurately in the absence of patients with ‘genuine’ pathology. However, the results of such trials using rigorous inclusion criteria may not necessarily be relevant to the vast majority of clinics where menstrual blood loss is not routinely measured. From a clinical point of view, it is probably more useful to recruit women on the basis of a subjective complaint of menorrhagia since this mirrors clinical practice and is likely to increase generalisability of the trial findings.

In studies of urinary incontinence it is well known that women who attend urodynamic clinics constitute a small proportion of the total number of women in the community suffering from urinary incontinence. Thus, studies may be conducted on a group of patients not representative of the target disorder. Explicit description of the eligibility criteria allows the readers to draw their own conclusions regarding the applicability of the data to their own specific contexts. Those performing secondary research can also use these data to assess heterogeneity between trials.

A specific problem associated with infertility trials is the question of how to deal with the male partner. Conventionally it is the woman who undergoes treatment, and it is she who is considered to be the participant in trials and subjected to recruitment, randomisation and follow-up. However, in trials where satisfaction, acceptability and costs are outcomes, it is perhaps appropriate to seek the male partners’ views as well.

An important aspect of the choice of the study population involves the effect of the participants on generalisability of the findings. Although study participants may meet eligibility criteria, participation is voluntary and volunteers may differ from the general population in terms of general health, co-interventions, educational level, motivation, and ability to follow a protocol. Ethnic minorities may be missed on account of unfamiliarity with the language of the questionnaires used. In pragmatic trials, it is particularly important that appropriate steps are taken in order to ensure that the participants are representative of the total eligible population. It may be useful to monitor refusals in order to document whether participants differ substantially from non-participants.

INTERVENTIONS

Due to its unique mix of medical and surgical workload, gynaecology offers a number of diverse interventions that need to be tested in the context of clinical trials. Some examples are shown in Table 24.2.

DEFINING OUTCOMES

For any trial, it is crucial to have a clear research question, i.e. a priori hypothesis and a clinically

Table 24.2. Types of interventions subjected to clinical trials in gynaecology

Intervention	Examples of trials
Packages of care	Information packages in use in general practice for appropriate treatment and referral in menorrhagia
	The value of guidelines in infertility for general practitioners
Surgical techniques	Hysterectomy versus endometrial ablation
	Different types of endometrial ablation, e.g. TCRE versus laser
Drug trial	Placebo versus tranexamic acid for menorrhagia
Comparison of different treatment modalities	Medical versus surgical termination of pregnancy
	Mirena IUS versus endometrial ablation for menorrhagia
	Expectant treatment versus *In vitro* fertilisation (IVF) for unexplained infertility
Laboratory techniques	*In vitro* fertilisation (IVF) versus intra-cytoplasmic sperm injection
	Alternative methods of cryopreservation of human embryos
Place of care	One-stop specialist clinic versus general clinic
	Outpatient versus inpatient endometrial ablation
Investigations	Effectiveness of methods of screening for *Chlamydia trachomatis*
	Hysterosalpingography as a test of tubal patency
	The post-coital test in the diagnosis of infertility
	Hysteroscopy in the diagnosis of menorrhagia

relevant primary outcome on which the power calculation is based. Outcomes of choice include those that are purely clinical, as well as others which may be patient centred or economic. The precise nature of the primary and secondary outcomes will depend on the type of trial and its clinical context. This may involve different levels of observation and analysis, incorporating the individual, the family and the community.[31]

CLINICAL OUTCOMES

Clinical outcomes are essential components of any clinical trial. These are the tools that measure whether the intervention works or not. They should be relevant to patients and meaningful to clinicians. Generally speaking, outcomes which represent critical events such as live births, deaths and repeat surgery are often more meaningful than outcomes involving measurements. For example, the proportion of women requiring blood transfusion after hysterectomy is a more clinically relevant outcome than the volume of blood lost during the procedure. The proportion of women in whom the uterus is empty at 24 hours may be a more meaningful outcome than the mean number of hours required for medical termination. In reproductive medicine, many clinical trials have tended to choose surrogate markers such as number of oocytes retrieved or fertilised as primary outcome rather than live birth or pregnancy. There may be logistical reasons for such a choice. The use of a surrogate endpoint can have the effect of reducing the length of a trial and the number of participants required. Using the example above, to detect a difference of 10% in live birth rate, from 30% to 40%, would require a minimum of 477 women in each group, assuming 90% power and 5% significance. However, to detect a difference of one-third of a standard deviation in the number of oocytes retrieved would require a minimum of 191 women per treatment group. It is essential when designing clinical trials that the primary outcome is the most clinically meaningful. Use of surrogate endpoints can reduce the sample size and costs, allowing a single centre to perform a trial that would otherwise require far higher levels of funding and a multi-centre approach. However, if surrogate endpoints are to be adopted in a clinical trial, the effect of the intervention on the surrogate must be highly predictive of the effect on the clinical endpoint.[32]

Explanatory trials usually rely on a single clinical outcome. For example, in a trial comparing drug treatments for menorrhagia, menstrual blood loss (in ml) may well be an appropriate primary outcome. Other physiological or biochemical outcomes such as haemoglobin level, volume urinary loss, extent of endometriosis visualised by laparoscopy, number of ovarian follicles seen on ultrasound scan and serum estradiol levels following ovarian stimulation may also be used in different situations. Unfortunately, they may not always correlate well with the clinically relevant outcomes – certainly from the patients' perspective. Often it is not sample size alone but also the length of follow-up that dictate use of a surrogate endpoint. Thus bone mineral density rather than the incidence of hip fracture may be chosen as a principal outcome in trials of hormone replacement therapy.[33]

Pragmatic trials usually require the evaluation of more than one outcome measure in order to come to a decision about the effectiveness, risks, costs and acceptability of an intervention. For example, in surgical trials of menorrhagia, outcomes should include one or more of satisfaction with treatment, menstrual flow, pain, premenstrual syndrome and period of recovery. Sometimes when the impact of a disease spreads beyond the individual to a wider group such as the family, general practitioners or carers, outcomes may need to be expanded to include a wider group. This may be relevant in trials of urinary incontinence or HRT. When designing a trial, it is important to classify the outcomes into primary and secondary.

QUALITY OF LIFE

Quality of life (QOL) is now accepted by most clinicians as an important outcome in clinical trials.[34] However, the term is sometimes used loosely and without a clear understanding of what it means.[35] Since QOL is considered to be a complex concept comprising physical, emotional and other dimensions, most questionnaires in common use not only assess the detailed aspects of QOL but also provide a summary score for overall health status.[36] Generic measures such as short-form health survey (SF-36)[37] broadly assess physical, mental and emotional health and can be used to compare conditions and treatments. Although the number of such instruments in current use is rapidly increasing, there is a remarkable level of consistency between them.[36]

Other methods include tools focusing on a single aspect such as pain or anxiety as well as individualised measures in which patients themselves define and rate the most important aspects of their QOL.[38] A number of condition-specific tools, which can be used either independently or to supplement generic measures, have been developed.[39] Examples include the King's College Questionnaire for Urinary Incontinence[40] and the Menstrual Distress Questionnaire.[41] The Endometriosis Health Profile-30[42] and The Menopause Rating Scale (MRS).[43]

A systematic review by Sanders *et al.*[44] showed that despite the plethora of instruments, the prevalence of reporting on QOL remains low, increasing from 1% in 1980 to 4% in 1997. There is also a general unwillingness to ask patients to supplement questionnaire-based data with personal responses, and lack of appreciation about the critical importance of response rates.

Patients themselves can find it difficult to distinguish between QOL and health status or to rate their health without a point of reference. At the same time, the effects of age and changing expectations need to be adjusted for when interpreting QOL scores. Overall, QOL offers a superior way of assessing treatment success in trials involving general gynaecology (such as menorrhagia, urinary incontinence, menopause, pre-menstrual tension) where interventions are targeted at women with benign but debilitating illnesses that compromise several key areas of day-to-day life. On the other hand, women seeking fertility treatment or abortion services are not necessarily unwell. The aim of treatment is to enhance their physical and mental well-being rather than correct a pre-existing deficit in health status. Existing instruments do not discriminate between these two broad groups and further refinements are needed with respect to assessing positive aspects of general and sexual health as opposed to the conventionally used negative aspects.[45] Meanwhile, simple global questions on self-reported health or QOL continue to be useful as prognostic measures for stratification of treatment allocation and as important outcome measures alongside purely clinical outcomes.

PATIENT SATISFACTION

There continues to be a general lack of agreement about the mechanisms which produce satisfaction, as well as the meaning of the word 'satisfaction' itself which has been defined as an 'evaluation based on the fulfilment of expectations'.[46] The conventional view is that satisfaction reflects the sum total of a number of patient-related factors, including expectations, characteristics and psychosocial determinants.[47] Over the past few years, patient satisfaction has become increasingly accepted as a measure of quality in health services and a valid outcome in RCTs.[48] This is particularly significant in the current climate of delivery of health care which aims to provide a patient-centred service with greater public involvement in planning.[49] The purpose of patient satisfaction measurement is to describe health care services from the patient's point of view, measure the 'process' of care and evaluate health care.[47] The particular strength of using satisfaction as an outcome is related to the unique circumstances of certain gynaecological trials such as those used for menorrhagia where not only the interventions but also the clinical outcomes may be dissimilar. In a trial of hysterectomy versus endometrial ablation, women would be expected to be amenorrhoeic following hysterectomy but not after ablation. Here, comparison of amenorrhoea rates is unlikely to be helpful in comparing the two groups, while satisfaction not only is a robust measure of treatment success, but also incorporates the sum total of a woman's experience of the alternative treatment arms including discomfort, recovery time, and side effects. A similar argument can be used to justify the use of the same outcome for trials comparing surgical and non-surgical treatment of urinary incontinence.

Despite their widespread use in clinical trials, assessment of patient satisfaction has been criticised on theoretical and methodological grounds and their practical use questioned.[50] Relatively few patients express open dissatisfaction with treatment.[51] Indeed satisfaction rates of 80% or more are reported by most hospital-based studies.[52] There is also little evidence to indicate that expressions of satisfaction result from the fulfilment of expectations; in some situations it is difficult to establish the fact that expectations exist at all. High satisfaction ratings do not necessarily mean that women have had good experiences in relation to the service as satisfaction may well make allowances for mitigating circumstances. If the aim is to provide women with a voice, it is important not to rely on satisfaction with treatment as a single outcome but to prioritise methods of accessing women's experience of interventions and the meaning and value they attach to them.[50] There are no off-the-shelf questionnaires that are completely satisfactory[53] and qualitative studies have demonstrated that high satisfaction rates cannot be taken as proof of positive experience. Many tools mentioned in the literature are not validated, while many expressions of satisfaction may not be evaluations at all.[54,55] Dissatisfaction may be more useful as a minimum level of negative experience and may be of potential use in benchmarking exercises. At the moment most clinical trials in gynaecology attempt to measure satisfaction using a number of direct and indirect questions. Some of these questions have been repeated at various points during follow-up to assess change in satisfaction rates over time. Despite the obvious shortcomings of the existing system, there has been an opportunity to refine and validate some of these questionnaires through repeated use in a series of related trials.[4] Acceptability has been measured by direct questions and by other tools such as the Semantic Differential technique in the context of menorrhagia and termination.[21–23]

In other areas such as infertility, satisfaction with treatment is more difficult to assess as the effect of the desired outcome (live birth) is predominant even where treatment is invasive or unpleasant. Conversely there is dissatisfaction with treatment where the outcome is failure to fall pregnant. Some attempts have been made in recent trials to specifically address separately satisfaction with 'treatment' as opposed to satisfaction with 'outcome'. This area is deserving of further study.

ECONOMIC EVALUATION

With the emergence of new methods of treatment comes an increasing awareness of the need to study not just the clinical effectiveness but also the cost-effectiveness of alternative treatments. Pragmatic clinical trials are the standard approach not only for evaluating interventions, but also for comparing costs.[56] The costs of treatments are usually estimated using information about the quantities of the resources used. For example, the resources used for hysterectomy include the staff time involved, the consumables used and the length of the subsequent inpatient stay. To estimate the cost of treatment, information about this resource use is combined with unit cost estimates, which provide a fixed monetary value to each cost generating item.[57] The total cost is then the weighted sum of quantities of resources used where weights are unit costs. Carrying out an economic evaluation alongside a randomised trial allows detailed information to be collected about the quantities used by each patient in the study. Such information allows a cost for each patient producing patient-specific cost data. This is turn reduces the extent to which comparison between the groups is based on assumptions about resource use. However, randomised trials are not necessarily the only way or necessarily the best way to address economic questions.[58] There is an important role for other methods, e.g. modelling.

In the context of RCTs, however, there is an urgent need to revise the way in which health economic outcomes are addressed within a clinical trial. While cost outcomes are generally regarded as secondary outcomes, the rationale for a formal sample size calculation with adequate power for the planned analysis is still relevant given the large variability in costs between individuals.[59] This is even more relevant where subsets are used for cost data for practical reasons. One review has identified an urgent need to improve the statistical analysis and interpretation of cost data in RCTs.[57] This is particularly relevant to the provision of descriptive statistics relating to costs. As cost data are typically skewed, the median can be interpreted as the typical cost for individuals. However, it is the mean cost that is important for policy decisions as it is this value, multiplied by the number of patients, which gives an estimate of the total cost of an intervention.

Table 24.3 provides some examples of outcome measures used in different types of gynaecological trials. A crude list such as this is useful, if only to illustrate the specific demands of different clinical areas. Overall, due to the limitations of using 'pure' clinical outcomes in benign gynaecology, 'satisfaction' and 'quality of life' (however defined) have found widespread acceptance as appropriate outcomes. In other areas such as infertility, 'satisfaction' is meaningless without the promise of live birth, while even the most invasive and uncomfortable treatment may be perceived to be entirely acceptable if it leads to pregnancy.

In general, even when relevant, purely clinical outcomes may lead to potential conflicts between the clinicians' and patients' points of view. A number of health state measures incorporating validated and reliable scales have been developed to address this very issue.[60] These may be generic or disease specific. Most pragmatic trials will use a number of outcomes from the above categories. At the same time it is best, in very large trials, to concentrate on a few simple outcomes, for reasons of convenience and efficiency.[61] There is also a statistical drawback to the use of multiple outcomes. The greater the number of outcomes, the higher the possibility of one of them reaching statistical significance on the basis of chance alone. It is important to consider the relevance of outcome measures to the stakeholders. It is thus important to predefine primary and secondary outcomes. The extent to which a trial changes practice will depend on the outcomes chosen.

Table 24.3. Outcomes in gynaecological trials

Clinical area	Outcomes	Comments
Infertility	• Live birth rate per couple • Live birth rate per treatment • Clinical pregnancy rate per couple • Clinical pregnancy rate per treatment • Biochemical pregnancy rate • Fertilisation rate • Implantation rate • Multiple pregnancy • Morbidity (e.g. ovarian hyperstimulation) • Costs	Although live birth per couple is the most robust outcome, it demands large sample sizes and a longer duration of follow-up. Live birth/clinical pregnancy rate per treatment is still used in many trials Multiple pregnancy and its effect on maternal and perinatal morbidity is increasingly being acknowledged as an important outcome of fertility trials
Menorrhagia	• Satisfaction • Acceptability • Quality of Life • Menstrual blood loss • Bleeding and pain scores • Morbidity • Repeat surgery • Haemoglobin level • Amenorrhoea rates • Costs	Satisfaction and quality of life are clinically more useful than objective measurement of menstrual blood loss or amenorrhoea rates, especially when trials compare treatments such as hysterectomy which guarantees amenorrhoea versus the Mirena intrauterine system or endometrial ablation which do not Satisfaction with treatment may not correspond to amenorrhoea rates Long-term follow-up is important in the evaluation of all new technologies
Urogynaecology	• Satisfaction • Acceptability • Quality of life • Symptom relief • Objective measurement of urinary loss • Surgical morbidity, repeat surgery • Length of hospital stay • Urodynamic assessment • Costs	Symptom relief and objective assessment of bladder function may not necessarily correspond with quality of life or satisfaction Long-term follow-up is necessary for effective evaluation of treatments
Hormone replacement therapy	• Menopausal symptoms • Quality of life • Hip fracture • Cardiovascular disease • Acceptability • Bone mineral density • Serum lipid profile • Side effects and morbidity	Historically, surrogate outcomes like lipid profile and bone density have been more popular than rates of cardiovascular disease or fracture
Termination of pregnancy	• Efficacy: evacuation of the uterus • Acceptability • Morbidity • Quality of life • Costs	Quality of life difficult to assess in the context of termination Long-term follow-up difficult

POWER AND SAMPLE SIZE CALCULATIONS

Before embarking on any clinical trial, due consideration should be given to ensuring that the study will have adequate power (usually 80% or 90%). Power is the probability that a study of a given size will detect as statistically significant a real difference of a given magnitude.[62] The sample size for each trial is usually based on the primary outcome. Although secondary outcomes are often investigated and subgroup analyses performed, the power of an RCT to provide conclusive answers to these may be limited. It is important to ensure that the study is designed to detect clinically important differences, if they exist. Conversely, if the statistical power is low, the results of the trial will be questionable as the numbers may have been too small to detect genuine differences. In general, a clinically worthwhile difference in the primary outcome should be identified as the point of reference for a sample size calculation. Intimate knowledge of the clinical area is crucial for this. For example, a 20% difference in satisfaction rate between two forms of treatments for incontinence may be considered to be clinically important. Conversely, against a background of low live birth rates, a difference of 5% to 10% may be enough to change clinical practice in an infertility trial. In infertility trials it is important that the sample size is based on the appropriate unit, e.g. couples rather than oocytes or cycles, where the unit of analysis is the couple.[63]

In determining the sample size attention should also be paid to the possibility of sample attrition and the need for any future subgroup analysis. For example, in abortion trials, a high non-response to follow-up should be anticipated and the sample size inflated accordingly.[22] In infertility trials, where it may be clinically important to assess the effect of the intervention in different clinical groups, a similar exercise will ensure meaningful subgroup analysis. However, detection of an interaction will require an increase in sample size; for example, to detect an interaction of the same magnitude as the main treatment effect with the same power will require the sample size to be inflated by a factor of four.[64] Subgroup analyses should be identified a priori and in general should be perceived as a hypothesis-generating exercise. Trials with small sample sizes have a high risk of failing to demonstrate a real difference (Type II error). This has been the case in gynaecological trials.[63] At the same time, aiming for unrealistically large sample sizes is counterproductive and possibly unethical if it means that a trial is abandoned due to failure of recruitment.

RANDOMISATION

In a clinical trial on women with menorrhagia, the outcome can be affected by participants' age, co-existent menstrual symptoms (dysmenorrhoea and premenstrual syndrome) and the presence of uterine fibroids. Randomisation involves allocating participants to groups such that individual characteristics do not influence the nature of the intervention. Any difference in outcome is therefore attributable to the treatment alone. Random allocation does not guarantee that the groups will be *identical* but it does ensure that any differences between them are due to chance alone.

The randomisation process must ensure that the random sequence of treatment allocations is concealed from those involved in recruiting patients to the trial and the treatment allocation is only revealed after a participant has been recruited to the trial. This can be achieved through the use of a central, computerised, randomisation system or placing the random allocations in sequentially numbered, sealed, opaque envelopes. In addition, it leads to treatment groups which are random samples of participants from the target population and thus makes valid the use of standard statistical tests.

While the simplest method of randomisation is tossing a coin, in practice this is not an accepted method of treatment allocation. The main reason for this is the lack of an audit trail that makes it difficult to confirm that the random allocation was done correctly. The random allocation should be determined in advance, preferably by using random numbers generated by a mathematical process. After the randomisation list has been prepared (by someone who will not be involved in recruitment), it must be concealed before being made available to researchers. Although the process of randomisation can occur at the recruitment point this is preferably done at long range, by telephone or even the internet. Alternatively, sequentially numbered opaque envelopes can be used. Differences in outcome between treatment groups are considerably larger in trials where allocation concealment is not strictly enforced as this produces a clear bias.

While simple randomisation techniques will, on average, allocate equal numbers to each arm, groups of different sizes can result. Block randomisation can be used to keep the numbers in each group very close at all times. In a trial of two alternative surgical treatments for menorrhagia we might want to ensure that each surgeon treats similar numbers of women by either method. Stratified randomisation produces a separate randomisation list for each surgeon (stratum) so that we get very similar numbers of patients receiving each treatment within each stratum. If envelopes are used, this may involve two separate lists of random numbers and two separate piles of sealed envelopes for each surgeon. With stratified randomisation we must always use blocks to ensure that there is balance of treatments within each stratum. While stratified randomisation can be extended to two or more stratifying variables there is a practical limit to the number of strata. Stratification by centre is standard practice in multi-centre trials.

Prognostic variables such as age, parity and duration of infertility are major determinants of outcomes such as live birth. In small trials random allocation may not provide adequate distribution of factors between the intervention groups. Here, it is still possible to achieve balance using minimisation, which is based on the concept that the next patient to enter the trial is allocated to whichever treatment which would minimise the overall imbalance between groups at any stage of the trial. Even in small trials this provides groups that are comparable across several prognostic factors. It is important to specify exactly which prognostic variables are to be used and to say how they are to be grouped. For example, age, previous pregnancy and duration of infertility are important prognostic factors for fertility. Minimisation in this context will require a statement about the actual age groups, e.g. less than 30 years and 30 or older. Minimisation is crucial in infertility trials where a clinically significant difference in live birth rates associated with alternative treatments is small and could easily be overpowered by the effect of prognostic factors such as age, parity and previous pregnancy.

A practical problem relating to randomisation concerns the emotive nature of some of the conditions under evaluation such as infertility or termination of pregnancy. Some women may be unwilling to accept the extra stress of participating in a trial over and above what is already a complex and psychologically challenging experience. There may also be compelling social reasons why women undergoing termination are less likely to opt for randomisation, and comply with trial protocols and follow-up arrangements. Infertile couples may be required to fund their treatment themselves. This could influence their decision to refuse to participate in a trial where the experimental arm (such as assisted hatching) is substantially more expensive than standard treatment, unless the trial organisers offer to absorb the extra costs. Often there is an imperative to provide ‘treatment’ at the request of the couples. This makes it difficult to recruit couples into a clinical trial where one of the options is ‘expectant management’.

CONCEALMENT OF ALLOCATION

The unpredictability of the randomisation process can only be successful if followed by allocation concealment, i.e. concealment of the sequence until patients have been assigned to their groups.[65] This ensures strict implementation of a random allocation sequence without foreknowledge of treatment assignments. Awareness of the next treatment allocation could lead to exclusion of certain women based on their prognosis because they would have been allocated to the perceived inappropriate group. For example, in a trial of unexplained infertility, women with a prolonged duration of infertility could be excluded if the next treatment allocation were known to be a 'no-treatment' arm. Adequate concealment would ensure that the decision to accept or reject a participant should be made and informed consent obtained without prior knowledge of the nature of the assignment.

Trials that use inadequate or unclear allocation concealment have tended to yield 40% larger estimates of effect compared with those which used adequate concealment.[66–68] Trials with poorly concealed allocation also generated greater heterogeneity in results, i.e. the results fluctuated extensively above and below the estimates from better studies.[66]

BLINDING

Double blinding seeks to prevent ascertainment bias, protects the sequence after allocation and cannot always be implemented.[1] As in the case with allocation concealment, lack of blinding may lead to exaggerated estimate effects of treatment. A survey of trials in gynaecology found that investigators could have used double blinding more often.[1] When used, methods of double blinding were poorly reported and rarely evaluated. It is recommended that authors provide adequate information about the methods used to ensure double blinding. This should include details such as the type of intervention (capsules/tablets), and efforts made to duplicate the characteristics of the treatment (taste, appearance, route of administration). In addition it is important to be explicit about the methods used to control the allocation schedule, such as location of the schedule during the trial, details of when the code was broken for analysis and the circumstances under which the code could be broken for individual cases (adverse reactions). Finally there should be a statement about the perceived success or failure of the double-blinding efforts.

EXCLUSIONS

Exclusions can occur due to eventual discovery about ineligibility, deviations from protocol, withdrawals or losses to follow-up. Exclusions before randomisation do not affect the internal validity of the trial but can compromise generalisability. For most pragmatic trials it is important to keep the eligibility criteria to a minimum. In practice it is unusual to find significant qualitative differences between women in trials and those in the general population. Exclusions after trial entry represent a further source of bias within an RCT as any drop-out over the course of the trial from those initially randomised participants is not likely to be random in nature. The accepted method of primary analysis in all cases is by 'intention to treat', i.e. analysis of patients in the originally assigned groups regardless of any breaches of protocol.[69] This can prove unnerving for clinicians, especially in the context of surgical trials. For example, in a trial comparing hysterectomy versus endometrial ablation many clinicians would find it difficult to accept results of analysis of amenorrhoea rates by intention to treat arguing that it is inappropriate to include hysterectomised women in the ablation group as this would lead to an overestimation of amenorrhoea rates. Investigators can also do secondary analyses, preferably pre-planned based on only those participants who fully complied with the trial protocol (per protocol) or who received a particular treatment irrespective of randomised assignment (analysis by treatment received). Secondary analyses are acceptable as long as researchers label them as such and highlight the non-randomised comparison groups. The advantage of randomisation is entirely lost when investigators exclude participants and in effect present a non-randomised comparison as the primary result, i.e. similar to a cohort study. Exclusions of participants can lead to misleading results.[70] Researchers sometimes exclude patients on the basis of outcomes that happen before treatment has begun, such as pregnancy in a couple with infertility. Although this may seem sensible inasmuch as the event of interest occurred independently of the treatment, the same argument could be used for excluding pregnancies in a no-intervention arm of the trial.

It is important to attempt to minimise exclusions and be explicit about those cases where exclusions occurred. This can be enforced by minimising the delay between randomisation and initiation of treatment. This can be particularly relevant to infertility trials where couples could fall pregnant before treatment can start or where the intervention is conditional to a set of clinical criteria. For example, in couples randomised to IVF or ICSI it may be more efficient to delay randomisation until after oocyte recovery so that women who have failed to respond to gonadotrophin stimulation are not included.

FOLLOW-UP

It is important to predetermine the length and type of follow-up for each trial. The precise circumstances and the time interval will depend on the nature of the trial. In fertility trials the traditional method was to express outcomes as pregnancy rates per cycle. This meant the duration of follow-up was brief. For more robust outcomes like pregnancy rate per woman, it may be necessary to extend the follow-up for 3–6 cycles depending on the nature of the treatment. A further nine months need to be added on to allow live birth per couple to be used as an outcome. For menorrhagia trials, 80% of re-treatments occur within two years, making this an acceptable duration for follow-up in the first instance. A prolonged period of follow-up of up to five years would be ideal as many women could expect the effects of their treatment to wane over time and long-term complications of therapy to surface. This would appear to be equally true for urogynaecology trials. For termination of pregnancy, follow-up has to be kept short as the loss to follow-up is high, as many women may not wish to be contacted at a remote period of time. For HRT trials, which genuinely wish to address crucial outcomes such as rates of fracture, cardiovascular disease, or Alzheimer's disease, follow-up may need to be extended to tens of years. This obviously raises significant ethical, logistic and financial issues which may well need to be taken into account whilst planning such trials.

DATA COLLECTION

Data in a trial are usually collected from sources such as case notes, local clinic databases and patient questionnaires. Occasionally interviews may be used to explore areas which are not capable of being probed adequately with questionnaires. General practitioners, local and national databases may also be accessed to obtain clinical information such as re-treatment rates or serious complications about patients who are lost to follow-up.

CONDUCT

Recruitment

To avoid recruitment bias, it is important to target all eligible women and record all refusals. It may be helpful to obtain some baseline clinical details about them in order to explore any major differences between participants and non-participants, which could affect the external validity of the trial.

Trial Coordination

Following informed consent, it is important to obtain baseline information by filling in datasheets or questionnaires prior to randomisation. Subsequent data collection should occur at the pre-specified times and an efficient system of timely reminders put into place. In pragmatic trials it is often important to distinguish those women who no longer wish to continue with the allocated treatment from those who wish to terminate their involvement with the trial and do not wish to be contacted for follow-up or have questionnaires sent to them. Hopefully the numbers in this latter group should be small but their wishes should be respected.

DATA ANALYSIS

This is an important aspect of the trial and errors here can lead to significant bias. As mentioned above, analysis should be by intention to treat. Each woman should be analysed as though she had received the intervention to which she had been randomised. This minimises any bias due to non-random removal of participants from the trial. The exception is explanatory trials, usually Phase I and II drug trials, where strict rules of exclusion for protocol violation apply. Occasionally it may be important from a clinical point of view to perform a separate analysis by treatment received or by design (e.g. equivalence trials). This should be clearly described as such and should be used to assess the primary outcome. Intention to treat can cause much consternation among clinicians, particularly in surgical trials where some outcomes may seem absurd – for example, continuing menstrual blood loss in women allocated to hysterectomy who did not undergo the operation but were analysed by intention to treat. A data analysis plan should be defined a priori, identifying the statistical analysis to be applied to the data. Infertility trials have potential for 'unit of analysis' error and in a review of 39 trials in this area, this error was identified in 32 (82%) studies.[63] Using live birth per couple as the primary outcome, with randomisation and analysis at the couple level, will protect studies from 'unit of analysis' errors.

PRESENTING RESULTS

Analysis should follow the data analysis plan as set out in the protocol and the CONSORT recommendations[71] should be observed. Particularly helpful is a trial chart which sets out in an explicit manner any exclusions or loss to follow-up. Results of subgroup analyses should be treated with caution and used mainly as hypothesis-generating exercises in most modest-sized trials. There should be a conscious attempt to limit discussion to the results generated by the trial and avoid speculation.

ETHICS OF TRIALS

The scientific rationale for conducting trials is collective equipoise. Clinicians need to be genuinely uncertain about the best treatment. In such a clinical situation, there should be no conflict between the interests of those participating in a trial and those who stand to gain in the future. The important issue is that participants are also in personal equipoise and give informed consent.

Despite awareness of its importance, there is evidence that some doctors do not seem to take informed consent as seriously as they should.[72] This may be because participants seem less willing to be randomised when they are given more preliminary data and made aware of any accumulating evidence of effectiveness. In many trials, a significant number of participants emerge from consultations expecting to benefit personally by their participation.

Some infertility-related procedures are described as 'licensed treatment' under the aegis of the Human Fertilisation and Embryology Authority (HFEA) in the United Kingdom. Clinical data pertaining to licensed treatments (including donor insemination, IVF and ICSI) are confidential and may not be revealed to researchers (including clinicians) who are not covered by the institutional HFEA treatment licence, without the explicit permission of the couple. This can create problems in accessing data, particularly follow-up data from notes or databases. Furthermore, trials involving manipulation of gametes and embryos need separate approval from the HFEA in addition to approval from the local ethics committee.

For all clinical trials, it is sensible from an ethical and financial point of view to have clear stopping rules as part of the original study design. An independent data monitoring committee should be available to review the results of interim analyses. Early stopping should only occur under pre-planned well-specified circumstances such as marked superiority or toxicity of one arm of the study which is greater than that originally hypothesised. Examples include stopping a trial evaluating the use of prophylactic antibiotics during hysteroscopic surgery where the control arm demonstrates a significantly higher rate of infection.[73] Alternatively, in a trial comparing a policy of single versus double embryo transfer during IVF (in order to prevent twin pregnancies) it may be appropriate to stop if the pregnancy rate in the single embryo group becomes unacceptably low.

CONCLUSION

Clinical trials in gynaecology have lagged behind those in other disciplines in terms of overall numbers as well as quality. There are few large multicentre trials, particularly surgical trials. The clinical population is heterogeneous and interventions under scrutiny diverse. Some treatments, such as those for infertility and unwanted fertility, target women (and their partners) who have specific reproductive health needs but are otherwise in good health. There is also a need for trials to be able to compare interventions that cross different treatment boundaries. Trialists need to design more pragmatic trials with clinically meaningful outcome measures. In gynaecology, these should be quality of life and satisfaction; in infertility, live birth rates per couple/woman. Finally, consideration should be given to collecting cost data where appropriate. This is often crucial in terms of planning gynaecological services which are effective, acceptable and affordable.

REFERENCES

1. Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ* (1996) **312**(7033): 742–4.
2. Out HJ, Schnabel PG, Rombout F, Geurts TB, Bosschaert MA, Coelingh Bennink HJ. A bioequivalence study of two urinary follicle stimulating hormone preparations: Follegon and Metrodin. *Hum Reprod* (1996) **11**(1): 61–3.
3. El-Refaey H, Rajasekar D, Abdalla M, Calder L, Templeton A. Induction of abortion with Mifepristone (RU 486) and oral or vaginal misoprostol. *New Engl J Med* (1995) **332**: 983–7.
4. The Aberdeen Endometrial Ablation Group. A randomised trial of hysterectomy versus endometrial ablation for the treatment of dysfunctional bleeding: clinical psychological and economic outcome at four years. *Br J Obstet and Gynaecol* (1999) **106**: 360–6.
5. Cooper KG, Jack SA, Parkin DE, Grant AM. Five-year follow up of women randomised to medical management or transcervical resection of the endometrium for heavy menstrual loss: clinical and quality of life outcomes. *B J Obstet Gynaecol* (2001) **108**(12): 1222–8.
6. Roland M, Torgerson DJ. Understanding controlled trials: what are pragmatic trials? *BMJ* (1998) **316**: 285.
7. Goverde AJ, McDonnell J, Vermeiden JPW, Schats R, Rutten FFH, Schoemaker J. Intrauterine insemination or in-vitro fertilisation in idiopathic subfertility and male subfertility: a randomised trial and cost-effectiveness analysis. *Lancet* (2000) **355**: 13–18.
8. Khan KS, Daya S, Collins JA, Walter SD. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertil Steril* (1996) **65**(5): 939–45.
9. McAlister FA, Strauss SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials. *JAMA* (2003) **289**: 2545–53.
10. Fender GRK, Prentice A, Gorst T, Nixon RM, Duffy SW, Day NE *et al* Randomised controlled trial of educational package on management of menorrhagia in primary care: the Anglia menorrhagia education study. *BMJ* (1999) **318**: 1246–50.
11. Morrison J, Carroll L, Twaddle S, Cameron I, Grimshaw J, Leyland A *et al* Pragmatic randomised controlled trial to evaluate guidelines for the management of infertility across the primary care-secondary care interface. *BMJ* (2001) **322**: 1–7.
12. Kerry SM, Bland JM. Sample size in cluster randomisation. *BMJ* (1998) **316**: 549.
13. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. *BMJ* (1998) **316**: 1455.
14. Donner A, Klar N. *Design and analysis of cluster randomization trials in health research*, 1st edn. London: Arnold (2000).
15. Aboulgar MA, Mansour RT, Serour JI, Amin YM, Kamal A. Prospective controlled randomised study of in-vitro fertilisation versus intracytoplasmic sperm injection in the treatment of tubal factor infertility with normal semen parameters. *Fertil Steril* (1996) **66**: 753–6.
16. Ruiz A, Remohi J, Minguez Y, Guanes P, Simon C, Pellicer A. The role of in vitro fertilisation and intracytoplasmic sperm injection in couples with unexplained infertility after failed intrauterine insemination. *Fertil Steril* (1997) **68**: 171–3.

17. Bhattacharya S, Hamilton MPR, Shabban M, Khalaf Y, Seddler M, Ghobara T *et al*. A randomised controlled trial of in-vitro fertilisation (IVF) and intra-cytoplasmic sperm injection (ICSI) in non-male factor infertility. *Lancet* (2001) **357**: 2075–9.
18. Torgerson D, Sibbald B. Understanding clinical trials: what is a patient preference trial? *BMJ* (1998) **316**: 360.
19. Brocklehurst P. Partially randomised patient preference trials. *Br J Obstet and Gynaecol* (1997) **104**(12): 1332–5.
20. Brewin CR, Bradley C. Patient preferences and randomised clinical trials. *BMJ* (1989) **299**: 313–5.
21. Henshaw RC, Naki SA, Russell IT *et al*. Comparison of medical abortion with vacuum aspiration; women's preferences and acceptability of treatment. *BMJ* (1993) **307**: 714–7.
22. Ashok PW, Kidd A, Flett GM, Fitzmaurice A, Graham W, Templeton A. A randomised comparison of medical abortion and surgical vacuum aspiration at 10–13 weeks gestation. *Hum Reprod* (2002) **17**(1): 92–8.
23. Cooper KG, Parkin DE, Garret AM, Grant AM. A randomised comparison of medical and hysteroscopic management in women consulting a gynaecologist for treatment of heavy menstrual loss. *Br J Obstet and Gynaecol* (1997) **104**: 1360–6.
24. Torgerson DJ, Klaber-Moffett J, Russell I. Patient preferences in randomised trials: threat or opportunity? *J Health Serv Policy* (1996) **1**(4): 194–7.
25. Piaggio G, Pinol APY. Use of the equivalence approach in reproductive health clinical trials. *Stat Med* (2001) **20**: 3571–87.
26. Driscoll GI, Tyler JPP, Hangan JT, Fisher PR, Birdsall MA, Knight DC. A prospective, randomized, controlled, double-blind, double dummy comparison of recombinant and urinary HCG for inducing oocyte maturation and follicular luteinization in ovarian stimulation. *Hum Reprod* (2000) **15**: 1305–10.
27. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ* (1996) **313**: 36–9.
28. Röhmel J. Therapeutic equivalence investigations: statistical considerations. *Stat Med* (1998) **17**: 1703–14.
29. Ebbutt AF, Frith L. Practical issues in equivalence trials. *Stat Med* (1998) **17**(15–16): 1691–1701.
30. Pandian Z, Bhattacharya S, Nikolaou N, Vale L, Templeton A. In-vitro fertilisation for unexplained subfertility (Cochrane Review). *The Cochrane Library*, Issue 4. Oxford: Update Software (2002).
31. Roland M, Torgerson DJ. Understanding controlled trials: what outcomes should be measured? *BMJ* (1998) **317**: 1075–80.
32. Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? *Ann Intern Med* (1996) **125**: 605–613.
33. Cranney A, Welch V, Adachi JD, Guyatt G, Krolicki N, Griffith L, Shea B, Tugwell P, Wells G. Etidronate for treating and preventing postmenopausal osteoporosis (Cochrane Review). The Cochrane Library, Issue 2. Oxford: Update Software (2002).
34. Patrick DL, Bergner M. Measurement of health status in the 1990s *Annu Rev Public Health* (1990) **11**: 1165–83.
35. Gill TM, Feinstein AR. A critical appraisal of the quality of life measurements. *JAMA* (1994) **272**(8): 619–26.
36. Fayers PM, Sprangers MAG. Understanding self-rated health. *Lancet* (2002) **359**: 187–8.
37. Ware JE, Sherbourne CD. The MOS 36-item short form health survey (SF-36). Conceptual framework and item selection. *Med Care* (1992) **30**: 473–83.
38. McGee H, Hannah M, O'Boyle CA, Hickey A, O'Malley K, Joyce CRB. Assessing the quality of life of the individuals: the SEIQOL with a health and gastroenterology unit population. *Pschol Med* (1991) **21**: 749–59.
39. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease specific quality of life in clinical trials. *Can Med Assoc J* (1986) **134**: 889–95.
40. Kelleher CJ, Cardozo LD, Khullar V, Salvatore S. A new questionnaire to assess the quality of life of urinary incontinent women. *Br J Obstet Gynaecol* (1997) **104**(12): 1374–9.
41. Moos RH. Typology of menstrual cycle symptoms. *Am J Obstet Gynecol* (1969) **103**: 390–402.
42. Jones G, Kennedy S, Barnard A, Wong J, Jenkinson C. Development of an endometriosis quality-of-life instrument: The Endometriosis Health Profile-30. *Obstetr Gynaecol* (2001) **98**(2): 258–64.
43. Schneider HP, Heinemann LA, Rosemeier HP, Potthoff P, Behre HM. The Menopause Rating Scale (MRS): comparison with Kupperman index and quality-of-life scale SF-36. *Climacteric* (2000) **3**(1): 50–8.
44. Sanders C, Egger M, Donovan J, Tallon D, Frankel S. Reporting on quality of life in randomised controlled trials: bibliographic study. *BMJ* (1998) **317**: 1191–4.
45. Graham WJ. Outcomes and effectiveness in reproductive health. *Soc Sci Med* (1998) **47**(12): 1925–36.
46. Williams B. Patient satisfaction: a valid concept? *Soc Sci Med* (1994) **38**: 509–16.
47. Sitzia J, Wood N. Patient satisfaction: a review of issues and concepts. *Soc Sci Med* (1997) **45**(12): 1829–43.

48. Bury M. Doctors, patients and interactions in health care. In: *Health and illness in a changing society*. London: Routledge (1997).
49. Department of Health. *The new NHS: Modern and Dependable*. London: HMSO (1998).
50. Williams B, Coyle J, Healy D. The meaning of patient satisfaction: an explanation of high reported levels. *Soc Sci Med* (1998) **47**(9): 1351–9.
51. Hopton JL, Howie JGR, Porter MD. The need to look at the patient in General Practice satisfaction surveys. *J Fam Pract* (1993) **10**: 82–7.
52. Williams SJ, Calnan M. Key determinants of consumer satisfaction in general practice. *J Fam Pract* (1991) **8**: 237–42.
53. McIver S, Meredith P. There for the asking: can the government's planned annual survey really measure patient satisfaction? *Health Serv J* (1998) **19**: 26–7.
54. Leimkuhler A, Muller U. Patient satisfaction – artefact or social fact. *Nervenarzt* (1996) **67**: 765–73.
55. Owens DJ, Batchelor C. Patient satisfaction and the elderly. *Soc Sci Med* (1996) **42**: 1483–91.
56. Drummond MF, Davies L. Economic analysis alongside clinical trials. Revisiting the methodological issues. *Int J Technol Assess Health Care* (1991) **7**: 561–73.
57. Barber JA, Thompson SG. Analysis and interpretation of cost data in randomised controlled trials: review of published studies. *BMJ* (1998) **317**: 1195–1200.
58. Fayers PM, Hand DJ. Generalisation from phase III clinical trials: survival, quality of life and health economics. *Lancet;* (1997) **350**: 1025–7.
59. Briggs A. Economic evaluation and clinical trials: size matters. *BMJ* (2000) **321**: 1362–3.
60. Bowling A. *Measuring disease: a review of disease specific quality of life measurement scales*. Buckingham: Open University Press. (1994).
61. Peto R, Collins R, Gray R. Large scale randomised evidence: large sample trials and overviews of trials. *J Clin Epidemiol* (1995) **48**: 23–40.
62. Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall (1991).
63. Vail A, Gardner E. Common statistical errors in the design and analysis of subfertility trials. *Hum Reprod* (2003) **18**: 1000–4.
64. Brooks ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analysis in randomised controlled trials: quantifying the risks of false positives and false negatives. *Health Technol Assess* (2001) **5**(33): 1–64.
65. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance not choice. *Lancet* (2002) **359**: 515–19.
66. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* (1995) **273**: 408–12.
67. Moher D, Pham B, Jones A *et al.* Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analysis? *Lancet* (1998) **352**: 609–13.
68. Juni P, Altman D, Egger M. Assessing quality of controlled trials. *BMJ* (2001) **323**: 42–6.
69. Pocock SJ. *Clinical trials: a practical approach*. Chichester: John Wiley & Sons, Ltd (1983).
70. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet* (2002) **359**: 781–5.
71. Bury M. CONSORT statement. Doctors, patients and interactions in health care. In: *Health and illness in a changing society*. London: Routledge (1997).
72. Edwards SJ, Lilford RJ, Hewison J. The ethics of randomised controlled trials from the perspectives of patients, the public and health care professionals. *BMJ* (1998) **317**: 1209–12.
73. Bhattacharya S, Parkin DE, Reid TMS, Abramovich DR, Mollison J, Kitchener HC. A prospective randomised study of the effects of prophylactic antibiotics on the incidence of bacteraemia following hysteroscopic surgery. *Eur J Obstet Gynaecol Reprod Biol* (1995) **63**: 37–40.

25

Early Pregnancy Termination

BEVERLY WINIKOFF[1] AND BATYA ELUL[2]
[1]Gynuity Health Projects, New York, NY 10010, USA
[2]Mailman School of Public Health, Columbia University, New York, NY 10034, USA

INTRODUCTION

Induced abortion refers to the intentional interruption of pregnancy before viability. Of the 210 million pregnancies estimated to occur worldwide each year, nearly one-quarter or approximately 50 million are believed to end in abortion.[1,2] Although this represents a global rate of 35 abortions for every 1000 women of reproductive age (15–44 years), abortion rates vary significantly by region. For example, in North America and Western Europe, rates are low, generally around 11–22 abortions per 1000 women of reproductive age, while in Eastern Europe, the rates are amongst the highest in the world, with about 90 abortions per 1000 women of reproductive age. In the developing world, abortion rates for Asia, Africa and Latin America range from 31 to 37 per 1000 women of reproductive age.[3]

Early pregnancy terminations generally involve abortions conducted during the first trimester of pregnancy or up to 12 weeks since the last menstrual period (LMP). The vast majority of induced abortions are conducted in the first trimester, the majority of which occur before nine weeks' gestation. In the United States and the United Kingdom, for example, 90% of abortions occur in the first trimester, and 60% are performed before nine weeks' gestation.[2,4]

Today, most early abortions are conducted using surgical techniques under local anaesthesia, but the use of medical abortion for early termination is growing rapidly worldwide. In the United States, for example, medical abortion accounted for only 0.3% of all early abortions in 1996,[5] but by 2005 this figure had increased to approximately 11%.[4,6] In several European countries, where medical abortion has been available for over 15 years, 50–60% of all first-trimester abortions are performed using medical methods.[7]

Large numbers of randomised controlled trials and observational studies of methods for early pregnancy termination were conducted in the late 1960s and 1970s when abortion was first legalised in many developed and several developing countries. These early studies focused largely

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

on the safety and efficacy of three primary surgical abortion techniques – manual vacuum aspiration (MVA), electric vacuum aspiration (EVA), and dilatation and curettage (D&C) – and, in some cases, compared aspects of the three techniques of early surgical abortion. The past 15 years have seen a resurgence in research on early abortion, due largely to the availability of medical abortion, a profoundly different method of early abortion. These more recent studies have compared not only the safety and efficacy of early surgical and medical abortion, but also the relative acceptability of each method. Studies conducted in the past decade have also examined modifications to the original medical abortion regimen, both with regard to dosing and to service-delivery factors. As surgical and medical abortion are two intrinsically different methods of abortion, much attention has also been given to methodological issues relevant to research on early pregnancy termination in recent years. This chapter reviews some of those issues.

BACKGROUND: METHODS OF EARLY PREGNANCY TERMINATION

Early pregnancy termination can be accomplished using several different methods that fall under the broad categories of surgical or medical procedures (Figure 25.1 and Table 25.1). Clinical and practical aspects of both surgical and medical abortion can be studied from the point of view of medical systems, providers and patients (Table 25.2).

SURGICAL ABORTION

Surgical abortion is the most commonly used method of early termination worldwide. Surgical abortion is an invasive procedure conducted by a trained medical practitioner and entails the physical removal of the products of conception (POC) from a woman's uterus. One of the three different early surgical abortion techniques – manual vacuum aspiration (MVA), electrical vacuum aspiration (EVA) and dilatation

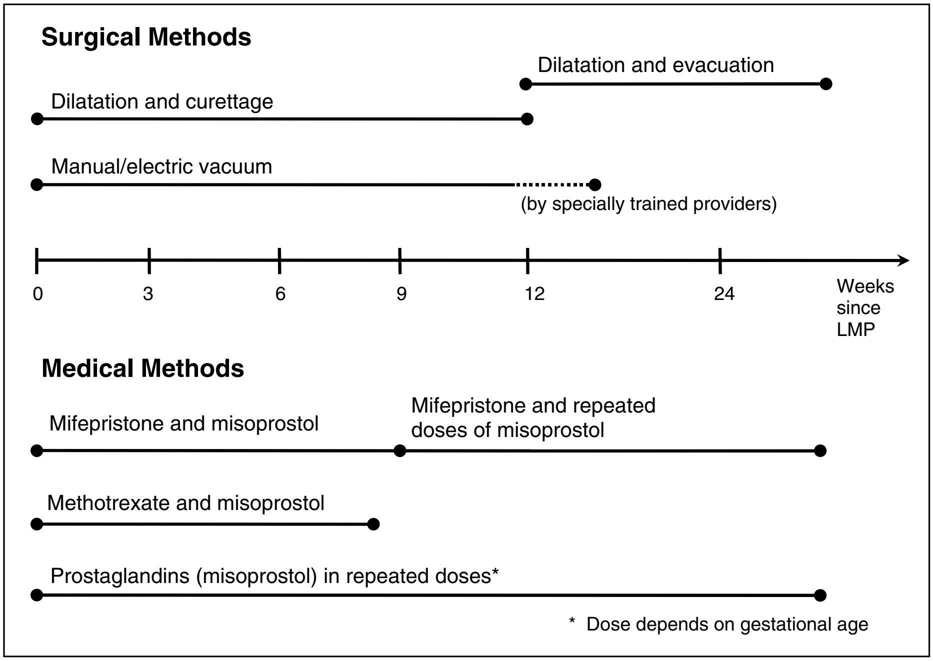

Surgical Methods
Dilatation and evacuation
Dilatation and curettage
Manual/electric vacuum
(by specially trained providers)
0
3
6
9
12
24
Weeks since LMP
Medical Methods
Mifepristone and misoprostol
Mifepristone and repeated doses of misoprostol
Methotrexate and misoprostol
Prostaglandins (misoprostol) in repeated doses*
* Dose depends on gestational age

Figure 25.1. Timing of early abortion methods (adapted from World Health Organization[8] and Cates and Ellertson[9]).

Table 25.1. Overview of surgical and medical abortion methods

			Surgical		
Method	Device	Mechanism of action	Success rate	Salient features	Possible drawbacks
Manual vacuum aspiration (MVA)	Portable aspiration syringe. Canulae range from 4 to 12 mm in diameter	Uterine contents are gently aspirated through a sterile canula into a handheld vacuum device in under 10 min	95–100%	Extremely safe; complications rare. Can be performed in an examining room and by a mid-level provider (requires special training and practice). Appropriate for low-resource settings and can be done quickly	Used with mild sedation and/or local anaesthesia
Electrical vacuum aspiration (EVA)	Electric aspiration pump. Canulae range from 4 to 12 mm in diameter	Uterine contents are gently aspirated through a sterile canula into an electric vacuum device in under 10 min	95–100%	Extremely safe; complications rare. Can be performed in an examining room and by a mid-level provider (requires special training and practice). Appropriate for low-resource settings and can be done faster than MVA	Used with mild sedation and/or local anaesthesia. Noise from the electric aspirator may make some women uncomfortable
Dilatation and curettage (D&C)	Uterine dilators and a sharp, sterile, metal curette	Uses a sharp curette to scrape and empty contents of uterus after cervical dilation. Cervical priming prior to dilation can be accomplished with misoprostol, other prostaglandins, mifepristone or laminaria		Extremely safe; physicians are most frequently trained to perform D&C	Used with mild or heavy sedation and/or local or general anaesthesia. Longer recovery time. Causes heavier bleeding than vacuum with 2–3 × higher rate of major complications. Requires a physician and sterile equipment. More expensive than MVA or EVA

			Medical		
Method	Drug	Mechanism of action	Success rate	Salient features	Possible drawbacks
Mifepristone and misoprostol	Mifepristone is an antiprogesterone. Misoprostol is a synthetic prostaglandin	Mifepristone terminates the pregnancy by inhibiting progesterone; it also primes the cervix. Misoprostol induces cervical dilation and uterine contractions, expelling the products of conception	95–98%	Extremely safe; avoids surgical intervention and anaesthesia. Is more natural and resembles menses. Woman has more control, can be undertaken in a home setting	Some protocols require extra visits to clinic. Some side effects such as cramping, bleeding, and nausea. Mifepristone not available or approved in many countries

(*continued overleaf*)

Table 25.1. (*continued*)

Method	Device	Mechanism of action	Medical Success rate	Salient features	Possible drawbacks
Methotrexate and misoprostol	Methotrexate is an antimetabolite of folic acid. Misoprostol is a synthetic prostaglandin	Methotrexate inhibits growth of rapidly dividing cells in the embryo and trophoblastic tissue. Misoprostol induces cervical dilation and uterine contractions, expelling the products of conception	90–95%	Extremely safe; avoids surgical intervention and anaesthesia. Is more natural and resembles menses. Woman has more control, can be undertaken in a home setting	Some protocols require extra visits to clinic. Some side effects such as cramping, bleeding, and nausea. Long duration – may take from 3–4 weeks to complete. Teratogenicity of methotrexate is an important concern for ongoing pregnancies
Misoprostol alone	Misoprostol is a synthetic prostaglandin	Misoprostol induces cervical dilation and uterine contractions, expelling the products of conception	85–90%	Extremely safe; avoids surgical intervention and anaesthesia. Is more natural and resembles menses. Woman has more control, can be undertaken in a home setting. Drug is widely available and at low costs	Some protocols require extra visits to clinic. More severe side effects and less effective than when used in conjunction with mifepristone or methotrexate

Table 25.2. Surgical vs. medical abortion – characteristics of each which may help the woman choose

Surgical	Medical
Quick	Avoids surgery and anaesthesia
More certain, procedure complete within minutes	Waiting, uncertainty, may take several days to complete
Woman is less involved	Woman has more control, more involved
Anaesthesia possible	Pain can be managed with analgesics
Invasive	More natural, like menses
Risk of complications such as uterine/ cervical injury or infection	Extra clinic visit depending on protocol used
Bleeding, cramping, nausea (actual or feared)	Bleeding, cramping, nausea (actual or feared)

and curettage (D&C) – is used, depending on the woman's gestational age, local medical practice and provider training. MVA and EVA are preferred and considered safer than D&C for early terminations.[9,10]

Because instruments need to be inserted into the uterus, cervical dilation is common before surgical abortions performed in the latter weeks of the first trimester. Cervical dilation reduces tearing and damage to the cervix, and lets the provider perform the abortion faster and with greater ease. Cervical dilation can be performed with mechanical dilators, osmotic dilators (e.g. laminaria), or with drugs (e.g. misoprostol 3–4 hours prior to the operation, mifepristone 36 hours prior, or vaginal gemeprost 3 hours prior). In some cases, however, cervical dilation is overused, increasing the length and cost of the procedure, as well as the potential for damage to the cervix if it is done too quickly or ineptly.[8,11]

Pain management is also an important component of surgical abortion. Local anaesthesia is recommended for early procedures and is generally favoured over general anaesthesia by both clinicians and women. General anaesthesia causes post-procedure grogginess and nausea, thus increasing the length of the recovery period. Local anaesthesia allows the woman to be alert during the procedure, although women can be given additional sedatives or tranquilisers to reduce anxiety, and analgesics to minimise post-procedure pain. In addition, local anaesthesia is frequently deemed safer and has been associated with less blood loss than general anesthesia.[12] For all types of surgical abortion, patients are usually asked to return for a follow-up visit.[8,13,14]

Manual Vacuum Aspiration

MVA empties the products of conception from the uterus using gentle suction from a sterile manual syringe. Often used for early surgical abortion, particularly in developing countries, MVA is highly effective, with success rates of 95–100%. The method is extremely safe and quick, taking 3–11 minutes on average to complete. As no electricity is required, MVA equipment can be used at almost any facility, regardless of the level of the health care system. The procedure can be provided by a trained mid-level provider.[8,10,13]

Electrical Vacuum Aspiration

EVA employs the same technique as MVA, differing only in the use of an electric pump (rather than a manual pump), which makes the procedure faster. Some women, however, dislike the noise of the pump. Like MVA, success rates for EVA range from 95 to 100%, and the method can be performed by trained mid-level providers. Electricity is a prerequisite for operating EVA equipment, however, which may make it unsuitable for extremely low-resource environments.[8,10]

Dilatation and Curettage

D&C, or sharp curettage, involves cervical dilation and utilises a sharp, sterile, metal curette to clean the walls of the uterus. It is usually performed beginning at six weeks' gestational age, although some providers prefer delaying use of this technique until eight weeks' gestational age. Although most gynaecologists receive training in D&C, it should only be used when other methods are unavailable. Compared with vacuum aspiration, the procedure is associated with more blood loss, more pain and higher rates of complication. D&C is frequently performed under general anaesthesia or local anaesthesia combined with heavy sedation, though it can also be performed using local anaesthesia alone. The procedure is frequently conducted in an operating room as opposed to an examination room. While mid-level providers have been trained to provide D&C in some settings, the method is most commonly performed by a trained general practitioner or gynaecologist.[8,10,15]

MEDICAL ABORTION

Early medical abortion, a non-invasive procedure based on the administration of drugs, is a relatively new alternative to surgical abortion. Medical abortion was first registered in France and China in 1988, has since been registered in an additional 33 countries, and now accounts for the majority of early abortions in some jurisdictions. In contrast to some methods of early surgical abortion, medical abortion does not require anaesthesia. Many women describe medical abortion as a more natural procedure which resembles menstruation and which they, as opposed to a physician, can control.[15–18]

As with surgical abortion, there are several different types, or regimens, of medical abortion. The only regimens approved by drug regulatory agencies and commercially marketed combine the use of the antiprogestin mifepristone and a prostaglandin, usually misoprostol. Medical abortion regimens using misoprostol alone or in combination with methotrexate, an antimetabolite, have also been developed but have not yet been registered in any country. The effect of all these approaches is similar to that of a spontaneous abortion. Bleeding occurs for an average of

9–11 days (similar to or heavier than a menstrual period), although in rare cases can continue for much longer.[8,9] Side effects may include nausea, pain from cramps, diarrhea, or vomiting.[13,17,19]

Mifepristone and Misoprostol

Mifepristone acts as a progesterone antagonist, causing the foetus to detach from the uterine walls. It is generally administered at a clinic by a clinician, mostly because the registering pharmaceutical companies and some regulatory agencies require restricted models of distribution. From a medical perspective, mifepristone is not a dangerous drug and certainly does not require distribution more onerous than normal prescription status would confer. Misoprostol, a synthetic prostaglandin registered for the prevention and treatment of gastric ulcers and available on prescription in over 80 countries, is usually administered at home or at the clinic 24–72 hours after mifepristone ingestion. It induces uterine contractions which lead to the expulsion of the POC.[9,17] Misoprostol is most commonly administered orally (400 or 600 mcg) but in some settings is administered vaginally (800 mcg).[15,20] A follow-up visit is usually required one to two weeks after mifepristone administration to confirm that a complete abortion has been achieved. In cases of ongoing pregnancy (about 1–3% of cases), surgical intervention is used to end the pregnancy; thus, clinics offering medical abortion must be prepared to offer surgical interventions or to refer patients requiring surgical interventions to another facility.[10] Mifepristone-based regimens have been approved for use up to 7–9 weeks' gestation depending on the country, although recent studies have evaluated variants of the original regimen for use in pregnancies from 9 to 12 weeks' gestation.[8,21]

Methotrexate and Misoprostol

This regimen follows a similar 2–3 clinic-visit protocol to the mifepristone – misoprostol regimen described above. Methotrexate is a folic acid antimetabolite and cytotoxic drug which stops rapidly dividing cells, such as those in the embryo, from growing. It is followed by misoprostol approximately 6–7 days later, which induces uterine contractions and leads to the expulsion of the POC.[17] The methotrexate – misoprostol regimen is usually used through eight weeks' gestation with a success rate of around 92%. As it may take as long as 3–4 weeks for a complete abortion, some women opt out of completing this regimen when presented with the option. In addition, teratogenicity is an important concern for ongoing pregnancies.[8,10,22]

Misoprostol Alone

Misoprostol, a widely available and inexpensive prostaglandin E1 analogue, causes strong uterine contractions which lead to the expulsion of the POC.[17,23] While it can be used alone when mifepristone and methotrexate are unavailable, it has lower effectiveness than when combined with the above drugs, generally in the range of 85–90% when used up to nine weeks' gestation.[22] Moreover, when misoprostol is used alone, women may experience more painful cramps, and, depending on dose and route of administration, intensified gastrointestinal side effects compared with when it is combined with mifepristone or methotrexate.

METHODOLOGICAL ISSUES: OUTCOME MEASURES

Three primary outcome measures have been assessed in clinical studies of early abortion: efficacy, safety and acceptability.

EFFICACY

Among the most important characteristics of an abortion method is its efficacy. Information on method efficacy drives regulatory decisions as to whether a specific method should be approved for use in a given country, provider decisions to offer a method, and patient decisions to select it.

The definitions, assessments and analytical approaches used to measure efficacy in clinical studies vary by method of abortion, and in the case of medical abortion, by study as well. For early surgical abortion, measurement of 'failure' is relatively straightforward and defined as any repeat surgical intervention for confirmed or suspected incomplete abortion or ongoing pregnancy. Incomplete abortions or ongoing pregnancies may be identified immediately following the surgical procedure through the examination of the POC. As it may be difficult to distinguish the POC from other materials obtained during the procedure in the earliest stages of pregnancy, in some cases, failed surgical abortions are identified when study participants return to the clinic complaining of symptoms suggestive of incomplete abortion (i.e. cramping, fever) or ongoing pregnancy (i.e. breast tenderness, morning sickness). In some studies, surgical abortion failures are subclassified as either incomplete abortions or ongoing pregnancies.[24,25] As surgical abortion is a discrete event, occurring over several minutes, the total failure rate or cure rate is computed by simple division and expressed as a proportion: the number of women receiving repeat surgical interventions divided by the total number of women receiving surgical abortions in the trial. (Strictly speaking, this proportion is the risk of failure. In the discrete case where the procedure occurs quickly, though, the risk approximates the rate.)

For early medical abortion, definitions and assessments of 'failure' have varied widely, as have the methods used to calculate total failure rates. In some cases, these variations have been so acute as to preclude comparisons across trials. For example, definitions of success in studies of mifepristone – misoprostol abortion range from achieving a non-viable pregnancy – either complete or incomplete abortion – within 48 hours after taking mifepristone[26] or within 24 hours after administration of the first dose of misoprostol,[27] to expulsion of the foetus even when surgical evacuation of the placenta was required,[28] to complete abortion without the need for surgical intervention.[24,25,29] While the later definition is perhaps the most common, the maximum waiting time used in this definition has varied greatly as well, from 24 hours to 30 days. As variations in waiting periods can significantly affect efficacy rates (because mifepristone medical abortion will eventually result in a complete abortion in almost all cases without recourse to surgical intervention), a follow-up period that best reflects feasible service-delivery scenarios should be used. In most cases, this will be a 7- or 14-day follow-up period.

Similarly, while nearly all published studies calculate medical abortion failure rates as a simple proportion (analogous to the approach used for surgical abortion), there has been considerable variation in the choice of denominator and even the choice of numerator. In some cases participants who do not follow the study protocol or are lost to follow-up are omitted entirely from the analysis and excluded from both the numerator and denominator of the failure rate[30] whereas in other studies, participants whose outcomes are unknown are counted as failures, and included in both the numerator and denominator of the failure rate.[31] Moreover, there is variation in the subclassification of failures. For instance, several studies subclassified failures as incomplete abortions or ongoing pregnancies while others have used a classification of incomplete abortion, ongoing pregnancy or required intervention for medical indications such as bleeding.[24,25,32]

Calculation of medical abortion failure rates is further complicated by the fact that women can opt out of the process before a determination of efficacy can be made and obtain a surgical abortion, and providers can perform surgical interventions that are not medically necessary. Winikoff *et al.*[33] proposed a definition of failure for medical abortion that accounts for these unique features of the method and for the fact that women chose medical abortion, at least in part, to avoid surgery: any surgical intervention for any reason whatsoever is considered a failure of the medical method. Using this definition, the total failure rate is the number of women requiring surgical intervention divided by the number of women using the medical method in the trial. Interventions that make up the numerator are further classified as user choice – medically unnecessary interventions that are performed at the patient's or providers' request before the study end – and method failures – interventions that are performed for incomplete abortion or ongoing pregnancy at the end of the study or to treat complications, such as profuse bleeding, during or at the end of the study.

Trussell and Ellertson[33] proposed an extension to this definition that accounts for whether women are able to or opt to follow the entire study protocol, from taking the drugs and returning for follow-up at the prescribed times, through the calculation of both perfect-use and typical-use failure rates. They suggest that non-compliant women only be excluded from the calculations of the perfect-use rates. Life table procedures, the standard tool for analysis of contraceptive failures, have also been proposed to account for women who begin the abortion procedure but may not finish it (due to their own decisions or those of their providers) and to depict the temporally drawn-out nature of medical abortion.[34] A reanalysis of data from six published studies using the life table approach showed that the traditional proportion method of efficacy calculation is biased upward when those lost to follow-up are excluded from the analysis and downward when they are included in the analysis.[35] While this approach produces an unbiased estimate of medical abortion efficacy rates, it has not been widely used to date.

SAFETY

Increasing access to a variety of methods of early abortion has renewed interest in the safety profile, and particularly in the relative safety profile, of both surgical and medical abortion. Increasingly, researchers report not only on serious adverse events, but also on the more routine side effects experienced by women undergoing abortions, including bleeding (length and quantity), cramping, pain and nausea. Side effects must be well understood so that proper training and counselling materials can be developed. Realistic expectations are essential for the method to be used by suitable providers and patients, and to minimise anxiety and unnecessary backup interventions. Additionally, as regulatory decisions and clinical protocols are typically based in part on the side-effect profile of the method, accurate documentation of the severity and incidence of side effects is critical.

Several different approaches have been used to obtain data on side effects experienced at different stages of the abortion procedure or process. Most commonly, data are gathered from patient reports or provider observations at each clinic visit. These data are generally used to calculate an overall incidence of any or specific side effects during the treatment period (i.e. the proportion that 'ever experienced' a side effect during the study). A few studies, however, have separated side effects reported during different stages of the abortion process or procedure.[29,36] For example, in the case of medical abortion, side effects experienced immediately after mifepristone administration are disaggregated from those experienced in the 48 hours between mifepristone and misoprostol administration, in the 2–4 hours following misoprostol use, and in the period between misoprostol administration and the follow-up visit. For surgical abortion, side effects experienced directly or shortly after the procedure are separated from those experienced in the period between the procedure and the follow-up visit.

In some studies, patients have been provided with side-effect diaries on which to record the occurrence of specific side effects on each day of the study.[24,25,37,38] Data from such diaries are used to calculate the mean total number of days each side effect is experienced as well as the range of days on which it is experienced. Visual analogue scales have also been used to document the severity of pain or quantity of blood loss during early abortion.[39,40] In a handful of studies, blood has been collected and median blood loss quantified.[41–43] In the initial studies of medical abortion in developing countries, haematologic measures were obtained before and after the abortion to determine whether perceptions of blood loss reflected clinically meaningful loss and whether the method was suitable for populations where anaemia is prevalent.[24,25,35]

While the measures used to collect data on side effects during early abortion are well developed, interpretation of those data is complicated by several factors. First, study participants are rarely questioned about baseline pregnancy symptoms such as nausea and vomiting and thus symptoms of pregnancy are often erroneously attributed to the abortion procedure, inflating the overall incidence of these side effects. Additionally, of concern in studies comparing medical and surgical abortion, medical abortion patients may report more side effects than surgical abortion patients simply because they are not anaesthetised during the procedure. Similarly, in the case of bleeding, medical abortion patients observe their blood loss while, in the case of surgical abortion, most of the fluid is extracted by the provider and rarely observed by the patient.

ACCEPTABILITY

In addition to examining efficacy and safety, clinical trials of methods of early pregnancy termination have examined acceptability of the method. Acceptability provides information on whether patients and, less commonly but not less importantly, providers will accept a new method of abortion or a modification to an existing method of abortion. Rather than being strictly an inherent quality of a method, acceptability is the result of an interaction among the values a patient holds, the patient's perceptions of the attributes of a particular method of early abortion, and the service-delivery system the patient encounters. If a method's perceived attributes correspond to a patient's values, she may consider the method desirable, preferable or acceptable. Anything affecting values or perceptions can therefore affect acceptability. Characteristics that may influence both values and perceptions include ethnicity, nationality, culture, class, education, personality and experience. Perceptions are also influenced by the inherent qualities of the method and the available alternatives.[44–46]

Most data on the acceptability of methods of early abortion come from interviews with patients before and after their abortions. Such an approach records a patient's thoughts about the method's attributes before and after use, allowing inferences about any changes that occur. Commonly used measures include overall satisfaction, ratings of side effects compared with expectations, best and worst features of the method, and whether the patient would use the method again or recommend it to others.[24,25,40,47,48]

OTHERS

Other outcomes assessed in clinical trials of early abortion include completion of surgical abortion using MVA without the need for general anaesthesia[49] or an overnight stay, and the proportion of patients who select home administration (as opposed to clinic administration) of misoprostol following mifepristone or methotrexate.

METHODOLOGICAL ISSUES: STUDY DESIGN

High-quality and reliable clinical trials are expected to conform to several standards of clinical care and research practice. Among the criteria that are used to define good-quality research are: a study population representative of some clinically meaningful and identified universe, proper informed consent procedures, adequate sample size, randomisation of treatments, and blinding, preferably of subject and clinician. At times, research on early abortion methods meets these standards only with difficulty, in part because of the nature of the services being studied and in part because of the methods of early abortion themselves.

STUDY POPULATION

Clearly, studies of abortion technology can only be carried out among women presenting to health care facilities for termination of pregnancy and thus are not representative of the universe of all pregnant women. Thus, results cannot necessarily be extrapolated to pregnant women who are not seekers of abortion in each setting where the studies are undertaken. Additionally, because of the very nature of recruitment into studies of new technologies, there may be unknown biases in the types of women who participate in early abortion studies. For example, studies of innovation in abortion techniques need to be carried out in centres where abortion services are already offered. If women present for the usual service, they may not want to bother with the extra trouble of study participation (i.e. informed consent, additional paperwork time, and possibly lab work) when they can get the treatment they came for without enrolling in a study. In this sense, it may be the attraction (or not) of the novelty being studied that determines the extent to which women will be willing to participate in a study of early abortion – and the extent to which the resulting study population differs from the population of all women seeking abortion. Additionally, when enrolment in a study allows women some chance of getting a desirable new abortion alternative (e.g. medical abortion where only surgery had been available before), there is an impetus for 'knowledgeable' but dissatisfied consumers (i.e. those who had used the surgical alternative previously and did not like it) to volunteer for the new alternative, regardless of whether the novelty is allocated by patient option or is randomised. Such 'dissatisfied users' may differ significantly from the overall population of abortion service users. Indeed, studies of medical versus surgical abortion have disproportionately attracted women who have had previous abortions (usually around 50%), suggesting that this dynamic may well be important.[50,51] Similar motivations may come into play if the new technology or service being offered is free and the standard service has out-of-pocket costs or co-payments for which the woman is responsible.[44]

INFORMED CONSENT

Ideally, the informed consent for a study should reliably give information about a new alternative – if that is a patient option – or, if the study is randomised, give equally extensive, fair and objective descriptions of both alternatives that the patient may receive. However, this proves difficult when one technology or service is already standard and another is new to the staff as well as the patients. Medical staff members responsible for recruitment are familiar with patient responses and possible side effects and complications of the treatments they are accustomed to offering, but their descriptions of novel approaches are often reflections of 'book learning' and can be sparse and stilted. Furthermore, they are often unable to answer patient questions with the same degree of assuredness and detail. Even if providers are well aware of the success rates and side effects of new therapies, they do not know if those figures will apply to the method in their own hands, putting them at a double disadvantage: they do not have personal experience with the method being studied and they also do not know if the information written into the informed consent from other sources will apply to their own clinical population.

SAMPLE SIZE

The sample size used for any study comparing methods, types or regimens of early abortion should be able to show that the postulated differences in the therapies under study, if found, are likely to be real and not chance occurrences. Even for descriptive studies, it is desirable to have a sufficiently large enough sample so that point estimates of effect and side effects have relatively narrow confidence intervals. Virtually no sample size in clinical studies, however, is large enough to reveal truly rare side effects of complications of any therapy, including early abortion technologies. Problems that occur a few times per hundred thousand treatments are impossible to predict on the basis of the kind of clinical studies that can be carried out in the real world.

In addition, early abortion technologies are so effective and so safe that it is very difficult to implement studies large enough to capture the small differences between therapies. This issue has been particularly vexing in identifying the best regimens of mifepristone combined with misoprostol for early abortion. The quest to discover ways to raise the success rate of an already highly effective therapy is seductive but frustrating. First, there needs to be clinical importance to the difference: with therapies that already have success rates about 95%, it is sometimes hard to argue that millions of dollars should be spent on research to show how to increase success to 97%. (To show that a new dosing regimen has a 2% point higher success rate than one now giving a 95% success rate, a comparative study with 80% power would require 1170 women in each arm.) On the other hand, the litigious climate surrounding many developed-country abortion services does put pressure on the providers to find the most reliable, safe methods with the least chance of leaving women with ongoing pregnancies or serious side effects.

One approach that has been advocated in this circumstance is the use of the 'equivalence trial' which postulates that within a stated difference in efficacy (2, 3, 4, ..., 10%, for example), two therapies can be considered essentially the same. The problem with this approach is, of course, that the two therapies may actually be that much different, and the postulated clinical indifference may not be the case for many providers. Indeed, the reports of equivalence studies occasionally create misunderstanding when the two therapies are reported to be 'not significantly different' without the caveat that they are not significantly different 'only if you don't care if they are, in fact, as much as 2, 3, 4, ..., 10% different' (see, for example, Creinin *et al.*).[52]

The issue of sample size (and composition) has been particularly relevant in studies seeking to define the gestation age limit for various medical abortion regimens. This issue is relevant since most early medical abortion regimens become less effective with advancing gestational age. However, the regimens do not stop working at any one gestational age, so the decline in efficacy is gradual. Since most studies enrol all-comers and therefore have a range of gestational ages below the cut-off (usually 49, 56 or 63 days' LMP), the success of the therapy in the resulting sample represents a weighted average of success rates at the various gestational ages. It is therefore almost impossible to compare regimens and results across studies, since only large sample sizes with good randomisation will be comparable for this important variable. Subgroup analysis is usually not very helpful to get good point estimates of week-by-week efficacy rates – and certainly not for day-by-day efficacy rates – since the samples for each week of gestation are inevitably too small to test significant differences. Indeed, in places where medical abortion is used frequently (and which therefore have substantial populations to study), the proportion of patients in the later gestational ages is usually small, since most women come early for abortion services. This natural tendency makes examination of efficacy in the 56–63 days' LMP range particularly difficult. It is not ethical to ask women to wait to have a later abortion so that these latter days can be studied more efficiently.

RANDOMISATION

In order to assure that the groups being compared in a clinical study are equivalent in the most possible ways, the standard procedure is random allocation to treatment. This is possible when one is studying variants of two similar treatments or aspects of a treatment that are not determinative of a woman's choice to use it. For example, one can easily randomise patients to different pain control regimens, use of different surgical techniques, or two different doses of medical abortion drugs. It is much more difficult (and perhaps not useful) to attempt to randomise treatments as different from each other as medical and surgical abortion.

We need to keep in mind the purpose of randomisation and the underlying research question that is implied: all things being equal, which of these treatments is more effective (or safer, or more acceptable, etc.)? This kind of research question is most appropriate when the answer will be used to select a therapy or service that will be the standard for all patients of a certain type. But we know in the world of abortion medicine that the different approaches to early abortion will be part of a range of choices available to women. So we are not trying to determine which is the 'best' treatment to offer to the whole population: since women will self-select these treatments ultimately, a randomised sample of the entire population of abortion-seekers might not be as representative of the women who will be exposed to one or the other technique as a population of those types of women who will self-select to each one.

Indeed, there are also practical reasons why randomisation fails to produce the hoped-for sample allocation. When a new method is being offered in comparison with a standard method, women may sign up for the research simply because they hope to receive the new method. If they are allocated to the old method, they may choose to drop out of the study, leaving a biased sample. Indeed, this kind of behaviour has been noted in previous attempts to randomise such comparisons.[53] Other studies attempting randomisation between two very different technologies have found that very few women do not already have an preference for one or the other.[54] Such pre-existing preferences may make it very difficult to recruit for a randomised study of two very different abortion methods.[18,55] Indeed, even the protocol rules will frequently mean than a randomised sample is no longer representative of the population seeking treatment: since women have to be eligible for all the treatments in order to be included in the study, there is a priori exclusion of all women who have any contraindications to any treatment – when, of course, in the real world such women will be assigned to the treatment to which they have no contraindications. In abortion studies, this problem is less severe than in many other research fields, since methods of early abortion are so safe that they have very few contraindications and therefore very few women are excluded from studies of either surgical or medical methods.

Randomisation also obscures the initial acceptability of a new technology, and sometimes we would very much like to know who would want to use it, what their reasons are for their choices, how they viewed the actuality of the therapy after receiving it, and whether they were surprised, disappointed or especially pleased by any aspects of the method. All of these questions are either impossible to answer or distorted by a randomised design. On the other hand, randomisation is an excellent tool in circumstances where its use aids rather than obfuscates the answers to the research question being asked. In the end, like all tools, it is best for the job for which it is intended.

BLINDING

Blinding is concealment of the therapy or intervention a person is receiving as a way to prevent changes in behaviour or attitudes that could influence the outcome. This blinding may be of the patient alone (single blinding), of the patient and the provider/researcher (double blinding) or of the patient, provider and entire research team (triple blinding). Indeed, such a strategy to diminish the chances of bias is considered the gold standard but is not always attainable. For example, it is virtually impossible to blind studies comparing surgical and medical methods of early abortion. Providers clearly know whether they are performing a surgical act or not, and, while patients can be blinded to drug ingestion by use of placebos, they cannot be blinded to surgery without sham procedures, which is neither practical nor ethical. Indeed, even with blinding of the initial intervention, patients would likely know which treatment they had received because of the very different course and timing of events between medical and surgical intervention.

While some types of abortion provision can be blinded, others cannot, even in studies dealing only with either medical or surgical abortion. For example, while the use of two different drugs for analgesia during surgery can be blinded, the use of 'verbal anaesthesia' or various forms of reassurance cannot be concealed. Neither can one conceal various aspects of counselling or information provision. For medical abortion, one can blind the dose of some medications (such as in studies of 200 vs. 600 mg mifepristone), but it is harder to blind certain other features of the regimen. For example, researchers are currently interested in whether sublingual or buccal administration of misoprostol is more effective following mifepristone, but conducting a blinded study in which a woman had to hold four pills under her tongue at the same time as four other pills in her cheek would be extremely difficult.

Finally, blinding may not be appropriate if acceptability of the different alternatives tested is important or if acceptability may influence the outcome variables being studied. In a blinded study, each person 'experiences' both therapeutic alternatives that are offered, so it is not possible to find out which way of doing things is more acceptable to women. Indeed, for such personal and elective services as abortion, the reactions of women may be key information in understanding the best way to design services. In addition, if there is any possibility that other outcomes may be influenced, even unconsciously, by acceptability, this information will not come out of a study that is blinded. For example, if the route of administration of misoprostol is related to a woman's willingness to wait out the procedure and not ask to have it interrupted, then a blinded design would not allow the researcher to see this effect, since all women would effectively be experiencing the emotional response to both routes of administration.

SUMMARY

With early abortion, one of the most common clinical procedures, ensuring access to safe, effective and acceptable methods, is a priority for researchers, advocates and policy makers interested in women's health. Clinical studies can play an important role in the refinement of existing methods and the development of new methods. Careful measurement and interpretation of outcomes in clinical studies of early abortion methods, however, merit further attention. Features intrinsic to abortion services and methods of early abortion themselves also make conforming to the basic principles of clinical studies difficult.

ACKNOWLEDGEMENTS

The authors are very grateful to Carolina W. Galvão for her research and formatting assistance.

REFERENCES

1. Hagele M. Sexual and reproductive health and rights in the European Union. *Entre Nous* (2005) **59**: 26–8.
2. Education for Choice Website. Statistics. Available at: www.efc.org.uk (2005).
3. Henshaw SK, Singh S, Haas T. The incidence of abortion worldwide. *Int Fam Plann Perspect* (1999) **25**(Suppl): S30–8.
4. Alan Guttmacher Institute Website. Induced abortion in the United States. Available at: www.agi-usa.org/sections/abortion.php (2005).
5. MacIsaac L, Darney P. Early surgical abortion: an alternative to and backup for medical abortion. *Am J Obstet Gynecol* (2000) **183**(Suppl): S76–83.
6. Danco Laboratories. Presentation at the meeting *Mifepristone Medical Abortion: Making Progress in the Face of Adversity*, sponsored by Gynuity Health Projects, New York, June 2 (2005).
7. Bracken H, Winikoff B. The state of medical abortion in Europe today. *Entre Nous* (2005) **59**: 7–9.
8. World Health Organization. *Safe Abortion: Technical and Policy Guidance for Health Systems*. Geneva: World Health Organization (2003).
9. Cates W, Ellertson C. Abortion. In: Hatchner R *et al*., eds, *Contraceptive Technology*. New York: Ardent Media (1998) 679–700.
10. McInerney T, Baird TL, Hyman AG, Huber AB. *A Guide to Providing Abortion Care*. Chapel Hill: Ipas (2001).
11. Flett GM, Templeton A Surgical abortion. *Best Pract Res Clin Obstet Gynaecol* (2002) **16**: 247–61.
12. Grimes DA, Schultz KF, Cates W, Tyler CW. Local versus general anesthesia: which is safer for performing suction curettage abortions? *Am J Obstet Gynecol* (1979) **139**: 1030–5.
13. Fuller L. Options for early abortion in the United States. *Resour Women's Health* (2001) **3**: 1–4.
14. Keder LM. Best practices in surgical abortion. *Am J Obstet Gynecol* (2003) **189**: 418–22.

15. Grimes DA, Creinin MD. Induced abortion: an overview for internists. *Ann Intern Med* (2004) **140**: 620–6.
16. Beckman LJ, Harvey SM, Satre SJ. The delivery of medical abortion services: the views of experienced providers. *Women's Health Issues* (2002) **12**: 103–12.
17. Creinin MD. Medical abortion regimens: historical context and overview. *Am J Obstet Gynecol* (2000) **183**(Suppl): S3–9.
18. World Health Organization Scientific Group *Medical Methods for Termination of Pregnancy*. Geneva: World Health Organization (1997).
19. Spitz I, Bardin C, Benton L, Robbins A. Early pregnancy termination with mifepristone and misoprostol in the United States. *New Eng J Med* (1998) **338**: 1241–7.
20. Newhall E, Winikoff B. Abortion with mifepristone and misoprostol: regimens, efficacy, acceptability and future directions. *Am J Obstet Gynecol* (2000) **183**(2): S44–53.
21. Ashok P, Hamoda H, Flett G, Kidd A, Fitzmaurice A, Templeton A. Patient preference in a randomized study comparing medical and surgical abortion at 10–13 weeks gestation. *Contraception* (2005) **71**: 143–8.
22. Pymar HC, Creinin MD. Alternatives to mifepristone regimens for medical abortion. *Am J Obstet Gynecol* (2000) **183**(Suppl): S54–64.
23. Borgatta L, Mullally B, Vragovic O, Gittinger E, Chen A. Misoprostol as the primary agent for medical abortion in a low-income urban setting. *Contraception* (2004) **70**: 121–6.
24. Winikoff B, Sivin I, Coyaji K, Cabezas E, Bilian X, Sujuan G, Ming-kun D, Krishna U, Eschen A, Ellertson C. Safety, efficacy, and acceptability of medical abortion in China, Cuba, and India: a comparative trial of mifepristone-misoprostol versus surgical abortion. *Am J Obstet Gynecol* (1997) **176**(2): 431–7.
25. Winikoff B, Sivin I, Coyaji K, Cabezas E, Xiao B, Gu S, Du M, Krishna U, Eschen A, Ellertson C. The acceptability of medical abortion in China, Cuba and India. *Int Fam Plann Perspect* (1997) **23**(2): 73–89.
26. Nuutila M, Toivonen J, Ylikorkala O, Halmesmaki E. A comparison between two doses of intravaginal misoprostol and gemeprost for induction of second trimester abortion. *Obstet Gynecol* (1997) **90**(6): 896–900.
27. Ho P, Ngai S, Liu K, Wong G, Lee S. Vaginal misoprostol compared with oral misoprostol in termination of second-trimester pregnancy. *Obstet Gynecol* (1997) **90**(5): 735–8.
28. Weeks A, Stewart P. The use of mifepristone in combination with misoprostol for second trimester termination of pregnancy. *Br J Fam Plann* (1996) **21**(2): 43–4.
29. Elul B, Ellertson C, Winikoff B. Coyaji K. Side effects of mifepristone-misoprostol abortion versus surgical abortion: data from a trial in China, Cuba, and India. *Contraception* (1999) **59**(2): 107–14.
30. Peyron R, Aubeny E, Targosz V, Silvestre L, Renault M, Elkik F, Leclerc P, Ulmann A, Baulieu E. Early termination of pregnancy with mifepristone (RU 486) and the orally active prostaglandin misoprostol. *New Engl J Med* (1993) **328**: 1509–13.
31. Sang GW, Weng LJ, Shao QX, Du MK, Wu XZ, Lu YL, Cheng LN. Termination of early pregnancy by two regimens of mifepristone with misoprostol and mifepristone with PG05 – a multicentre randomized clinical trial in China. *Contraception* (1994) **50**: 501–10.
32. World Health Organization Task Force on Post-Ovulatory Methods of Fertility Regulation. Termination of pregnancy with reduced doses of mifepristone. *Brit Med J* (1993) **307**: 532–7.
33. Winikoff B, Ellertson C, Clark S. Analysis of failure in medical abortion. *Contraception* (1996) **54**(6): 323–7.
34. Trussell J, Ellertson C. Estimating the efficacy of medical abortion. *Contraception* (1999) **60**: 119–35.
35. Hedley A, Ellertson C, Trussell J, Turner A, Aubeny E, Coyaji K, Ngoc N, Winikoff B. Accounting for time: insights from a life-table analysis of the efficacy of medical abortion. *Am J Obstet Gynecol* (2004) **191**: 1928–33.
36. Harper C, Winikoff B, Ellertson C, Coyaji K. Blood loss with mifepristone-misoprostol abortion: measures from a trial in China, Cuba and India. *Int J Gynecol Obstet* (1998) **63**: 39–49.
37. Blum J, Hajri S, Chelli H, Ben Mansour F, Gueddana N, Winikoff B. The medical abortion experience of married and unmarried women in Tunis, Tunisia. *Contraception* (2004) **69**: 63–9.
38. Elul B, Hajri S, Ngoc N, Ellertson C, Slama C, Pearlman E, Winikoff B. Can women in less-developed countries use a simplified medical abortion regimen? *Lancet* (2001) **357**: 1402–5.
39. Banerjee K, Kriplani A, Kumar V, Rawat KS, Kabra M. Detecting fetomaternal hemorrhage after first-trimester abortion with the Kleihauer-Betke test and rise in maternal serum alpha-fetoprotein. *J Reprod Med* (2004) **49**: 205–9.
40. Clark S, Blum J, Blanchard K, Galvao L, Fletcher H, Winikoff B. Misoprostol use in obstetrics and gynecology in Brazil, Jamaica, and the United States. *Int J Gynecol Obstet* (2002) **76**: 65–74.
41. Prasad R, Choolani M, Roy A, Ratnam S. Blood loss in termination of early pregnancy with

mifepristone and gemeprost. *Aust NZ J Obstet Gynecol* (1995) **35**(3): 329–31.

42. Rodger M, Baird D. Blood loss following induction of early abortion using mifepristone (RU 486) and a prostaglandin analogue (gemeprost). *Contraception* (1989) **40**: 439–47.
43. Chan Y, Ho P. Blood loss in termination of early pregnancy by vacuum aspiration and by combination mifepristone and gemeprost. *Contraception* (1993) **47**: 85–95.
44. Winikoff B. Acceptability of medical abortion in early pregnancy. *Fam Plann Perspect* (1995) **27**: 142–8.
45. Winikoff B, Coyaji K, Cabezas E, Coyaji B, Krishna U, Concepcion O, Eschen A, Sivin I, Brady M. Studying the acceptability and feasibility of medical Abortion. *Law Med Health Care* (1992) **20**(3): 195–8.
46. David HP. Acceptability of mifepristone for early pregnancy interruption. *Law Med Health Care* (1992) **20**: 188–94.
47. Winikoff B, Ellertson C, Elul B, Sivin I. Acceptability and feasibility of early pregnancy termination by mifepristone-misoprostol. *Arch Fam Med* (1998) **7**: 360–66.
48. Ngoc N, Winikoff B, Clark S, Ellertson C, Am K, Hieu D, Elul B. Safety, efficacy and acceptability of mifepristone-misoprostol medical abortion in Vietnam. *Int Fam Plann Perspect* (1999) **25**(1): 10–14.
49. Hamoda H, Flett GM, Ashok PW, Templeton A. Surgical abortion using manual vacuum aspiration under local anaesthesia: a pilot study of feasibility and women's acceptability. *J Fam Plan Reprod Health Care* (2005) **31**: 185–8.
50. Westfall JM, Kallail KJ. Repeat abortion and use of primary care health services. *Fam Plann Perspec* (1995) **27**: 162–5.
51. Fisher WA, Singh SS, Shuper PA, Carey M, Otchet F, MacLean-Brine D, Dal Bello D, Gunter J. Characteristics of women undergoing repeat induced abortion. (2005) *Can Med Assoc J* (2005) **172**: 637–41.
52. Creinin M, Potter C, Holovanisin M, Janczukiewicz L, Pymnar H, Schartz J, Meyn L. Mifepristone and misoprostol and methotrexate/misoprostol in clinical practice for abortion. *Am J Obstet Gynecol* (2003) **188**(3): 664–9.
53. Rosen AS, Von Knorring K, Bygdeman M. Randomized comparison of prostaglandin treatment in hospital or at home with vacuum aspiration for termination of early pregnancy. *Contraception* (1984) **19**: 423–35.
54. Ashok PW, Kidd A, Flett GM, Fitzmaurice A, Graham W, Templeton A. A randomized comparison of medical abortion and surgical vacuum aspiration at 10–13 weeks gestation. *Hum Reprod* (2002) **17**: 92–8.
55. Creinin M, Fox M, Teal S, Chen A, Schaff E, Meyn L. A randomized comparison of misoprostol 6 to 8 hours versus 24 hours after mifepristone for abortion. *Obstet Gynecol* (2004) **103**: 851–9.

26

Maternal and Perinatal Health

JOSÉ VILLAR,[1] ALLAN DONNER,[2] EDGARDO ABALOS[3] AND LIANA CAMPÓDONICO[3]

[1]Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford OX3 9DU, UK
[2]Dept. of Epidemiology and Biostatistics, Faculty of Medicine and Dentistry,
University of Western Ontario, London, Canada
[3]Centro Rosarino de Estudios Perinatales (CREP), Rosario, Argentina

INTRODUCTION

Randomised controlled Trials (RCTs) are now well accepted as the least biased study design for evaluating clinical and surgical treatments, screening methods and preventive nutritional or educational interventions. Their use has increased in recent years in medicine in general and in the area of perinatology in particular, with results often summarised in the form of systematic reviews and meta-analyses. Randomised trials conducted during pregnancy and the perinatal period to evaluate either preventive interventions or treatments have a series of methodological, ethical and logistical issues that must to be considered separately from standard RCTs in medicine.[1]

Our aim here is to discuss these methodological issues based on the experience gained by the World Health Organization (WHO) in conducting large multicentre RCTs and their corresponding systematic reviews (Table 26.1 and Figure 26.1). Trials specifically related to the prevention and treatment of pre-eclampsia/eclampsia are conducted in the context of the WHO 'Global Program to Conquer Preeclampsia',[2] all others are part of the 2004–2009 'WHO Programme of Work for Maternal and Perinatal Research' presently being implemented.[3]

SYSTEMATIC REVIEWS BEFORE NEW TRIALS

Before preparing protocols and implementing RCTs, we have adopted the policy of completing systematic reviews of the available evidence in order to justify the need for a new trial. Systematic reviews are reviews of the evidence on a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant primary research.[4] These concepts have been extended by our programme to include basis science evidence.[5] predictive factors and early markers[6] as well as the more standard systematic review of previously conducted randomised trials testing

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

Table 26.1. RCTs conducted with leading participants of the WHO Maternal and Perinatal Research Network up to 2005

Trial topic	Centres	Participants	Status
Antenatal care	5	24 678	Published (2001)
Prevention postpartum haemorrhage	9	18 530	Published (2001)
Treatment pre-eclampsia (Magpie trial)*	28	10 141	Published (2002)
Reduction of unnecessary caesarean section	5	149 206	Published (2004)
Evaluation of Reproductive Health Library	2	76 053	Submitted (2005)
Primary prevention of pre-eclampsia (calcium trial)	7	8 400	Submitted (2005)
Long-term follow-up Magpie trial*	19	3 375	Submitted (2005)
Long-term follow-up of calcium trials	2	800	Submitted (2005)
Screening and treatment of asymptomatic bacteriuria	4	18 000	Ongoing
Primary prevention of pre-eclampsia (antioxidants trial)	4	4 150	Ongoing
Treatment of postpartum haemorrhage	5	1 400	Ongoing
Secondary prevention of pre-eclampsia (treatment moderate hypertension)	6	2 000	In preparation

* Coordinated by another institution.

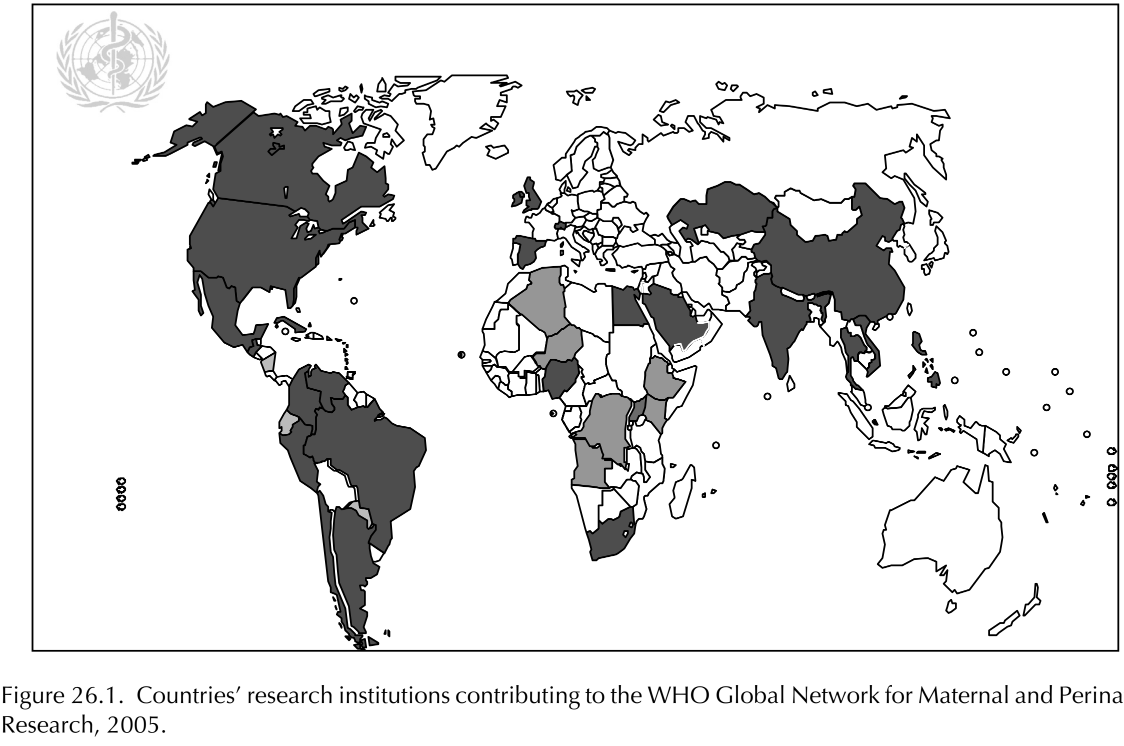


Figure 26.1. Countries' research institutions contributing to the WHO Global Network for Maternal and Perinatal Research, 2005.

similar research questions.[7] All available information is then critically evaluated following a predesigned protocol with the aim of providing justification for the new research project.

Statistical methods for pooling results (meta-analysis) may or may not be used in the review, but the possibility of heterogeneity in the results is always examined and evaluated. Identification of subgroups, where there is a differential effect, is important for the preparation of the protocol, particularly in selecting the study population to be included in the new trial.

Such a research policy provides us with the strongest background information for all planned maternal and perinatal WHO trials. When these trials are completed, the new data are incorporated into the original systematic review and published in conjunction with the results of the new trial, allowing readers to obtain a complete and updated evaluation of all available evidence in conjunction with the new trial.[8] We strongly encourage researchers who may be considering the design of a new randomised trial to always regard the preparation of a systematic review as an integral part of the protocol development.

TAXONOMY OF CLINICAL TRIALS IN MATERNAL AND PERINATAL HEALTH

Different types of RCT's have been used in perinatology according to the nature of the research question and the specified objectives. We classify them either as pragmatic evaluative trials or as explanatory trials. Pragmatic trials are usually conducted in the context of existing clinical services with the routine clinical care minimally affected. In these trials the primary outcomes usually tend to be indications of severe morbidity or mortality. Although outcomes of diseases of unknown aetiology, such as pre-eclampsia or neonatal mortality, are routinely collected, they are still relatively rare (prevalence between 1.0% and 5%) and large numbers of pregnancies have to be surveyed. Trials evaluating modifications of health services or preventive strategies are included in this category.

Conversely, explanatory trials are usually smaller, in which very detailed information is collected in addition to routine clinical and laboratory data. Pharmacokinetic and intermediate markers that usually precede the clinical manifestation of the disease are evaluated. Furthermore, a special form of these two types of trials is the continuation of the original trial, maintaining the original randomisation scheme, but conducting follow-up until adolescence. Table 26.2 presents examples of such classification using trials recently conducted by the WHO Maternal and Perinatal Research Programme.

As can be seen, these trials can range from single medical interventions of pharmacological agents given only once in time (e.g. the postpartum period) to interventions administered early in pregnancy with the corresponding follow-up period (usually seven months maximum), including data collected on intermediary variables (explanatory trials), to complex manipulations of health services introducing new protocols for aiding the clinical decision-making process, or different types of continuing medical education interventions (pragmatic trials of health service interventions).

Regardless of the type or length of intervention, trials during pregnancy have in common their short period of implementation, as compared with trials for treatment of chronic conditions. Moreover, the subjects are mostly young, motivated (particularly if the intervention involves preventive strategies for the foetus), healthy women in which the primary outcomes involve two individuals: the mother and the foetus–newborn, the latter exposed to treatments only via the mother.

In some instances, the effect could be beneficial to one individual, e.g. reduce maternal blood pressure, but detrimental to the other (e.g. reduce foetal growth). Trade-off in these effects often provides the framework for the proper interpretation of these trials. Long-term postnatal follow-up evaluating outcomes on both mother and child are also now often recommended. Here again, a trade-off may exist between beneficial (e.g. a reduction in neonatal mortality) and harmful outcomes (e.g. an increase in neurological sequelae in children).

WHY DO WE NEED LARGE SIMPLE TRIALS IN PERINATOLOGY?

It is well accepted that in modern obstetrics and perinatology, even in poor populations,

Table 26.2. Taxonomy of clinical trials in maternal and perinatal health using the Experience of the WHO Maternal and Perinatal Research Network

Pragmatic trials		
Health service interventions		Introduction of the new WHO antenatal care model (cluster randomisation) Introduction of mandatory second opinion before caesarean section (cluster randomisation) Introduction of the reproductive health library to change practices (cluster randomisation)
Treatment regimen evaluations		
	Preventive	Prevention of postpartum haemorrhage with misoprostol Prevention of pre-eclampsia with calcium supplementation Prevention of pre-eclampsia with vitamin C and E supplementation
	Therapeutic	Treatment of pre-eclampsia with injectable magnesium sulphate Treatment of postpartum haemorrhage with oral misoprostol plus injectable uterotonics Treatment of mild hypertension in pregnancy with antihypertensive drugs
Explanatory trials		
Type of outcome	Clinical	Misoprostol dose effect evaluation for prevention of postpartum haemorrhage and side effects
	Laboratory	Timing of umbilical cord cutting and haemoglobin levels of the newborn and during infancy
	Microbiology	Screening, diagnosis and treatment of asymptomatic bacteriuria
	Ultrasound	Effect of calcium supplementation on foetal growth evaluated by ultrasonography
Long-term follow-up for other than the primary outcomes of the original trial		
Monitoring side effects		Long-term follow-up of the effect of the treatment with magnesium sulphate in pregnancy on child growth and development
Monitoring long-term beneficial effects		Long-term follow-up of the effect of calcium supplementation in pregnancy on blood pressure of the offspring

the best that can be realistically expected for most new treatments or modified regimens is a moderate effect on reducing severe morbidity and mortality. For example, a reduction of 20% in the rate of an adverse outcome, such as preterm delivery or neonatal mortality, is often considered to represent a major effect. We therefore have to design trials that can reliably discriminate between moderate treatment effects, yet have important clinical or programmatic value. Therefore, in our field the planning process inevitably leads to large trials: that is, trials with sufficient numbers of primary events to detect relevant differences in severe morbidity and mortality.

Some treatments may have large effects on secondary outcomes or on intermediate mechanisms or markers of the pathophysiology of a given condition. For example, antihypertensive drugs readily lower blood pressure during pregnancy. However, effects on severe maternal morbidity and perinatal mortality have been difficult to document, and similarly for rather small negative side effects that may characterise foetal growth.[9] This builds the case for proper evaluation of regimens for treating mild to moderate hypertension during pregnancy in order to prevent pre-eclampsia. However, after years of routine use of antihypertensive medication in pregnancy, the corresponding definitive trial is yet to come, underscoring the complexity of implementing such trials during pregnancy.

Large trials have to be simple in order to recruit a large number of pregnant women over a reasonable period of time, to allow results of trials, when available, to still be relevant, and to perform trials at an affordable cost. We have found that the complexity of a trial may be a barrier to recruitment, can interfere with clinical practice, may encourage participants to leave the study and, finally, may restrict the generalisability of the results (external validity). This principle also extends to setting women's eligibility criteria for the trial. If the inclusion/exclusion criteria are complex or based on criteria not widely used in routine antenatal care, the trial will result in recruiting a relatively narrow group of pregnancies, again reducing the external validity of the results.

We always aim to adopt and integrate trials into existing clinical practice. For example, consider the screening strategies for selecting women for inclusion in the WHO misoprostol trial during the third stage of labour, a trial designed to prevent postpartum haemorrhage. This study, which recruited over 18 000 postpartum women within three years, included only four questions and involved no laboratory testing or complex clinical examinations.[10] The WHO calcium supplementation trial required only two inclusion criteria and recruited 8300 pregnant women starting early in pregnancy (<20 weeks) within a two-year period following antenatal care of participating centres. This trial is the largest trial available on this topic and had the power to explore mortality outcomes associated with pre-eclampsia seldom done before.[11]

These large trials have the additional advantage that if randomisation is correctly conducted and treatment allocation adequately concealed, the baseline characteristics will be very well balanced between groups, therefore reducing the burden of data collection as only data on important risk factors need to be collected. Additionally, data analysis is inevitably simpler with less statistical modelling required. We operate on the principle that it is usually preferable to collect 10 times less data on 10 times more patients.

Then, are very large trials having unrestrictive protocols and simple data collection systems the solution to all our clinical and public health research problems in perinatology? Although this is undoubtedly a useful research strategy, we should also be aware of its limitations. For example, large, simple trials may produce negative results if compared with routine care, because of the likely use of non-trial therapy (co-intervention). Moreover, 'too simple' treatment protocols may end up with a too weak intervention; and simplified data collection forms could increase misclassification of outcomes less objective than mortality. These issues need to be carefully considered (Table 26.3).

There are two special types of randomised trials that are increasingly used to evaluate interventions given to pregnant women which require, in general, a larger sample size than other types of trials: cluster randomised trials and trials aiming to demonstrate equivalence between treatments. We discuss these in more detail below.

CLUSTER RANDOMISED TRIALS

We have conducted a pragmatic intervention trial[12,13] that can be used to illustrate two relatively new approaches that may be taken to the delivery of interventions and design of trials in perinatal health. These include the randomisation of clusters rather than individuals to different intervention groups and the use of the randomised consent design proposed by Zelen in 1990.[14]

The purpose of the WHO antenatal care trial[12,13] was to compare the effect of two programmes of routine antenatal care on the health of mothers and newborn babies. These programmes include medical and non-medical interventions. One of the programmes consisted of the 'best standard treatment' as offered in well-developed antenatal care clinics, with the other consisting

Table 26.3. Some advantages of simple pragmatic trials

Feasible to recruit really large numbers of people	Simple eligibility criteria Simple trial entry procedures
Conducted within the existing health services	Intervention feasible without additional staff or technology Data collection based on what is likely to be available
Considerably less expensive than more complex studies	
Minimal additional work for already busy clinicians	
Encourages participants to stay in the study	
More complete and better quality data	
Simpler data management	
Results relevant to clinical practice in a wide range of settings	

only of tests, clinical activities and follow-up actions that have been scientifically demonstrated to be effective in improving outcomes for mothers and newborns, the ‘reduced’ or ‘new model’. All patients attending antenatal care for the first time after the start of the study period at all of the selected clinics were included for follow-up. The primary foetal outcome for the study was the rate of low birth weight (below 2500 g), with the primary maternal outcome defined as the rate of a predefined morbidity index including three severe morbidities (pre-eclampsia, severe anaemia and severe urinary tract infection) all of them related to antenatal care.[12]

It is now well recognised that the randomisation of intact clusters of individuals rather than individuals themselves to intervention groups can lead to a substantial loss of statistical power, or equivalently to a reduction in total effective sample size. The loss in power can be quantified by the trial ‘design effect’ or ‘variance inflation factor’ (IF), which may be defined as the ratio of the variance of the estimated intervention effect obtained under cluster randomisation to that obtained under individual randomisation. Equivalently, it may be regarded as the ratio of the required sample sizes in the two designs needed to preserve the same degree of statistical power.

The magnitude of IF increases with both the degree of within-cluster similarity, as quantified by the intracluster correlation coefficient (ICC), as well as with the average cluster size. For a design randomising clusters of fixed size m to each of two or more intervention groups, with no stratification on baseline risk factors, the design effect is given by $\mathrm{IF} = 1 + (m - 1)$ ICC. The ICC may be most simply interpreted as the standard Pearson product-moment correlation between any two observations in the same cluster. If $\mathrm{ICC} = 0$, the responses within clusters are totally independent, whereas if $\mathrm{ICC} = 1$ one member of the cluster provides the same amount of information as do all members. Negative values of the ICC are usually considered implausible in cluster randomisation trials.

For clusters typical of those in the antenatal care trial, the value of the ICC is generally small and positive. However, when accompanied by relatively large cluster sizes the resulting design effect may be considerable. For example, if the value of the ICC is 0.01, a cluster size of 100 will yield a design effect very close to 2.0, implying that the trial size must be doubled in order to preserve the same degree of statistical power as obtained under individual randomisation.

Standard methods for sample size estimation and analysis are generally not applicable under cluster randomisation, since a positive value of the ICC can rarely be ruled out. In particular, a sample size calculation that ignores the clustering in the design may lead to an underpowered trial (increased Type II error) while an analysis that ignores the clustering may lead to spurious statistical significance (increased Type I error).

Detailed discussion of these issues is given by Donner and Klar.[15]

Despite the expected loss in statistical efficiency, it was very natural in the antenatal trial to randomise clinics rather than individuals to the intervention and control groups. Several factors made this decision inevitable, even though many of the clinics were quite large, some of them serving as many as several hundred patients. The first factor was based solely on logistical grounds. Having set up a programme for implementing the new intervention, it would have been administratively inconvenient and potentially awkward to administer it to some patients in a clinic but not to others. The recruitment of physicians to the trial would also likely be enhanced if they were not required to treat some patients differently than others. However, the main reason from an evaluation perspective for choosing to randomise clinics was to avoid the risk of contamination that could result if the same staff were to treat both experimental and control group patients.

Having made this decision, it was important to recognise that a positive value of the ICC could be expected with respect to both the foetal and maternal outcomes, i.e. it was reasonable to expect some between-clinic variation with respect to these outcomes. For example, suppose the characteristics of patients attending a given clinic were related to age or other demographic characteristics of the clinic staff. To the extent these characteristics are also related to the trial endpoints, a clustering effect will be introduced within clinics. In addition, the outcomes on two or more patients treated by the same staff member could share the influence of that staff member's style of practice.

It was initially assumed that the average number of patients entered per clinic in the trial would be about 450. However, an advance estimate of the ICC was also needed in order to plan the final trial size. Obtaining such estimates is often a difficult task if pilot data or published values from trials randomising the same unit and measuring similar outcomes are not readily available. Fortunately preliminary data from one trial site indicated that the ICC for low birth weight might be taken as 0.001, yielding an expected design effect of about 1.45. Thus the 'effective sample size' per clinic in this trial was expected to be about $450/1.45 = 310$.

The reduction in effective sample size for a cluster randomisation trial also implies that the chance of imbalance on baseline risk factors is greater than in an individually randomised trial allocating the same total number of individuals. Thus some form of stratified randomisation is often adopted in such trials, unless the number of randomised units (clusters) is very large and the target population fairly homogeneous.

Some degree of stratification was regarded as essential in the WHO antenatal trial, since it enrolled 53 clinics across four very diverse sites, both culturally and demographically. Apart from stratification by study site (country), further stratification by clinic size was also considered desirable. This was partly because it would assure reasonable balance between the total number of individuals randomised to each trial arm. However, it was also judged that clinic size could at least partially serve as a surrogate for selected baseline characteristics that were potentially related to outcome, such as socioeconomic status and place of residence.

Ultimately each of the four participating study sites contributed from 12 to 17 clinics, yielding 12 000 study subjects in each arm of the study. Clinic size, categorised as small, medium or large, was adopted as a secondary stratification variable in the allocation scheme.

Note that since the sample size calculation for this trial ignored the stratified allocation, this calculation can be regarded as conservative to the extent that the stratification led to some increase in overall precision. The final sample size calculation also took into account a possible loss to follow-up rate of 10%, adding a further layer of protection.

EQUIVALENCE TRIALS

The motivation behind most RCTs is to demonstrate a 'positive' result, whereby one intervention is found to be superior to another. However, in some trials, the efficacy of a new intervention is evaluated in a context where it is already known to have substantial advantages in terms of ease of administration, cost or safety as compared with the standard intervention. In this case the objective of the trial may be to show that the new intervention is no less efficacious than the standard, rather than to demonstrate its clear-cut superiority. These trials are referred to as 'equivalence trials'.

As pointed out by Jones *et al.*,[16] conventional significance tests have little relevance in equivalence trials. This is because failure to detect an intervention effect may largely be a result of poor statistical power, and therefore does not imply equivalence. Moreover, in very large trials an effect which is detected as statistically significant may not have clinical relevance, i.e. the observed difference may correspond instead to a conclusion of 'practical equivalence'.

Confidence intervals around the point estimate (either relative risk or risk difference), on the other hand, have a very natural role to play in interpreting the results of an equivalence trial. If a range of practical equivalence can be prespecified, an investigator can examine if the confidence interval for the intervention effect lies entirely within this range. If so, equivalence is demonstrated in the sense that effects having clinical relevance have been ruled out to a reasonable degree of certainty. Otherwise equivalence cannot be claimed.

A principal objective of the antenatal trial was to evaluate an alternative to the standard model of care that was designed to be both lower in cost and less inconvenient for pregnant women and their families. Under these conditions, the new model would therefore be worth adopting provided it was 'practically equivalent' to the standard model with respect to efficacy.

The proportion of newborns experiencing low birth weight, the primary foetal outcome variable for the trial, was expected to be about 0.10. It was agreed by the investigators that an increase in this rate up to 0.12, representing an intervention odds ratio of approximately 1.2, would be consistent with the conclusion that the new programme of antenatal care is as equally effective as the standard programme. Thus a claim of practical equivalence would require that the upper 95% confidence limit for the intervention odds ratio not exceed 1.2. After adjusting for clustering effects and the impact of relevant baseline covariates, the upper limit of the 95% confidence interval for the intervention odds ratio associated with low birth weight was in fact 1.15,[12] thus allowing a claim of practical equivalence with respect to this outcome variable. We concluded that the provision of the new model seems not to affect perinatal outcomes and could be implemented.[12]

THE WHO ANTENATAL TRIAL AS AN EXAMPLE OF THE ZELEN DESIGN

Resistance to randomisation is often claimed as one reason that many clinical trials fail to accrue a sufficient number of patients. Zelen [14,17] proposed the 'randomised consent' design as a means of increasing accrual in those clinical trials for which patient recruitment is perceived to be handicapped by the need to administer informed consent prior to randomisation. This strategy is particularly relevant to open trials, where the treatment is known to researchers and participants.

In the Zelen design, eligible patients are randomised to either the control or intervention group before obtaining informed consent. Only those patients assigned to the intervention group are informed of the trial and asked for consent. However, those patients refusing consent are still followed up (without treatment) in accordance with the intent-to-treat principle that requires the primary statistical analysis to include all randomised patients. Although this is acknowledged to create some dilution of treatment effect, if a large number of patients in the intervention group do not provide consent, it is hoped that this disadvantage will be offset by the increased ability of the investigator to accrue a larger number of patients. Patients assigned to the control group are assumed to receive standard treatment consistent with their usual medical care and, if only routine data are used, may not be aware that they are participating in a trial.

The Zelen design has proven somewhat controversial for individual randomisation trials, in part for ethical concerns about half of the patients not knowing that their medical treatment has been determined by chance, and in part for methodological concerns raised by the possible dilution of the treatment effect. However, a version of this design arises very naturally, and sometimes inevitably, in the planning of many cluster randomisation trials. This is because such trials often require the investigators to consider obtaining consent at two distinct levels. Thus at the cluster level, consent must be obtained by a key decision maker (e.g. physician, principal, mayor) for permission to randomise the cluster (e.g. practice, school, community) for which they are legally responsible to one of two intervention groups. Assuming that such permission has been granted, randomisation of that cluster may then take place without the individual member's knowledge or consent. Informed consent, when obtained at all at the individual level, can therefore only be for permission of the subjects to be followed up and for their data to be used in the analysis.

Informed consent in the antenatal trial was requested at the cluster level from the director or medical chief in all antenatal clinics and at the individual level only from women attending clinics randomised to the intervention group. Women attending the control group clinics received 'the standard treatment' offered in those clinics. These data were therefore used no differently than they would be routinely by clinical departments or other authorities in the analysis of health outcomes data. In this sense, the antenatal trial, as well as many other cluster randomisation trials, has implicitly used a variant of the Zelen design initially proposed mostly for use in individually randomised therapeutic trials.

HETEROGENEITY WITHIN MULTICENTRE TRIAL RESULTS

We have observed very often in multicentre randomised trials that there is heterogeneity of results among patient subgroups, including trial centres. If such heterogeneity has been incorporated in the hypothesis before the trial was conducted with a substantive biological basis, the appropriate tests of interaction are conducted. Unfortunately, on many occasions, no such prior knowledge is available. If the trial results are overall 'negative', considerable tension may be produced among investigators, particularly when only some individual subgroups produce 'positive' results.

What do we do then with this intra-trial heterogeneity? We first explore the type of heterogeneity, e.g. qualitative versus quantitative. The latter is to be expected in multicentre trials considering the large variation in populations and staff enrolled in these studies. The former, however, means that the treatment under study has some type of interaction with, for example, the place where medical care is provided or the characteristics of the study population (i.e. it is effective in some settings and does more harm than good in others). Furthermore, the observed heterogeneity can be tested statistically against the null hypothesis of homogeneous results, across study sites, but such statistical tests have very low statistical power.

We recommend that all sources of heterogeneity of trial results should be explored and attempts made to understand the underlying cause. For example, we have recently been confronted with such heterogeneity during the analysis of a large multicentre randomised trial testing the hypothesis that oral 600 μg of misoprostol (a prostaglandin E1 analogue) with strong uterotonic effect can reduce severe blood loss and postpartum haemorrhage during the routine management of the third stage of labour. When used prophylactically, it was expected to be equivalent to the standard preventive regimen of 10 IU of injectable oxytocin, in terms of postpartum blood loss of 1000 ml or more, which was considered one of the two primary outcomes.[10]

In this trial, we observed statistically significant heterogeneity among centres for this primary outcome ($p = 0.02$). The relative risk of having blood loss $\geq$1000 ml for a woman in

Table 26.4. Rate and blood loss of 1000 ml or more in the postpartum period according to treatment group and study centre of the WHO randomised trial of misoprostol in the third stage of labour[10]

Study site	Misoprostol *n*/*N* (%)	Oxytocin *n*/*N* (%)	Direction of effect
Argentina	96/1358 (7%)	49/1361 (4%)	Misoprostol worst
China	18/1093 (2%)	10/1098 (0.9%)	Misoprostol worst
Egypt	3/1708 (0.2%)	6/1703 (0.4%)	Misoprostol better
Ireland	15/221 (7%)	9/225 (4%)	Misoprostol worst
Nigeria	36/785 (5%)	40/783 (5%)	Equal
South Africa	56/1405 (4%)	51/1409 (4%)	Equal
Switzerland	17/173 (10%)	16/177 (9%)	Equal/misoprostol worst
Thailand	57/900 (6%)	24/899 (3%)	Misoprostol worst
Vietnam	67/1570 (4%)	57/1572 (4%)	Equal

the intervention arm ranged from 0.50 (very protective) to 2.37 (very harmful) across centres (Table 26.4).

We explored the route of administration of oxytocin (the standard treatment) intravenously versus intramuscularly as a possible source of the heterogeneity, as the route of administration did in fact vary across centres. However, we found that this factor did not influence the observed pattern of results. Multivariate analysis including other variables, as well as sensitivity analysis excluding some centres, was conducted, but, as is often the case, we could not ultimately explain the heterogeneity. Therefore the recommendations from this trial focused on the overall effect, but centre-specific incidence rates were published without any statistical testing.[10]

We advise that researchers should usually resist presenting separate results from individual centres included in a multicentre randomised trial. In this regard, it should be considered that the probability of at least one centre out of nine showing an effect reversal is close to 80%.[18]

SEVERE MORBIDITY INDICES AS OUTCOMES IN RANDOMISED TRIALS

The outcomes of trials (often called endpoints or events) are defined as the components of patients' clinical and functional status after an intervention has been applied.[19] Results of a pragmatic trial usually provide information about the effectiveness of interventions in their usual clinical settings and, as such, the choice of outcomes measured in clinical trials is an important design consideration within the availability of routinely collected data.

The primary outcome is the event determined by the precise question that needs to be answered.[20] It is normally the outcome that indicates whether or not the trial provides evidence at an acceptable level of confidence that the treatment is efficacious. The trial sample size is estimated taking into account the incidence of the primary outcome and the margin expected to be reduced.

For pragmatic trials in perinatology, primary outcomes should be those that are likely to be meaningful to several stakeholders (e.g. clinicians, policy makers, patients) who may have to make decisions about future implementation of the intervention. Although the major endpoints of interest could be relatively rare, the interventions can affect two individuals (the mother and the baby), so data on efficacy and safety need to be collected for both.

Trials examining treatments that are expected to have an effect on mortality and major morbidity for either the mother or the baby could adopt a primary composite outcome measure that includes mortality along with other severe but non-fatal endpoints (often called 'near-miss' cases). Composite outcomes have practical sample size implications, since a composite endpoint is expected to have a higher event rate than any of the component endpoints. This strategy also makes possible the study of conditions that are considered markers of the severity of the primary event. For example, in a recent pre-eclampsia trial, we used the presence of at least one of the following: eclampsia, abruptio placentae, renal failure or HELLPS syndrome to construct a composite severity index.[11]

In the WHO antenatal care Trial,[12] the primary foetal/neonatal outcome was low birth weight (<2500 g) among singleton births. However, the primary maternal outcome was the maternal morbidity index defined as the presence of at least one of the following severe conditions for which antenatal care is believed to be effective: pre-eclampsia or eclampsia during pregnancy or within 24 h of delivery, severe postpartum anaemia, and severe urinary tract infection including pyelonephritis. Several other standard maternal and perinatal events, as well as the individual components of the maternal morbidity index, were considered as secondary outcomes. This index included the three conditions with the same weight (unweighted composite index). Although there have been suggestions that some type of weights could be incorporated in such an index, we consider any such weighting would be essentially arbitrary and simply add confusion to the task of interpreting the results.

In general, therefore, composite outcomes are usually made up of several severe events, any one of which clearly constitutes a negative result. Sometimes, these events share pathogenic mechanisms (such as pre-eclampsia and eclampsia), and sometimes they are more independent (such as pyelonephritis and postpartum anaemia). However, each of the events compiled in a composite outcome should clearly have clinical relevance, the rate of occurrence of each individual component should be similar, as well as the expected magnitude of the effect of the intervention.

Furthermore, there is always the possibility that the treatment could benefit one component of the index, but increase the risk of another. This is possible, for example, when maternal and newborn outcomes are included in a composite index. We believe that such a combination should be avoided.

A problem with the interpretation of composite outcomes (especially if they show a statistically significant difference) is the risk of assuming that the reported overall benefit can be extended to each one of their individual components. It has been shown that the effect of treatment on the mortality component of composite outcomes can differ from its effect on other outcomes. Moreover, it is known that some outcomes are inadequately individually reported by authors.[21] Composite outcomes have to be interpreted as a ‘package’, and their individual components should therefore always be reported as separate, secondary outcomes.

COORDINATING COMPLEX MULTICENTRE TRIALS DURING PREGNANCY

We will discuss here issues that need to be considered for the coordination and management of large multicentre trials.

BEFORE STARTING THE TRIAL

Before the implementation of a trial, all the local bureaucratic and legal issues in each participating country that are necessary for protocol approval, i.e. Ministry of Public Health, National Ethics Committees, Local or Hospital Ethics Committees, Drug Regulating Bodies, etc., have to be considered. These procedures may be quite lengthy, so they have to be made explicit in the protocol timelines. Also, some of these steps have important budgetary implications.

The set-up of a central coordinating unit is very important. It has to be easily identifiable and available for contact to all collaborators through a telephone line, a dedicated email address and a web page. Central coordination plays a major role in large multicentre trials supervising all the activities and supporting collaborators to solve everyday problems. Regional and local coordinating units could also be important in very large trials, since regional coordinating units can facilitate local capacity building and the development of human resources. With this philosophy, WHO has organised a perinatal research network in more than 35 countries in America, Africa and Asia (Figure 26.1).

These regional coordinating units have played a major role in running high-quality, pragmatic RCTs and epidemiological studies in developing countries. In the ongoing multinational WHO Global Data System for Maternal and Perinatal Health, for example, 235 hospitals from 16 countries in Africa and the Americas have collected accurate data from more than 165 000 deliveries. This was achieved with considerable effort by the 525 collaborators from each region in only three months of data collection.

Several activities need to be considered before starting recruitment in a multicentre trial. Introducing the trial's protocol at the hospitals and clinics involves the preparation and distribution of materials directed to care providers and patients. In multinational trials, translation into local languages is essential. Training activities, including all personnel involved at any stage of the trial (nurses, midwifes or doctors for recruiting patients, data collectors, etc.), have to be considered in the trial budget and timelines. For example, before starting recruitment in the WHO RCT of calcium supplementation for the prevention of pre-eclampsia among low-calcium-intake women,[11] nurses, midwives and doctors at the participating clinics and hospitals were trained on blood pressure measurement techniques, based on the WHO document 'A practical guide on how to measure blood pressure and test for proteinuria'.[22] Note that this is not contrary to the spirit of the pragmatic nature of the trial, as these guidelines are universally recommended in maternity care, but with limited standardisation and quality control.

DURING THE IMPLEMENTATION OF THE TRIAL

Our experience in conducting large multinational trials has shown that some issues are more efficiently resolved when they are continuously monitored. We conducted weekly monitoring of recruitment during the conduction of the calcium supplementation trial previously mentioned; in Argentina, for example, local coordinators detected problems at one clinic and corrective measures were implemented immediately. Periodic meetings at the coordinating units are also important in order to follow up the performance of the trial. This strategy allows the coordinators to detect early treatment compliance problems, or unexpected high rates of losses to follow-up. For example, if the leading health care provider of the best recruiting centre in a trial suddenly goes on sick leave, alternative arrangements must be made as soon as possible (e.g. recruiting more centres, increasing recruitment in the already existing ones).

Periodic visits to the centres by principal investigators help to maintain enthusiasm and commitment of the local staff, allowing problems to be more easily detected and resolved. It is also encouraging for people involved to receive periodic progress reports on their activities, as well as documented changes in the overall trial profile. Unfortunately, many principal investigator meetings are conducted at the central unit, or at hotels near to airports. This strategy, while convenient, is not good for the overall trial enthusiasm and for the improvement of skills of local staff. We encourage people to have their meetings at local clinics, where the trial is conducted, and thus more easily contribute to local capacity-building efforts. Table 26.5 lists some management actions that could have an impact in the trial success.

Sometimes women recruited into a trial need to be followed up after hospital discharge postpartum. As postdelivery routine medical records are not well developed in many countries, it is necessary to develop strategies for patient follow-up at home, according to the specific objectives of the trial. At recruitment it has to be considered whether contact details other than those routinely collected are necessary, such as alternative addresses and phone numbers (e.g. from relatives, work, school, etc.). Moreover, some women may need extra visits at the hospital or clinic, or home visits. Those activities, if

Table 26.5. Checklist for randomised trial management

At the preparatory and early recruitment stage	Be sure to get ethics approval and all-trial related bureaucratic procedures resolved Select and train all personnel required for the trial Distribute all study data forms and related materials Consider piloting forms and procedures
During the recruitment phase	Periodically visit the recruiting centres Maintain commitment and enthusiasm Monitor recruitment rates, compliance and follow-up Check the quality of the data Be available! Respond to queries immediately, set up alternative strategies to deal with unexpected problems
At the analysis stage	Keep the collaboration together 'teamwork' Resolve queries and inconsistencies early
At the end of the trial	Organise a final collaborators' meeting to share results before publication Acknowledge all collaborators' participation (in publications, certificates, etc.)

planned in advance and considered in the budget, are feasible even in urban areas located in developed or developing countries.

For example, more than 3200 children between 9 months and 6 years of age exposed *in utero* to magnesium sulphate (nearly 70% from developing countries) were approached and evaluated recently by us in a long-term follow-up trial for the prevention of eclampsia among women with pre-eclampsia.[23] Figure 26.2 shows how the families' addresses were placed on a city map at recruitment for a more effective use of resources when home visits were planned for the postdelivery follow-up.

Finally, wherever the data are physically entered and processed, the data management system should include the detection and resolution of inconsistencies and errors as soon as these problems are detected. Periodic listing of such errors should be sent to the local coordinators for confirmation or correction. The closer these reports are from the time of the event, the easier the resolution of the query. Online, real-time data entry and management at the clinic or maternity unit level, as a highly efficient tool, is therefore now routinely incorporated to all our randomised trials.

GLOBAL PARTNERSHIPS: THE WAY FORWARD

We have presented some methodological and practical reasons for preferring large randomised trials in perinatology. We have also discussed how to organise and implement them. We have noted that the recruitment of sufficiently large numbers of subjects requires multicentre international collaboration. Developing a trial protocol with input from a diverse group of people requires searching for compromises that every participating centre is able and willing to accept. Protocol flexibility needs to strike a balance between allowing continuation of routine local practices, and maintaining methodological quality in trial procedures. Efficient communication is crucial, in order to resolve any problems without delay, but may also present challenges. Our new online system is now available in most places and has facilitated coordination and data management.

Consent for participation in trials raises further issues, particularly as different countries will have different procedures and norms. Moreover, many new legal barriers have been introduced in some Western countries for clinical

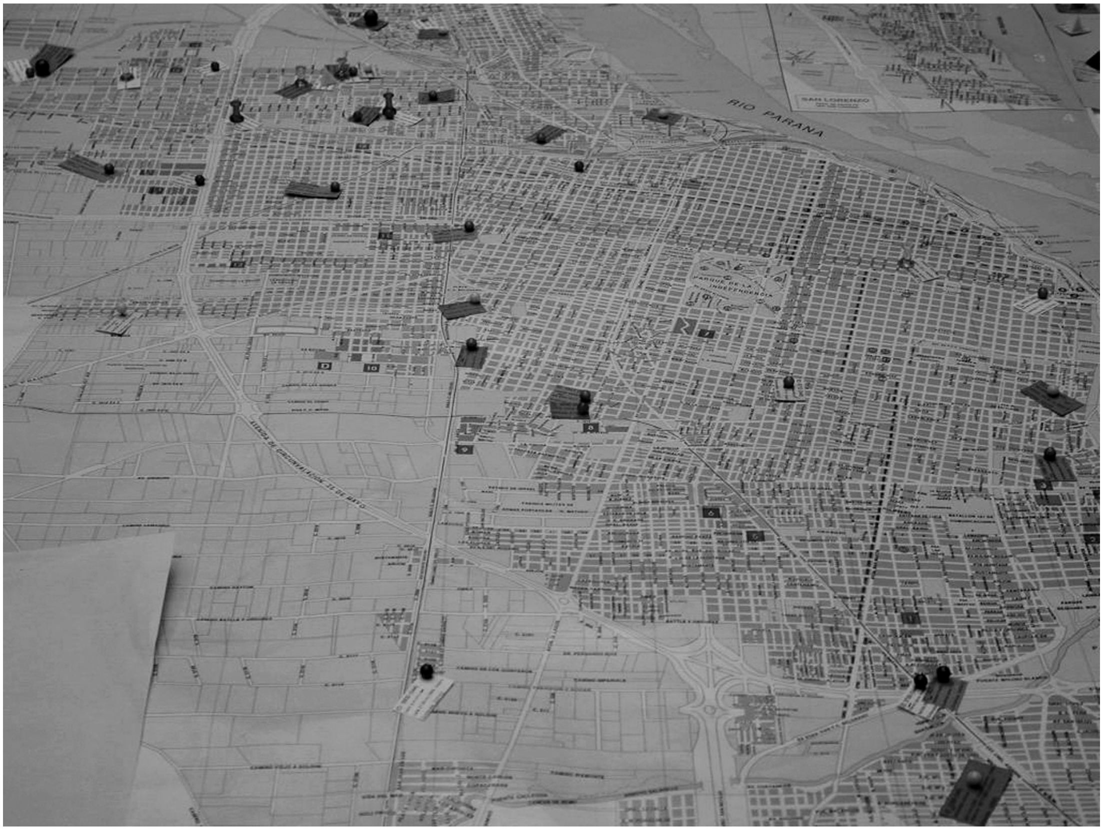


Figure 26.2. Locations of families to be followed up were geographically placed on a city map before starting home visits after hospital discharge.

and epidemiological research. Most of these difficulties can be overcome through regular consultation that is sensitive to local norms, values and beliefs. Partnerships based on mutual trust and respect are essential for the success of large collaborative trials, avoiding 'safari research'-like projects. There is a growing demand for strong evidence from high-quality randomised trials in maternal and perinatal health. Our experience demonstrates that international collaboration offers a feasible, enjoyable and productive strategy for addressing priority questions of global importance.

REFERENCES

1. Villar J, Carroli G. Methodological issues of randomized controlled trials for the evaluation of reproductive health interventions. *Prev Med* (1996) **25**(3): 365–75.
2. WHO Global Programme to Conquer Preeclampsia/Eclampsia. Department of Reproductive Health and Research, World Health Organization, Geneva (2002).
3. World Health Organization: Maternal and Perinatal Research Five Years Programme of Work: 2004–2009. Department of Reproductive Health and Research, World Health Organization, Geneva (2003).
4. Villar J, Mackey ME, Carroli G, Donner A. Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models. *Stat Med* (2001) **20**: 3635–647.
5. Mignini L, Latthe P, Villar J, Kilby M, Carroli G, Khan K. Mapping the theories of preeclampsia: the role of homocysteine. *Obstet Gynecol* (2005) **105**: 411–25.

6. Conde-Agudelo A, Villar J, Lindheimer M. World Health Organization systematic review of screening tests for preeclampsia. *Obstet Gynecol* (2004) **104**: 1367–91.
7. Villar J, Carroli G, Khan-Neelofur D, Piaggio G, Gülmezoglu M. Patterns of routine antenatal care for low-risk pregnancy (Cochrane Review). The Cochrane Library, Issue 3. Oxford: Update Software (2004).
8. Carroli G, Villar J, Piaggio G *et al*. WHO systematic review of randomised controlled trials of routine antenatal care. *Lancet* (2001) **357**: 1565–70.
9. Abalos E, Duley L, Steyn D, Henderson-Smart D. Antihypertensive drug therapy for mild to moderate hypertension during pregnancy (Cochrane Review). The Cochrane Library, Issue 4. Oxford: Update Software (2004).
10. Gülmezoglu M, Villar J, Ngoc NT *et al*. WHO multicentre randomised trial of misoprostol in the management of the third stage of labour. *Lancet* (2001) **356**: 689–95.
11. World Health Organization. Calcium supplementation for the prevention of pre-eclampsia among low calcium intake women. A randomized controlled trial. WHO, Geneva (2002).
12. Villar J, Ba'aqeel H, Piaggio G *et al*. WHO randomized trial for the evaluation of a new model of routine antenatal care. *Lancet* (2001) **357**: 1551–64.
13. Villar J, Bakketeig L, Donner A. *et al*. The WHO Antenatal Care Randomised Controlled Trial: rationale and study design. *Paediatr Perinatal Epidemiol* (1998) **12**: 27–58.
14. Zelen M. A new design for randomized clinical trials. *New Engl J Med* (1979) **300**: 1242–5.
15. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research* London: Arnold (2000).
16. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *Br Med J* (1996) **313**: 36–9.
17. Zelen M. Randomised consent designs for clinical trials: an update. *Stat Med* (1990) **9**: 645–56.
18. Senn S. *Statistical issues in drug development*. Chichester: John Wiley & Sons, Ltd (1999) 187–206.
19. *The Cochrane Handbook Glossary*, Version 4.2.4. The Cochrane Library, Issue 2. Oxford: Update Software (2005).
20. The tactics of performing therapeutic trials. In: *Clinical Epidemiology*, 3rd edn. Sackett DL, ed. http://www.nsite.ca/sackett/uploads/collections/93338888.pdf (accessed May (2005).
21. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *J Am Med Assoc* (2003) **289**: 2554–9.
22. World Health Organization. A practical guide on how to measure blood pressure and test for proteinuria HRP/RHR/WHO. WHO, Geneva (2001).
23. Magpie Trial Follow Up Study Management Group. The Magpie Trial Follow Up Study: outcome after discharge from hospital for women and children recruited to a trial comparing magnesium sulphate with placebo for pre-eclampsia [ISRCTN86938761]. *BMC Pregnancy Childbirth* (2004) **4**(5).

RESPIRATORY

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green

27

Respiratory Medicine

ANDERS KÄLLÉN
Jullovsvagen 13, Lund, Sweden

AIRWAYS DISEASES

Airway obstruction is a common and important feature of some respiratory diseases. It can be acute, 'semi-chronic' (e.g. due to cancer) or chronic. The chronic obstructive airway diseases can be divided into whether the obstruction is reversible or not. In the former case the patient usually has asthma, in the latter case chronic obstructive pulmonary disease, abbreviated COPD. These disease concepts lack precise definitions, and the division is only meant as a first approximation. Both diseases are inflammatory diseases mainly of the lower respiratory tract: in asthma there is an inflammatory process mainly in the central airways, whereas the inflammation of COPD is predominantly peripheral with progressive destruction of lung tissue. Inflammation in the upper respiratory tract, i.e. rhinitis, is characterised by both acute and chronic conditions, the most distinctive being seasonal hayfever.

MEDICAL BACKGROUND

Asthma

From a clinical point of view, asthma presents itself by recurrent breathlessness, cough or wheeze caused by variable or intermittent narrowing of the intra-pulmonary airways. The severity of these symptoms has a wide range, from very mild intermittent with symptoms only upon provocation, to severe persistent with large impact on daily life. There is no precise definition of asthma, and therefore the prevalence is hard to establish. We know, however, that it is commoner in children than in adults, and more common in boys than in girls.[1] A figure for children around 10% and half that for adults is probably close to reality in most of the Western world. There is, however, a definite regional inhomogeneity with regions with much higher prevalence and regions where the disease is rare. Most epidemiological studies seem, however, to agree that the prevalence is rising.[2]

The high prevalence of asthma gives a poor prediction of the impact of the disease on the community, because the overwhelming majority of asthmatics are mild sufferers with symptoms confined to wheezing after exercise or breathlessness in association with an upper respiratory tract infection. At the other end of the spectrum, asthma is a crippling, life-threatening disease with acute severe attacks requiring emergency

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

room treatment. In the Western world, about 1% of adults and 2% of children need medical attention for asthmatic symptoms.[3,4]

Many factors are known to cause narrowing of intra-pulmonary airways. The sensitivity to such stimuli varies between individuals but under normal circumstances the concentration of such substances is too low to produce symptoms in healthy subjects. Asthmatic patients are more or less characterised by a high sensitivity to such stimuli, a phenomenon called *non-specific bronchial hyperresponsiveness*. The most common cause of non-specific bronchial irritation is exercise and, for many, this may be the only manifestation of their asthma.

It is essential to make a clear distinction between this non-specific hyperresponsiveness and the allergic reactions. Allergy is an immunological reaction to a specific environmental agent. Hyperresponsive bronchi, in addition to responding in an exaggerated fashion to exogeneous stimuli, will also respond in an enhanced fashion to inflammatory mediators released in the bronchial wall as a result of an allergic reaction. Thus a trivial allergic reaction in a hyperresponsive bronchus may provoke a large bronchoconstrictive response. There is little, if any, relationship between the degree of atopy and non-specific hyperresponsiveness. Instead the degree of non-specific hyperresponsiveness is associated with the degree of inflammation in the respiratory tract.[5]

Asthma may start at any age. When starting during childhood and adolescence it is likely to be associated with atopy, as compared with when symptoms start later in life. Most asthmatic patients have perennial symptoms, but a minority show a seasonal variation, sometimes confined to periods with airborne pollen, sometimes to the winter months. Thus different asthmatics may have symptoms during different periods of the year, with long periods of absolute or relative relief between attacks of varying severity.

In general, asthma carries a favourable prognosis because the bronchial inflammation does not usually cause permanent tissue damage. However, in a subgroup of subjects, irreversible bronchial obstruction develops later in life.[6]

Rhinitis and Nasal Polyposis

The upper respiratory tract is to some extent like terminal bronchioli without smooth muscles. Instead the nose has venous sinusoids and the major reason for obstruction of the upper airway tract is vasodilation of capacitance vessels and oedema while secretion can contribute. Another difference to the lower tract is that stimulation of nervous irritant receptors in the nose results in sneezing, which is the cleaning reflex of the upper airways corresponding in a way to coughing, which is the cleaning reflex of the lower airways.

Inflammation in the upper respiratory tract, rhinitis, presents as one or more of the symptoms nasal congestion, rhinorrhea (i.e. runny nose), sneezing and itching. Chronic inflammatory conditions can in predisposed individuals result in benign protrusions of nasal polyps into the nasal cavity, polyposis.

Rhinitis can be allergic or non-allergic, where the latter is characterised by the presence of symptoms of varying severity. Allergic rhinits can be seasonal as hay fever (SAR = Seasonal Allergic Rhinitis), or perennial. In the latter case the symptoms can be due to continuous exposure to allergens like the house dust mite, or may present itself intermittently as episodes triggered by allergens like grass pollen. Despite the common inflammatory denominator for allergic rhinitis and polyposis, there is no evidence that the two conditions are closely linked, or that allergy plays a major role in the aetiology of nasal polyposis.

Rhinits is a very common disease, but surprisingly little is known about its epidemiology. The nose has a filter function, and is therefore exposed to a much larger amount of inhaled allergens per cm^2 than bronchi, especially when the allergens are large. SAR is due to airborne plant pollen. From a clinical point of view the most widely distributed ones are those of grasses, but some tree pollens, including birch and olive tree, are also important, as is ragweed. It is important to note that the pollen season for an individual plant species varies from one country or region to another. Also, whereas the season for airborne pollen is limited to perhaps half a year in temperate zones, in warmer climates it is so long that what seem to be perennial symptoms may be provoked by multiple and sequential seasonal allergies.

Patients with nasal polyposis suffer from a series of symptoms, just as in rhinitis, but in particular nasal blockage and an absence of smell. The prevalence is not known (there are few epidemiological studies) but is probably in the range of 1–4%. The diagnosis of polyps requires appropriate inspection of the nasal cavities by a trained physician.

COPD

Chronic obstructive pulmonary disease (COPD) is characterised by long-term, in general progressive, irreversible obstruction of the flow of air out of the lungs. To a large extent it is comprised of two related diseases:

1. Chronic bronchitis, whose clinical definition is productive cough (from bronchial secretion) on most days for three months/year for two consecutive years. The mucus hypersecretion comes from hypertrophied bronchial glands and increases the risk of bacterial lung infections.
2. Emphysema, which has a pathological definition with enlargement of the alveoli due to the destruction of the walls between them. These walls contain elastic fibres, so their destruction reduces the elasticity of the lung, leading to collapse, and thus obstruction, of airways.

The disease entities asthma, chronic bronchitis and emphysema are in no way mutually exclusive: a given patient can have symptoms from more than one. The definitions of the last two do not imply that the patient has airway obstruction, so not everyone with these diseases has COPD.

COPD is believed to affect more than half a billion people worldwide, causing perhaps 3 million deaths annually. When diagnosed, this is often in a relatively late stage of the disease, with less than 50% of lung function remaining, so the majority of cases are at any specified point in time undiagnosed. The prevalence of diagnosed COPD is about 5%, and is increasing.

Pathologically COPD is a disease with periferal inflammation (thus rather a bronchiolitis than bronchitis) with progressive lung tissue destruction. In the Western world, by far the most important factor responsible for COPD is smoking it has been said to be responsible for up to 90% of COPD patients.[7] However, only about 15–20% of all cigarette smokers develop COPD. The mechanism seems to be that cigarette smoke attracts cells (neutrophils, macrophages and cytotoxic T-cells as opposed to eosinophils and T-helper cells in asthma) to the lungs that promote inflammation, and these are stimulated to release elastase, an enzyme that breaks down the elastic fibres in lung tissue. Normally the lungs are protected against this enzyme by the elastase inhibitors, among them α_1-antitrypsin, which is produced in the liver (congenital deficiency of this enzyme is another, but rare, causation for emphysema). Air pollution has been suspected to have similar effects as smoking, but it is unclear to what extent that is an important aetiological factor for COPD. Also, there is a high COPD incidence in women in Asia attributed to cooking fumes.[8]

The typical COPD patient has been smoking 20 or more cigarettes a day for more than 20 years and presents with a chronic cough, shortness of breath (dyspnea) and frequent respiratory infections. If the underlying disease is mainly emphysema, shortness of breath may be the only symptom. Initially the dyspnea only comes during physical exercise, but as the disease progress it occurs already on minimal exertion. For the patient with chronic bronchitis dominating, the major symptoms are chronic cough and sputum production. The sputum may be clear but is usually coloured and thick as bacterial colonisation is common.

Acute exacerbations are key clinical features of COPD. There is presently no universal definition of what an exacerbation in COPD is, but vaguely it is a sustained worsening of the patient's condition, beyond normal day-to-day variation. Exacerbations are more common the more advanced the disease is, and they have considerable impact on quality of life and the need for health care resources. They are often triggered by a viral infection, to which a bacterial colonisation is added which requires antibiotic treatment.

It is part of the ageing profile that lung function deteriorates. In non-smokers without respiratory disease we expect the lung function index FEV_1 (see below) to decline by 20–30 ml per year, beginning at an age around 35 years. This rate of decline is larger for smokers, dependent on amount of smoking, and also a key component in the natural history of COPD. After smoking cessation this accelerated decline in FEV_1 shifts towards what is found in non-smokers. In late stages of COPD weight loss, measured as body mass index (BMI), is a good predictor of survival, and comorbidity in terms of pulmonary hypertension (cor pulmonale) almost regularly develops, with symptoms such as ankle swelling.

CURRENT TREATMENTS

The respiratory tract has a limited repertoire of responses to irritation or other stimulation. In the nose vasodilation leads to decreased airway calibre and nasal blockage. The bronchi may change their calibre or alter the amount of glandular secretion produced, leading to obstruction. There is oedema, hyperaemia and cellular infiltration of the wall of the tract. Afferent nerves may signal information to the brain stem to produce sneezing (upper tract) or cough or the sensation of breathlessness (lower tract). The relative importance of these factors varies between individuals, and different drugs interfere with different factors.

Drugs for chronic obstructive respiratory diseases are given either systemically, as tablets, or by local administration using an inhalation device. When it comes to inhaled products, it is important to note that a treatment consists of two objects, a drug to be delivered to the body and an inhalation device used for this deliverance. We will not discuss devices here, only drug classes. It is, however, important to understand that the amount of drug delivered to the airways may vary considerably from one inhalator to another.[9] The same might be true of the distribution pattern within the lungs, with potential consequences for the effectiveness of the treatment.

Bronchodilator Drugs

There are three basic groups of bronchodilator drugs: β_2-agonists (today by far the largest), xanthines and anticholinergics.[10] Their modes of action differ somewhat. We discuss each class of drugs separately.

β_2-agonists bind to the β adrenergic receptor and stimulate the intracellular accumulation of the signal substance cAMP (cyclic adenosine monophosphate). There are now three known types of β-adrenergic receptors in the human body: stimulation of the β_1-receptor causes cardiac stimulation and intestinal inhibition, whereas stimulation of the β_2-receptor results in bronchodilation, vasodilation, stimulation of skeletal muscles and uterine contractile inhibition. A third type, β_3-receptors, causes lipolysis.[11]

The development within this drug class has been towards more and more potent and selective β_2-agonists. The first generation of drugs were short-acting with a duration of action of, at most, 4–6 hours. Lately a few long-acting drugs, with duration of action superseding 12 hours, have been introduced.[12]

β_2-agonists are of benefit to the majority of asthmatic patients because of the bronchodilator property; rapid-acting ones are often given as rescue medication for relief of symptoms. The drug class does, however, have actions other than smooth muscle relaxation that may contribute to their long-term therapeutic effect in asthma and motivate their use in COPD: they stimulate mucociliary function in the airways, restore normal clearance of bronchial secretion and inhibit microvascular permeability in the airways leading to decreased mucosal oedema.

Side effects are a consequence of the binding to receptors in tissues and organs outside the lung: tremor by binding to receptors in skeletal muscle, tachycardia by binding to receptors in the heart (this problem was reduced as the drugs became highly selective, but there are some β_2-receptors in the heart as well) and hypokalemi, due to a redistribution from extra- to intracellular spaces. In general tolerance develops rapidly to the extrapulmonary effects, so these are usually mild or absent in patients, though individual variation in the sensitivity can make the use of these drugs impossible in the occasional patient.

Xanthines, the most well-known member of which is caffeine but the most widely used one as treatment for airway obstruction is theophylline, are potent smooth muscle relaxants by acting directly in the intracellular messenger cAMP. Thus they have about the same pharmacological actions as β_2-agonists. But since they act intracellularly and not by binding to a receptor on the cell surface, the effect is more generalised and the side effects are somewhat different and potentially more serious than those of β_2-agonists. The most important ones relate to the gastrointestinal, cardiovascular and central nervous systems. At the start of treatment with oral theophylline, most patients will experience some caffeine-like symptoms including irritability and nausea, symptoms which usually fade away after a few days. For that reason, however, treatment is usually initiated in subtherapeutic doses and progressively increased over a period of 1–2 weeks.[13]

The serious side effects, in contrast to the caffeine-like ones, are well correlated to plasma concentrations. In clinical practice theophylline concentration in plasma has to be monitored and dose adjusted so that the plasma concentrations lies within a therapeutic window. Because of this the use of xanthines has diminished over the last 10 years as alternative treatment has become available.

Anticholinergic drugs have been used since ancient times for the treatment of asthma. The use of various plant derivatives has evolved through synthetic atropine to more selective bronchodilating anticholinergic agents with fewer side effects than atropine.[14]

The bronchodilating effect of this drug class is due to the drugs' antagonism of the binding of acetylcholine (from the vagal nerve) to the muscarinic receptors of bronchial smooth muscle. These drugs are particularly used in treating reversible airway obstruction in COPD.

The side effects of anticholinergic agents are due to the blockade of muscarinic M_2-receptors in other organs and include dryness of the mouth, blurred vision, urine retention and difficulty in micturition, tachycardia, flushing and lightheadedness.

Corticosteroids

That glucocorticosteroids (GCSs) have a therapeutic effect on asthma, rhinitis and other anti-inflammatory diseases has been known for a long time and is due to their being human-made analogues of an endogenous anti-inflammatory steroid–cortisol. Cortisol is in a way nature's own remedy for inflammation: if we remove the adrenal glands inflammatory reactions are greatly exacerbated. Regulation of endogenous cortisol is complex, involving the hypothalamic–pituitary–adrenal (HPA) axis. During a severe inflammatory response, elevated levels of cytokines stimulate centres in the brain, leading to an increase of cortisol, in the circulation thereby attenuating the inflammatory response. It is now believed that even at normal levels, endogenous hormones will regulate inflammation. The GCS mode of action is by binding to a glucocorticoid receptor within the cell's cytosol. When used for treating asthma, for example, GCSs lead to a reduction of airway inflammation, mucous hypersecretion and airway reactivity while restoring the integrity of the airways.[15]

Originally GCSs were given systemically as asthma treatment. There are, however, well-known side effects of high doses of oral GCSs over a long time that limit that usage. These include, but are not limited to, osteoporosis, hypertension, adrenal insufficiency and Cushingoid features as well as growth retardation in children. Concern about these side effects diminished the use of oral GCS as an asthma treatment. Inhaled GCSs have improved the benefit/risk ratio. Since administration is aimed directly at the site of inflammation, lower doses can be used, giving lower GCS concentrations in plasma with largely negligible systemic side effects as a result. Inhaled GCSs are now widely accepted as first-line anti-inflammatory therapy for asthma,[16] and an oral course for about 10 days of a GCS is the preferred way of treating exacerbations in many parts of the world.

To evaluate the long-term side effects of inhaled GCSs is difficult. They are rare at doses given in asthma treatment, so large numbers of patients and long-term clinical studies are needed. Some information can be gained by studying the endogenous cortisol levels. As already mentioned, the endogenous cortisol level is controlled by the highly complex HPA-axis. Introduction of exogenous GCS in the plasma will affect this axis and lead to a suppression of the endogenous cortisol levels, the degree of which is determined by the plasma concentrations and the potency of the drug. Thus, the degree of suppression is a measure of the amount of active (on the HPA-axis) exogenous GCS in the body.

Other Drugs

Vasoconstrictors are used extensively in rhinits. Topical α-agonistic sympatomimetics effectively and promptly alleviate the nasal blockage. They have no effect on rhinorrhea, nasal itch or sneezing.[17]

Antihistamines are used for rhinitis, mainly as rescue medication. Their main effect is to block peripheral H_1-receptors, which limits vasodilatation in the nasal mucosa. They have an effect on nasal itching, sneezing and discharge, but little or no effect on nasal congestion and blockage.[18]

Disodium cromoglycate (DSCG) and nedocromil sodium DSCG has been used as a prophylactic anti-asthma drug, mainly by children and young adults.[19] To be effective it should be administered four times per day. Originally its mechanism of action was proposed to be stabilisation of mast cells, though that is probably not the case. Taken immediately before exposure, DSCG affects the asthmatic reactions induced by various stimuli. However, after discontinuation of long-term treatment, DSCG seems not to have modulated the bronchial hyperresponsiveness or the underlying inflammatory reaction.

Nedocromil sodium is another non-steroidal substance with anti-inflammatory properties *in vitro*. It acts as an inhibitor at several levels of neurogenic inflammation in asthma. Clinical studies have demonstrated improvements in airway functions, including a reduction of bronchial hyperreactivity, but it does not protect against maximal airway narrowing, which is an important feature of inhaled corticosteroids.

Both DSCG and nedocromil are remarkably free from side effects. They are also used for rhinitis, with effects similar to antihistamines.

Leukotriene modifiers The cysteinyl leukotrienes are products of the arachidonic acid metabolism with effects that mimic many features of asthma, e.g. by increasing eosinophil migration, mucus production, airway wall oedema and causing bronchospasm. Oral leukotriene receptor antagonists, to be administered once or twice daily, are available along with an oral leukotriene synthesis inhibitor, which has to be administered four times daily.

Leukotriene modifiers improve airway function and decrease the need for additional maintenance and rescue asthma therapies. Leukotriene modifiers also attenuate bronchospam induced by allergens, cold air, salicylates and exercise. In patients with chronic, persistent asthma, results from clinical studies indicate that inhaled corticosteroids have a more consistent and greater average effect than antileukotriene dugs.[20]

The long-term safety of leukotriene modifiers is still not clear. Some patients reducing their oral corticosteroids when treatment with antileukotriene drugs has been initiated have developed a special type of vasculitis called Churg–Strauss syndrome. However, it might be the unmasking of a pre-existing condition and not induced by the leukotriene modifier per se.

MEASUREMENT SCALES

When measuring the status of the chronic obstructive airways disease in a subject we can either rely on data obtained at a visit to the clinic, or monitor the patient over a longer period by daily recordings at home.

At a visit to the clinic the primary focus is usually to obtain an objective, indirect, measure of airways narrowing–a lung function test. Such a test measures some functional index of the airway calibre in some kind of experimental setting. We will discuss various such indices and experimental settings and what they measure.

In the latter case, long-term daily recordings, we provide the patient with a diary card and usually ask him or her to record twice-daily information relating to symptoms of the disease under study. In addition patients are often given a device to obtain an objective measurement of lung function at home, traditionally a peak flow meter.

These two approaches are in no way mutually exclusive–in a long-term study we can make experimental manouevres of the first kind. As an example there is virtually no long-term asthma trial that does not measure FEV_1 on visits to the clinic. However, for the present discussion we consider experimental approaches and diary card approaches separately, except that single FEV_1 measurements at the clinic will be discussed along with diary cards.

Lung Function Measurements

Airway narrowing leads to an increased resistance to the airflow. The airway resistance can be measured directly with body plethysmography (in a 'body-box'), an expensive and rather complicated piece of equipment. Another way of measuring lung function is by flow measurements, which use a much more inexpensive apparatus, a spirometer. However, spirometric measurements, to be discussed below, depend not only on airway resistance, but also on lung volumes.

When doing a spirometric manouevre the patient takes a maximum deep breath and then exhales as rapidly as possible as much as possible. The spirometer records the exhaled volume as a function of time, $V(t)$. From this curve (Figure 27.1) a number of spirometric indices can be obtained. The most widely used measure is the forced expiratory volume in

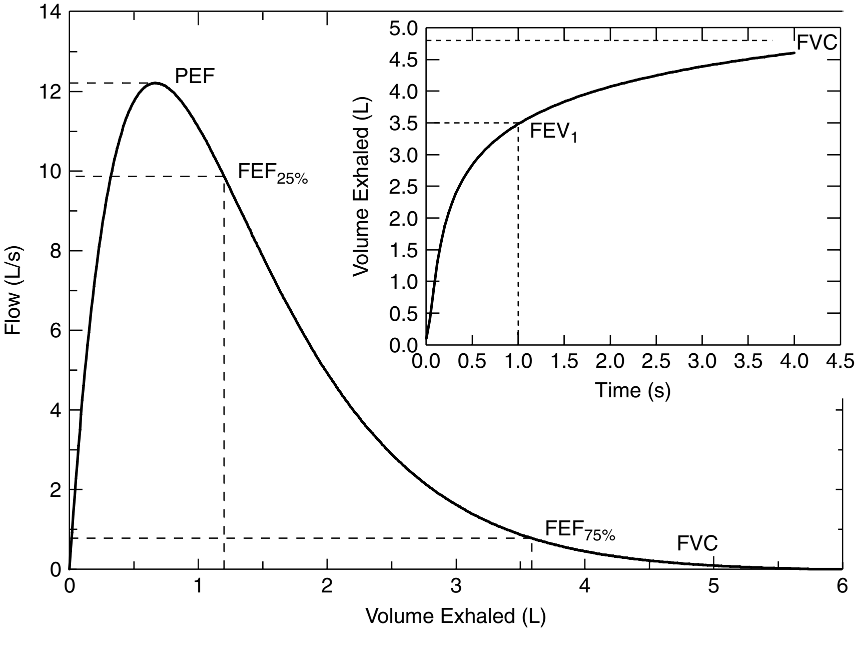


Figure 27.1. Illustration of some spirometric measurements.

1 second ($V(1)$, denoted FEV_1), followed by the forced vital capacity ($V(\infty)$, denoted FVC). If the expiratory effort has been markedly inadequate it is usually obvious from the trace.

By calculating numerically the derivative of $V(t)$ we obtain the expiratory airflow. Its maximum value, which usually occurs within 100 ms, is the peak expiratory flow (PEF). Often the spirometric result is shown by plotting the flow against the volume exhaled. From this curve we can identify both PEF and FVC (but not FEV_1), but can also define new measurements, like $\text{FEF}_{25\%}$, which is the flow when 25% of the FVC has been exhaled. Another measure of current interest is $\text{FEF}_{25-75\%}$, which is the amount of volume expired per second when the exhaled volume increases from 25% to 75% of FVC. It is considered to measure effects in the small airways.

A full spirometric manouevre consists of the measurement of the inspiratory part also. The inspiratory vital capacity (IVC) is a measure of the functional residual capacity (FRC) and is an important measure in COPD patients.

If performed correctly, the spirometric test is highly reproducible but somewhat effort dependent. Different parameters are effort dependent to different degrees, e.g. FEV_1 is less dependent than FVC, since it only needs maximum effort for 1 s. The direct measurement of airway resistance (R_{aw}), which is done in the body box, is effort independent, but has a poor reproducibility. Since the spirometry has a good reproducibility, and uses fairly simple and portable equipment, it is most useful for clinical purposes. In special situations, however, the assessment of resistance might be preferable.

PEF is much more effort dependent than FEV_1, but it can also be measured by much cheaper apparatus than a spirometer. Such a peak flow meter is often provided to the patients for self-monitoring at home. Instructions are then given to fill in a diary card and to contact the health care service when PEF has dropped for a few consecutive days below prespecified levels. In the same way, PEF can be monitored with this simple device in a long-term study by recording, often twice daily, in a diary card.

There is a diurnal variation in FEV_1 and other lung function measurements. It is therefore important, when comparing different such measurements obtained at different visits to the clinic for the same patient, that these measurements are taken at approximately the same time of the day.

Lung function measurements can be followed in order to assess effects, but also to characterise disease severity. However, lung function is a function of gender, age and 'size of patient'. Therefore a lung function parameter cannot be judged on an absolute scale–an FEV_1 of 2.4 L means different things for a young, tall boy and an old, tiny woman. A measure of disease severity would be the ratio of the actual FEV_1 and the would-be, and unmeasurable, FEV_1 the patient should have without the obstructive airway disease. As a remedy for the latter various predicted formulae have been obtained for different lung function parameters. Thus, a key disease severity parameter is the FEV_1 in per cent of predicted normal, for both asthma and COPD. There are a number of such formulae available, generally depending on demographic variables like race, gender, age and height.[21] It should be emphasised, however, that these measures cannot be anything but rather approximate ones, since the predicted normal values are not exact counterparts to the unknown lung function without disease! If the lung function is between 80 and 120% of the predicted normal value, it is in general considered to be 'normal'.

Another disease characteristic obtained from lung function measurements is the reversibility. This is an index obtained from a very simple single dose monitoring experiment (see below): we measure FEV_1, give a rapid-acting β_2-agonist and wait 30 minutes (typically) and measure FEV_1 again. The classical reversibility is then obtained as

$$\text{reversibility} = 100 \times \frac{\text{FEV}_1(\text{after}) - \text{FEV}_1(\text{before})}{\text{FEV}_1(\text{before})}$$

A value in excess of 15% was previously considered indicative of reversible airways obstruction, though later guidelines[22] use 12%. The basis for these numbers is somewhat unclear–it is probably chosen in order to be 'certain' that there is an effect: the variability in FEV_1 is such that a numerical increase per se is not definite proof of an improved lung function.

Upper Airway Function Tests

There are also upper airway function tests similar to the lung function indices discussed above. They are, however, much less used, since symptom scores are considered of overriding importance in rhinitis studies. Resistance can be estimated by two different techniques: posterior rhinomanometry in which values are obtained by probes placed in the mouth, and anterior rhinomanometry in which a device in the nose is used. Less complicated and expensive methods for assessing nasal patency rely on the measurement of peak nasal flow either on inspiration (PNIF) or expiration (PNEF). We do not discuss these methods in any further detail.

DESIGNS FOR EXPERIMENTAL ASTHMA TRIALS

For asthma studies, there are a number of experimental designs to measure various aspects of the therapeutic effect based on objective lung function measures. For this section, let E denote an index of lung function. In most real-life cases this is FEV_1, but the discussion is not restricted to this case.

We can group the designs into two groups: either the response after administration(s) of a study drug is followed, which can be done by time or by increasing doses, or the protective effect of the study drug to some provocation is assessed.

Single Dose Monitoring

This type of experiment is simple. Consider one individual on one occasion when this experiment is performed. We first take a baseline measurement, E_0, give the study drug and then follow lung function at predetermined time points after study drug administration. This provides us with an approximation of a response curve $E(t)$, where we use $E(0) = E_0$ (though technically it was obtained at a time point $t < 0$). From this curve a number of measures can be obtained for further analysis. The two most important measures derived from the curve $E(t)$ are:

1. The average level, defined as the area under the curve (of the polygonal approximation we have observed to the response curve) divided by observational time. This we denote by E_{av}.
2. The maximal level, $E_{\max}$.

We can also compute $t_{\max}$, the time at which $E_{\max}$ occurred. This is sometimes useful. Other potential measures are related to the concepts of responders, 'onset of action' and 'duration of effect'. Tradition has it that for FEV_1, effect is declared at a time point t if there is a 15% increase compared with baseline at that time. Based on such a concept, we can define the time of onset as the time point (if any) at which the polygonal curve cuts the line $E = 1.15 \times E_0$ for the first time (for rapid-acting bronchodilators one usually adds the restriction that this should occur within 30 minutes). The ending of effect occurs at the time point on the polygonial approximation which is followed by at least two observations below the line $E = 1.15 \times E_0$, provided that two measurements were taken. If only one was taken, that will suffice, and if no measurement was found below the line, censor the end of effect to the last measured time. The duration is then the time from the onset of action to end of effect.

The main problem with these definitions of onset and duration of action is not the arbitrary number 15%, but the fact that effect is measured by relating to baseline. This is not appropriate, since lung function has a clear diurnal variation. It might be a reasonable approximation for a few hours, the perceived time of clinical efficacy of a short-acting β_2-agonist, but will produce a incorrect result if used for a longer period. In fact, there are studies in which a patient receiving placebo as treatment has had a definite increase in lung function already on the first measurement after treatment administration (changes in the means–not individual spurious events), so the use of baseline as a reference when declaring effects should be very much questioned.

A related problem is to define responders. As the name suggests, a responder is a subject who responds to the treatment. Traditionally this has been decided based on the maximal increase from baseline. The discussion above implies that this is not necessarily a good way to go. To actually measure effect, we need to relate the measurements to the measurements obtained without drug administration. However, since asthma is not a stable disease, these must be taken simultaneously. And this is impossible! In clinical trials we do not really need this concept at all, except for descriptive purposes. We will return to the question of duration in the discussion of an example.

One lesson, however, from the discussion is that the effect for many of these variables is often measured clinically as per cent change. This means that

$$\Delta\text{effect} = \Delta E/E \approx \Delta \ln(E)$$

which by integration motivates why many lung function indices should be analysed on a logarithmic scale. We analyse these types of trials with multiplicative models, which is justified by this observation.

Challenge Tests

A challenge test is similar to the single dose monitoring test, except that most of the monitoring takes place after a provocation of some kind. Two important cases of challenge are as an exercise, either a treadmill test or using a cycle ergometer, and an allergen to which the patient is allergic. A baseline measurement E_0 is taken, often after administration of the study drug. Then the provocation is done and lung function followed. In most cases there are two phases in the reaction found. First there is an immediate reaction with bronchoconstriction within minutes which lasts 1–2 hours. Several hours later there is a delayed reaction with a much slower and sustained time course.

Typically an exercise test is followed only during the immediate reaction, the actual existence of a delayed reaction is controversial. The protective effect of the study drug can be measured by maximal decrease in lung function from baseline

$$\text{Index}_{EIB} = 100 \times (E_0 - E_{\min})/E_0$$

and we only need to follow the patient until we know he or she has attained the low turning point, whereafter the patient is given a high dose of a bronchodilator in order to restore lung function. Because of the intrinsic variability in lung function measurements, spurious local minima can occur in the measurement series–it is important that the investigator has certified that the global minimum has occurred before stopping. The most common lung function measurement here is again FEV_1. It should be noted that a better definition of the index would be $\text{Index}_{EIB} = 100 \times E_{\min}/E_0$, since then the analysis could be done on the multiplicative scale as discussed above!

For the allergen challenge test we are more interested in the whole response for 10–12 hours, in order to study both the immediate reaction and the late reaction. The immediate reaction (EAR = early asthmatic reaction) is an episode of acute bronchoconstriction which peaks between 10 and 20 minutes after inhalation and resolves within 1.5–2 hours. The late reaction (LAR = late asthmatic reaction) is probably an inflammation-mediated bronchoconstriction which starts about 3 hours after allergen inhalation and does not resolve for many hours. Allergen challenge tests are potentially dangerous, and are therefore not much favoured as a mode of studying asthma.

If they are, we need to measure FEV_1 repeatedly during the first hour, and then more sparsely during the next 7–8 hours (perhaps ones an hour). The EAR is most often defined as the maximum per cent reduction in FEV_1 (from baseline) occurring in the first hour after challenge, whereas the LAR is defined as the maximum percent reduction in FEV_1 (again from baseline) occurring between 3 and 7 hours after challenge. Alternatively we compute the area under the curve for the first hour and for the period between 3 and 7 hours after challenge and use that as an efficacy measure in much the same way as for the single dose monitoring experiment.

Hyperresponsiveness Studies

The level of airway responsiveness to a non-specific stimulus is measured by exposing subjects to the stimulus and measuring the response. There are a number of dialects of this test, by varying the selected stimulus, the mode of administration of it and the method of assessing the response.

The most commonly used stimuli are methacholine and histamine, though small doses of allergens can also be used. Methacholine and histamine produce similar responses, but the latter has more side effects and can only be administered safely in concentrations up to 32 mg/ml, whereas methacholine can be used safely in concentrations up to 256 mg/ml (these numbers should be compared with the clinical definition of hyperresponsiveness, which is that the provocation dose (PD_{20}, see below) is ≤ 8 mg/ml). The stimulus is administered from an aerosol, which can be done in different ways. Suffice it here to note that one can do it with or without a dosimeter which controls the dose. Response is generally measured either as FEV_1 or as airway resistance (or its inverse, conductance).

Technically the subject first inhales saline and then inhales progressively increasing, often doubled, doses of the stimuli from the aerosol at 3 minute intervals. There is a measurement after each dose administration, so we can consider the response to be a function of the last concentration or dose given. In both cases the saline inhalation produces the baseline value. From this dose–response curve (I call it that, though sometimes it is a concentration–response curve) different characteristics can be computed. A general dose–response curve is sigmoidal in shape and well approximated by a log–linear portion over most of its response range. We can, however, not clinically obtain information on much more than the lower part of this dose–response curve, which means that traditional measures for dose–response curves (ED_{50} and slope, see below) are not usually estimatable. We can think of the effect of the drugs as a parallel shift of the response curve so that if a given response is obtained with dose D without the drug, it takes dose ρD (with, hopefully, $\rho > 1$) to get that response with the drug. To estimate ρ in this type of study we fix a level, expressed as per cent decrease in the response, and estimate the dose of stimuli needed to obtain that level. The dose which gives a decrease of $x\%$ in the response variable is denoted PD_x (or PC_x if we do not control doses). For FEV_1 we usually compute PD_{20}, whereas for R_{aw} a higher percentage can be used.

The actual algorithm for estimation of PD_x can vary. The following suggestion is justified by this description of the dose–response curve.

1. If there is a dose with less than $x\%$ decrease followed by a dose with more than $x\%$, log–linear interpolation (of $\log D$ vs. response) is done.
2. If the first dose provoked a fall in excess of $x\%$, we cannot do log–linear interpolation. In that case we do a linear interpolation back to baseline and obtain a dose corresponding to a fall of $x\%$ from this. However, we never go back more than to half the first dose given.
3. If the last dose produced a fall of less than $x\%$, we extrapolate log–linearly, but only up to twice the highest dose given.

What to use as dose can also be discussed. If we do not control the dose, we must use the concentration given. If we control the dose, we can choose to use cumulative doses or last dose without much difference to the final results, when provocation doses are compared (because, for a geometric series, the sum is essentially proportional to the next dose, and we compare ratios). In general the use of the cumulative dose seems to be favoured.

The measure PD_{20} is not limited to the possible interpretation discussed above (as the relative dose potency ρ of the stimuli). If the effect is due to changes in both position and shape, the measure can still be used. For epidemiological purposes another index has been introduced, the two-point slope, which is the per cent decline from baseline to last dose, divided by last dose. Though this index has a clear interpretation (as per cent decrease per unit dose), this interpretation is wrong since the decrease is not linear with dose – instead it is virtually zero until it becomes linear with log dose.

Note It has been suggested that you cannot estimate PD_x if there has not been a fall of $x\%$. Technically this means that we should note it as missing. This might be sensible for caring for the patient, where this is perceived as no hyperresponsiveness. However, for a clinical study, where treatments are compared, it is imperative to do an estimation. Setting it to missing means that the analysis loses the information that a high dose is needed to achieve the specified decrease!

EXPOSURE STUDIES ON RHINITIS

For allergic rhinitis there are two study designs of the experimental type available. Both are exposure studies, one in the natural season, one in an artificial season:

1. The Park study. In this study the subjects are exposed to pollen over a period of 1–2 days by walking around in a park. There are two main problems with this type of study – it is highly dependent on season and the patients often find it very boring.
2. The experimental Nasal Allergen Challenge Artificial Season model. In this type of study the subjects are artificially exposed to pollen for some period. This can be either as spray application for a few consecutive days, or in an Environmental Exposure Unit (EEU) in which subjects are exposed to pollen in a special room for, typically, 3 hours on a number of consecutive days. In this room there is a flow of air to which the pollen is added and evenly distributed in the air by fans.

In both cases we measure nasal symptoms as the outcome variable. Both these studies are parallel groups in design, but effects can often be demonstrated with rather small patient numbers.

EXERCISE TESTS IN COPD

Since a progressive decline in physical fitness is the main characteristic of COPD, exercise tests are useful for a proper evaluation of treatment effects in these patients. In these tests exercise can be done as walking, running (treadmill tests) or bicycling. The basic design of the test can be, in its purest form,

1. to determine the maximal workload sustainable, or
2. to fix the workload at some level, and determine the endurance time.

An example of the first kind is to measure the distance walked in a prespecified time, like 6 or 12 minutes. The second-kind counterpart would be to fix (individually) the pace at which walking should be done and then measure time walked. It is believed that the second kind of experiment is more relevant in the study of COPD – that it correlates better with breathlessness and disability. The first kind of test is probably much influenced by attitude and expectation. We should also note that the second kind of test should provide a lower metabolic and respiratory stress than the first one and that the limiting factor in an exercise test does not have to be physical fitness – COPD patients may well fail due to muscular fatigue before general fatigue.

In practice many tests used constitute a compromise between these two approaches: a specific time schedule is designed so that the workload is held fixed for some fixed time, then increased for ‘a step’ for another period of time etc. Typical cycle-ergometry and treadmill tests have this design, as has the so-called shuttle walking test in which the patient walks at a given pace for one minute, then increases the pace for another minute etc., all according to a well-defined protocol. The natural outcome of these experiments is an endurance time, though for some cycle-ergometry tests you could alternatively use the total workload (but these should be heavily correlated).

For a comparison of some exercise tests in COPD, see Oga *et al.*[23]

In conjunction with these tests, measurements of breathlessness are usually done. There are different tests available. A much used dyspnoea score is the Modified Borg Scale,[24] in which dyspnoea is scored on a 0–10 scale before and after the exercise test. Alternatively one can use a visual analogue scale with the same effect. An alternative scale is the Transitional Dyspnoea Index,[25] for which we first rate three factors (functional impairment, magnitude of task and magnitude of effort) on baseline, each on scales 0–4 (well, 4–0 actually–the scale is reversed), and then rate the changes over the exercise directly on a scale from −3 to +3.

LONG-TERM CLINICAL STUDIES WITH DIARY CARDS

Diary Card Studies in Asthma and Rhinitis

In a diary card study, the patient is provided with a diary card to fill in various information about the status of his or her disease under investigation, often twice daily. For most asthma studies, the patients also measure PEF. It is important that the patient uses the same peak flow meter throughout the study, since different brands have different scales, and there is a considerable within-brand variability as well. In addition to this, some symptom scoring is requested. This can be either an overall assessment of symptoms, or assessments of specific symptoms, like wheezing, shortness of breath and cough for asthma. Finally, for asthma studies, the use of rescue medication, usually a short-acting β_2-agonist, should be entered into the diary card. With the increased use of information technology paper-based diary cards are more and more being replaced with electronic counterparts, which has the potential benefit of monitoring when the recordings are done. Some such devices can also contain a spirometer, which makes it possible to replace the somewhat variable PEF measurement with the more accepted FEV_1 measurement. The fact that the electronic device can be programmed, so that it only accepts data obtained at the time point when it should be obtained, increases the validity of these types of data. The FEV_1 measurements recorded with a portable spirometer should be more valid than PEF data obtained by a peak flow meter and manually recorded on a paper-based diary card. Our discussion will primarily relate to the old paper-based diary cards with a concomitant peak flow meter. We leave it to the reader to assess potential changes that occur for electronically acquired data.

In terms of basic design we have two types of long-term clinical studies in asthma:

1. Studies in which treatments are fixed throughout the period under investigation. An arm in such an study might be, for example, budesonide Turbuhaler 200 μg b.i.d.
2. Studies in which the treatment is not fixed throughout the period under investigation. In such studies we can either vary the dose of the investigational product, or vary the dose of some concomitant treatment. One typical such study has an arm in which treatment is initiated with a high dose of a given GCS, which is then reduced according to some scheme until the patient is no longer controlled on the present dose. A variant is the steroid sparing studies, in which a fixed dose of some investigational treatment is given throughout the study period and concomitant with this treatment some GCS is given, the dose of which is then reduced in steps. For inhaled GCS, oral steroid sparing studies have been performed in this way; for other anti-inflammatory drugs like leukotriene modifiers, inhaled steroid sparing studies are relevant.

The usage of the diary card data varies between these two types of data. In studies with fixed treatments they define the primary efficacy variables, whereas in studies with varying treatment doses, dose changes are conditioned on the diary card variables and these therefore act only as control variables.

Diary card data in a long-term clinical study often represent a considerable amount of data, as measured in megabytes on disk. The number of megabytes, however, does not truly reflect the information value. Data are not obtained in a very controlled fashion. Morning values are generally considered slightly sharper than evening values, since sleep is comparatively similar among patients and data should be obtained and recorded immediately after waking up.

For that reason, peak flow obtained in the morning is often considered the primary variable of interest in a long-term diary card study on asthmatics. The day-to-day variability, for a symptomatic asthmatic, can be considerable. However, using the mean of all values over a prespecified period, as long as possible, generally provides us with a measure that has proven to be useful in many clinical studies. An alternative efficacy measure is to use FEV_1 measurements obtained at visits to the clinic. Though each individual FEV_1 measurement so obtained is much more reliable than a single PEF measurement, the overall mean over a treatment period of daily recorded PEF measurements obtained in the morning is, in our experience, a more efficient variable for demonstrating differences between treatments in lung function. Since most treatments are mainly symptomatic, integrated measures over time are the relevant 'endpoint' measures.

When using FEV_1 obtained at visits to the clinic as the primary variable in a long-term clinical trial, we must take the diurnal variation of lung function into account. Thus it is important that FEV_1 is measured at approximately the same time of the day on each visit. To obtain maximal efficiency we also need to schedule the patients for visits to the clinic early in the morning (around 8 a.m.), with approximately the same argument as given for peak flow morning measurements above. The possibility of enforcing this will very much determine the effectiveness of the FEV_1 variable in discriminating between treatments.

For rhinits, the symptoms recorded are nasal blockage, rhinorrea, sneezing and/or itchy nose which sometimes are combined into the nasal index score, which is the sum of them. In addition to this, eye symptoms are recorded as a secondary variable. The most readily available objective measure in the clinical trial setting is either the Peak Nasal Inspiratory Flow (PNIF) or Peak Nasal Expiratory Index. Of these the PNIF parameter seems to be the most discriminative.

The data in diary cards can be used in different ways to compute variables for use in statistical comparisons. As already indicated, in fixed dose studies period mean values are often computed, not only of PEF measurements, but also of symptom scores and of the use of rescue medication. Because of the intrinsic variability in the underlying disease it is important to compute means over long periods, preferably the full treatment period. This means that, for some drugs at least, the mean will contain data from a period of onset of action, though the effect of that will be minor in long-term studies.

Mean values of symptom scores do not seem very meaningful to most clinicians in assessing the actual response. An alternative is to compute the percentage symptom-free days, which is somewhat simpler to interpret clinically. Similarly it might be useful to compute the percentage days with no rescue medication.

When many symptoms are recorded individually, one approach to the analysis of the data is to compute the sum of the symptoms (as the nasal index score), but an alternative is to analyse them simultaneously in a multivariate analysis.

Even more useful, sometimes, is to collect data within day, or adjacent days, to form new measures. A simple such measure for asthma is to define a patient to have control of the asthma, if there are no symptoms and the patient did not use rescue medication. The percentage of such days with asthma control can be a useful summary measure for some patient populations, typically rather mild ones. A variant of this is to define mild exacerbations, or episodes, of asthma from diary cards by looking for worsening of lung function and/or increase in rescue consumption and/or symptoms. The exact criteria for such episodes probably need to be adjusted to the patient population under study, and to the study design. In order to avoid spurious events, it might be a good idea to define an event to have occurred for two consecutive days in order to be labelled an episode. From an analysis point of view we can analyse time to first such exacerbation or the percentage episode-free days.

One final note on response data in studies on asthmatic patients, especially PEF, and the disease asthma in general: when we interpret diary card data, obtained over a longer period, we must interpret them on a group mean level. A discussion of individual responders is virtually meaningless. A responder refers to a patient that responds to the investigational treatment. This cannot be assessed on the basis of diary card data, since the underlying disease is, by definition, varying–what seems to be a clear response could well be a period of good asthma control totally unrelated to drug effect (in some cases a study effect) and the converse. This is obvious once one has inspected placebo data in a long-term study. However, this does not exclude that one can define responders according to some criteria and compare per cent responders between treatments, since that is a comparison on group level.

So far we have considered studies with fixed treatments during the investigational period. In a study which tries to identify the minimal dose on which the patient is controlled, the obvious endpoint for analysis is this minimal dose. This is true irrespective of whether the dose in question refers to the investigational drug or to some concomitant drug (as in the oral sparing studies alluded to above).

More explicit examples of this will be demonstrated below.

Long-Term Studies in COPD

The most important goals in the management of COPD, and therefore ultimate drug targets, are the prevention of disease progression, relief of symptoms, improvement of exercise tolerance and improvement of health-related quality of life. Exercise tests have already been discussed, so we can group the effects to investigate in long-term studies into three groups:

1. Signs and symptoms/Quality of Life
2. Exacerbation rates
3. Rate of decline in FEV_1.

The first two of these focus on the symptomatology of COPD, whereas the last focuses on modification of the progressive lung destruction.

Changes in COPD symptomatology can be measured in different ways. By signs and symptoms above, I mean that we collect information on things like night sleep, breathlessness, coughing and chest tightness on a more or less daily basis, using diary cards. Additional information can be obtained from quality of life questionnaires that should be filled in at the start of the study and at least at study termination. Quality of life questionnaires are further discussed below. In addition to this, the patient can measure PEF at home as a measure on lung function, or we can do spirometry on visits to the clinic and measure FEV_1.

The data in diary cards can be used in different ways to compute variables for use in statistical comparisons. We typically compute means over periods for comparison. Mean values of symptom scores do not seem very meaningful to most clinicians in assessing the actual response. An alternative is to compute the percentage symptom-free days, which is somewhat simpler to interpret clinically. Similarly it might be useful to compute the percentage days with no rescue medication.

When many symptoms are recorded individually, one approach to the analysis of the data is to compute the sum of the symptoms, but an alternative is to analyse them simultaneously in a multivariate analysis. Studies of this kind do not need to be very long if the treatment has an immediate effect. Three to six months may well suffice.

One of the key features of COPD is the regular occurrence of exacerbations of the disease, needing health care resources. To identify these and measure a treatment effect on the exacerbation rates is therefore one key objective of a clinical programme in COPD. Studies of this type should typically be at least a year in duration.

When studying the rate of decline in FEV_1 we try to demonstrate that we actually interfere with the destructive process in the lungs with our treatment. Today that seems to require studies at least three years long, in order to get reliable estimates of the rate of decline. In the future the same objective may be possible to resolve by imaging techniques, in which pictures of the lung are analysed for the quantification of emphysema.

More explicit examples of this will be demonstrated below.

QUALITY OF LIFE

Asthma and COPD are chronic disorders that can place considerable restrictions on the physical, emotional and social aspects of the lives of patients. Assessments of the patients' own perception of the impact of the condition on their life, of general well-being, is known as measurement of quality of life. Quality of life may be useful for assessing the degree of morbidity, e.g. in order to evaluate the health economic impact of diseases in the community. It is assessed by questionnaires that include a large set of physical and psychological characteristics assessing the general functioning and well-being in the context of life style. Quality of life scales are either general and not specifically designed for patients with any particular disease, or they are more specific disease related but, as of to day, in general not applicable to the general population due to cultural differences.

General health status scales such as the Sickness Impact Profile with 136 items[26] have been proposed. A compromise between lengthy questionnaires and single-item measures of health has also been proposed. The Nottingham Health Profile with 45 items and SF-36 (a Measures of Sickness short-form general health survey) are now widely used and validated. The SF-36 Health Status Questionnaire is based on 36 items selected to represent eight health concepts (physical, social and role functioning; mental health; health perception; energy/fatigue; pain; and general health).[27] Its quality of life scales have been shown to correlate with the severity of asthma, but it has yet to demonstrate any superiority over the simpler, diary-card-based symptom scores for demonstrating effect in clinical trials.

For COPD the St George's Respiratory Questionnaire has gained importance in later years. It is perhaps the most comprehensive questionnaire for the evaluation of quality of life in airways diseases and allows for direct numerical comparisons to be made among patients, study populations and therapies, and has sensitivity when applied to mild as well as severe disease.[28] It was developed by Paul Jones at St George's Hospital in London in 1990 and is designed to measure impact on overall health, daily life and perceived well-being. The measure consists of 50 (76 responses) items that produce three domain scores and one overall score. The domains are: symptoms (severity and frequency), activity (that cause or are limited by breathlessness) and impact (on social life and psychological disturbances caused by the airways disease).

One related instrument should be discussed here, though not considered by some a quality of life questionnaire since it does not measure impact on life. It is the Clinical COPD Questionnaire (CCQ, abbreviated so because it was originally called the COPD Control Questionnaire) which is a recently developed instrument according to what is presently considered best practice in the field. It is a questionnaire including 10 items divided into three domains: symptoms (4 items), functional state (4 items) and mental state (2 items). The symptoms domain includes the four items *Short of breath at rest/Short of breath doing physical activities/Did you cough/Did you produce phlegm.* There are two versions available, one for use at visits to the clinic with a recall period of one week and one diary version to be completed every day.

CLINICAL TRIAL METHODS

HOW TO AVOID BIAS

Blinding

Most effect measurements of the respiratory diseases discussed here are influenced to a non-negligible degree by the patient's expectations. One typical example of this is that in some double-blind, placebo-controlled single dose trials measuring bronchodilation, there is an immediate response in lung function also in the placebo group, which probably is due to (false) expectations. The classical methods to avoid expectation bias, blinding and randomisation, are therefore important. A clinical study in this area should follow a double-blind approach in which study drugs are prepacked in accordance with a suitable randomisation schedule, and supplied to the trial centre(s) labelled only with the subject number and the treatment period so that no one involved in the conduct of the trial is aware of the specific treatment allocated to any particular subject, not even as a code letter.[29] The code should not be broken until all decisions concerning data validity have been taken and documented.

Many studies in the respiratory area concern inhalation products, where there are not only, say, two different drugs involved, but also two different inhalers (or perhaps one drug in two different inhalers). To maintain blinding in those situations one often needs to resort to the 'double-dummy' technique. This means that for each inhaler there has to be two clones: one with active drug and one with placebo. On each inhalation occasion, the subject has to inhale not only from the inhaler with active substance, but also from the other inhalers, but containing placebo. This might lead to a large number of inhalations per occasion, which in turn might lead to incomplete compliance. Note that the use of different inhalers implies a consideration of the order in which these should be taken. Carryover effects from one type of inhaler might dictate which should be taken first/last, whereas in other situations a balanced scheme might be called for.

Rhinitis studies pose a special problem in terms of blinding because the double-dummy technique is not considered appropriate – there is a fear that additional placebo material may clear the airways of drug so that the response is different with and without simultaneous placebo administration. This is a problem mainly when two different drugs inhaled through different devices are to be compared. The partial remedy that is most often used is to include a placebo group, and let half the group have one device and the other half the other device. That way, at least, the patient does not know whether he or she gets the active drug or not.

Open labelled studies might be acceptable for some systemic effect studies where the outcome variable is the plasma concentrations of some marker, or in long-term safety studies.

Other Sources of Bias

Another way to risk selection bias, also in a randomised, double-blind study, is to exclude data obtained on treatment. To exclude patients on data obtained prior to first dose cannot in itself produce bias. However, the prognostic factors for respiratory trials, like FEV_1 in per cent of predicted normal and reversibility, are only estimates of time-varying entities and are not precise enough that we can actually claim that a patient violating some inclusion criteria is not necessarily an appropriate patient for the trial. They are essential in order to focus the investigator on the appropriate patient population, but once a violation to the protocol criteria has emerged it might be appropriate to use the patient in the statistical analysis. With Senn,[30] I consider the protocol a guide to the physician, not the statistician.

Protocol violations after the first dose should not in general invalidate the patient for the statistical analysis. If such a protocol violation is confounded with treatment effects, their omission might bias the result. However, the fact that they are protocol violations might in itself imply that the measurements are improper measurements, which is another type of bias. This problem is illustrated in respiratory trials by the use of rescue medication.

During an asthma or COPD trial patients are usually provided with short-acting β_2-agonists to use as rescue medication. This means that some measurements of lung function and symptoms will be influenced by this add-on therapy, and the validity of those measurements (as direct treatment measures) will therefore be questioned. If treatments have the same effect they pose no problem, since they should occur with similar frequency in different groups. However, a more effective treatment is expected to have less consumption of rescue medication. As a consequence there is a bias towards no effect by including those measurement when computing period means, for instance. If we ignore them (i.e. consider them to missing) we introduce a bias in the same direction, since we only count the days when the patient was, relatively speaking, symptom free. There seems to be no easy way out of this dilemma, and the approach we have taken is to ignore the additional information on recently taken rescue medication for the analysis, but instead plot, descriptively only, for each day the proportion of patients that take rescue close enough to peak flow measurement. Hopefully, and this is usually the case, the main result and this graph give the same message on effect. At least this approach should be conservative.

TRIAL DESIGN CONSIDERATIONS

From a bird's-eye perspective, there are two types of responses for respiratory diseases:

1. Immediate responses that disappear within a short period of time. This includes the fast bronchodilating effect of β_2-agonists, responses to various provocations and specific systemic effects that are measured by markers in the blood (like plasma cortisol for GCSs and serum potassium for β_2-agonists). Many of these studies are single dose studies.
2. Long-term studies addressing effects on symptoms or average lung function of the underlying condition.

Crossover Trials

For the first type of studies, the crossover design is well suited. In such a study each subject is randomised to a sequence of treatments, and acts as his or her own control for treatment comparisons. In many cases this is attractive, because the within-subject variability is smaller than the between-subject variability, which means that a smaller study is required for the same power, as compared with a parallel group study. Numerous variations exist, e.g. trials in which each subject receives only a subset of the treatments studied (incomplete design), and trials where the same treatment is repeated within a subject.

However, there are caveats with crossover studies. The primary caveat is the possible presence of carryover effects (in fact, non-equal carryover effects), which might bias treatment comparisons. When deciding if a crossover design is appropriate for a particular study, we therefore must convince ourselves, beforehand, that we can get rid of possible carryover effects by separating the various treatment periods with wash-out periods during which no treatment is given. For a single dose short-acting β_2-agonist trial a wash-out period need in general only be a few days. A trial which studies cortisol depression after short periods of GCS administrations should have wash-out periods of 1–2 weeks, though shorter ones are acceptable in single dose studies.

When periods in crossover trials contain repeated dosing over a few weeks, and the actual experiment is performed at the end of such a period, it is often unnecessary to have drug-free wash-out periods between periods. But to take that step, one must make plausible that taking a new treatment directly after another does not by itself have any effect on the variables to be analysed.

Parallel Group Trials

Since the treatment for respiratory diseases in general is to achieve a prolonged improvement of the underlying condition, the most important trials need to extend over longer periods. The most natural design for them is the parallel group design, where the subjects are randomised to one of a number of arms, each arm being allocated a different treatment. These treatments will include the investigational product at one or more doses, possibly including a placebo (dose $= 0$) arm, and possibly some active control treatments at one or more doses.

HANDLING OF MISSING VALUES

The experimental types of studies are often very difficult to perform clinically. To get good-quality data, it is extremely important that the investigator and staff have good knowledge of and experience in this type of study. It is in general better to do such studies in one or two experienced centres with many patients, as compared with using many centres with fewer patients, despite the fact that the study might take longer to perform. With those premises missing values are, in our experience, a negligible problem since in general experimental sequences are complete. When missing values occur, they are due to discontinuations between experimental sessions and as these are few, there is no problem in analysing the resulting unbalanced study.

It is a completely different issue with long-term clinical studies. Here, not only patients do discontinue treatment, but there are also missing observations during the period. For the fixed treatment trial the purpose is in general to achieve a steady state, at a group level, on the treatment and then compare the level of the measured variable in steady state between groups. For patients reaching steady state we therefore have in general a number of data points measuring the steady-state level; in the case of diary cards, quite a few. The efficacy variable is then usually a summary statistic of these data points, like a period mean of a diary card variable. Missing values during this period do therefore not constitute a major problem–we take the mean of the available data.

Sometimes a long-term clinical study contains experimental procedures, like a methacholine provocation test – or just spirometry at the clinical visit, at least FEV_1. This is then done at least once pre-randomisation and then again only a few times on treatment, in particular on the last protocol visit. The effect variable should not be defined as the change from baseline to last protocol visit, but as the change from baseline to the last visit on treatment the patient attended. Specifically, instead of analysing the change in log PD_{20} from visit 2 to visit 8, we define the efficacy variable as the change from visit 2 to the last visit on treatment, which might be visit 4, 6 or 8 in a particular study. Technically this is equivalent to what is called the last value carried forward, or the last value extended, principle, but there is no need to use that label if we define the efficacy variable appropriately.

This still leaves us with one key problem: what if we do not have any efficacy measurements on treatment to use? To avoid that problem in diary card studies, it is often better to define the full treatment period as the period over which to compute summary statistics. At least that provides an effect measurement for each individual who has started to fill in the diary cards. The drawback is that the period mean for one patient can be the mean of 90 data points, whereas for another it is the mean of only a few data points. The next step is in general to analyse these period means with an ANOVA, and then the information of the precision of the computed mean is lost. On balance, it seems better to have an effect measurement on each patient.

For data obtained on clinical visits, the risk of having no measurement at all on treatment is not negligible. The omission of such patients from the analysis means that there might be a potential bias in the end result. To understand this, assume that there are no withdrawals in group A, but half the patients withdraw from group B because of insufficient efficacy. The remaining patients in group B are then the ones who needed less treatment. That patient group has a corresponding subgroup in group A of approximately the same size (as a consequence of proper blinding and randomisation procedures), but group A also contains another subgroup of patients, corresponding to the ones that dropped out from group B. The remaining groups are therefore not really comparable, and inference drawn from available data might be misleading!

However, there is no simple, trustworthy, remedy for this. Our approach is to use available data for the analysis, hoping that the potential bias is conservative (which it probably is in most cases for respiratory trials). However, if there is a large difference in withdrawal rates between the groups, it is logical to do the primary analysis on withdrawal data to assert group differences.

When describing diary card data, daily mean value curves by treatment are useful. When computing these mean values, missing values pose great problems in that raw mean values can produce very misleading impressions of group behaviour. To see why, consider a placebo arm in a diary card study in which the patients with worsening symptoms drop out progressively (the worse the symptoms, the earlier they drop out). As the patients with low response values drop out, the group mean will increase, so the temporal behaviour of the mean values will indicate that the placebo group increases in effect with time. However, this effect is due solely to withdrawals!

To avoid this culprit when plotting the temporal behaviour of variables some kind of imputation of data is needed, in order to keep the denominator the same when computing mean values within a treatment group. The simplest such imputation is to use the last value extended (LVE) approach, in which the last value for a withdrawal is extended to later time points. Using this principle, the mean values plotted can be interpreted as follows: the mean at time t is the mean of the last recorded measurements up to and including time t. When using this principle for diary card variables like PEF it is often better not to take only the last measurement, but rather the mean of a few measurements. More sophisticated approaches based on some kind of multiple imputation technique for missing data can also be considered, but the add-on value of doing so is probably very small for the average study in respiratory diseases.

MULTIPLE COMPARISONS

A respiratory trial usually contains a number of effect variables, and often also a number of different treatments. Thus there are multiple comparisons to be done. This poses a major problem, because of the risk of overemphasising fluke significances due to many comparisons.

To handle the many effect variables we therefore have to predefine which one is to be considered the primary one. It is from the result on this variable that the overall statistical conclusion from the study can be drawn. In general one study can have a few different objectives that are not closely related (like efficacy and safety), and then a primary variable for each objective should be appropriate. However, it is probably too statistical an approach to focus only on the primary variable when trying to understand the results of a clinical trial. No variable fetches all aspects of a respiratory disease, and the approach should be to select the most sensitive variable as the primary variable, to decide on the overall conclusion, but then a number of secondary variables should be described so that one gets an overall picture of what is ‘going on’.

When it comes to the problem of multiple treatment comparisons, the study logic should be structured in terms of well-spelt-out objectives. To prove efficacy might mean one comparison; to estimate a relative dose potency another analysis. With precisely formulated questions the multiplicity problem here should at least diminish substantially. This approach will be illustrated in what follows.

SAMPLE SIZE DETERMINATIONS

In order to certify that a proposed study is of an appropriate size, a sample size justification is needed in the protocol. It cannot be justified ethically to submit a number of patients to a study protocol if there is no hope whatsoever of demonstrating what you want to demonstrate. Similarly, if the study is heavily overdimensionalised we have put an unnecessary number of patients at whatever risk the study can carry with it, and that is not ethical too. However, sample size determination is there to ethically justify the study in advance – it has no consequences when the results are obtained.

In the respiratory area many test hypotheses are stated in terms of mean values, and for such variables the sample size is (essentially) proportional to the ratio $(\sigma/\Delta)^2$, where σ is the residual standard deviation and Δ is the mean difference we do not want to miss. More details are given in Chapter 2. When using a multiplicative model for a variable, these entities refer to the logarithm of the variable in question. Note that σ means different things in a crossover trial and in a parallel group trial – in the former case it refers to a within-patient variability (more exactly $\sqrt{2} \times$ the residual standard deviation of the ANOVA) and in the latter to a between-patient variability. Also, what is relevant is the residual standard deviation from the proposed ANOVA, which might contain a baseline adjustment.

The following table shows some typical values of the sample size parameters that can be used for asthma trials. Each example will be discussed in more detail after the table.

Increase in	Design	σ	Δ	$(\sigma/\Delta)^2$
PEF morning (L/min)	PG	40–45	10–20	4–20
Symptom score (0–3 scale)	PG	0.4–0.5	0.05–0.15	6–100
FEV_1 (L)	PG	0.4–0.5	0.05–0.1	15–100
ln (AUC FEV_1) (L)	XO	0.07–0.10	0.05–0.10	0.5–4
$^{2}\log(\mathrm{PD}_{20})$	XO	0.9–1.1	1–2	0.8–4

Here the range is not a range–the lower number for σ represents an optimistic number, the larger number a conservative one. Similarly, for Δ the range is more of a typical range for which to dimensionalise, not a range on what can be obtained. To obtain numbers (per group in the parallel group case) from this we should multiply by approximately 25.

For the crossover measurements of the table, we just note that the AUC refers to AUC-based average over the full period and that for that variable the pre-dose FEV_1 value is used as covariate in the analysis. For the PD_{20} case no baseline covariate is used.

For the parallel group measurements we use baseline covariate. For PEF morning a baseline is obtained as the mean value over a number of measurements, typically 1–2 weeks, and then the effect variable as the mean of 1–3 months of data. The shorter the periods, the larger the residual standard deviation. Similarly, for FEV_1 the table refers to a measurement at both baseline and end of treatment, but the treatment value could well be a mean of a number of measurements. Moreover, the FEV_1 data refer to the situation when visits to the clinic are spread out over the morning, the European style, as discussed earlier. With visits scheduled at 8 a.m. precisely larger effects can be expected.

Concerning symptom scores, these too are obtained from period means of diary card data, and relate to a typical asthma study. Changes in symptom scores are often small in studies of asthmatics with mild–moderate severity, since they do not have many symptoms on entry. In rhinitis studies a combination of symptom scores is often done. If we use the TNS discussed earlier we typically have a standard deviation of about 1.3 and effect sizes of 0.5–1, giving a $(\sigma/\Delta)^2$ of 2–8. Typically, therefore, rhinits studies can be smaller than asthma studies.

For COPD, exacerbation rates may be more important as the outcome variable. A rate of one exacerbation per year can be used in sample size calculations.

PHASE I/II STUDIES

EFFICACY STUDIES

In terms of efficacy, not much can be done in a Phase I trial. These trials, mainly concerned with tolerability and pharmacokinetics, give no real clue to whether a new drug actually works or not. Note that in general a respiratory drug must be very well tolerated to be useful, since there are so many efficacious and safe drugs on the market.

When trying to establish that a new drug is effective or not and to estimate clinically relevant doses, the approach differs between the drugs that have more or less immediate effect on lung function, typically bronchodilators, and the ones that work more indirectly on lung function, via the inflammatory process, as GCSs. For bronchodilators we can use small-scale experimental studies, whereas for anti-inflammatory drugs we typically need long-term studies from the very start.

To establish efficiency is conceptually simple: all we need to do is to show that a given dose of the drug is superior to placebo. There is, however, no true placebo treatment in long-term asthma or COPD trials–at a minimum the patients need to be provided with a short-acting β_2-agonist to be used as rescue medication. All new drugs are therefore studied on top of some baseline treatment, which in most cases is not very constant. For example, a GCS treatment is taken in addition to rescue medication. It potentially becomes a problem when you want to introduce a new rescue treatment, which is to be taken when needed, and you need to document long-term effects.

The next question, possibly posed in the same study, is which is the relevant dose span for further investigation? For a dose–response study (including the simple one with only placebo and one dose) the choice of patients is important. The majority of asthmatics, the mild ones, do not need much therapy and because of the relative imprecision in the measurement tools, many patients will have no measurable response in many of the tests discussed. For a bronchodilator study of the crossover type, where the design contains a number of experimental days when response is followed after a dose, some bronchoconstriction is needed in order to see an effect. By definition the bronchoconstriction varies with time for an asthmatic, so the measured response will depend not only on the dose of the bronchodilator, but also on the degree of airway narrowing on the particular day the experiment is done. Thus it is difficult to assess efficacy for the individual patient, which we handle in clinical studies by considering means. But more importantly, in order to achieve dose–response we need sufficiently many patients with sufficiently many days on which they can respond. The selection of such patients is not easy!

For long-term diary card studies we have a similar problem. The effect measures are rather noisy, and we generally need somewhat larger studies to measure a signal through all the noise. In parts of the world with a widespread health care system, most asthmatics are rather well treated. In particular the use of GCS already in fairly mild asthma in Europe, Australia and Canada seems to have made the majority of asthmatics more or less symptom free for most of the time. That in turn means that the traditional diary card might have difficulty in catching any responses.

From the foregoing discussion it should be clear that the magnitude of response in a particular variable is very difficult to assess: if it is small, is it because the patients studied did not have much room for improvement or was it because the drug was of minor efficacy? The only way, it seems, to actually assess the degree of response is to compare it with something we know, by experience, to work–to include an active control in the clinical trial. In my personal view, a clinical dose–response trial in asthmatics without an active control has very little information value. Also note that we should not need placebo in order to prove efficacy – it should suffice if we could prove that there is a dose–response relationship (this is in fact the point with the expression ‘show dose–response’: to prove that the drug has a pharmacological effect). This does not rule out (if the slope is positive) that small doses have a negative effect and larger doses a positive effect, something that should be borne in mind when interpreting the results.

What can be done in a placebo-controlled dose–response study without an active control is to estimate the minimal effective dose (MED). This addresses the question: ‘Which is the lowest dose, of those studied, that is proven effective?’ and tests, in a recursive manner to control significance level, the doses from highest to lowest with placebo. There are a number of algorithms available that can be used, but the key point about MED is that it is the lowest dose that *from this study* we can claim is effective. Thus the result depends heavily on the size of the study and choice of patient population, a property the average physician probably would not like MED to have. Thus there is great danger in using MED as defined here for decision making. In my view MED is more likely to lead to false decisions than correct ones.

The information we actually want from a dose–response study is the shape of the dose–response curve, to allow us to pick the ‘best’ dose. Not a detailed shape, but a simple approximation which can be used to derive insights. As long as there is a monotonous dose–response curve, a more complicated description than the one provided by the formula

$$E = E_0 + E_{\max}/[1 + (\mathrm{ED}_{50}/D)^{\gamma}] \qquad (27.1)$$

is rarely needed. This formula contains four parameters: E_0 is the baseline level of the response variable, corresponding to placebo, $E_{\max}$ is the maximal effect attainable, ED_{50} is the dose required to obtain 50% of this and, finally, γ is a sensitivity parameter which measures how much the response changes with changes in dose. The shape of this function is a sigmoidal curve with the extremely important property that over much of the range (say from $E_0 + 0.2E_{\max}$ to $E_0 + 0.8E_{\max}$) it can be well approximated by a straight line $E = a + b \ln D$. A description of such a dose–response curve should be the purpose of the dose–response trial, not to discuss the individual doses that were actually chosen to be used in the study.

Identifying the dose–response curve, however, does not give you a hint on how well the treatment compares with competing treatments. For that purpose it is wise to include an active control in the trial. We can then use the dose–response curve to estimate the dose of the new drug that produces the same effect as the active control does, hopefully with confidence limits.

Example: Bronchodilation

The bronchodilating effects of two long-acting β_2-agonists, we call them A and B, each with its own inhalation device, were compared by giving single dose administrations, followed by repeated measurements of FEV_1 over a 12-hour period. The following five treatments were studied in a randomised, double-blind, double-dummy crossover study: 6, 12 and 24 μg of drug A, 50 μg of drug B and placebo.

In Figure 27.2 we plot the geometric mean values, expressed as per cent increase from the baseline measurement taken prior to treatment administration. The reason for plotting geometric, and not arithmetic, means is that results are often

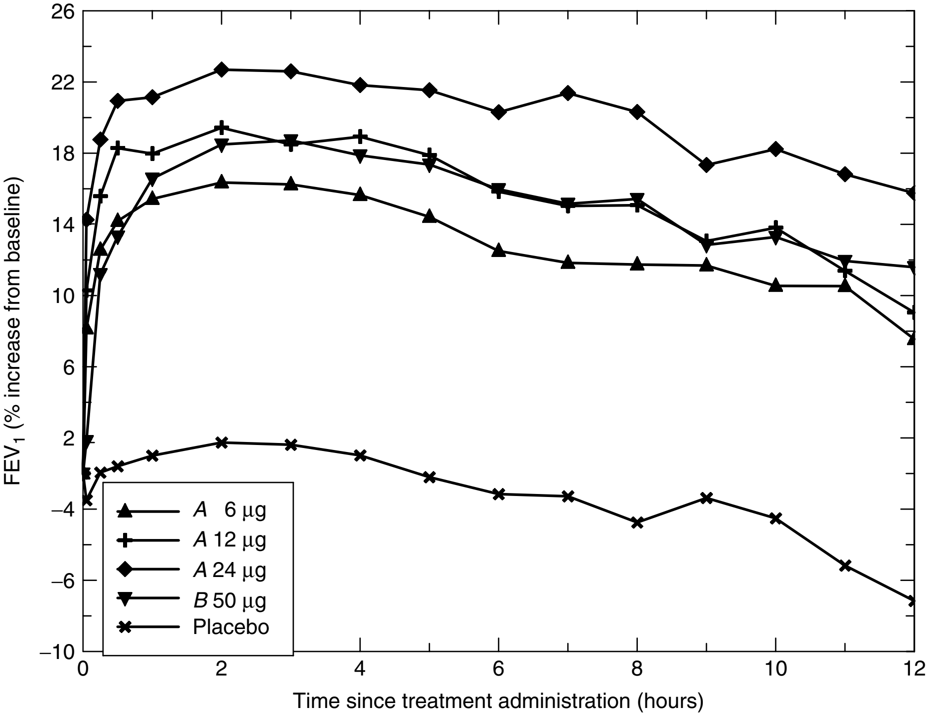


Figure 27.2. Geometric mean values, expressed as per cent increase from the baseline measurement, of FEV_1 measurements over 12 hours for individual treatments.

to be expressed as per cent increases, and then data should be analysed on a logarithmic scale so geometric means are therefore more natural than arithmetic ones. As a consequence, differences are unnatural entities to discuss and should be replaced by ratios.

To actually analyse the data we want some overall summary statistic that includes both the maximal effect and the duration of response, and we use the area-based average $\mathrm{FEV}_{1,av}$ over 12 hours. If we want to keep the idea of analysing on a multiplicative scale, we need to compute the area all the way down to zero. Alternatively, we could integrate over the baseline measurement, but then the area might be negative and we would be forced to do the final analysis on the original scale. We have chosen the former approach.

The ANOVA model uses the model

$$\ln(\mathrm{FEV}_{1,av}) = \text{patient}+\text{treatment}+\text{period} + \ln(\mathrm{FEV}_{1,base})$$

To have baseline as a covariate in a single dose study is essential. If we observe lung function over a short time period, baseline is very important and we could probably just as well use $\mathrm{FEV}_{1,av}/\mathrm{FEV}_{1,base}$ as the effect variable. However, when we observe over a longer time period, the influence of baseline should diminish and after many hours could probably just as well be ignored. By using it as a covariate, we get a reasonable compromise between these two extremes.

Based on the results from this analysis we can address various questions:

1. Which doses of drug *A* can we claim to be effective? To find this out we compare them, from highest to lowest dose, with placebo. Here is the result in tabular form:

Treatment	Mean ratio	95% confidence limits
24 μg of *A*	1.214	1.177, 1.252
12 μg of *A*	1.176	1.140, 1.212
6 μg of *A*	1.154	1.119, 1.190

Mean ratio relates to placebo. We see that the mean effect is 15–21% larger than it is for placebo, and the confidence intervals clearly show that all these comparisons were statistically significant. So we can claim that 6 μg is an effective dose of drug *A*, without compromising the significance level (see the discussion on MED).

2. Which dose of drug *A* has the same mean effect as the reference treatment, 50 μg of drug *B*? To do this, we fit (weighted linear regression to keep track of the uncertainties of the means−see Källén and Larsson[31]) a straight line to drug *A* means vs. log dose and estimate the dose that has the same mean effect as the reference treatment. This is illustrated in Figure 27.3. The actual dose estimate was 9.3 μg with 95% confidence limits 3.4−19 μg. As a consequence we find that 24 μg of drug *A* as a single dose has greater bronchodilating effect over 12 hours than 50 μg of drug *B*.

3. What about the duration of the effect? Looking at Figure 27.2 we see from the placebo curve indications of the diurnal variation that is known for lung function. To define duration by identifying when individual curves cross a line, say, 15% above baseline does not seem appropriate – if the placebo curve drops, you might still have a good response even when you are back to baseline. A more statistically sound approach would be to rephrase the question as 'is there still an effect after 12 hours?' and then compare the treatments to placebo at that time point. This is done in the following table:

Treatment	Mean ratio	95% confidence limits
24 μg of drug *A*	1.246	1.187, 1.309
12 μg of drug *A*	1.176	1.120, 1.234
6 μg of drug *A*	1.168	1.112, 1.226
50 μg of drug *B*	1.200	1.144, 1.260

Again mean ratio relates to placebo and we have responses that are between 17% and 25% larger than that after 12 hours. Thus all active treatments clearly have a duration in excess of 12 hours.

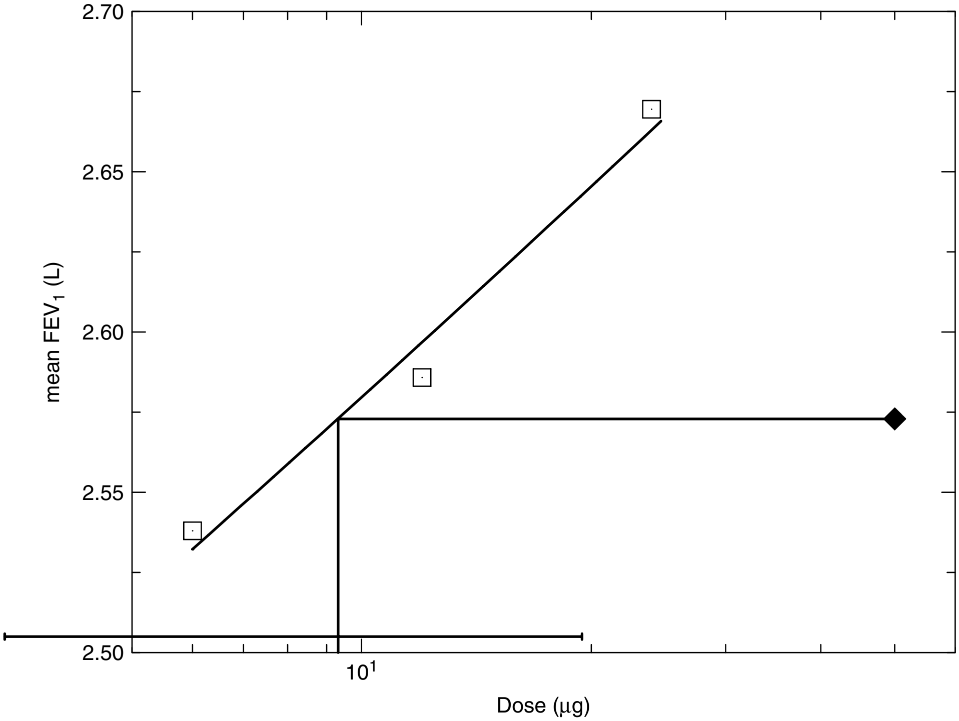


Figure 27.3. Treatment mean values for 12-hour average FEV$_1$ with fitted log–linear dose–response curve for drug *A* and estimation of D_{eq} relative to 50 μg of drug *B*.

SYSTEMIC EFFECTS

Effects of antiasthma drugs are in general not confined to the lungs. Both β_2-agonists and GCSs have receptors on/in cells throughout the body. Since the drugs are cleared through the bloodstream they will therefore have systemic effects (albeit perhaps not measurable). In contrast to the antiasthmatic effects, these effects can be measured both in healthy volunteers and in patients.

GCSs are synthetic cousins to an endogenous, anti-inflammatory, substance, cortisol. Given this, we can compare the pharmacodynamic systemic effects of different GCSs by comparing their effects on endogenous cortisol levels. This has the added advantage over drug plasma concentrations that it accounts for differences in potency in decreasing plasma levels of cortisol (which is done by negative feedback on the HPA-axis). It is important to state at this point that we do not study endogenous cortisol levels because they themselves represent a dangerous side effect. They are studied because they are sensitive markers for the ‘amount’ of exogenous GCS in the body!

With this model in mind we can use cortisol in plasma as an index of the systemic burden of therapeutically given GCS. By measuring the effect on cortisol we get a rather direct measure of the overall potency and concentration in plasma of the GCS. We can, however, not measure it time point by time point and compare with measurements without drug, since the level of cortisol is determined as a balance between production and elimination (with a half-life of about 1.5 hours) and the GCS acts on the production side. We therefore need to study a longer period. The cortisol levels in plasma have a diurnal rhythm which is very pronounced, so the most appropriate study to do is to give repeated doses of the GCS until a new steady state has been reached and then measure P-cortisol for 24 hours. A typical schedule is every other hour. The most useful variable to study is the area under curve for those 24 hours. At steady state, when there is a 24-hour periodicity, this is proportional to the amount produced during 24 hours.

The actual clinical consequences of the levels attained are very hard to assess. They are surrogate measures of the long-term effects, but as such they should provide useful relative information on different GCSs, though we can never expect effects on P-cortisol to be a perfect predictor of, say, relative risk of osteoporosis. The distribution pattern in the body of the GCS might be of some consequence for this.

Example: Comparison of Plasma Cortisol Dose–Response Curves

We want to compare two inhaled steroids (with inhalation devices), call them A and B, in terms of their degree of suppression on the plasma cortisol level. The comparison is done in healthy volunteers in a randomised, open, seven-way crossover study. Each treatment period consists of four days, and there was a wash-out period of at least two days between each such treatment period. Each steroid was given in three doses: 200, 400 and 1000 μg b.i.d. for A and and 200, 375 and 1000 μg b.i.d. for B. The seventh treatment was a placebo treatment. Blood samples were measured every second hour during the last 24 hours in each treatment period (10 p.m. to 10 p.m.). In all 21 healthy volunteers participated in the study and all completed all treatment periods.

The effect of the fact that the study is open is hard to assess. If the administration of one of the drugs is associated with more stress than the administration of the other, this might bias the result. However, this seems unlikely, and doing the study as open has the benefit that fewer inhalations are required on each occasion.

To analyse the study, we first do an ANOVA. It is done on the logarithm of the concentrations with standard factors for a crossover study: subject, treatment and period. As a first presentation of the results we can compare all active treatments to placebo:

Treatment	Mean ratio(%)	95% conf. limits	P
200 μg of A	97.4	73.3, 129.3	0.85
400 μg of A	94.1	70.9, 125.0	0.67
1000 μg of A	70.0	52.7, 92.9	0.014
200 μg of B	76.4	57.6, 101.5	0.063
375 μg of B	53.5	40.3, 71.1	0.00003
1000 μg of B	9.1	6.9, 12.1	<0.00001

Here the mean ratio is presented as a percentage. For instance, 76.4% for 200 μg of drug B means that there is a suppression of $100 - 76.4\% = 23.6\%$.

This result does not tell us much about how the drugs compare. To do that we can fit parallel non-linear dose–response curves to the mean effect data, adjusting for precision by using a weighted nonlinear regression.[31] We assume for this analysis that, given enough steroids, the cortisol levels go down to zero. The result is graphically shown in Figure 27.4.

With the appropriate parametrisation here, we obtain that the relative dose potency is estimated to 3.7, with 95% confidence limits 2.7 and 6.4. Thus, in terms of depressing cortisol levels, B is estimated to be about four times more potent than A (remember that a letter stands for a GCS plus a device–change the device and this relation might change). Or put in other words: to achieve the same average depression in cortisol, we can use a four times larger dose of A than of B.

Having obtained this result, the immediate question is: ‘how relevant is this result to the target group of the drug–the asthmatic patient?’ We cannot extrapolate these results to patients ‘as is’. There is a basic difference between a healthy volunteer and an asthmatic: the latter has an ongoing inflammatory process. This means

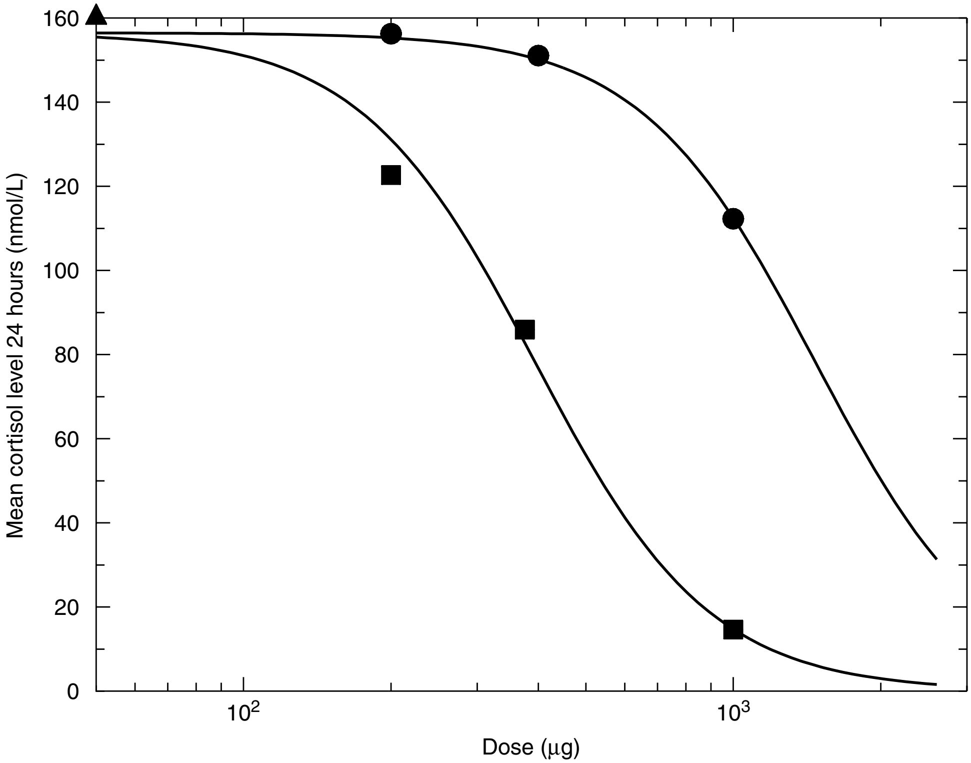


Figure 27.4. Estimated dose–response mean value curves for treatments *A* (to the right) and *B* (to the left).

that the dynamic system regulating cortisol is disturbed (compared with healthy) and we can expect smaller absolute effects of a given dose of the GCS. So a typical patient might have a larger ED_{50} than a typical healthy volunteer. By the same token we can expect different patients to vary considerably in this respect.

However, there is no reason to expect that different GCSs should act differently in patients and in healthy volunteers, i.e. there is no reason to claim that the relative effect of two GCSs, as measured by the potency ratio, ρ, should differ between patients and healthy volunteers, or between different categories of patients for that matter. If such differences turn out to be the case, the reason must be that the systemic dose differs between healthy volunteers and patients, and then, most likely, between patients of different degrees of severity in their disease.

PHASE III STUDIES

DOCUMENTING EFFICACY AND SAFETY

Most drugs for obstructive airway diseases are given for maintenance treatment, and the main point to document is the level of disease control of the proposed treatment. At the same time it is important to document the adverse event profile, since most of the drugs in this therapeutic area are considered very safe, and safety must not be an issue with a new drug. Thus the pivotal confirmatory trials for the airway diseases asthma, COPD and rhinitis are parallel group, diary card studies typically spanning from a month up to a year.

Asthma Trials

For asthma the typical study length for an efficacy study seems to be three months, though occasionally longer studies are needed. As already discussed, there is a continuous scale of severity for each of the respiratory diseases. The severity of asthma can be classified into a few groups, each of which has its recommended medical treatment. Both the classification and the recommended treatment are, like the disease under study, somewhat varying with time. The following classification is meant only to be indicative of what type of criteria are used for such classification:

Intermittent: These patients have normal lung function between occasional exacerbations with symptoms at most once a week. FEV_1 in per cent of predicted normal should be $\geq 80\%$.

Mild persistent: Now symptoms appear weekly, but not many times a day and exacerbations may affect both activity and sleep. FEV_1 in per cent of predicted normal should be $\geq 80\%$.

Moderate persistent: Symptoms appear daily, and affect both activity and sleep. There are night-time symptoms weekly and FEV_1 is within 60–80% of predicted normal.

Severe persistent: These patients have continuous symptoms, frequent exacerbations and night-time symptoms, which limit physical activity. FEV_1 in per cent of predicted normal is $\leq 60\%$.

This classification borrows from the GINA classification.[32] However, that classification also uses a concept of variability which we do not discuss.

To describe the patient's disease severity in a clinical trial we use data obtained at a visit to the clinic before randomisation. In addition, a diary card is provided for a run-in period to assess symptoms and use of rescue medication. Inclusion in a study is often based on these measurements. A more practically oriented classification of the severity of asthma is based on the use of GCS in the patient's regular treatment: no GCS (intermittent–mild), inhaled GCS up to 400 μg/day (mild persistent), inhaled GCS in the range 400–1000 μg/day (moderate) and inhaled GCS $\geq$ 1000 μg/day (severe). So the daily dose of background GCS treatment can be used as an indicator of asthma severity. Another classification is based on PC_{20}.[33]

The best way to use the information in diary cards might vary between patient groups. As already explained, the traditional use of diary card data is to assess changes in period means. This is expected to work best in patients with moderate asthma. In the intermittent group one should not expect effects of any considerable magnitude because the lung function is close to normal, and the patient is, for most of the time, symptom free. In the group of the most severe patients, patients often have obtained an irreversible component to the disease and therefore show little improvement in lung function. Instead the focus for studies in severe patients might be on how a new treatment can substitute for an old treatment without compromising the patient, e.g. how much oral steroids can be spared by taking inhaled GCS, or how much inhaled steroids can be spared when taking a concomitant leukotriene modifier.

The main objective of the confirmatory trials is to show efficacy, and thus requires a placebo control. This was discussed earlier. Concerning doses, for the majority of drugs for airway diseases, there is not one dose that is appropriate for all patients. Instead a range of doses has to be justified and documented in the clinical programme. The general discussion on MED and dose–response earlier is relevant here too.

Example: A Diary Card Study with Fixed Treatment Arms

The main outcome variable in a diary card trial with fixed treatment arms is some kind of summary measure (typically a mean value) over a longer period, presumed to represent, at a group level, for most part a steady-state situation. We also need a corresponding measure during the baseline/run-in period for the statistical analysis.

Figure 27.5 shows the estimated daily means of a three-month clinical trial in asthma. The study was a multicentre, double-blind, double-dummy study of three months, treatment with investigational drugs. There were 52 centres

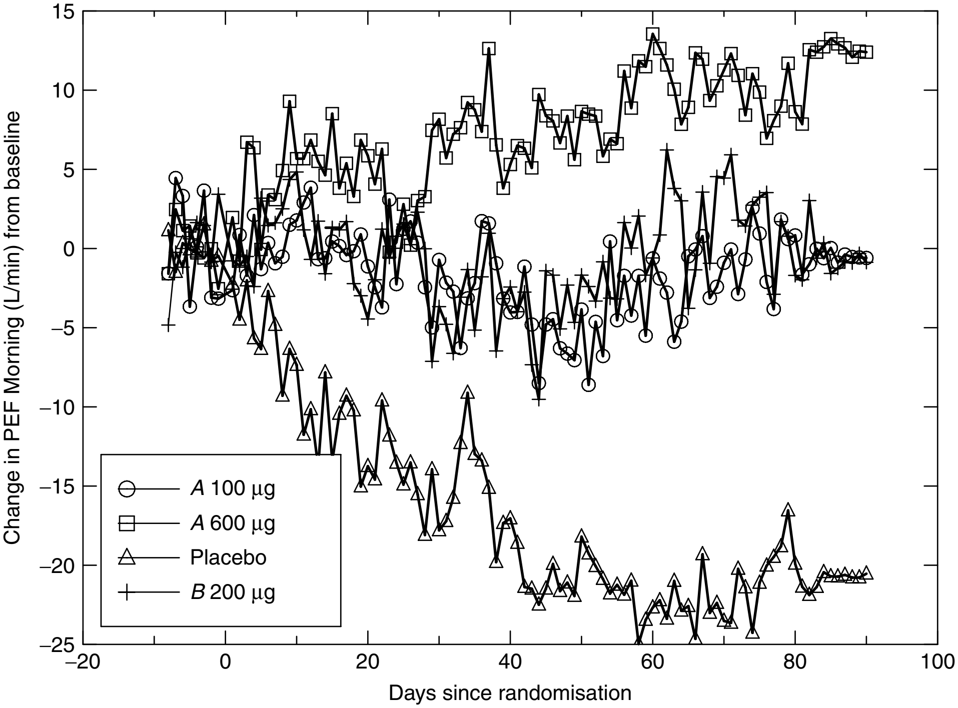


Figure 27.5. Estimated daily mean values of the change from baseline in PEF morning. See text for computational details.

in seven different countries worldwide and the randomised treatments were placebo, two dose levels, 100 and 600 µg daily dose, of a GCS *A* and one dose level, a daily dose of 200 µg of the GCS *B*. In all 547 patients were enrolled, of which 472 were randomised and 383 completed the study. There were twice as many withdrawals in the placebo group than in the other groups. The mean age was 44 years, the mean FEV_1 in per cent of predicted normal was 70% and the mean reversibility was 24%. All patients were on inhaled steroids when entering the study – the mean daily dose was 850 µg ranging from 500 to 1500. Thus the population must be characterised as being moderate–severe.

To plot the temporal behaviour of the effect of the four treatments, simple mean values are expected to produce a bias towards no effect. At least when comparison is done with placebo, in this group there were more withdrawals and many of these can be expected to be due to reasons that are correlated to (lack of) treatment effects. Some weeks into the treatment period, raw mean values for this group will therefore mainly include patients that are not in desperate need of GCS treatment. This will introduce a selection bias which will partly hide the effect of the treatments. To avoid this problem, that different days will contain different patients, we have preprocessed the data when plotting Figure 27.5 to make sure that all days contain data from all patients. The way this is done is as follows:

1. Linear interpolation is done in order to impute all missing values between the first and the last recorded day for each patient.
2. If the patient withdrew from the study before the 90th treatment day, the mean of the last three recorded days is extended from the last recorded day plus one to day 90 post-randomisation.

Then daily means are computed by treatment. In Figure 27.5 an additional operation has been made: for each treatment we compute a baseline by taking the average of the run-in means and subtract that from all daily means. This is done in order to highlight changes, since that is what the analysis focuses on.

The statistical analysis of the data uses the period means from the individual patients: the mean is computed for the run-in period, which is used as a baseline, and then for the treatment period (from the day of randomisation and onwards). The change from baseline is used as the effect variable, and the analysis is an ANOVA with treatments and centre as factors and baseline as the covariate. The following table shows the adjusted mean values for the effect variable for the four treatment groups, adjusted to a common baseline value (the mean over the full study population)

Treatment	Mean	SEM	95% confidence limits
Placebo	−10.5	3.3	−17.0, −4.0
A 100 μg	−0.3	3.3	−6.8, 6.2
A 600 μg	7.4	3.3	1.0, 13.9
B 200 μg	2.0	3.4	−4.7, 8.8

From this analysis we also find that there is a statistically significant, negative, dependence of the change in PEF to the baseline PEF and that the estimated residual standard deviation was 35.9 L/min. As is common for these kinds of data, the explanatory power of the analysis is small: only 8% of the variability is explained by the model.

Other diary card variables are often analysed the same way as was shown for PEF morning above, by first computing individual period means. Symptom scores and rescue medication are, however, variables for which the average value is not necessarily easy to interpret. Symptom scores are really ordered categorical data, and even though the average gives a hint on the amount of symptoms, it is not clear that, for example, $(2 + 2)/2$ means the same as $(1 + 3)/2$. For rescue medication, the problem is mainly that we use the mean as a measure of location for a distribution which, for some patients, might be skew. Also the distribution of period means over patients may well be skew.

For symptom scores and rescue medication it is therefore often useful to compute the percentage symptom-free days, or the percentage days with no rescue, instead of period means. At least the former is often for mild patients a more efficient measure than the corresponding period mean. And it is clinically easier to interpret. This idea can be carried one step further: we can introduce the concept of an ‘asthma-controlled day’, as one in which there are no asthma symptoms and no rescue medication was needed. The percentage of such days is often a useful variable, at least in studies on mild–moderate asthmatics. Modifications to a ‘well-controlled’ day for more severe patient populations is possible.

A Dose Reduction Study

To compare the efficacy of two GCSs, a randomised, double-blind, parallel group study with two treatment arms (one for each GCS) was designed. The objective is to estimate the relative dose potency by starting each arm on a high dose of the GCS, treat for some weeks, step down the dose and treat for some weeks, make a further step down, etc. This is done until the asthma is no longer controlled. This way we obtain for each patient the lowest dose on which the patient had asthma control (the one previous to the last one). This we call the MED (Minimal Effective Dose) for the patient.

In such a study the diary card variables per se are not of independent use; they should not be compared between groups, except possibly the data on the highest dose. Instead it is expected that the mean values are similar in the two arms over the treatment period; what varies is the underlying dose producing those effects. The effect variable of interest is the MED, which is to be compared between the groups.

The nature of MED is such that the best way of analysing it is not immediate. On one hand it will be rather discrete in nature, with only a few possible levels. On the other hand, the most informative way of expressing the result is to say how much more was needed on average for one as compared with the other (i.e. the MED for A was on average 125% that for B). We advocate for that reason that MED is analysed under a multiplicative ANOVA model, i.e. after log transformation of the dose, provided that MED > 0 in all cases. The most appropriate way to do this is to regard the data as interval censored data for the analysis.

One final comment on the design of long-term asthma trials: in many instances, especially for dose–response studies, it is informative for the interpretation of the results to relate the observed effects to the 'highest effects attainable'. This can be done within a study, so that the patients are put on a heavy treatment, consisting of a high dose of a GCS and a long-acting β_2-agonist (or whatever is considered necessary to get the patient in as good a condition as possible), during run-in, in a period after a run-in period or by adding on a period at the end of the study with a similar treatment. The purpose of this is to be able to quantify the response in terms of what can actually be achieved in the patients under study. If we put this reference period at the end of the study, we must make certain that all patients, including withdrawals, pass it in order to avoid having problems with a selection bias. If we put this reference period before randomisation, we might carry over effects into the randomised treatments with their potential problems. But having such a period as reference often helps in the interpretation of the results.

Rhinitis Trials

Classification of rhinitis patients into groups according to severity is lacking. The accepted division is between occasional and continuous expression of symptoms, i.e. between seasonal allergic rhinitis and perennial rhinitis. The rhinits symptoms are the same, so the measurements, notably symptom scores, are the same. As already indicated, symptoms are often recorded in diary cards for blockage, runny nose, sneezing/itchy and eye symptoms, and the sum of the first three makes up the Combined Nasal Symptom Score or Nasal Index Score, which is a useful primary variable for clinical trials.

The difference between seasonal and perennial rhinitis lies more in the study design/conduct. For perennial rhinitis the situation is similar to that for asthma or COPD in that the patient can start the trial at almost any time. For hay fever, however, the study must be conducted over a rather short period of pollen exposure. What makes these trials more difficult is that ideally the patients should be included during the onset of the pollen season to get baseline data, then followed over one to several pollen peaks with treatment. The intensity of the rhinitis is dependent on pollen counts in the air, and lack of treatment effects can well be due to insufficient pollen exposure. Therefore concomitant collection of pollen data is not only useful but almost necessary when trying to understand lack of effect in such studies.

COPD Studies

Most drugs for COPD are given for maintenance treatment, and the main point to document is the level of disease control of the proposed treatment. At the same time it is important to document the adverse event profile, since most of the drugs in this therapeutic area are considered very safe, and safety must not be an issue with a new drug. Thus the pivotal confirmatory trials for COPD are parallel group studies typically spanning from a month up to a year.

For COPD there are a number of staging systems for the severity of the disease, all of them based on FEV_1. In general COPD is classified as mild, moderate and severe disease, as in the following BTS classification:[34]

Mild: This is what is called the smoker's cough. The patient has $FEV_1 > 60\%$ of predicted normal, no breathlessness and is in general unknown to the health care system.

Moderate: The patient has breathlessness on exertion and FEV_1 in the range 40–60% of predicted normal.

Severe: The patient has breathlessness in everyday activities and $FEV_1 < 40\%$ of predicted normal.

What is to be proved in a clinical programme for a COPD drug depends on what the claim of the drug is. We can crudely divide effects on COPD into two groups: symptomatic effects and disease-modifying effects. The natural history of COPD is one of an accelerated progressive decline in lung function leading up to a distressful, premature, death. This decline in lung function leads to progressive symptoms and diminished exercise endurance. Symptomatic effects relate to the alleviation of symptoms and improvement of quality of life, whereas disease-modifying effects are effects that lessen the decline rate in lung function.

A drug with symptomatic effects on COPD should work on a rather short timescale. It should lead to improved symptoms, fewer exacerbations and better performance on exercise tests. Many drugs that were originally anti-asthma drugs have been tried, and licensed, for COPD indication. Part of their effect may well be due to the reversible component that many COPD patients have in their disease–in other words, an anti-asthma effect within the COPD. In order to claim effects upto and above this, studies have been performed in which one tries to exclude patients with reversible components by using an exclusion criterion on patients with a reversibility test above 15% of baseline. To claim that short-term effects seen in the population are due to non-anti-asthma effects because of such an exclusion criterion is obviously not correct–a patient with reversible airways obstruction can well have a reversibility of less than 15% on a particular occasion. Since COPD is a disease affecting small airways, it seems more logical to base short-term effects on measurements related to these airways, as opposed to FEV_1. However, regulatory requirements make FEV_1 the primary efficacy variable in COPD studies–at least as of this writing–though emphasis is put on the symptoms and/or exercise tests also. In fact the CPMP guidelines require two primary efficacy variables in COPD studies: one should be FEV_1 and the other a symptom score.[35]

A COPD study for a drug with primarily symptomatic effects is in general a long-term study, probably with diary cards. Prevention of exacerbations is perhaps the most important aspect of COPD treatment, so a six-month study is the minimum. The following example illustrates how exacerbation can be analysed and the points to consider when doing it.

Example: Analysis of COPD Exacerbations

A one-year multicentre, multi-country, double-blind, randomised clinical trial with two treatments, an active and placebo, was performed with the study of exacerbations as the primary objective. Exacerbations were defined in this study as either a hospitalisation and/or a course of oral steroid. For each exacerbation a starting date and an end date was identified. One way to describe the data is given in Figure 27.6. In the left panel we plot the cumulative mean (over number of patients) number of exacerbations up to a time point versus that time. We see there that the placebo group have more exacerbations than the active group. The right panel shows the difference between these curves, together with (pointwise and approximative) 95% confidence intervals. Already this simple description gives a good indication of a statistically significant difference in the mean number of exacerbations for the two treatments. Note that the computation of confidence limits needs to take into account that exacerbations within patients are dependent. A reasonable model for this dependency is that each patient has a constant risk of experiencing an exacerbation, but that this risk (rate of exacerbations) differs between patients, both because differences in disease severity and because the environment differ for patients.

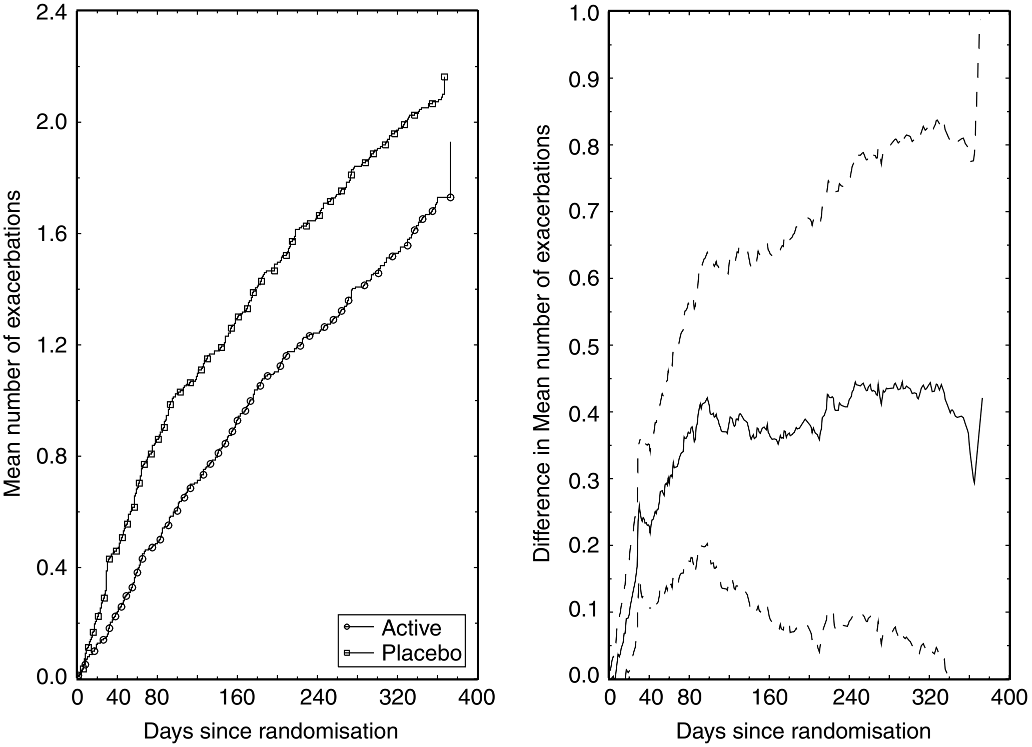


Figure 27.6. Mean cumulative number of exacerbations, and the mean difference with pointwise 95% confidence intervals.

The two classical modes of analysis are not to use all exacerbations in the analysis, but rather to summarise patient data into either time to first exacerbations (after randomisation) or total number of exacerbations. The former can be analysed with a Cox proportional hazards model, and for the latter we can assume that the number of exacerbations for a patient follows a Poisson process, but since the rate differs between patients we need to do the analysis using a model that allows for this heterogeneity (it shows up as an overdispersion if you run a Poisson model on the data). One such model is the negative binomial. In this analysis, we want to estimate the rate of exacerbations, so we need to use the logarithm of observational time as an offset in the analysis. The following table shows the treatment results for this study (using a model that allows for treatment and country as factors):

Variable	Hazard ratio	95% C.I.
Time to first event	0.71	0.55, 0.90
Total number	0.69	0.52, 0.91

These two analyses agree well: they both estimate that the rate of exacerbations is reduced (compared with placebo) to about 70% of the original one when treated with the active treatment, and in both cases we can claim that there is a reduction. For these data, if we analyse the total number of exacerbations with a Poisson model, we find an overdispersion of 2.9, strongly indicating that the heterogeneity we deduced above is present in data, and which motivates the use of the negative binomial instead.

We have, however, not used all information about exacerbations in this analysis. We know the start time and duration for each of these, and would like to incorporate that in the analysis. Further, we need to account for the different exacerbation rates between patients (as we did when analysing the total count) in this analysis, which means that, compared with using only time to first, including time to second, third, etc., adds much less to precision than if these times represented new patients. An analysis can be done as a Cox regression with (shared gamma) frailty, which is analogous to the negative binomial approach to total count data. The result is that the estimated hazard is 0.70 with 95% confidence limits 0.52, 0.96. For both this model and the negative binomial model, the estimated variance for the gamma distribution that describes heterogeneity among patients was 1.8.

A COPD drug which claims disease-modifying properties has a heavier burden of proof on it. The effect of disease modifying is that the rate of decline in lung function is reduced. To prove this, we need to do long-term studies over 3–5 years, or perhaps more, in which lung function is measured repeatedly. The statistical analysis should focus on the rate of decline, which could be done using a linear mixed effects model.

THERAPEUTIC EQUIVALENCE

One of the challenges for drug development is to prove that a new treatment is therapeutically equivalent to a reference treatment. In the area of respiratory diseases this problem appears in two different settings: when we want to register a new formulation, most often a new inhaler, and in market claims of equality of two treatments. The background and motivation of these differ somewhat, so we discuss them separately.

Bioequivalency of Two Devices

Bioequivalency refers to a specific problem. Assume that a drug is delivered as a tablet or in some other form, such that it must pass through the bloodstream before reaching its site of action. Then the plasma concentration profiles of the drug define the clinical effect. This reasoning is the rationale for the bioequivalence concept: to prove that two formulations are bioequivalent, we show that the plasma concentrations profiles are sufficiently similar. From that we can then logically infer that the therapeutic effects are sufficiently similar to have the same therapeutic effect. This is in general a rather straightforward problem, requiring only small pharmacokinetic studies.

For many years there has been a well-defined method to establish bioequivalency in this situation. We reduce the general question of similar plasma concentration curves to key measures of rate and extent of absorption, including AUC. The ratio of two AUCs measures the relative bioavailability of the two formulations and the requirement is that the ratio of the means (analysed under a multiplicative model) should have confidence limits within 80–125%.

For inhaled respiratory products, however, the site of action, the lungs, lay prior to the bloodstream (i.e. when the drug appears in the blood it has in general had its desired pharmacological effect). Thus plasma concentrations cannot predict the effect by pure logic! Equal delivered dose does not logically imply the same effect for different inhalers, because different inhalers could deposit the drug in different parts of the lungs. For that reason, to bridge from one inhalation device to another is not necessarily a simple case of measuring plasma concentrations. As of this writing there is substantial confusion on how to proceed with bioequivalence studies for inhalers. We will discuss some aspects of the problem here.

The first aspect is that what you inhale are particles, and these will be deposited differently depending on size. To give an equivalent effect, we therefore need an equivalent *in vitro* performance of the two inhalers. For the rest of the discussion we assume that *in vitro* data are similar for the two inhalers.

The basic assumption is that since what appears in the bloodstream does not have to have passed the site of action, systemic exposure, as measured by drug concentrations in the circulation, is not necessarily sufficient to conclude efficacy. However, similar systemic exposure should be sufficient to deduce similar systemic effects, and therefore reduce much of the question of bioequivalent side effects in the classical bioequivalence method.

Logically, if we measure blood concentrations and conclude that the systemic exposure is the same, the outstanding question is whether the drug has been delivered to the site of action. For a nasal spray it is hard to see how it can fail to do so. For a nasal spray, therefore, to require anything beyond *in vitro* data and pharmacokinetic data might be overkill. For an orally inhaled drug the situation is more complicated.

If we consider a bronchodilator as an example, we need the drug to hit the receptors of the contracted muscles. To check that the drug has hit them, we can do a pharmacodynamic study, e.g. a single dose study in which FEV_1 is followed for a number of hours, or a bronchoprovocation study if that is preferred. A suggested design is to study two or three doses of each inhaler, in order to see not only that the response is similar, but also that there is a similar sensitivity to changes in doses. These are, relatively speaking, simple studies to perform.

The next question is what should the decision rule be for bioequivalency for such a study, i.e. when are two inhalers considered to be similar? There seems to have been two approaches in use over the last few years. One is to use the word comparability. This is, for good reasons, not well defined and essentially means that there is a dose–response relationship on each device and that, numerically, the mean on each dose level is similar in the eye of the regulator. Thus there is no true statistical decision plan associated with the study and it is not used as proof per se, only as supportive information to what *in vitro* and pharmacokinetic data provide. The other approach is strictly statistical. In this case we should compute the relative dose potency with confidence limits. At present, in the case of bioequivalency for pMDIs for albuterol (salbutamol in the United States), the FDA requires that the 90% confidence limits for this parameter should be contained in the interval 2/3–1.5. The justification for these limits is not clear, but they imply that the mean effect is so similar for the two pMDIs that they could be switched on the market.

When it comes to deciding bioequivalency for orally inhaled anti-inflammatory drugs the problem is even more difficult, since there are, as of today, no designs available that can provide the kinds of answers that were discussed for bronchodilators, for a reasonable price. The problem is that in most studies the dose–response appears rather flat, so very large studies are needed for the type of decision plan that was indicated above.

Marketing Therapeutic Equivalence

The other aspect of therapeutic equivalence is to show that a new treatment is as effective as an old one, whereas it has some other benefits compared with it.

Proving that two treatments are equivalent has, however, a long history in the context of medicine. The traditional way was to misuse the P-value technology – if we could not demonstrate a difference ($P > 5\%$) the treatments were equal. This is obviously wrong (it is similar to 'not proven' in court, which means just that, not that one is proven innocent), which is by now acknowledged by most, but not all, workers in the field. A theoretically valid approach became legitimised by the ICH guidelines,[29] which define an algorithm borrowed from the original bioequivalency concept for plasma concentrations. First you prespecify some limits (corresponding to the 80–125% limits above) and if your 95% (*sic!*) confidence interval for the mean difference is contained within this prespecified interval, you can declare therapeutic equivalence. This approach is, however, complicated when your effect scale has no obvious interpretation, such as a lung function scale or a symptom scale. To be a sensible approach, the prespecified limits must imply that clinicians do consider such a small difference to be of no clinical consequence – the predetermined limits must be agreed on. It is hard to foresee that this can actually be done in the field of respiratory medicine, since many effect changes mean different things depending on what population is studied.

There is, however, a sensible, though expensive, approach available – the one used for demonstrating bioequivalency of bronchodilators. The key there is to translate from the effect scale to the dose scale, by studying dose–response. Almost all drugs in the respiratory area (though there are exceptions) are available in multiple doses, and when two treatments (drug plus inhaler) are to be compared, at least one of them can in general be varied on some kind of dose scale.

The simplest such design is as follows. To prove that dose a of a treatment A is therapeutically equivalent to a treatment B (dose need not be specified), we can study doses $a/2$ and $2a$ of A (not dose a!) and treatment B. By assuming a linear dose–response relationship (versus log dose) for A, we can estimate the dose of A that has the same effect as treatment B. Illustrations of this, with more than two doses of A, were presented earlier. Now, if the 90% confidence limit for the dose of A that has the same effect as treatment B lies between $a/2$ and $2a$ it is reasonable that dose a of A is equivalent from a clinical point of view to treatment B from an efficacy point of view. This is because half the dose has less effect and twice the dose more effect. Implicitly this assumes that the clinical response to a suboptimal dose is to double it, which is what is done with most drugs in the respiratory area. In general, to draw the conclusion of therapeutic equivalence, large studies might be needed.[36]

Often one tries to establish the therapeutic equivalence in one clinical study and compare benefits in the other. Basically I do not think this is the way to go about solving this kind of problem–in most cases it is probably a problem that should be discussed in terms of therapeutic ratio, as in the next section.

The Real Issue–The Therapeutic Ratio

Dose–response studies provides us, at best, with doses for further investigation. However, whether a drug ends up as being superior or inferior to what is on the market is not determined by what dose it is given in. What is the point in halving the nominal dose if you get twice as many adverse effects?

The appropriate measure here is the therapeutic ratio. The therapeutic ratio for a drug relates the positive effect to the negative effect. To understand this, assume that effects, both positive and negative, are measured on a scale from 0 to 100. Also assume, for the time being, that the dose–response curves for positive and negative effects are parallel. Then a therapeutic ratio of 2 for this drug means that twice as large a dose is needed to get the same negative effect as positive effect. Thus we can define the therapeutic ratio for drug A as $\text{TR}(A) = \text{ED}_{50}(\text{side effect})/\text{ED}_{50}(\text{effect})$. If we have two drugs, we can define the relative therapeutic index of A to B as $\text{TR} = \text{TR}(A)/\text{TR}(B)$, which is equivalent to the ratio $\rho(\text{effect})/\rho(\text{side effect})$, where ρ is the relative dose potency for the two drugs with respect to the indicated effect. This means that we can estimate not only the relative therapeutic index, but also confidence intervals to the estimate, by assessing the relative dose potencies. Moreover, we can define the therapeutic index as a ratio of relative dose potencies without assuming that the effect and the side-effect curves are parallel. To be meaningful it only requires that the dose–response curves for efficacy are parallel and those for the side effect are parallel. We can estimate the therapeutic index by combining effects from two studies, one on efficacy and one on side effect, but it is better still to obtain all the information in one study.

The first problem to solve is the precise definition of outcome variables, both positive and negative. Different results can be obtained by using different outcome variables. It is therefore important that the precise objective is spelt out and the outcome variables related to this. The following example illustrates this.

Example: Estimating the Relative Therapeutic Index

In order to assess the relative usefulness of a long-acting β_2-agonist, call it A, and a short-acting one, which we call B, we want to compare one topical effect, bronchodilation, and one systemic effect, suppression of serum potassium. The study was of crossover design with single-dose administrations and serial measurements of both variables were taken.[37] In order to be able to get meaningful estimates of the relative therapeutic index we need many patients, relatively speaking, and in order to obtain a simpler study we choose to measure the maximal effect on each parameter and compare it.

Thus a randomised, double-blind, six-period crossover study was designed with the following single dose treatments: placebo, 6 μg, 24 μg and 72 μg of drug *A* and 200 μg and 1800 μg of drug *B*. Each treatment period consisted of a single dose administration which was followed for four hours and from each experimental sequence the maximal FEV_1 value and the minimal S-potassium value were extracted for statistical analysis.

Figure 27.7 demonstrates the main result. We see (period and baseline-adjusted) treatment means together with straight line approximations to the dose–response curves for each variable. In order to make the results more interpretable, mean values (and lines) are expressed as a percentage of the mean value for placebo. From these straight lines we can estimate the relative dose potency, as discussed above, for each variable separately:

Variable	ρ	95% confidence limits
FEV_1	147	65, 534
S-potassium	60	41, 91

Thus we see that in terms of efficacy, the long-acting drug *A* is almost 150 times more efficient than the short-acting *B*. On the side-effect side, *A*

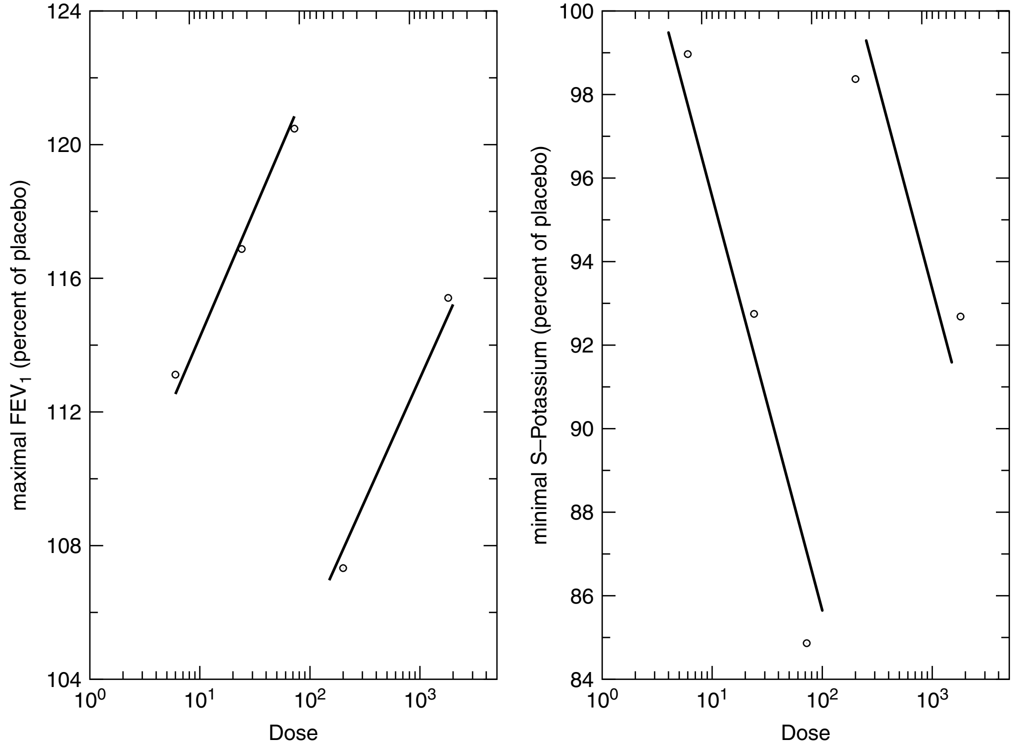


Figure 27.7. Adjusted mean values for each treatment and outcome variable.

is 60 times more potent, so from these data we see that the relative therapeutic index is estimated to $146.6/59.75 = 2.4$. To obtain confidence limits is somewhat involved since we need to take into account the covariation of the two variables. How to do this is outlined in Källén.[38] The result is that the relative therapeutic index is estimated to be 2.5 with (approximate) 95% confidence limits 1.02 and 9.0 ($p = 0.046$).

The conclusion from this is that in terms of the variables of this analysis treatment A is estimated to be 2.5 times 'better', but it is certified that it is 'better' than treatment B. Of course, this result is not better than the data. From Figure 27.7 we see that the lowest dose of each drug has a very small average effect on serum potassium. It could (and should) be questioned if these doses really are on the log–linear part of the dose–response curve. We can repeat the analysis by incorporating the doses of 24 and 72 μg of A and 1800 μg of B for serum potassium.

OTHER ISSUES

PHASE IV

Much Phase IV work focuses on comparisons with competitor products in order to demonstrate the advantages of the new product. This has been discussed in previous chapters and will not be repeated here. In addition, special safety issues might have to be addressed, which might call for large-scale studies in order to study some rare event.

PHARMACOECONOMICS

Asthma and COPD are costly illnesses–the costs of asthma can account for as much as 1–1.5% of all resources in the health care sector. The costs are, however, unevenly distributed among patients; it is not uncommon for 10% of the most severe asthmatics (usually patients with uncontrolled asthma) to account for over 50% of the total cost. There should thus be room for a significant cost reduction by improving disease control.

The costs can, basically, be divided into three categories (usually only the first, although sometimes also the second category is measured):

1 *Direct costs*, defined as health care resources consumed, include costs associated with drugs, devices, consultations with physicians, emergency room visits and hospitalisation.
2. *Indirect costs*, defined as lost productivity, include time off work or school, either patient or relative, premature retirement and death. Indirect costs may account for up to 50% of the cost of asthma.
3. *Intangible costs*, contain factors related to quality of life (grief, fear, unhappiness, pain, etc.).

Within the direct costs, drugs and general practitioner visits can, crudely, be considered costs of managing controlled asthma, whereas emergency room and hospital costs can be assumed to relate to treatment failure. Assuming this to be about 75% of the total cost, the major part of the costs of asthma appear to be a result of inadequately controlled disease. Thus, the goal is to get control of the asthma, which (for example) can be done by patient education[39] and prophylactic therapy. As an example of the latter, it has been demonstrated that the introduction of high-dose inhaled steroids in patients with severe asthma reduced the number of days of hospitalisation by 80%.[40] It should also be noted that a part of drug costs consists of rescue medication, like short-acting β_2-agonists, and is in itself a sign of the disease not being adequately under control, and that an increase in prophylactic therapy, like inhaled steroids, might decrease these costs.

So, the economics of asthma informs us that a large population of mild-to-moderate asthmatics has a low daily cost. For some of these patients, the disease becomes uncontrolled and costly, with hospitalisation and time off work or school, possibly progressing into early retirement or death. Some of these cases are probably due to bad compliance with treatment regimens. International guidelines therefore stress that prophylactic treatment should be introduced at an early stage in asthma treatment, resulting in an increase in drug and general practitioner costs, but hopefully leading to reduced hospitalisation costs and indirect costs. As an observation on this topic, it can be mentioned that the cost of the avoidance of one admission to hospital will pay for about three years of treatment with inhaled steroids.[11]

Apart from collecting data on costs, it is also important to measure the individual's quality of life and the effectiveness of various interventions (where effectiveness, as compared with efficacy, ideally should measure the effects of an intervention in clinical practice and also in units the patients care about). This is perhaps especially important if a new medical treatment increases the total costs, as it then becomes important to relate these extra costs to the additional effectiveness gained. Currently recommended effectiveness variables include the number of symptom- and episode-free days. In cost-effectiveness analyses, if both costs and effectiveness are higher for one of the treatments, the difference in costs is divided by the difference in effectiveness to obtain a cost-effectiveness ratio. This ratio is, for example, expressed as: compared with treatment *A*, treatment *B* costs \$*x* per symptom-free day gained. That ratio thus gives support in answering the question of whether the additional effectiveness (or quality of life or disease control) can justify the extra costs.

Furthermore, it is important to keep in mind that costs often are country-specific, e.g. in the case of an international clinical trial, and some adaptation is needed before translating results from one country to another. This is so because the outcome of a new treatment will depend on the local medical tradition, drug pricing and the unit costs of other health care resources, and a number of social conditions like the labour market. In general, though, as the interest for 'value for money' and cost-containment in the health care sector grows, the importance of these kinds of evaluations is likely to increase. From clinical trials it is thus important to measure these kinds of variables (both costs and effectiveness), which can then be used as the basis for a cost-effectiveness analysis.

REFERENCES

1. Beasley R. Worldwide variation in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis and atopic eczema. *Lancet* (1998) **351**: 1225–32.
2. Chadwick DJ, Cardew G. *The rising trends in asthma, Ciba Foundation*. Chichester: John Wiley & Sons, Ltd (1997).
3. Peat J, Li J. Reversing the trend: reducing the prevalence of asthma. *J Allergy Clin Immunol* (1999) **103**: 1–10.
4. Sears MR. Epidemiology of childhood asthma. *Lancet* (1997) **350**: 1015–20.
5. Barnes PJ. The changing face of asthma. *Q J Med* (1987) **63**: 359–65.
6. Lange P, Parner J, Vestbo J, Schnohr P, Jensen G. A 15-year follow-up study of ventilatory function in adults with asthma. *New Engl J Med* (1998) **339**: 1194–1200.
7. Sherman CB. The health consequences of cigarette smoking. Pulmonary diseases. *Med Clin N Am* (1992) **2**: 355–75.
8. Buist SA. Smoking and other risk factors. In: Murray JF, Nadel JA, eds, *Textbook of Respiratory Medicine*, 2nd edn. Philadelphia: WB Saunders (1994) 1259–87.
9. Selroos O, Pietinalho A, Riska H. Delivery devices for inhaled asthma medication Clinical implications of differences in effectiveness. *Clin Immunother* (1996) **6**: 273–99.
10. Selroos O. Bronchial asthma, chronic bronchitis and pulmonary parenchymal diseases. In: Morén F, Dolovich MB, Newhouse MT, Newman SP, eds, *Aerosols in Medicine. Principles, diagnosis and therapy*, 2nd edn. Amsterdam: Elsevier Science (1993) 261–89.
11. Bartnes PJ. Beta-adrenergic receptors and their regulation. *Am J Respir Crit Care Med* (1995) **152**: 838–60.
12. Löfdahl CG, Chung KF. Long-acting β_2-adrenoceptor agonists: a new perspective in the treatment of asthma. *Eur Respir J* (1991) **4**: 218–24.
13. Rall TW. Drugs used in the treatment of asthma: the methylxanthines, cromolyn sodium, and other agents. In: Gilman AG, Rall TW, Nies AS, Taylor P, eds, *Goodman and Gilman's The pharmacological basis of therapeutics*, 8th edn. New York, Pergamon Press (1990) 618–37.
14. Gandevia B. Historical review of the use of parasympatholytic agents in the treatment of respiratory disorders. *Postgrad Med J* (1975) **51**(Suppl 7): 13–20.
15. Brattsand R, Selroos O. Glucocorticosteroids. In: Page CP, Metzger WJ, eds, *Drugs and the Lung*. New York: Raven Press (1994) 101–220.
16. The British guidelines on asthma management. *Thorax* (1997) **52**(Suppl 1): S1–21.
17. Howarth PH. The medical treatment of chronic rhinitis. In: Busse WW, Holgate ST, eds, *Asthma and Rhinitis*. Boston: Blackwell Scientific (1995) 1415–28.

18. Simons FER. Antihistamines. In: Mygind N, Naclerio R, eds, *Allergic and non-allergic rhinitis. Clinical Aspects*. Copenhagen: Munksgaard (1993) 123–36.
19. Bernstein JA, Bernstein IL. Cromones. In: Barnes PJ, Grunstein MM, Leff AR, Woolcock AJ, *Asthma*. Philadelphia: Lippincott–Raven (1997) 1647–66.
20. Barnes NC. Effects of antileukotrienes in the treatment of asthma. *Am J Respir Crit Care Med* (2000) **161**: S73–6.
21. Quanjer PH, Tammeling GJ, Cotes JE, Pedersen OF, Peslin R, Yernault JC. TI: lung volumes and forced ventilatory flows. Report Working Party: Standardization of Lung Function Tests. European Community for Steel and Coal. *Eur Respir J* (1993) **6**(Suppl 16): 5–40.
22. ATS guidelines.
23. Oga T, Nishimura K, Tsukino M, Hajiro T, Ikeda A, Izumi T. The effects of oxitropium bromide on exercise performance in patients with stable chronic obstructive pulmonary disease. A comparison of three different exercise tests. *Am J Respir Crit Care Med* (2000) **161**: 1897–901.
24. Borg GAV. Psychophysical bases of perceived exertion. *Med Sci Sports Exercise* (1982) **14**: 377–81.
25. Mahler DA, Wells CK, Feinstein AR. The measurement of dyspnea. Contents, interobserver agreement, and physiologic correlates of two new clinical indexes. *Chest* (1984) **85**: 751–8.
26. Berger M *et al*. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* (1981) **19**: 787–805.
27. Stewart AL, Hays RD, Ware JE Jr. The MOS short-form general health survey. Reliability and validity in a patient population. *Med Care* (1988) **26**: 724–35.
28. Jones PW, Quirk FH, Baveystock CM. The St George's Resp questionnaire. *Resp Med* (1991) **85**: 25–31.
29. ICH Harmonised Tripartite Guideline. Statistical principles in clinical trials. *Stat Med* (1999) **18**: 1905–42.
30. Senn S. *Statistical Issues in Drug Development*, Statistics in Practice. Chichester: John Wiley & Sons, Ltd (1997).
31. Källén A, Larsson P. Dose response studies: how do we make them conclusive? *Stat Med* (1999) **18**: 629–41.
32. Global Initiative for Asthma. National Institutes of Health. National Heart, Lung, and Blood Institute Publication 95-3659 (1995).
33. Woolcock AJ. Therapies to control the airway inflammation of asthma. *Eur J Respir Dis* (1986) **69**(Suppl 147): 166–174.
34. BTS guidelines for the management of chronic obstructive pulmonary disease. *Thorax* (1997) **52**(Suppl 5): S1–28.
35. Committee for proprietary medicinal products (CPMP). Points to consider on clinical investigation of medicinal products in the treatment of patients with chronic obstructive pulmonary disease (COPD). London: EMEA (1998).
36. Källén A, Larsson P. On the definition of therapeutic equivalence. *Drug Inf J* (2000) **34**: 349–54.
37. Rott Z, Böcskei C, Poszi M, Juhasz G, Larsson P, Rosenborg J. On the relative therapeutic index between formoterol Turbuhaler and salbutamol pressurized metered dose (pMDI) inhaler in asthmatic patients. *Eur Respir J* (1998) **12**(Suppl 28): 324s.
38. Källén A. A note on therapeutic equivalence and therapeutic ratio with application to studies in respiratory diseases. *Drug Inf J* (2001) **35**: 1495–1505.
39. Liljas B, Lahdensuo A. Is asthma self-management cost-effective? *Patient Educ Counsel* (1997) **33**: 97–104.
40. Adelroth E, Thompson S. Advantages of high-dose inhaled budesonide. *Lancet* (1988) **i**: 476.

SURGERY & ANAESTHESIA

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

28

Anaesthesia and Pain

PAUL S. MYLES[1] AND JOHN RIGG[2]
[1]Department of Anaesthesia & Pain Management, Alfred Hospital, Melbourne, Australia
[2]Department of Anaesthesia & Perioperative Medicine, Armadale Health Service,
Armadale, Western Australia, Australia

INTRODUCTION

Modern anaesthesia began in the 1840s, when dentists, surgeons and other practitioners became aware of the pain-relieving and soporific properties of nitrous oxide ('laughing gas'), ether and chloroform. The era of painless surgery had begun.[1,2] Early anaesthesia practitioners had limited knowledge of physiology, physics or chemistry, and so learned by trial and error. Some, such as the evolving epidemiologist John Snow, had a special interest in the basic sciences and applied such principles to their clinical practice. Clinical observation, experience and documentation of successful and difficult cases led to a body of knowledge based on anecdote, selective case series and personal stature within various medical communities. Such knowledge was often presented and discussed at local meetings, and some published.[2] Early anaesthesia was dominated by inhalational anaesthesia, with some questioning of variations in delivery devices, and, later, choice of vapour: ether or chloroform.[2,3]

Intravenous anaesthetic drugs were introduced in the early 1920s, and muscle relaxants in the late 1930s. The latter development allowed a reduction in the need for 'deep' anaesthesia, but mandated techniques of tracheal intubation and artificial ventilation. A vast array of new, potent, inhalational and intravenous anaesthetic drugs and other muscle relaxants followed.

Decisions regarding choice of anaesthetic technique, drug combinations, use of various airway devices, intravenous fluid therapy, and equipment, most often were guided by clinical experience, observation and, in some cases, effective marketing by industry.

The specialty of anaesthesia has always had an emphasis on patient monitoring, with ongoing development of equipment to deliver anaesthetic gases and oxygen in a safe and dependable manner. The close proximity of the anaesthetist to the patient, and the immediacy of clinical responses to changes initiated by drug administration, instilled an appreciation of monitoring and safety. Clinical experience and suspected adverse consequences associated with anaesthesia led to a number of large-scale investigations.[3–5] These were among the first of many subsequent large cohort studies designed to monitor adverse effects

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

and mortality associated with anaesthesia, including specific committees that deal with deaths under anaesthesia.[6,7] Extrapolation of the recommendations of these committees to other countries is constrained by differences in reporting, extent of specialist training, use of nurse anaesthesia and standards of monitoring.

In the late nineteenth century suspected excess mortality associated with chloroform anaesthesia led to a number of official investigations – the Hyderabad Chloroform Commissions.[3] Perhaps the most influential survey of deaths in the operating theatre occurred with a large multicentre survey, conducted by Beecher and Todd from 1948 to 1952.[4] Nearly 600 000 anaesthetics were administered in 10 surveyed hospitals in the United States, with an overall anaesthetic mortality of about 1:1500. One of the strongest risk factors was inclusion of the new muscle relaxant drug, curare. In the next decade the National Halothane Study, conducted in 34 hospitals in the United States and including more than 850 000 anaesthetics, found that halothane hepatitis had an incidence of 1:10 000.[5] Repeat administrations of halothane were a clear risk factor. These were the first of many subsequent large cohort studies designed to monitor adverse effects and mortality associated with anaesthesia

In Australia, a committee investigating deaths under anaesthesia reviewed all deaths associated with anaesthesia in the state of New South Wales, from 1984 to 1990.[6] The committee reviewed 1503 deaths from 3.5 million surgical procedures. In 60%, death was considered to be inevitable and in 4%, fortuitous. Factors under the control of the anaesthetist contributed to the fatal outcome in 11% of cases ($n = 172$). Factors under the control of the surgeon contributed to the fatal outcome in 28% of cases. In this period, deaths in which factors under the control of the anaesthetist contributed to the fatal outcome occurred at a rate of 1 in 20 000 operations. More recent figures suggest a reduction in anaesthetic mortality, with an incidence of about 1 in 80 000.[7,8]

Clearly, improvements in care are difficult to identify in such retrospective analyses, but anaesthetic and other interventions are known to reduce surgery-related complications or those secondary to pre-existing patient co-morbidity.[9]

Increasingly complex surgical procedures, performed more frequently in older and sicker patients, demand more active clinical interventions by contemporary anaesthetists. Risks of adverse outcomes are greater, and new drugs and technologies introduce the possibility of additional complications. Contemporary anaesthetic practice extends beyond the operating theatre and intensive care unit to include pain medicine, perioperative medicine including pre-anaesthesia assessment and possible modification of therapy, and simulator training.

CLINICAL RESEARCH METHODS

Anaesthesia research, because of its reliance on drug administration and equipment function, has focused mostly on an understanding of drug pharmacology, respiratory and cardiovascular physiology, the metabolic consequences of anaesthesia and surgery, and patient monitoring and safety. Laboratory studies, elucidation of mechanisms of action, and small clinical studies have dominated the anaesthetic research agenda. In addition, large audits and other surveys have documented current practices, and assisted the identification of important clinical features associated with anaesthesia and surgery. Case reports have been useful in some circumstances. Such observational research methods have their strengths, but also have important limitations.[10,11]

The randomised controlled trial is universally accepted as being the best method of testing drug efficacy, because the randomisation process minimises bias and enhances the reliability and validity of results and conclusions.[11,12–16] As in other areas of medicine, the clinical trial has been shown to provide the most reliable evidence of benefit (or harm). But many trials are too small to detect potentially useful treatments;[17,18] these, and observational studies, may miss an opportunity to identify clinically meaningful effects that can be more reliably identified with large randomised trials.[19–23] Thus the hierarchy of evidence-based medicine is very relevant to anaesthesia and pain research and its clinical applications.[9,11]

Outcome variables used in anaesthesia research range from surrogate measures of true outcomes (disability or death), to pain control, nausea, vomiting, quality of recovery, patient satisfaction and hospital length of stay. Because anaesthesia has a strong record of safety, with very low rates of true morbidity and mortality, such low event rates have important implications for trial design. Because most improvements in anaesthesia are modest and incremental, large numbers of patients need to be studied in order to have adequate statistical power to detect a clinically important difference.[12–16] Consequently, anaesthesia trials are often designed to focus on surrogate measures; these endpoints occur more frequently and so lower sample sizes can provide adequate statistical power in these circumstances.

SURROGATE OUTCOME MEASURES

The use of surrogate, or intermediate, outcome measures in anaesthesia is widespread.[11,24–26] An example of this approach is studies using measures of myocardial ischaemia as a primary endpoint.[27–30] Perioperative myocardial ischaemia can be measured with ST-segment analysis of the ECG; this is very simple using modern anaesthetic monitors that have in-built automated software. A more useful approach is to use Holter monitoring, which can be continued into the postoperative period when most episodes of ischaemia occur.[27] Other investigators have used the pulmonary artery catheter,[28] intraoperative transesophageal echocardiography,[29] and troponin flux.[30] Very few anaesthesia studies have used myocardial infarction as a primary endpoint, presumably because its incidence is very low (<2%). Such studies would require a sample size greater than 2000 to provide sufficient statistical power and reliable estimates of effect.

Other common surrogate endpoints in anaesthesia trials include blood pressure or hypotension (for organ failure), cardiac output (for cardiac failure), urine flow, serum creatinine or creatinine clearance (for renal failure), arterial blood gases or pulmonary mechanics (for respiratory failure), arrhythmias (for cardiac arrest), carotid Doppler flow (for stroke), cognitive deficit (for brain injury and dementia), and analgesic medication usage or pain scores (for unrelieved pain).

MECHANISTIC STUDIES AND ANAESTHETIC DRUG POTENCY

Since the introduction of general anaesthesia in the 1840s, clinicians have been interested in techniques to administer the variety of drugs used.[2] Physiology, pharmacology and physics have remained key basic sciences. The immediate effects observed during administration of anaesthesia – pain relief, loss of consciousness, loss of breathing control, muscle paralysis – lead to questioning as to how such drugs work, and how they can be more safely administered. New developments are tested in the laboratory, and are then applied to clinical settings. The focus of attention remains mechanisms of action. Clinical outcomes are of secondary importance. Such mechanistic studies have been the foundation of anaesthesia research, and continue to inform current research practices.

The site of action of general anaesthetics is poorly understood.[31] Animal and other laboratory studies have identified the γ-amino-butyric acid ($GABA_A$) receptor as a key factor, but although most classical general anaesthetics produce hypnosis (unconsciousness) and amnesia and suppress motor responses to noxious stimuli, their actions on other physiological responses vary. The suppression of nociceptive motor responses is mediated primarily within the spinal cord, whereas hypnosis and amnesia are mediated within the brain.[31] Thus a definition of depth of anaesthesia is problematic, with anaesthetic potency traditionally defined in terms of nociceptive motor responses.[32] However, recent advances in processed EEG monitoring have led to new methods of quantifying hypnotic depth.[23] This new field, which has seen the introduction of numerous technologies in recent years, has been associated with increased research focused on responses and recovery from anaesthesia, and avoidance of awareness.

OUTCOME STUDIES IN ANAESTHESIA

Anaesthesia has a limited causative role in most serious outcomes after surgery.[24,26] For this reason many anaesthesia outcome studies focus on early recovery characteristics and measures of comfort (relief from pain, nausea and vomiting).[26] Nevertheless, there have been some large trials in anaesthesia that have focused on major morbidity and mortality.[20–23,33–36] Some of these are described in more detail later in this chapter.

INFORMED CONSENT AND ETHICS

The ethical principles underlying randomised trials include autonomy, equipoise and beneficence. Autonomy requires a clear understanding of the purpose of the study, the possible risks and benefits, and the ability of the subject to choose whether to participate or not. Potential subjects should be allowed to consider their participation in a comfortable environment, with enough time to weigh up options and have any discussion with relatives or friends. These requirements can be challenged in anaesthesia or pain research, where the unfamiliar surroundings of a hospital, the anxiety of surgery and anaesthesia, painful conditions, dependence on clinicians (who may also be the researcher), and the limited time available before surgery limit the consent environment. Another difficult area is the enrolment of critically ill patients in the emergency department or intensive care unit.

Equipoise infers that the clinician and patient have no particular preference or reason to favour one treatment over another.[37,38] Newly developed drugs and equipment offer promise, and so both patients and clinicians may believe they are superior to established therapies. The conflicting roles of the anaesthetist, as researcher and clinician, are sometimes difficult to resolve in this situation.[31–43] The patient must also provide informed consent freely, a process requiring adequate disclosure of information, competency and understanding, and self-determination.[37–41,44]

Patients approached before elective surgery are often anxious, and feel vulnerable, and this can restrict their ability to process new information and so make a truly informed decision about participation.[37,38] Prerandomisation is a process whereby the patient is first randomised to a particular treatment group and then approached (with knowledge of his or her treatment allocation) for informed consent.[45] Zelen originally proposed this method to exclude patient notification and consent if the patient had been allocated to standard treatment.[45] In view of the unsatisfactory conditions that frequently exist for provision of informed consent before surgery, this approach may be useful in some clinical settings.

Examples of prerandomisation in anaesthesia and critical care trials can be found in the literature.[39,46] Truog has discussed some of these issues with reference to a recent trial of extracorporeal membrane oxygenation in neonatal respiratory failure.[47] Here, 39 neonates with severe pulmonary hypertension and respiratory failure were randomly allocated to extracorporeal membrane oxygenation (ECMO) or conventional medical therapy (CMT, $n = 35$). In the first phase of the study, 4 of 10 babies in the CMT group died and 9 of 9 babies in the ECMO group survived. Randomisation was halted after the fourth CMT death, as planned before initiating the study, and the next 20 babies were treated with ECMO (Phase II). Of the 20, 19 survived. The overall survival of ECMO-treated infants was 97% (28 of 29) compared with 60% (6 of 10) in the conventional treatment group, $P = 0.01$.

Myles *et al.*[39] investigated five alternative randomisation schedules in 770 patients in the immediate preoperative period and measured the resultant recruitment rates. Methods studied included one-sided informed consent, prerandomisation and the consumer principle (where the doctor or patient could modify the randomisation ratio to more than or less than 1:1). The overall recruitment rate was about 55%. Alternative methods of randomisation did not result in any significant difference in recruitment rates, and so offer little or no advantage over conventional randomisation and informed consent in this study. Interestingly, the preoperative anxiety level of the patient, as measured by a 100 mm analogue scale, did not affect recruitment rates, consenters 36^{27} mm vs. non-consenters 40^{30} mm, $P = 0.21$. Factors associated with improved recruitment rates included older patient age, English as a first language and staff gender. Female research staff had lower rates of patient recruitment when compared with males.[39]

There is concern that the environment in which consent for anaesthesia research is sought may be coercive. Tait *et al.*[43] highlighted some of these issues in a survey of parents considering recruitment of their child into a clinical trial. They studied the factors that parents consider in consenting to their child's participation in anaesthesia research. The study sample consisted of 246 parents who had been approached for permission to allow their child to participate in a clinical anaesthesia study. Parents were asked to complete a questionnaire detailing the reasons for their decision to consent or decline their child's participation. Perceived risk and the importance of the study were the primary factors in the parents' decisions to consent or decline. Only 2.8% of non-consenters strongly considered a lack of privacy as a deciding factor; 15% stated that they had insufficient time in which to make a decision, and none felt pressured. Furthermore, only 3% of consenters strongly considered an obligation to consent. This study, supported by other research,[39,41,42] suggests that the hospital setting and preoperative period are generally acceptable to patients approached to consider participation in perioperative clinical studies.

Other options for obtaining patient consent include surrogate (or next-of-kin) consent, especially with emergency conditions, critical illness, or with children. A common approach used to deal with these challenging situations is to use deferred consent,[48] where the patient is enrolled and studied, and approached at a later time during his or her their recovery for consent. This introduces the possibility of survivor bias, in that those who have made a good recovery may be more likely to consent, and crucial information about major complications or deaths may be missed.

MEASUREMENT OF PAIN

Clinical practice guidelines recommend frequent measurement of pain intensity in order to optimise treatment.[49] Despite pain being acknowledged as a multidimensional experience, it is common for it to be assessed with one of several unidimensional scales.[50,51] The visual analogue scale (VAS) has become a standard measurement tool in both pain research and clinical practice,[50–64] but is also used to measure other subjective experiences such as preoperative anxiety, postoperative nausea, and patient satisfaction after surgery and anaesthesia.

Pain intensity is classically rated on a 100 mm horizontal line, and this is then measured from the left boundary as a VAS score in millimetres – see Figure 28.1. The VAS score has been shown to correlate well with acute pain levels,[50–52,56,59–62] although it is recognised that it has a measurement error of about 15 to 20 mm.[52,59] Despite this, the simplicity for patients, clinicians and researchers, as well as the reliability and reproducibility of the VAS, make it a widely used and valid form of pain measurement.

There has been some controversy in the literature regarding which statistical tests should be used when analysing VAS data.[53,55,56] Most researchers treat VAS scores as numerical data.[53,55,64] Mantha *et al.*[55] surveyed the anaesthetic literature and found that approximately 50% of studies used parametric tests. Dexter and Chestnut[53] used a multiple resampling (of VAS data) method to demonstrate that parametric tests had the greater power to detect differences among groups. Myles *et al.*[56] have demonstrated that the

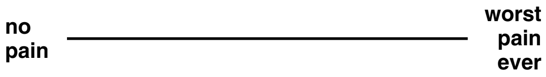


Figure 28.1. A 100 mm visual analogue scale (VAS) for pain. The subject is asked to rate his or her pain intensity with a mark at any point along this line. The pain VAS score is calculated by measuring the distance from the left edge (in mm). Severe pain is generally accepted if VAS > 70 mm.

VAS has linear scale properties in patients with mild, moderate and severe pain after surgery, and concluded that the VAS score can be considered as ratio data for statistical analysis and interpretation. Thus, a change in the VAS score represents a relative change in the magnitude of pain intensity for all patients with acute pain. Thus, if a treatment leads to a change in VAS score from 60 mm to 30 mm, then this represents a 50% reduction in pain intensity.

Previous authors[57] have suggested that a 50% reduction in baseline VAS represents a meaningful analgesic effect. Others have found that a change in VAS of about 20 mm represented satisfactory pain relief in patients after surgery,[52] or that the minimal change in acute postoperative pain rated on an 11-point scale that is clinically significant is 20%.[58] Others have studied the relationship between the pain VAS and morphine requirements in patients after surgery.[55,56]

A pain VAS score is an indirect, surrogate measure of pain. It is unclear what value or change in VAS determines what is an acceptable level of pain to patients.[51,52,57,58] The importance of pain control to patients probably differs with that presumed by the doctors and nurses caring for them. One study surveying the perceptions of patients and their family members, nurses and doctors found significant differences in many items defining the postoperative experience.[64] Patients rated the importance of not having any severe pain after surgery less than nurses and doctors (0–100 scale), 63 vs. 78 vs. 72, respectively, $P = 0.042$; and avoidance of moderate pain 64 vs. 89 vs. 65, respectively, $P = 0.014$. Thus the value placed on complete pain control may be overemphasised, with patients being satisfied with less complete analgesia as long they can avoid severe or uncontrolled pain. This finding has implications for pain research, as it suggests that small changes in a VAS, or modest changes in mild or moderate pain, are unimportant to most patients, and do not justify their use as study endpoints.

MEASURES OF HEALTH STATUS, QUALITY OF RECOVERY AND PATIENT SATISFACTION

As described earlier in this chapter, clinical anaesthesia research has focused mostly on surrogate or true adverse events.[65–68] However, with improved safety and quality of anaesthesia and surgery, and the greater expectations of patients, more emphasis is now being placed on other features of postoperative recovery.[24] Quality of recovery,[65,69,70] and patient satisfaction,[71–73] are two such indices. Long-term quality of life is also relevant and can be measured in the months and years after anaesthesia and surgery.[70]

Research into these aspects of health status has centred on psychometric testing; these address measures of validity, reliability and responsiveness.[74–77] The health status instruments being developed can be used for discrimination, prediction and/or evaluation.[74] It is the intended purpose of the instrument that determines how it should be evaluated and it is advisable that a specific instrument be selected for a specific purpose. Predictive indices are intended to identify patients whose health status changes. Evaluative indices are used to measure a change in health status and so their responsiveness is an essential characteristic.[74]

There are limited studies in anaesthesia assessing patient satisfaction.[78–81] Satisfaction with health care is usually very high (>85%). Consequently it is difficult to identify a representative sample of patients dissatisfied with care without studying large populations.

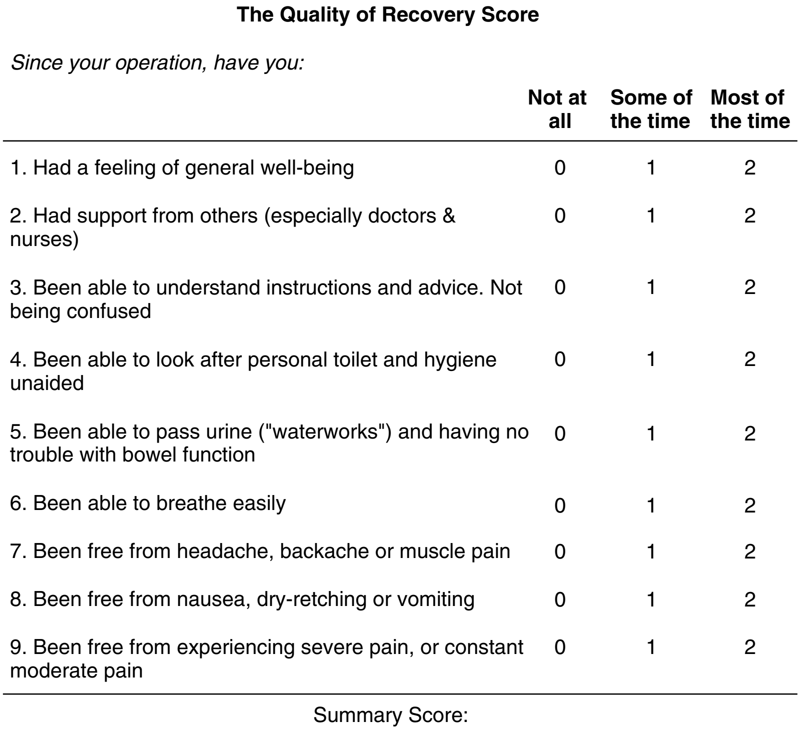
In a retrospective analysis of an ongoing database, the records of 10 811 patients were reviewed to rate their satisfaction with care on the day after surgery.[72] The overall level of satisfaction was 96.8%; 246 (2.3%) patients were ‘somewhat dissatisfied’, and 97 (0.9%) were ‘dissatisfied’ with their anaesthesia care. Patients who were dissatisfied were younger, 49 (19) years vs. 54 (20) years, $P < 0.0005$, and had shorter duration of anaesthesia, 2.2 (1.9) hours vs. 2.4 (2.2) hours, $P = 0.018$. After adjustment for patient and surgical factors, there was a strong relationship between patient dissatisfaction and the occurrence of postoperative pain, nausea and vomiting, and other complications. Not surprisingly, awareness under anaesthesia, although rare, was strongly associated with patient dissatisfaction.

Minor postoperative complications are important to patients and represent an area for potential improvement in anaesthetic, surgical and nursing care.[65,69,78–82] Clinical trials testing interventions to reduce pain, postoperative nausea and vomiting (PONV), and other complications may result in improvements in the quality of postoperative recovery, and patient satisfaction.

In an analysis of 5672 adult surgical patients, Myles *et al.*[80] found that the number of complications after surgery and anaesthesia was associated with lower rates of patient satisfaction. Dexter *et al.*[73] developed a valid and reliable measure of patient satisfaction after minor surgical procedures, and found that unrelieved pain was associated with lower patient satisfaction.

QUALITY OF RECOVERY

Myles *et al.*[65] described the development and psychometric evaluation of a nine-item patient-orientated quality of recovery score ('QoR Score') after anaesthesia and surgery (Figure 28.2). The QoR Score has good validity, reliability and clinical acceptability in patients undergoing many types of surgery,[65,69] and is related to patient satisfaction with anaesthesia.[80] Of 5672 adult surgical patients, those that were satisfied had higher scores than those who were dissatisfied, median QoR Score 16 vs. 13, $P < 0.0005$.[80] Thus the QoR Score can be used as a global measure of early postoperative health status, and as a proxy measure of patient satisfaction.



The Quality of Recovery Score

Since your operation, have you:

	Not at all	Some of the time	Most of the time
1. Had a feeling of general well-being	0	1	2
2. Had support from others (especially doctors & nurses)	0	1	2
3. Been able to understand instructions and advice. Not being confused	0	1	2
4. Been able to look after personal toilet and hygiene unaided	0	1	2
5. Been able to pass urine ("waterworks") and having no trouble with bowel function	0	1	2
6. Been able to breathe easily	0	1	2
7. Been free from headache, backache or muscle pain	0	1	2
8. Been free from nausea, dry-retching or vomiting	0	1	2
9. Been free from experiencing severe pain, or constant moderate pain	0	1	2

Summary Score:

Source: Copyright, Lippincott, Williams and Wilkins.

Figure 28.2. A nine-item quality of recovery score – the QoR Score. Patients are interviewed hours or days after surgery and asked to circle the most appropriate responses[65,80].

There is also a more comprehensive 40-item questionnaire, the QoR-40, which rates each item on a 5-point scale, and has five dimensions of recovery: (i) emotional state, (ii) physical comfort, (iii) psychological support, (iv) physical independence and (v) pain control.[69,70,71,80] The QoR-40 has been evaluated in cardiac and neurosurgical patients.[69,83] The QoR-40 had excellent responsiveness in the days after surgery. Lower scores were associated with longer duration of surgery, postoperative complications and increased length of stay in hospital.[69,70] These data support construct validity, predictive ability and responsiveness of the QoR-40. Other instruments have also been developed.[71,73] Such numerical scoring instruments allow quantification of patients' early postoperative health status – their quality of recovery – and are useful indices of outcome in those patients who do not suffer serious complications.

Quality of life measures have also been used.[70,83,84] The short-form health survey, SF-36, is a 36-item health status questionnaire measuring eight dimensions of quality of life: physical functioning, role limitations due to physical problems, bodily pain, social functioning, mental health, role limitations due to emotional problems, vitality–energy/fatigue, and general health perception.[85] Each dimension has a possible score of 0 (poor health) to 100 (excellent health). The SF-36 has been used to assess QoL after anaesthesia and cardiac surgery,[70,81] and ICU care.[84]

TRIAL DESIGN ISSUES IN ANAESTHESIA

CONDUCTING CLINICAL TRIALS IN ANAESTHESIA

Important adverse outcomes after surgery are rare. For example, the incidence of stroke, renal failure or death after vascular or coronary artery surgery is 2% to 4%. In order to detect a moderate, but clinically important, difference between groups, many thousands of patients need to be studied.[12,15,16] Such trials are not commonly undertaken in anaesthesia, intensive care or pain medicine. For example, the average sample size of 208 trials investigating possible efficacious treatments in traumatic brain injury was 82.[18]

Nevertheless, there have been some excellent examples of large trials in anaesthesia.[20–23,33–36] In some of these the investigators selected a high-risk group in order to increase the number of adverse events in the study; this reduced the number of patients required to achieve adequate statistical power.

STATISTICAL POWER

Many trials are too small to detect a clinically important difference.[11,17,18] During the design phase of a proposed trial, consideration of the primary endpoint (is it clinically important?) and an estimation of sample size are required. These factors, and an estimation of study power, are now required by journals to be included in published reports of anaesthesia research.

TECHNICAL AND SKILL FACTORS

Most drug trials compare two or more treatments, and all other aspects of care are presumed to be equivalent. The intervention – drug formulation, dose, route of administration – is standardised, and estimates of effect can be attributed to the study drug (alone). This approach needs to be re-evaluated in perioperative research, where the technical ability of the individual performing the study intervention will vary. For example, a study investigating the benefits of a pulmonary artery catheter,[86,87] or regional blockade,[35] may be dependent on the skill and knowledge of the anaesthetist and/or other staff providing care. This source of variation can confound a trial.

For example, the failure rate of epidural block in surgical and obstetric practice is about 12% to 17%.[88,89] An experienced group of anaesthetists, performing a large number of epidural blocks for general and vascular surgery, studied 1014 patients receiving epidural fentanyl/bupivacaine infusions for analgesia after surgery.[89] They found that the patient's pain relief was rated as good to excellent on 83% of postoperative visits – that is, on 17% of occasions the patient had inadequate pain relief. Mechanical problems, including dislodgment of the catheter, accounted for 19% of infusion discontinuations within the first 72 h. Such variation adds additional confounding to a study investigating the effects of epidural blockade, and may obscure a beneficial effect. Variations in the application of a proven intervention may limit generalisability of the results.

Pulmonary artery catheterisation provides an indirect measure of left ventricular preload (filling pressure) and cardiac output. Knowledge of these variables helps direct therapies in patients undergoing major cardiovascular surgery and in those who are critically ill postoperatively. A study of 146 intensivists given such data from critically ill patients were asked to make management choices.[86] There was significant heterogeneity in treatment choices among intensivists. This suggests that an evaluation of such a monitor, or therapies guided by such monitoring, need to consider variations in practice. One method of dealing with such heterogeneity is to have strict treatment protocols, or at the very least, collect data defining the variations that may exist so that statistical adjustment (or modelling) can be applied.

SEQUENTIAL ANALYSIS

Sequential analysis is a valid and efficient research design method to detect a significant difference between groups as early as possible.[90] This approach creates, in advance, stopping boundaries that are evaluated throughout the conduct of a trial. Classically these are statistical comparisons done after each pair of observations is assessed. These methods, as in other areas of medical research, are uncommonly used in anaesthesia, but examples can still be found.

Patients undergoing total hip replacement are at risk of postoperative thromboembolism, for which heparin prophylaxis is recommended therapy. However, such patients often receive spinal anaesthesia and are therefore at risk of inadvertent spinal haematoma, especially if anticoagulated. Spinal anaesthesia may also reduce deep vein thrombosis (DVT). Thus heparin may not be necessary in these circumstances, and the risk of epidural haematomata can be minimised. Samama *et al.*[91] studied the benefits and risks of administering heparin during spinal anaesthesia in patients undergoing total hip replacement. They did a randomised, double-blind trial comparing low-molecular-weight heparin (LMWH) with placebo for 10 days after surgery. Efficacy was assessed by venography on day 10, using sequential analysis. The sample size estimation was based on a 40% reduction in the incidence of DVT, from 25% to 15%. Using values of alpha of 0.05 and beta of 0.20, it was calculated that 500 patients would need to be studied (250 in each group). Data were analysed every 50 patients, and a maximum final sample size was set at 750 patients (i.e. 15 analyses). A sequential analysis using the simple triangular test procedure was incorporated in the design.[90] Here two calculations are made: a test for the difference between groups, and another of the information available in the trial at that time. A plot of the two statistics is done, using them as the coordinates in the sequential design for the analysis. The trial was stopped early.[91] There was a significant reduction in the incidence of DVT in the LMWH group when compared with the placebo group, 14% vs. 37%, $P = 0.002$. No spinal haematomata or gross neurological sequelae were observed during the study. However, this study was underpowered to detect an increase in risk of spinal haematoma or paraplegia (reported incidence about 1:20 000), and so we have no meaningful information about this very serious, but rare, complication.

INTERIM ANALYSIS

The classic method of sequential analysis incorporates continuous or multiple looks at the accumulating data. Contemporary interim analysis includes a planned, but limited, number of looks at the data. In either case there is an increased likelihood of finding a significant difference between groups purely by chance, but this can be accommodated by adjusting for repeat testing.[90,92,93] Interim analyses are an integral part of most large trials, including those in anaesthesia.[20,23,35,94]

Poldermans *et al.*[95] did a randomised, unblinded, multicentre trial to assess the effect of perioperative beta-blockade on the incidence of cardiac death and non-fatal myocardial infarction in high-risk patients within 30 days after major vascular surgery. High-risk status was determined by reversible ischaemia being demonstrated using stress echocardiography, in which 1351 patients were screened, but only 112 (of 173 eligible) were recruited to the trial. Enrolled patients were randomly assigned to receive standard perioperative care or standard care plus perioperative beta-blockade with bisoprolol. As part of the study design, an interim analysis was planned after enrolment of the first 100 patients. In accordance with the O'Brien and Fleming criteria, the protocol specified that the trial would be stopped if there was a significant difference, using an alpha value of 0.001. The trial was stopped after the first interim analysis, with a combined incidence of cardiac events in the standard-care group of 34% vs. bisoprolol group 3.4%. Despite there being few trial events (11 deaths, 9 myocardial infarctions), this trial has been very influential in supporting perioperative beta-blockade in patients at risk of postoperative myocardial infarction.

CROSSOVER TRIALS

It may be difficult to detect a significant difference between groups in trials when the observation of interest is subject to substantial variation. This is common in perioperative studies where variations in surgery type and extent and anaesthesia techniques, patient co-morbidity and other factors can all affect outcomes. Crossover trials can equalise some of these factors.[96,97] In general, crossover trials have been underutilised in anaesthesia research, but some examples can be found in the anaesthetic and intensive care literature.[98,99]

For example, the role of epinephrine as a vasoconstrictor affecting local anaesthetic efficacy was clearly demonstrated with a crossover design. Local anaesthetic drugs such as ropivacaine have their duration of effect increased by the addition of epinephrine, whose vasoconstrictor effects delay absorption from the site of action. Niemi *et al.*[100] did a prospective, randomised, double-blind crossover trial, testing the addition of epinephrine, 2 μg/mL, to a thoracic epidural infusion of ropivacaine and fentanyl in 12 patients after major surgery. The main outcome measure was pain intensity evaluated on a VAS. Pain increased when epinephrine was omitted from the epidural infusion, $P < 0.001$. During the study period without epinephrine (>3 h), pain intensity was unacceptable despite rescue analgesia. After restarting the epidural mixture with epinephrine, pain was reduced and sensory blockade restored. The mixture with epinephrine was associated with less nausea and facilitated early mobilisation. Thus the addition of epinephrine to an epidural containing ropivacaine markedly improved the quality of pain relief.

IV cannulation can be painful and distressing for some patients, particularly children. Cutaneous analgesia can be provided by a eutectic mixture of local anaesthetics (EMLA) cream, but EMLA cream takes 60–90 min to take effect and this limits its usefulness. In a randomised, double-blind, placebo-controlled crossover trial of 42 subjects, the speed of onset of EMLA cream was determined after brief pretreatment of the underlying skin with low-frequency ultrasound.[101] Four treatments were compared: ultrasound pretreatment followed by application of 1 g EMLA or placebo cream for 5 min, 10 min, 15 min and 60 min without ultrasound pretreatment as positive control. Pain was assessed by pin-prick testing. Pain VAS scores and patient preference for EMLA or placebo cream were measured at each time point. Based on both pain scores and patient preference, efficacy was achieved in the EMLA groups as compared with placebo at all time points. After ultrasound pretreatment and then 5, 10 or 15 min after EMLA cream application, pain scores and overall preference were statistically indistinguishable from EMLA cream application for 60 min (without ultrasound pretreatment).

Low-frequency ultrasound pretreatment appears to be safe and effective in producing rapid onset of EMLA cream.

N-OF-1 TRIALS

N-of-1 trials are a good method of testing the specific benefits of a proposed new treatment in an individual patient.[102] The results are not intended to be generalised to other patients. This trial design may be useful when optimising an anaesthetic technique for a patient requiring repeated surgical procedures, for long-stay patients in the ICU in order to optimise a sedative or analgesic regimen, or in patients with chronic pain.

Eide and Stubhaug[103] examined whether ketamine, an anaesthetic and analgesic drug known to block *N*-methyl-D-aspartate (NMDA) receptors, was able to relieve glossopharyngeal neuralgia in a 56-year-old woman with unremitting severe pain that had lasted for seven years. Ketamine was evaluated with the *N*-of-1 trial approach, using a double-blind design. The optimal oral dose, 60 mg administered six times per day, was first used in an open dose-escalating trial. This was followed by a period in which the patient received either oral ketamine or placebo during 10 two-day periods. Ketamine administration was associated with marked pain relief, as shown by statistically significant pain relief and reduction of pain intensity.

DRUG DEVELOPMENT STUDIES

PHASE I/II STUDIES

Early phase studies in anaesthesia and pain are usually initiated and sponsored by pharmaceutical companies as part of drug development programmes.

Scott *et al.*[104] compared the acute central nervous and cardiovascular effects of the local anaesthetics ropivacaine and bupivacaine in 12 healthy male volunteers, using a randomised, double-blind design. In each case, the study drug was administered via an IV infusion at a rate of 10 mg/min up to a maximal dose of 150 mg. Study subjects were first familiarised with the central nervous system (CNS) toxic effects of local anaesthetics by receiving a preliminary IV infusion of lidocaine. The infusions of ropivacaine and bupivacaine were given at least seven days apart. CNS toxicity was identified by the CNS symptoms and the subjects were told to request that the infusion be stopped when they felt definite but not severe symptoms of toxicity such as perioral numbness, lightheadedness, or tinnitus. In the absence of definite symptoms, the infusion was stopped at a maximal dose of 150 mg. Cardiovascular changes in conductivity and myocardial contractility were monitored using ECG and echocardiography. Ropivacaine caused less CNS symptoms and was 25% less toxic than bupivacaine in regard to the dose tolerated.

The safety of spinal (intrathecal) preservative-containing neostigmine was assessed in a Phase I study after the sole manufacturer of the preservative-free solution ceased production.[105] Earlier studies of the preservative-free preparation had been found to produces analgesia in humans in a variety of pain models. After preclinical toxicity screening in animals and US Food and Drug Administration approval, 12 volunteers received spinal neostigmine 10 μg, 30 μg or 100 μg, containing the preservatives methyl- and propyl-parabens. This preparation produced dose-dependent analgesia, nausea, weakness and sedation similar to the preservative-free preparation. This study supports further investigation of the safety and efficacy of spinal preservative-containing (generic) neostigmine.

PHASE III STUDIES

Phase III efficacy trials are widespread in anaesthesia and pain research. Many are designed and conducted by pharmaceutical companies, testing investigational drugs in specifically targeted areas. But there are also many independent studies being done.

A Phase I study of spinal neostigmine was presented in the preceding section. Chung *et al.*[106]

studied the analgesic efficacy and safety of spinal neostigmine, spinal morphine, and their combination in patients undergoing caesarean section under spinal anaesthesia. This Phase III study randomly allocated 79 term parturients into four groups to receive isotonic sodium chloride solution 0.2 mL (control group), neostigmine 25 μg, morphine 100 μg, or the combination of spinal neostigmine 12.5 μg and morphine 50 μg with spinal 0.5% bupivacaine 12 mg. Postoperative analgesia was provided by IV patient-controlled analgesia (PCA) using fentanyl and ketorolac. Compared with the control group, the time to first PCA use was significantly longer in the neostigmine group ($P < 0.001$), with lower 24 h analgesic consumption ($P < 0.001$). Nausea and vomiting were the most common side effects of spinal neostigmine (74%). Analgesic effectiveness was similar between the neostigmine and morphine groups. Compared with the neostigmine group, the combination group had significantly prolonged analgesic effect and reduced 24 h PCA consumption ($P < 0.05$) with less severity of nausea and vomiting ($P = 0.058$). Compared with the morphine group, the combination group tended to have prolonged times to first PCA use ($P = 0.054$) with a lower incidence of pruritus ($P < 0.03$).

Recombinant human-activated protein C (drotrecogin alfa activated) has anti-thrombotic, anti-inflammatory and profibrinolytic properties.[107] In a Phase III pharmaceutical company-sponsored trial, using a randomised, double-blind, placebo-controlled multicentre design, 1690 patients with severe critical illness due to sepsis were randomly allocated to receive an IV infusion of either placebo or drotrecogin alfa activated for 96 hours. The primary endpoint was all-cause mortality at 28 days. The subsequent mortality rate was 31% in the placebo group and 25% in the drotrecogin alfa activated group; relative risk reduction 19% (95% CI: 6.6–31), absolute risk reduction 6%, $P = 0.005$. The incidence of serious bleeding was higher in the drotrecogin alfa activated group (3.5% vs. 2.0%, $P = 0.06$).

Previous studies have demonstrated reduced postoperative morphine requirements and/or improved pain relief when non-steroidal anti-inflammatory drugs are administered in conjunction with PCA. Plummer *et al.*[108] did a randomised, double-blind trial aimed at determining whether these effects could be achieved with a sustained-release ibuprofen formulation given preoperatively, so that it could avoid the need for oral administration during the early postoperative period. The study also aimed to determine whether a reduction in morphine administration was associated with reduced opioid side effects. They enrolled 115 patients undergoing lower abdominal gynaecological surgery, and these were randomly allocated to receive either ibuprofen, 1600 mg ($n = 57$), or placebo ($n = 58$) preoperatively and again 24 h after the first dose. Patients were assessed every 4 h up to 24 h postoperatively. Those receiving ibuprofen reported significantly less pain at rest, $P = 0.023$. Patient opinions of the efficacy of their pain-relieving medication ($P < 0.001$) and quality of sleep ($P = 0.036$) favoured ibuprofen. The results demonstrated improved efficacy with no increase in side effects when sustained-release ibuprofen is used as an adjunct to morphine PCA.

PHASE IV STUDIES

Patients undergoing surgery and anaesthesia are at risk of many complications. These range from postoperative nausea and vomiting, to myocardial ischaemia and infarction renal and respiratory failure, and chronic pain states.

Nikolajsen *et al.*[109] investigated whether post-amputation stump and phantom pain in the first year after lower-limb amputation is reduced by preoperative epidural blockade with bupivacaine and morphine. In a randomised, double-blind trial, 60 patients scheduled for amputation were randomly allocated to receive epidural bupivacaine (0.25% 4–7 mL/h) and morphine (0.16–0.28 mg/h) for 18 h before and during the operation (blockade group), or epidural saline (4–7 mL/h) and oral or intramuscular morphine (control group). All patients had general anaesthesia for the amputation and were asked about stump and phantom pain after one week and then after 3, 6, and 12 months by two independent examiners. Study endpoints were rate of stump and phantom pain, intensity of stump and phantom pain, and consumption of opioids. After one week, 14 (52%) patients in the blockade group and 15 (56%) in the control group had phantom pain (difference 95% CI: −31 to −23, $P = 0.9$). There were no significant differences between groups at all later time points; at 3 months: 14 (82%) vs. 10 (50%), 95% CI: 4% to 61%, $P = 0.09$; at 6 months: 13 (81%) vs. 11 (55%), 95% CI: −3% to 55%, $P = 0.2$; and at 12 months: 9 (75%) vs. 11 (69%), 95% CI: −27 to 40, $P = 1.0$. The authors concluded that perioperative epidural blockade before amputation and continued into the postoperative period does not prevent phantom or stump pain. Thus there was no evidence of pre-emptive (preventative) analgesia in this setting.

Acupuncture at the P6-point of the wrist has been used to treat nausea and vomiting. Gan *et al.*[110] evaluated the efficacy of electro-acupoint stimulation, ondansetron versus placebo for the prevention of PONV. Patients undergoing major breast surgery under general anaesthesia were randomly allocated into active electro-acupoint stimulation (A), ondansetron 4 mg IV (O), or sham control (placement of electrodes without electro-acupoint stimulation – placebo [P]). The incidence of nausea, vomiting, rescue antiemetic use, pain, and patient satisfaction with management of PONV, were assessed at 0, 30, 60, 90, 120 min, and at 24 h. The primary endpoint was defined as complete response (no nausea, vomiting, or use of rescue antiemetic), and was significantly more frequent in the active treatment groups compared with placebo both at 2 h (A 77% vs. O 64% vs. P 42%, respectively; $P = 0.01$) and 24 h postoperatively (A 73% vs. O 52% vs. P 38%, respectively; $P = 0.006$). The need for rescue antiemetic was less in the treatment groups (A 19% vs. O 28% vs. P 54%; $P = 0.04$). Patients in the treatment groups were more satisfied with their management of PONV compared with placebo. When used for the prevention of PONV, electro-acupoint stimulation or ondansetron was more effective than placebo with greater degree of patient satisfaction, but electro-acupoint stimulation seems to be more effective in controlling nausea, compared with ondansetron. Interestingly, the authors found that stimulation at P6 also has analgesic effects.

CLINICAL TRIAL GROUPS

There are a number of productive trial groups in anaesthesia and intensive care medicine. These include groups formed to conduct specific trials, or discipline-focused groups such as the Canadian Critical Care Clinical Trials Group, the Multicentre Study of Perioperative Ischemia (McSPI) Group, the Outcomes Research Group, the Australian and New Zealand Intensive Care Society Clinical Trials Group, and, more recently, the Australian and New Zealand College of Anaesthetists Trials Group. These and other investigators have conducted a number of large trials, often studying high-risk groups to maximise statistical power.

PRAGMATIC TRIALS IN ANAESTHESIA

Pragmatic trials test for effectiveness in routine clinical practice. They include large numbers of patients, are done in a variety of settings (usually multicentre), and test simple interventions used by clinicians who may or may not have research expertise. They therefore represent ‘real-world’ patients and clinical practice.[12–16] Tunis *et al.*[111] have argued that reliable evidence is essential to improve health care quality, and that this depends on large trials done in such settings; they are often referred to as pragmatic or practical clinical trials. This need has also been recognised in anaesthesia.[112]

The ability of epidural anaesthesia and postoperative analgesia to decrease the incidence of death and major complications after major surgery has been debated for many years. Park *et al.*[36] did a multicentre randomised trial of 1021 patients, comparing epidural with alternative regimens, and measured death and major complications during and for 30 days after abdominal surgery. Overall, there was no significant difference in the incidence of death and major complications between the two groups. A subgroup analysis identified a possible benefit in patients undergoing aortic surgery, with a lower rate of death and major complications (22% vs. 37%), but this was not observed in later studies.[113] Rigg *et al.*[35] did a similar multicentre clinical trial in 915 high-risk patients undergoing major abdominal surgery. They found no difference in their combined major morbidity and mortality endpoint at 30 days, epidural 57% vs. control 61%, $P = 0.30$. Mortality at 30 days was low in both groups, epidural 5.1% vs. control 4.3%. Eight categories of morbid endpoints were examined: respiratory, cardiac, renal, hepatic, gastric, pancreatic, haematological and inflammatory. One of these – respiratory failure – occurred less frequently in patients managed with epidural techniques (23% vs. 30%, $P = 0.03$).

Inadvertent hypothermia during and after major surgery is associated with some adverse outcomes.[114–116] It leads to peripheral vasoconstriction and decreased subcutaneous oxygen tension, as well as impairment of white cell function, bacterial killing and wound healing. Kurz *et al.*[21] did a randomised, double-blind (patients, surgeons) trial in 200 patients undergoing colorectal surgery, comparing routine intraoperative care (the hypothermia group) with maintenance of normothermia using a forced air warming device. They evaluated surgical wounds daily for up to two weeks and also cultured suspected infected wounds. The mean (SD) intraoperative core temperature was 34.7 (0.6)°C in the hypothermia group and 36.6 (0.5)°C in the normothermia group, $P < 0.001$. Wound infections were more common in the hypothermia group, 18/96 patients (19%) vs. normothermia, 6/104 (6%), $P = 0.009$. The duration of hospitalisation was prolonged by more than two days (20% increase) in the hypothermia group ($P = 0.01$). This trial demonstrates that hypothermia delays healing and predisposes patients to wound infections; this may lead to a longer hospital stay.

Antiplatelet therapy prevents venous thromboembolism in a variety of medical settings, especially in various high-risk groups.[117] Rodgers *et al.*[33] did a large multicentre randomised trial of 17 044 patients undergoing surgery for hip fracture or elective hip arthroplasty. The study treatment was 160 mg daily aspirin or placebo, started preoperatively and continued for 35 days. Endpoints were mortality and in-hospital morbidity up to day 35. Among the patients with hip fracture, aspirin therapy was associated with a 43% (95% CI: 18%–60%) reduction in pulmonary embolism, $P = 0.002$, and a 29% (95% CI: 3%–48%) reduction in symptomatic DVT, $P = 0.03$. Confirmed pulmonary embolism or DVT occurred less commonly in aspirin-treated patients compared with control, 1.6% vs. 2.5%, $P = 0.0003$. Deaths due to bleeding were few (aspirin 13 vs. placebo 15), but there was an excess of six postoperative transfused bleeding episodes per 1000 patients in the aspirin group, $P = 0.04$. Rates of venous thromboembolism were lower in hip arthroplasty patients, but similar beneficial effects were seen. This study supports the routine inclusion of aspirin as a useful measure to reduce thromboembolic complications after hip surgery.

Awareness is an uncommon complication of anaesthesia, but devastating to the patient, and affects about 1 in 1000 patients undergoing surgery with general anaesthesia.[72] Some patients, such as those undergoing cardiac surgery, caesarean section, rigid bronchoscopy, and those undergoing surgery after major trauma, are at increased risk (incidence about 1%). Advances in the ability to process the electroencephalogram, such as bispectral index (BIS) monitoring, can provide a more reliable measure of anaesthetic depth and so may assist titration of anaesthetic drugs to minimise the possibility of intraoperative awareness. Myles *et al.*[23] did a randomised, double-blind, multicentre trial in 2463 patients at high risk of awareness, comparing anaesthesia-guided BIS monitoring, a type of processed EEG, with routine care. Patients were assessed by a blinded observer for awareness at 2–6 h, 24–36 h and 30 days after surgery. An independent committee, blinded to group identity, assessed each report of awareness. The primary outcome measure was confirmed awareness under anaesthesia at any time. There were 2 reports of awareness in the BIS-guided group and 11 reports in the routine care group, relative risk reduction 82%; 95% CI: 17% to 98%, $P = 0.022$. This trial demonstrated the effectiveness of BIS monitoring in reducing the risk of awareness in patients undergoing relaxant general anaesthesia.

In 1998 a systematic review and meta-analysis of trials concluded that albumin solutions, when compared with saline, were associated with excess mortality in critically ill patients.[118] This study generated a lot of controversy and the findings were disputed by many experienced ICU clinicians. This led to the ANZICS Group conducting a multicentre, randomised, double-blind trial.[34] They compared the effect of fluid resuscitation with albumin or saline on mortality in 6997 critically ill patients. The primary endpoint was all-cause mortality at 28 days. There were 726 deaths in the albumin group, as compared with 729 deaths in the saline group, relative risk 0.99; 95% CI: 0.91 to 1.09, $P = 0.87$. There were no significant differences between the groups in the mean (SD) numbers of days spent in the ICU (6.5 (6.6) in the albumin group and 6.2 (6.2) in the saline group, $P = 0.44$), days spent in the hospital (15 (9.6) and 16 (9.6), respectively; $P = 0.30$), days of mechanical ventilation (4.5 (6.1) and 4.3 (5.7), respectively; $P = 0.74$). The study found no evidence that albumin is associated with excess mortality or major morbidity in ICU patients. This finding contradicts the conclusions of the earlier meta-analysis.[118]

Intracranial aneurysm surgery can result in new postoperative neurological deficits which may affect long-term outcome. Hypothermia is commonly used during such surgery in a belief, based on animal studies and uncontrolled human studies, that it has a neuroprotective effect. Todd *et al.*[20] did a multicentre, randomised, double-blind trial in patients with subarachnoid haemorrhage requiring craniotomy and aneurysm clipping to determine whether active intraoperative cooling improves outcome. They enrolled 1001 patients in 30 centres, and randomly allocated them to intraoperative hypothermia (target temperature 33°C, using surface cooling techniques) or normothermia (target 36.5°C). Patients were assessed postoperatively and examined approximately 90 days after surgery, at which time their neurological status was determined using the validated Glasgow Outcome Score. Additional measures included the NIH Stroke Scale, Barthel's Index and the Rankin Disability Score. There were no significant differences in the duration of ICU stay, total length of hospitalisation or discharge destination (home, other hospital, dead). At final follow-up, 66% of hypothermic patients ($n = 499$) had a Glasgow Outcome Score of 1 (good outcome) vs. 63% of normothermic patients ($n = 501$); odds ratio 1.14, 95% CI: 0.88 to 1.48, $P = 0.32$. Mortality at the time of follow-up in the two groups was 5.8% and 6.4%, respectively. There were no significant differences in other outcome measures. Importantly, and consistent with previous studies,[21] there was some evidence of increased infection risk, with higher rates of postoperative bacteraemia in the hypothermic group (5.0% vs. 2.6%, $P = 0.05$). Thus intraoperative hypothermia did not improve neurological outcome following craniotomy in intracranial aneurysm patients.

SYSTEMATIC REVIEWS AND META-ANALYSES

Although few large trials have been done in anaesthesia and pain medicine, meta-analysis of small trials has allowed sufficiently large samples to be used to estimate treatments effects for many interventions.

Perioperative myocardial ischaemia is a typical surrogate measure in anaesthesia and critical care studies, occurring in 20% to 40% of patients at risk of cardiac morbidity. It is assumed to represent true cardiac morbidity. Nishina *et al.*[119] did a systematic review of randomised trials that tested the efficacy of clonidine in reducing perioperative myocardial ischaemia. They retrieved 28 studies, but could only include seven relevant randomised trials ($n = 664$) in their meta-analysis. The pooled odds ratio was 0.49, 95% CI: 0.34 to 0.71.

There were insufficient data to estimate the effect on myocardial infarction. The authors concluded that these findings justify conducting a definitive study to test the benefits of clonidine.

The possible benefits of perioperative clonidine therapy have been confirmed recently. Wallace *et al.*[120] did a randomised double-blind trial in 190 surgical patients with known or suspected coronary artery disease. Patients received clonidine, 200 μg (oral plus skin patch), or matched placebo the night before surgery and on the morning of surgery. The skin patch provided a continuous administration of study drug for four days and was then removed. Patients underwent any one of several types of major non-cardiac surgery. Myocardial ischaemia was detected with Holter ST-segment monitoring of the ECG for 7 days.[27] The incidence of perioperative myocardial ischaemia was significantly reduced with clonidine compared with placebo, 14% vs. 31%, $P = 0.01$. Interestingly, the authors followed up patients for two years and in a secondary analysis found that clonidine improved survival, 85% vs. 71%; relative risk 2.33, 95% CI: 1.12 to 4.76, $P = 0.035$. Such a benefit would be important if it were confirmed in subsequent trials.

There has been long-standing uncertainty about possible adverse effects of epidural analgesia in women during labour; in particular that it may slow the progress of labour and increase the need for forceps delivery or caesarean section. Low concentrations of local anaesthetic solutions reduce motor blockade and so, theoretically at least, should minimise pelvic and abdominal muscle weakness during the second stage of labour. Liu and Sia[121] did a systematic review and meta-analysis of randomised trials comparing low-concentration epidural infusions with parenteral opioids. They identified seven relevant trials. As has been demonstrated in many small trials, epidural analgesia had greater analgesic effectiveness, with more women randomised to receive epidural analgesia with adequate pain relief. There was no evidence that epidural analgesia increased the risk of caesarean section ($n = 2962$), odds ratio 1.03, 95% CI: 0.71 to 1.48. However, epidural analgesia was associated with a longer second stage of labour, weighted mean difference 15 min, 95% CI: 2 to 28 min. This was associated with an increased risk of instrumental vaginal delivery (three trials, $n = 1092$), odds ratio 2.11, 95% CI: 0.95 to 4.65.

Meta-analysis has some limitations, including publication bias, duplicate publication, heterogeneity, and inclusion of historical (out-dated) studies.[122,123] These have cast doubt on the validity of certain meta-analyses, most notably the previously described albumin versus saline for the resuscitation of the critically ill.[118]

Von Elm *et al.*[124] evaluated 141 systematic reviews (129 337 subjects) in anaesthesia. Of these, 56 authors (40%) confirmed that they had identified duplicate publications. Sixty articles were published twice, 13 three times, 3 four times and 2 five times. The prevalence of covert duplicate articles – that is, those without a cross-reference to the main article – was 5.3%. Of the duplicates, 34 (33%) were sponsored by the pharmaceutical industry, and 66 (64%) had authorship that differed partly or completely from the main article. The median delay in publication between main articles and duplicates was one year (range, 0–7 years). These findings suggest that duplicate publication is unlikely to be detected during the peer-review process.

Earlier in this chapter we referred to several clinical trials investigating the effects of epidural analgesia on postoperative morbidity and mortality.[35,36] Rodgers *et al.*[125] did a systematic review of all trials comparing neuraxial blockade (epidural or spinal) with alternative regimens. They identified 141 relevant trials including 9559 patients. Mortality was reduced by about a third in patients allocated to neuraxial blockade: 103 deaths/4871 patients vs. 144/4688 patients, odds ratio 0.70, 95% CI: 0.54 to 0.90, $P = 0.006$. Neuraxial blockade reduced the odds of DVT by 44%, pulmonary embolism by 55%, transfusion requirements by 50%, pneumonia by 39% and respiratory depression by 59% (all $P < 0.001$). Although there was limited power to assess subgroup effects, the proportional reductions in mortality did not clearly differ by surgical group, type of blockade (epidural or spinal), or in those trials in which neuraxial blockade was combined with general anaesthesia compared with trials in which neuraxial blockade was used alone. The authors concluded that neuraxial blockade reduces postoperative mortality and other serious complications. The validity of the conclusions of this meta-analysis is under question, after the findings of two subsequent large multicentre randomised trials which found minimal benefits associated with the use of epidural anaesthesia.[40,48]

COST-EFFECTIVENESS STUDIES

Elliott *et al.*[126] compared the cost-effectiveness of several general anaesthetic agents in adult and paediatric day surgery populations. The anaesthetic regimens studied included propofol, isoflurane and sevoflurane (single drug, or in combination). Their major adverse endpoint of interest was PONV, a frequent cause of unplanned admission following day surgery. They randomly allocated 1063 adult and 322 paediatric surgical patients to one of four (adult) or two (paediatric) anaesthesia groups. Total costs were calculated from individual patient resource use to seven days postdischarge. Incremental cost-effectiveness ratios were expressed as cost per episode of PONV avoided. In adults, variable secondary care costs were higher for propofol induction and propofol maintenance (propofol/propofol, $P < 0.01$) than other groups and lower in propofol induction and isoflurane maintenance (propofol/isoflurane, $P < 0.01$). In both studies, predischarge PONV was higher if sevoflurane/sevoflurane ($P < 0.01$) was used compared with use of propofol for induction. In both studies, there was no difference in postdischarge outcomes at day 7. Sevoflurane/sevoflurane was more costly with higher PONV rates in both studies. In adults, the cost per extra episode of PONV avoided was £296 (propofol/propofol vs. propofol/sevoflurane) and £333 (propofol/sevoflurane vs. propofol/isoflurane).

Cheng *et al.*[127] analysed cost data from a clinical trial comparing two opioid regimens for cardiac surgery. They included an analysis of perioperative complications, time to tracheal extubation, duration of ICU and hospital stay, and resource utilisation of nursing staff. The trial had compared a fentanyl/isoflurane/propofol regimen with a remifentanil/isoflurane/propofol regimen in 304 cardiac surgical patients. The extubation times, ICU stay and hospital discharge times were not significantly different between groups. Other resource utilisation data were also equivalent. Further exploratory analysis identified increasing preoperative risk scores and age (> 70 years) as being significant factors associated with postoperative complications, length of stay and resource utilisation. The new (more expensive) opioid drug, remifentanil, does not seem to offer any advantages in this setting.

BIS monitoring has become more widely used in recent years, following several trials that had demonstrated greater accuracy with titration of general anaesthesia and faster patient recovery.[23] Liu[128] did a systematic review and meta-analysis that included 11 trials and 1380 patients, and found that use of BIS monitoring significantly reduced anaesthetic drug use by 19% (95% CI: 11%–27%), reduced the incidence of nausea/vomiting by 23% (95% CI: 1%–44%), and reduced time in the recovery room by 4 (95% CI: 1–7) min. Cost analysis using pooled costs to reflect North America, Europe and Asia indicated that use of BIS monitoring increased the cost per patient by US$5.6 because of the acquisition cost of BIS electrodes. Thus although use of BIS monitoring may lead to a small reduction in anaesthetic drug consumption, and risk of PONV and recovery room time, there was still a net cost associated with BIS monitoring. This analysis did not include possible savings associated with a reduction in awareness,[23] such as disability associated with post-traumatic stress disorder or litigation.

CLINICAL PRACTICE GUIDELINES

Clinical practice, or consensus, guidelines are intended to provide an expert, evidence-based approach to clinical practice.[129] They have much to offer. Practice guidelines are available for several anaesthetic procedures.[49,130–135]

Guidelines have been developed to assist with the preoperative evaluation of patients at risk of coronary events after surgery.[134] These were first produced in 1996, following several opinion-based commentaries and/or narrative reviews. They provide a framework for considering cardiac risk in patients undergoing non-cardiac surgery. The guidelines are based on trials and observational studies retrieved from electronic searches of the medical literature from 1995 to 2000, but also the expert opinions of 12 committee members representing various disciplines of cardiology, nuclear medicine, vascular medicine, vascular surgery and anaesthesia. One recommendation of the guidelines is that preoperative intervention is rarely necessary simply to lower the risk of surgery unless such intervention is indicated irrespective of the preoperative context. The purpose of preoperative evaluation is not simply to give medical clearance but rather to perform an evaluation of the patient's current medical status; make recommendations concerning the evaluation, management and risk of cardiac problems over the entire perioperative period; and provide a clinical risk profile that the patient, primary physician, anaesthetist and surgeon can use in making treatment decisions that may influence short- and long-term cardiac outcomes.

PONV remains an important and common complication of surgery.[72,78–80] It is feared by patients,[71,79,82] is a limiting factor in the early discharge of day surgery patients, and is a leading cause of unplanned hospital admission. PONV leads to increased recovery room time, nursing care and potential hospital admission – all factors that may increase total health care costs. Equally important are the high levels of patient discomfort and dissatisfaction associated with PONV.[72,79,82] Among high-risk patients, the incidence of PONV can be as frequent as 70% to 80%.

Gan *et al.*[135] have developed evidence-based guidelines for the management of PONV. An expert committee reviewed the medical literature on PONV and produced guidelines for management that were meant to be 'valid, reliable, clinically applicable, flexible, and clear'. The panel defined the following goals for the guidelines: (1) identify the primary risk factors for PONV in adults and children, (2) reduce the baseline risks for PONV, (3) identify the optimal approach to PONV prevention and therapy in various patient populations, (4) determine the optimal choice and timing of antiemetic administration, and (5) identify the most effective antiemetic monotherapy and combination therapy regimens. An evidence rating scale (I to V) was used to grade the study design. The quality of the data was judged by the panel, which determined whether the recommendation was good, fair or insufficient. For example, they concluded that a logistic regression analysis used to identify risk factors for PONV would be in level IV. However, information emerging from that study may be judged as high level ('A') by the panel.

The panel agreed that not all patients should receive PONV prophylaxis. They outlined why patients at low risk for PONV are unlikely to benefit from prophylaxis, because their baseline risk was low, and the subsequent NNT exceeded 10. Drugs for PONV prophylaxis should be used for patients at moderate risk of PONV. Double and triple antiemetic combinations were recommended for patients at high risk of PONV.[135]

EVIDENCE-BASED MEDICINE IN ANAESTHESIA

The rationale and practice of evidence-based medicine (EBM) has been described in detail,[136,137] yet are sometimes criticised.[138–140] Constructive and thoughtful appraisals of EBM and contrasting views of its importance and relevance in contemporary anaesthesia practice are also available.[7,9,12,14,139] Myles *et al*[9] estimated the proportion of anaesthetic interventions in routine practice that were supported by evidence in the literature. They surveyed their hospital practice, and found that 96.7% of clinical decisions were evidence-based of which 32% were supported by randomised trials (levels I and II).

These results are similar to recent studies in other specialties. A review of all publications in the five top-ranked (by impact factor) anaesthesia journals for the year 2000 found that about 20% were randomised trials.[141]

There are circumstances where a small, tightly controlled trial may be preferable to larger trials.[142,143] But if clinical research in anaesthesia and pain medicine is to provide more reliable and valid evidence to inform therapeutic decision making in the future, it is clear that a greater proportion of such research needs to be large, high-quality, multicentre randomised trials.[13,15,16,111] This is becoming more widely accepted among anaesthetists and pain medicine practitioners.[14,16,21–26,49,112,131,135,138]

REFERENCES

1. Sykes WS. The beginning of things. In: Sykes WS, *Essays on the first hundred years of anaesthesia*. Melbourne: Churchill Livingstone (1992).
2. Vinten-Johansen P, Brody H, Paneth N, Rachman S, Rip M. *Cholera, Chloroform, and the Science of Medicine: A Life of John Snow*. New York: Oxford University Press (2003).
3. Ali M, Ramachari A. History of anaesthesia and Hyderabad Chloroform Commissions. *Bull Indian Inst Hist Med Hyderabad* (1989) **19**: 47–61.
4. Beecher HK, Todd DP. A study of the deaths associated with anesthesia and surgery: based on a study of 599,548 anesthesias in ten institutions 1948–1952, inclusive. *Ann Surg* (1954) **140**: 2–35.
5. National Halothane Study. Possible association between halothane anesthesia and postoperative hepatic necrosis. *JAMA* (1966) **197**: 775–88.
6. Warden JC, Borton CL, Horan BF. Mortality associated with anaesthesia in New South Wales, 1984–1990. *Med J Aust* (1994) **161**: 585–93.
7. Mackay P. Safety in anaesthesia in Australia. A review of anaesthetic related mortality 1997–1999. *Australia and New Zealand College of Anaesthetists* (2002).
8. Spence AA. The lessons of CEPOD. *Br J Anaesth* (1988) **60**: 753–4.
9. Myles PS, Bain DL, Johnson F, McMahon R. Is anaesthesia evidence-based? A survey of anaesthetic practice. *Br J Anaesth* (1999) **82**: 591–5.
10. MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *Lancet* (2001) **357**: 455–62.
11. Rigg JR, Jamrozik K, Clarke M. How can we demonstrate that new developments in anesthesia are of real clinical importance? *Anesthesiology* (1997) **86**: 1008–11.
12. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* (1984) **3**: 409–20.
13. Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. *Lancet* (2001) **357**: 373–80.
14. Rigg JRA, Jamrozik K, Myles PS. Evidence-based methods to improve anaesthesia and intensive care. *Curr Opin Anaesthesiol* (1999) **12**: 221–7.
15. Peto R, Baigent C. Trials: the next 50 years. Large scale randomised evidence of moderate benefits. *BMJ* (1998) **317**: 1170–1.
16. Myles PS. Why we need large randomised studies in anaesthesia. *Br J Anaesth* (1999) **83**: 833–4.
17. Kelly MJU, Wadsworth J. What price inconclusive clinical trials? *Ann R Coll Surg Engl* (1993) **75**: 145–6.
18. Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I. Size and quality of randomised controlled trials in head injury: review of published studies. *BMJ* (2000) **320**: 1308–11.
19. Cooper DJ, Myles PS, McDermott F, Murray LJ, Laidlaw J, Cooper G, Tremayne AB, Bernard SS, Ponsford J, for the HTS Study Investigators. Pre-hospital hypertonic saline resuscitation of patients with severe traumatic brain injury – a randomized controlled trial. *JAMA* (2004) **291**: 1350–7.
20. Todd M, Hindman BJ, Clarke WR, Torner JC, for the IHAST Investigators. Mild intraoperative hypothermia during surgery for intracranial ancurysm. *New Engl J Med* (2005) **352**: 1345–52.
21. Kurz A, Sessler DI, Lenhardt R, The Study of Wound Infection and Temperature Group. Perioperative normothermia to reduce the incidence of surgical-wound infection and shorten hospitalization. *New Engl J Med* (1996) **334**: 1209–15.
22. Greif R, Akca O, Horn EP, Kurz A, Sessler DI. Supplemental perioperative oxygen to reduce the incidence of surgical-wound infection. Outcomes Research Group. *New Engl J Med* (2000) **342**: 161–7.
23. Myles PS, Leslie K, Forbes A, McNeil J, Chan M, for the B-Aware Trial Group. Bispectral index monitoring to prevent awareness during

anaesthesia: the B-Aware randomised controlled trial. *Lancet* (2004) **363**: 1757–63.

24. Lee A, Lum ME. Measuring anaesthetic outcomes. *Anaesth Intensive Care* (1996) **24**: 685–93.
25. Fisher DM. Surrogate outcomes: meaningful not! *Anesthesiology* (1999) **90**: 355–6.
26. Horan B. Evidence based medicine and anaesthesia: uneasy bedfellows? *Anaesth Intensive Care* (1997) **25**: 679–85.
27. Jain U, Body SC, Bellows W, Wolman R, Mangano CM, Mathew J, Youngs E, Wilson R, Zhang A, Mangano DT. Multicenter study of target-controlled infusion of propofol-sufentanil or sufentanil-midazolam for coronary artery bypass graft surgery. Multicenter Study of Perioperative Ischemia (McSPI) Research Group. *Anesthesiology* (1996) **85**: 522–35.
28. Ramsey SD, Saint S, Sullivan SD, Dey L, Kelley K, Bowdle A. Clinical and economic effects of pulmonary artery catheterization in nonemergent coronary artery bypass graft surgery. *J Cardiothorac Vasc Anesth* (2000) **14**: 113–18.
29. Smith JS, Cahalan MK, Benefiel DJ, Byrd BF, Lurz FW, Shapiro WA, Roizen MF, Bouchard A, Schiller NB. Intraoperative detection of myocardial ischemia in high-risk patients: electrocardiography versus two-dimensional transesophageal echocardiography. *Circulation* (1985) **72**: 1015–21.
30. Kendall JB, Russell GN, Scawn ND, Akrofi M, Cowan CM, Fox MA. A prospective, randomised, single-blind pilot study to determine the effect of anaesthetic technique on troponin T release after off-pump coronary artery surgery. *Anaesthesia* (2004) **59**: 545–9.
31. Campagna JA, Miller KW, Forman SA. Mechanisms of actions of inhaled anesthetics. *New Engl J Med* (2003) **348**: 2110–24.
32. Eger EI 2nd, Saidman LJ, Brandstater B. Minimum alveolar anesthetic concentration: a standard of anesthetic potency. *Anesthesiology* (1965) **26**: 756–63.
33. The PEP Trial Collaborative Group. Prevention of pulmonary embolism and deep vein thrombosis with low dose aspirin: Pulmonary Embolism Prevention (PEP) trial. *Lancet* (2000) **355**: 1295–1302.
34. Finfer S, Bellomo R, Boyce N, French J, Myburgh J, Norton R, SAFE Study Investigators. A comparison of albumin and saline for fluid resuscitation in the intensive care unit. *New Engl J Med* (2004) **350**: 2247–56.
35. Rigg JRA, Jamrozik K, Myles PS, Silbert BS, Peyton PJ, Parsons RW, Collins KS, MASTER Anaesthesia Trial Study Group. Epidural anaesthesia and analgesia and outcome of major surgery: a randomised trial. *Lancet* (2002) **359**: 1276–82.
36. Park WY, Thompson JS, Lee KK. Effect of epidural anesthesia and analgesia on perioperative outcome: a randomized, controlled Veterans Affairs cooperative study. *Ann Surg* (2001) **234**: 560–71.
37. Schafer A. The ethics of the randomised clinical trial. *New Engl J Med* (1982) **307**: 719–24.
38. Freedman B. Equipoise and the ethics of clinical research. *New EngI J Med* (1987) **317**: 141–5.
39. Myles PS, Fletcher HE, Cairo S, Madder H, McRae R, Devonshire D, Cooper J, Hunt JO, Richardson J, Machlin H, Morgan EB, Moloney J, Downey G. Informed consent and recruitment for clinical trials in the immediate preoperative period. *Anesthesiology* (1999) **91**: 969–78.
40. Myles PS, Tan N. Reporting of ethical approval and informed consent in clinical research published in leading anesthesia journals. *Anesthesiology* (2003) **99**: 1209–13.
41. Mingus IVIL, Levitan SA, Bradford CN, Eisenkraft JB. Surgical patient's attitudes regarding participation in clinical anesthesia research. *Anesth Analg* (1996) **82**: 332–7.
42. Montgomery JE, Sneyd JR. Consent to clinical trials in anaesthesia. *Anaesthesia* (1998) **53**: 227–30.
43. Tait AR, Voepel-Lewis T, Siewart M, Malviya S. Factors that influence parents' decisions to consent to their child's participation in clinical anesthesia research. *Anesth Analg* (1998) **86**: 50–3.
44. Meisel A, Roth LH. What we do and do not know about informed consent. *JAMA* (1981) **246**: 2473–7.
45. Zelen M. A new design for randomised clinical trials. *New Engl J Med* (1979) **300**: 1242–5.
46. O'Rourke PP, Crone RK, Vacanti JP, Ware JH, Lillehei CW, Parad RB, Epstein M. Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn: a prospective randomized study. *Pediatrics* (1989) **84**: 957–63.
47. Truog RD. Randomized controlled trials: lessons from ECMO. *Clin Res* (1993) **40**: 519–27.
48. Miller BL. Philosophical, ethical, and legal aspects of resuscitation medicine. I. Deferred consent and justification of resuscitation research. *Crit Care Med* (1988) **16**: 1059–62.
49. National Health and Medical Research Council. Acute pain management: scientific evidence. NHMRC (1998).
50. Revill SI, Robinson JO, Rosen M, Hogg MIJ. The reliability of a linear analogue for evaluating pain. *Anaesthesia* (1976) **31**: 1191–8.

51. Campbell WI, Patterson CC. Quantifying meaningful changes in pain. *Anaesthesia* (1998) **53**: 121–5.
52. DeLoach LJ, Higgins MS, Caplan AB, Stiff JL. The visual analogue scale in the immediate postoperative period: intrasubject variability and correlation with a numeric scale. *Anesth Analg* (1998) **86**: 102–6.
53. Dexter F, Chestnut DH. Analysis of statistical tests to compare visual analog scale measurements among groups. *Anesthesiology* (1995) **82**: 896–902.
54. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. *Ann Rheum Dis* (1978) **37**: 378–81.
55. Mantha S, Thisted R, Foss J, Ellis JE, Roizen MF. A proposal to use confidence intervals for visual analog scale data for pain measurement to determine clinical significance. *Anesth Analg* (1993) **77**: 1041–7.
56. Myles PS, Troedel S, Boquest M, Reeves M. The pain visual analogue scale: is it linear or non-linear? *Anesth Analg* (1999) **89**: 1517–20.
57. Moore A, Moore O, McQuay H, Gavaghan D. Deriving dichotomous outcome measures from continuous data in randomised controlled trials of analgesics: use of pain intensity and visual analogue scales. *Pain* (1997) **69**: 311–15.
58. Cepada MS, Africano JM, Polo R, Alcala R, Carr D. What decline in pain intensity is meaningful to patients with acute pain? *Pain* (2003) **105**: 151–7.
59. Aubrun F, Langeron O, Quesnel C, Coriat P, Riou B. Relationships between measurement of pain using visual analog score and morphine requirements during postoperative intravenous morphine titration. *Anesthesiology* (2003) **98**: 1415–21.
60. Ludington E, Dexter F. Statistical analysis of total labor pain using the visual analog scale and application to studies of analgesic effectiveness during childbirth. *Anesth Analg* (1998) **87**: 723–7.
61. Collins SL, Moore RA, McQuay HJ. The visual analogue pain intensity scale: what is moderate pain in millimetres? *Pain* (1997) **72**: 95–7.
62. Bodian CA, Freedman G, Hossain S, Eisenkraft JB, Beilin Y. The visual analog scale for pain: clinical significance in postoperative patients. *Anesthesiology* (2001) **95**: 1356–61.
63. Philip BK. Parametric statistics for evaluation of the visual analog scale. *Anesth Analg* (1990) **71**: 710.
64. Duncan PG, Cohen MM. The literature of anaesthesia: what are we learning? *Can J Anaesth* (1988) **3**: 494–9.
65. Myles PS, Hunt JO, Nightingale CE, Fletcher H, Beh T, Tanil D, Nagy A, Rubinstein A, Ponsford JL. Development and psychometric testing of a quality of recovery score after general anesthesia in adults. *Anesth Analg* (1999) **88**: 83–90.
66. Forrest JB, Cahalan MK, Rehder K, Goldsmith CH, Levy WJ, Strunin L, Bota W, Boucek CD, Cucchiara RF, Dhamee S *et al*. Multicenter study of general anesthesia. II. Results. *Anesthesiology* (1990) **72**: 262–8.
67. Cohen MM, Duncan PG, DeBoer DP, Tweed WA. The postoperative interview: assessing risk factors for nausea and vomiting. *Anesth Analg* (1994) **78**: 7–16.
68. Webb RK, Currie M, Morgan C, Williamson JA, Mackay P, Russell WJ, Runciman WB. The Australian incident monitoring study: an analysis of 2000 incident reports. *Anaesth Intensive Care* (1993) **21**: 520–8.
69. Myles PS, Weitkamp B, Jones K, Melick J, Hensen S. Validity and reliability of a postoperative quality of recovery score: the QoR-40. *Br J Anaesth* (2000) **84**: 11–15.
70. Myles PS, Hunt JO, Fletcher H, Solly R, Kelly S, Woodward D. Relationship between quality of recovery in hospital, and quality of life at three months after cardiac surgery. *Anesthesiology* (2001) **95**: 862–7.
71. Tong D, Chung F, Wong D. Predictive factors in global and anesthesia satisfaction in ambulatory surgical patients. *Anesthesiology* (1997) **87**: 856–64.
72. Myles PS, Williams DL, Hendrata M, Anderson H, Weeks AM. Patient satisfaction after anaesthesia and surgery: results of a prospective survey of 10,811 patients. *Br J Anaesth* (2000) **84**: 6–10.
73. Dexter F, Aker J, Wright WA. Development of a measure of patient satisfaction with monitored anesthesia care. *Anesthesiology* (1997) **87**: 865–73.
74. Gill TM, Feinstein AR. A critical appraisal of the quality-of-life measurements. *JAMA* (1994) **272**: 619–26.
75. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chron Dis* (1985) **38**: 27–36.
76. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chron Dis* (1987) **40**: 171–8.
77. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* (1989) **27**: S178–89.
78. Myles PS, Hunt JO, Moloney JT. Postoperative 'minor' complications. Comparison between men and women. *Anaesthesia* (1997) **52**: 300–6.

79. Orkin F. What do patients want? - preferences for immediate postoperative recovery. *Anesth Analg* (1992) **74**: S225.
80. Myles PS, Reeves M, Anderson, Weeks AM. Measurement of quality of recovery in 5672 patients after anaesthesia and surgery. *Anaesth Intensive Care* (2000) **28**: 276–9.
81. Hunt JO, Hendrata M, Myles PS. Quality of life after cardiac surgery. *Heart Lung* (2000) **29**: 401–11.
82. Moerman N, van Dam FSAM, Oosting J. Recollections of general anaesthesia: a survey of anaesthesiological practice. *Acta Anaesthesiol Scand* (1992) **36**: 767–71.
83. Leslie K, Troedel S, Irwin K, Pearce F, Ugoni A, Gillies R, Pemberton E, Dharmage S. Quality of recovery from anesthesia in neurosurgical patients. *Anesthesiology* (2003) **99**: 1158–65.
84. Lamer C, Harboun M, Knani L, Moreau D, Tric L, LeGuillou JL, Gasquet I, Moreau T. Quality of life after complicated elective surgery requiring intensive care. *Intensive Care Med* (2004) **30**: 1594–601.
85. McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. *Med Care* (1994) **32**: 40–66.
86. Jain M, Canham M, Upadhyay D, Corbridge T. Variability in interventions with pulmonary artery catheter data. *Intensive Care Med* (2003) **29**: 2059–62.
87. Sandham JD, Hull RD, Brant RF, Knox L, Pineo GF, Doig CJ, Laporta DP, Viner S, Passerini L, Devitt H, Kirby A, Jacka M, the Canadian Critical Care Clinical Trials Group. A randomized, controlled trial of the use of pulmonary-artery catheters in high-risk surgical patients. *New Engl J Med* (2003) **348**: 5–14.
88. Pan PH, Bogard TD, Owen MD. Incidence and characteristics of failures in obstetric neuraxial analgesia and anesthesia: a retrospective analysis of 19,259 deliveries. *Int J Obstet Anesth* (2004) **13**: 227–33.
89. Scott DA, Beilby DS, McClymont C. Postoperative analgesia using epidural infusions of fentanyl with bupivacaine. A prospective analysis of 1,014 patients. *Anesthesiology* (1995) **83**: 727–37.
90. Whitehead J. *The design and analysis of sequential clinical trials*, 2nd edn. Chichester: Ellis Horwood (1992) 80–92.
91. Samama CM, Clergue F, Barre J *et al.* Low molecular weight heparin associated with spinal anaesthesia and gradual compression stockings in total hip replacement surgery. *Br J Anaesth* (1997) **78**: 660–5.
92. McPherson K. Statistics: the problem of examining accumulating data more than once. *New Engl J Med* (1974) **290**: 501–2.
93. Pocock SJ. Statistical and ethical issues in monitoring clinical trials. *Stat Med* (1993) **12**: 1459–69.
94. The Acute Respiratory Distress Syndrome Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *New Engl J Med* (2000) **342**: 1301–8.
95. Poldermans D, Boersma E, Bax JJ, Thomson IR, van de Ven LL, Blankensteijn JD, Baars HF, Yo TI, Trocino G, Vigna C, Roelandt JR, van Urk H. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. *New Engl J Med* (1999) **341**: 1789–94.
96. Louis TA, Lavori PW, Bailar JC 3rd, Polansky M. Crossover and self-controlled designs in clinical research. *New Engl J Med* (1984) **310**: 24–31.
97. Woods JR, Williams JG, Tavel M. The two-period crossover design in medical research. *Ann Intern Med* (1989) **110**: 560–6.
98. Ngan Kee WD, Lam KK, Chen PP, Gin T. Comparison of patient-controlled epidural analgesia with patient-controlled intravenous analgesia using pethidine or fentanyl. *Anaesth Intensive Care* (1997) **25**: 126–32.
99. Myles PS, Leong CK, Weeks AM, Buckland MR, Bujor M, Smith JA, Rabinov M, Esmore DS. Early hemodynamic effects of left atrial administration of epinephrine after cardiac transplantation. *Anesth Analg* (1997) **84**: 976–81.
100. Niemi G, Breivik H. Epinephrine markedly improves thoracic epidural analgesia produced by a small-dose infusion of ropivacaine, fentanyl, and epinephrine after major thoracic or abdominal surgery: a randomized, double-blinded crossover study with and without epinephrine. *Anesth Analg* (2002) **94**: 1598–1605.
101. Katz NP, Shapiro DE, Herrmann TE, Kost J, Custer LM. Rapid onset of cutaneous anesthesia with EMLA cream after pretreatment with a new ultrasound-emitting device. *Anesth Analg* (2004) **98**: 371–6.
102. Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The n-of-1 randomized controlled trial: clinical usefulness. Our three year experience. *Ann Intern Med* (1990) **112**: 293–9.

103. Eide PK, Stubhaug A. Relief of glossopharyngeal neuralgia by ketamine-induced N-methyl-aspartate receptor blockade. *Neurosurgery* (1997) **41**: 505–8.

104. Scott DB, Lee A, Fagan D, Bowler GM, Bloomfield P, Lundh R. Acute toxicity of ropivacaine compared with that of bupivacaine. *Anesth Analg* (1989) **69**: 563–9.

105. Eisenach JC, Hood DD, Curry R. Phase I human safety assessment of intrathecal neostigmine containing methyl- and propylparabens. *Anesth Analg* (1997) **85**: 842–6.

106. Chung CJ, Kim JS, Park HS, Chin YJ. The efficacy of intrathecal neostigmine, intrathecal morphine, and their combination for post-cesarean section analgesia. *Anesth Analg* (1998) **87**: 341–6.

107. Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A *et al*. Efficacy and safety of recombinant human activated protein C for severe sepsis. *New Engl J Med* (2001) **344**: 699–709.

108. Plummer JL, Owen H, Ilsley AH, Tordoff K. Sustained-release ibuprofen as an adjunct to morphine patient-controlled analgesia. *Anesth Analg* (1996) **83**: 92–6.

109. Nikolajsen L, Ilkjaer S, Christensen JH, Kroner K, Jensen TS. Randomised trial of epidural bupivacaine and morphine in prevention of stump and phantom pain in lower-limb amputation. *Lancet* (1997) **350**: 1353–7.

110. Gan TJ, Jiao KR, Zenn M, Georgiade G. A randomized controlled comparison of electro-acupoint stimulation or ondansetron versus placebo for the prevention of postoperative nausea and vomiting. *Anesth Analg* (2004) **99**: 1070–5.

111. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* (2003) **290**: 1624–32.

112. McPeek B. Inference, generalizability, and a major change in anesthetic practice. *Anesthesiology* (1987) **66**: 723–4.

113. Peyton PJ, Myles PS, Silbert BS, Rigg JRA, Jamrozik K, Parsons RW, and the MASTER Anesthesia Trial Study Group. Perioperative epidural analgesia and outcome after major abdominal surgery in high risk patients. *Anesth Analg* (2003) **96**: 548–54.

114. Schmied H, Kurz A, Sessler DI, Kozek S, Reiter A. Mild hypothermia increases blood loss and transfusion requirements during total hip arthroplasty. *Lancet* (1996) **47**: 289–92.

115. Frank SM, Beattie C, Christopherson R, Norris EJ, Perler BA, Williams GM, Gottlieb SO. Unintentional hypothermia is associated with postoperative myocardial ischemia. *Anesthesiology* (1993) **78**: 468–76.

116. Leslie K, Sessler DI, Bjorksten AR, Moayeri A. Mild hypothermia alters propofol pharmacokinetics and increases the duration of action of atracurium. *Br J Anaesth* (1995) **80**: 1007–14.

117. Collaborative overview of randomised trials of antiplatelet therapy-II: maintenance of vascular graft or arterial patency by antiplatelet therapy. Antiplatelet Trialists' Collaboration. *BMJ* (1994) **308**: 159–68.

118. Cochrane Injuries Group. Human albumin administration in critically ill patients: systematic review of randomised controlled trials. *BMJ* (1998) **317**: 235–40.

119. Nishina K, Mikawa K, Uesugi T, Obara H, Maekawa M, Kamae I, Nishi N. Efficacy of clonidine for prevention of perioperative myocardial ischemia: a critical appraisal and meta-analysis of the literature. *Anesthesiology* (2002) **96**: 323–9.

120. Wallace AW, Galindez D, Salahieh A, Layug EL, Lazo EA, Haratonik KA, Boisvert DM, Kardatzke D. Effect of clonidine on cardiovascular morbidity and mortality after noncardiac surgery. *Anesthesiology* (2004) **101**: 284–93.

121. Liu EH, Sia AT. Rates of caesarean section and instrumental vaginal delivery in nulliparous women after low concentration epidural infusions or opioid analgesia: systematic review. *BMJ* (2004) **328**: 1410–15.

122. LeLoerier J, Gregoire G, Benhaddad A, Lapierre J, Derderian Fl. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New Engl J Med* (1997) **337**: 536–42.

123. Egger M, Smith GD. Misleading meta-analysis. Lessons learned from "an effective, safe, simple" intervention that wasn't. *BMJ* (1995) **310**: 752–4.

124. Von Elm E, Poglia G, Walder B, Tramer MR. Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *JAMA* (2004) **291**: 974–80.

125. Rodgers A, Walker N, Schug S *et al*. Reduction of postoperative mortality and morbidity with epidural or spinal anaesthesia: results from overview of randomised trials. *Br Med J* (2000) **321**: 1493–7.

126. Elliott RA, Payne K, Moore JK, Harper NJ, St Leger AS, Moore EW, Thoms GM, Pollard BJ, McHugh GA, Bennett J, Lawrence G, Kerr J, Davies LM. Clinical and economic choices in anaesthesia for day surgery: a prospective randomised controlled trial. *Anaesthesia* (2003) **58**: 412–21.

127. Cheng DC, Newman MF, Duke P, Wong DT, Finegan B, Howie M, Fitch J, Bowdle TA, Hogue C, Hillel Z, Pierce E, Bukenya D. The efficacy and resource utilization of remifentanil and fentanyl in fast-track coronary artery bypass graft surgery: a prospective randomized, double-blinded controlled, multi-center trial. *Anesth Analg* (2001) **92**: 1094–102.
128. Liu SS. Effects of Bispectral Index monitoring on ambulatory anesthesia: a meta-analysis of randomized controlled trials and a cost analysis. *Anesthesiology* (2004) **101**: 11–15.
129. Sniderman AD. Clinical trials, consensus conferences, and clinical practice. *Lancet* (1999) **354**: 327–30.
130. American Society of Anesthesiologists Task Force on Management of the Difficult Airway: practice guidelines for management of the difficult airway. *Anesthesiology* (1993) **78**: 597–602.
131. Practice Guidelines for Pulmonary Artery Catheterisation: a report by the American Society of Anesthesiologists Task Force on Pulmonary Artery Catheterisation. *Anesthesiology* (1993) **78**: 380–94.
132. ACC/AHA Task Force Report. Special report: guidelines for perioperative cardiovascular evaluation for noncardiac surgery. Report of the American College of Cardiology/American Heart Association Task Force on practice guidelines (Committee on Perioperative Cardiovascular Evaluation for Noncardiac Surgery). *J Cardiothorac Vasc Anesth* (1996) **10**: 540–52.
133. Practice guidelines for obstetrical anesthesia: a report by the American Society of Anesthesiologists Task Force on Obstetrical Anesthesia. *Anesthesiology* (1999) **90**: 600–11.
134. Eagle KA, Berger PB, Calkins H, Chaitman BR, Ewy GA, Fleischmann KE *et al.* ACC/AHA Guideline Update for Perioperative Cardiovascular Evaluation for Noncardiac Surgery – Executive Summary. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Anesth Analg* (2002) **94**: 1052–64.
135. Gan TJ, Meyer T, Apfel CC, Chung F, Davis PJ, Eubanks S, Kovac A, Philip BK, Sessler DI, Temo J, Tramer MR, Watcha M. Consensus guidelines for managing postoperative nausea and vomiting. *Anesth Analg* (2003) **97**: 62–71.
136. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based Medicine: How to Practice & Teach EBM*. Melbourne: Churchill Livingstone (1997).
137. Sacket DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*, 2nd edn. Boston: Little Brown (1991) 283–302.
138. Editorial. Evidence-based medicine, in its place. *Lancet* (1995) **346**: 785.
139. Horan BF. Evidence-based medicine and anaesthesia: uneasy bedfellows? *Anaesth Intensive Care* (1997) **25**: 679–85.
140. Goodman NW. Anaesthesia and evidence-based medicine. *Anaesthesia* (1998) **53**: 353–68.
141. Lauriston J, Moller AM. Publications in anesthesia journals: quality and clinical relevance. *Anesth Analg* (2004) **99**: 1486–91.
142. Powell-Tuck J, McRae KD, Healy MJR, Lennard-Jones JE, Parkins RA. A defense of the small clinical trial: evaluation of three gastroenterological studies. *BMJ* (1986) **292**: 599–602.
143. Sackett DL, Wennberg JE. Choosing the best research design for each question. *BMJ* (1997) **315**: 1636.

29

General Surgery

DR OLGA JONASSON[1]* AND DOMENIC REDA[2]
[1]University of Illinois College of Medicine, Chicago, USA
[2]Department of Veterans Affairs Cooperative Studies Program Coordinating Center,
VA Hospital, Hines, Illinois, USA

BACKGROUND

In July 2005, the cover of *The Lancet* carried the following quote: 'If everything has to be double-blinded, randomized, and evidence-based, where does that leave new ideas?'[1]

In 1987 laparoscopic cholecystectomy was introduced to the American general surgery world on the floor of the commercial exhibits of the American College of Surgeons' Annual Clinical Congress. It was seen as a great new idea and soon weekend courses in how to perform this new procedure had been developed, often provided by for-profit groups or the manufacturer. Surgeons interested in learning the technique and instrumentation were given didactic sessions and then practised for a day or two on porcine models. Given the immediate public demand for 'minimally invasive laser surgery', many hospitals accepted this brief training as adequate and purchased the expensive instrumentation needed for surgeons completing these courses to offer the procedure to patients. The majority of patients did well, but many were seriously harmed, especially in the early experience with laparoscopic cholecystectomy. The incidence of injury to the common bile duct was excessive, leading often to the need for repeated operations and a significantly diminished health-related quality of life.[2,3] Other injuries, especially to vascular structures, led to early discontinuation of laser use. The cost to the well-being of patients was large, and it was several years before the usefulness of this procedure was accepted, following a period of trial and error during which the technique was perfected, perhaps at the expense of the patients. Enthusiasm and competition had far outpaced the need for evidence of efficacy and safety, systematic education, and controlled safe dissemination into practice. By the time a trial could be developed, surgeons and patients had accepted laparoscopic cholecystectomy as the standard of care.[4–6] Today laparoscopic cholecystectomy has become, by far, the predominant surgical procedure for cholelithiasis, but its introduction into practice was precipitous.

* Dr Olga Jonasson died in August 2006 whilst this book was in production.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

WHY PERFORM CLINICAL TRIALS IN SURGERY?

Many new procedures have been introduced by their proponents at clinical meetings and through publication in clinical journals, only to disappear from practice when others are unable to reproduce the proponent's success.[7] The list of failed, discredited and obsolete surgical procedures is long (Table 29.1). Often, lack of benefit to patients or actual patient harm has ensued. Cost-effectiveness and health-related quality of life are rarely measured. As put by Wennberg in his testimony to the US Congress, 'We need a way to assure the American people that the needed evaluations of clinical theory are done in a timely way, before plausible but wrong ideas get institutionalized into everyday practice of medicine.'[8]

Unlike new drugs or devices that can be introduced into practice only if approved by regulatory authorities, new surgical procedures are nearly totally unregulated. The only obstacle a surgeon faces when developing a new technique is the hospital's operating room restrictions. If the necessary equipment is made available through purchase or provided by the manufacturer, and the hospital is convinced that the surgeon has a reasonable approach with potential patient benefit, few if any additional controls are imposed. Even the Institutional Review Boards are not usually involved and protocols are often not required. Outcomes are not documented in a systematic manner, and new ideas can be tried out – in patients – with no real plan for study or follow-up.

Table 29.1. Some examples of discredited surgical procedures

Ligation of internal mammary arteries (for coronary insufficiency)
Vineberg procedure (for coronary insufficiency)
Whitehead procedure (for haemorrhoids)
Nephropexy
Halsted radical mastectomy (for breast cancer at all stages)
Uterine suspension
Kidney decapsulation (for hypertension)
Sympathectomy (for hypertension)

WHEN SHOULD A CLINICAL TRIAL IN SURGERY BE CONSIDERED?

So, when is a clinical trial indicated in surgery? Trialists invoke the principle of equipoise to justify when a trial is appropriate. Equipoise exists when the available evidence and opinions of surgeons and patients about the efficacy of a new technique in comparison with the current standard of care are in balance, i.e. there is no clear indication of the superiority of either the new or the standard approach. In addition, trials of emerging techniques that are undergoing rapid refinements, as well as procedures that are already in widespread use, would be difficult to evaluate in the context of a randomised trial. Patient willingness to participate and the surgeon's willingness to enrol patients must also be considered.

Therefore, what is the justification for a surgical trial? Clearly, subjecting all new procedures to a clinical trial as a prerequisite to incorporation into routine practice is impractical. We should demand, however, that *major* innovations be accompanied by evidence from a carefully designed clinical trial before placing patients at risk (Table 29.2). It is only in a randomised, multicentre clinical trial that a major innovation can be tested, complications measured, and outcomes assessed in a manner that provides the highest validity to the results obtained. The trial must show that the new technique is at least as good as the one it replaces, and that it is safe in the hands of the average practising surgeon. In the course of conducting the trial, surgeons are systematically educated in the surgical technique, and multicentre trials tell us that the intervention under assessment can be safely generalised to routine practice. 'First, do no harm.'

FORCES OPPOSING CLINICAL TRIALS IN SURGERY

To serve this purpose, a means is needed by which *important* new ideas can be identified. One measure of importance is the predicted

Table 29.2. "Statement on Issues to Be Considered Before New Surgical Technology is Applied to the Care of Patients." American College of Surgeons, Statement 23. Reprinted with permission from *Bulletin of the American College of Surgeons* Vol. 80, No. 9, Pages 46–47, September 1995

As new technology is developed and made available for use, its worth is measured on the basis of the value and safety it confers for patients. Part of this process of evaluation should include a comparison with existing and proven technologies that deal with similar clinical problems. Indications for application of the technology should be carefully defined and patients selected using clear criteria. The qualifications of those who propose to use the new technology must be carefully assessed, verifying that the individual has had comprehensive education and experience in the management of the disease process for which the technology is applied, has acquired the necessary technical skills, and is competent to recognize and manage any complications resulting from use of the new technology. Questions that might be raised by surgeons and health care institutions as a new technology is introduced include the following:

1. Has the new technology been adequately tested for safety and efficacy?

In consideration of this question, data are developed based on careful, controlled clinical trials and observations. It may be that the initial evaluation of complex and novel technologies would be best accomplished in a few specialized centers where expertise and support facilities are available. The results of the trials should be published in peer-reviewed scientific literature so that the medical community at large has the benefit of studying the trial in detail before deciding to adopt and use the new technology.

2. Is the new technology at least as safe and effective as existing, proven techniques?

There is enormous public pressure brought to bear to reduce the extent of invasive surgical procedures in favor of the use of minimal access techniques. When assessing the new technology, safety becomes a major consideration. Although attractive to patients, some minimal access procedures may carry risks greater than the established counterpart, even the risk of sudden death or disability as, for example, might occur during performance of endovascular procedures in major arteries. Newer procedures may be less effective than their standard counterpart, and sometimes the early benefits of lessened hospital stay or reduced postoperative pain are outweighed by earlier recurrence of the disease process. Only with careful clinical trials and published observations can a conclusion be reached that the new technology is an acceptable and effective substitute for an established alternative.

3. Is the individual proposing to perform the new procedure fully qualified to do so?

Many skills are highly specialized. The mere acquisition of a skill is not the only criterion by which to measure qualifications. In order to determine and apply proper indications for a procedure and to select the appropriate patients for application of the technology, comprehensive knowledge of the disease process and experience in management of patients with the disease is essential. Prompt recognition and management of complications can only be achieved when the individual or team member is fully qualified in all aspects of treatment of the disease.

4. Is the new technology cost-effective?

"Cost-effectiveness" is a product of many factors: the costs of developing and testing the new technology; the actual costs of the equipment associated with the new technology; the costs of the support services needed to safely provide the service; the costs of educating health care professionals in the use of the technique; the benefits of lessened hospital stay and period of disability; and the likelihood of a good long-term outcome and lower long-term health care costs. Each of these factors must be taken into account when making a determination that the new technology offers cost benefits when compared with alternatives, either established treatments or nonsurgical options. It becomes necessary to establish and maintain a database in which outcome measures will include cost analyses.

number of patients subjected to the procedure. The magnitude of potential benefits and risks should also be considered. In the example of laparoscopic cholecystectomy, many millions of patients worldwide could benefit by improved technology. In other examples, such as the surgical treatment of Parkinson's disease by bilateral deep brain stimulation with implantable electrodes, the risks are so large that proof should be obtained before subjecting many more patients to the procedure.

Then, as our example so vividly demonstrates, a trial must be begun early and completed promptly, so that a potentially beneficial procedure can be introduced without delay. As new procedures emerge, there is often a rapid refinement of technique in the earliest cases. Optimally, a trial should be begun at the point of equipoise, and when the technique has become relatively stable but its use has not become widespread. This requires a commitment from those with these important new ideas to avoid publicity and premature endorsement and work instead towards development of a clinical trial. Payers, such as Medicare, should take an active role in expecting evidence that the new idea is safe and clinically effective, and that it is also cost-effective.

A number of steps are needed before a surgical trial is begun. Potential surgeon–participants should be surveyed to establish the need for the trial, and identify that equipoise is achievable. A multidisciplinary team including surgeons, statisticians and trial managers should be put together to design the trial, outcome measures must be defined in a clinically meaningful way, funding must be found, and eventual impact on practice should be assessed. Cost-effectiveness and health-related quality of life[9–11] measurements should be included. Identifying new procedures that should be subjected to a clinical trial and its funding and implementation must, therefore, be a voluntary collaborative effort between industry, surgeons, researchers, hospitals, payers and the public. Widespread education about the benefit of trials for important innovations will be needed.

BARRIERS TO CLINICAL TRIALS IN SURGERY

New ideas have flourished in the specialty of surgery, but clinical trials are rarely performed. In an informal survey of the calendar year 2004 of the five journals most widely read by general surgeons, 31 'randomised clinical trials' were found. Of these, one-third are multicentre trials with good trial design, but only two were reports of trials conducted in the United States. Conformance with CONSORT guidelines[12] is not uniformly required for publication in surgery journals.

Why are well-designed multicentre surgical clinical trials so uncommon in surgery, especially in the United States? A number of factors have been invoked. Industry aggressively promotes new products and often provides access to equipment and courses to learn the new procedure. Accompanying publicity and public awareness of something purported to be new and improved drives demand. Health care providers compete to be the first in the area to offer the new procedure, and aggressively market this service. Public demand then drives dissemination, a phenomenon seen most recently in 'minimally invasive surgery'. Clinical trials and gathering of evidence are bypassed because there is no incentive to delay dissemination.

A major hurdle in accomplishment of a clinical trial in surgery is the high cost of a multicentre trial. Some trials include the cost of hospitalisation for the operation. The costs of other trials involving established procedures concern the personnel needed for data collection, management and analysis of the data, at costs of several million dollars. When funding is sought from public peer-reviewed funding agencies such as the National Institutes of Health, an institute may be reluctant to spend much of its total available grant funds on a single trial, preferring to support several smaller basic research proposals instead. In addition, the competition for public funding is intense and the review processes difficult to successfully negotiate. Revisions to applications in response to reviewers' criticisms are usually required, imposing delays of years before the trial can begin. Timeliness of introduction may be lost. Funding from industry is an alternative, but this carries its own risks of design and data ownership.[13]

Clinician resistance to delay and the requirements of a clinical trial can become a major barrier. The prestige and competitiveness of being the first in the area to provide the new procedure are strong economic incentives for haste and impatience with the long wait involved before a clinical trial is begun and completed. There are important disadvantages, however, to rapid dissemination. Surgeons, who have not learned the required new skills such as laparoscopy during the structure of a residency programme, often experience a long learning curve.[14] During the learning process, poor outcomes may result.

Institutional Review Board procedures, especially for surgery trials, may prove difficult to navigate. The IRB is charged with protecting patient rights and safety, and takes this responsibility seriously, especially when examining proposals dealing with unproved invasive procedures. Informed consent documents are closely scrutinised to assure that volunteers for the trials are fully aware of the risks they are asked to accept, and that they know the purposes of the trial and the potential for benefit to others. Increased awareness of the paramount importance of IRB protections has lengthened reviews, and made approvals more difficult to achieve. The time from IRB submission of a clinical trial proposal to its approval may be many months. When the proposal involves more than one site, each institution's IRB must give approval. The delays imposed by this critically important process can also prove to be a barrier to timely commencement of the trial of an important new surgical procedure.

THE UNIQUE NATURE OF SURGICAL TRIALS

Clinical trials of surgical procedures are driven by the potential importance and the nature of the procedure itself. Blinding of patient or surgeon is usually impossible, but when possible it is critical to identify the placebo effect, known to be powerful in surgical procedures. In one simple trial examining the amount of pain and time to recovery from laparoscopic cholecystectomy compared with the then-standard of care, open cholecystectomy, it was necessary to apply blood-stained abdominal wound dressings to the laparoscopic patients to mimic the wounds of open patients.[15] In a trial of arthroscopic surgery for osteoarthritis of the knee, patients were randomised to arthroscopic debridement and a sham procedure. Patients in the placebo arm had similar outcomes to those receiving the actual procedure.[16] Placebo-controlled procedures are ethically controversial and the subject of heated debate, especially when the procedures are invasive and dangerous, such as the neurosurgical trials of stem cell implantation for Parkinson's Disease,[17] or invasive cardiac procedures. The information from trials of these potentially dangerous but valuable procedures is vital if the procedures are to be recommended.

When a surgical clinical trial is planned, it follows a preliminary experience of one surgeon in one institution. The proponent has a special interest in seeing that the procedure is successful and develops a high degree of skill in patient selection and in performance of the procedure. The institution's staff also become invested in seeing that the programme succeeds.

If found to be *efficacious* in the hands of this surgeon and the institution's team, it is necessary to prove that the new procedure is *effective* in the hands of other surgeons working in other institutions. A multicentre trial can prove, or disprove, this point but single-center trials cannot. Although it is important to select sites and surgeons for multicentre trials that will be able to accomplish their roles and follow the rules of the trial, it is also useful to choose sites and surgeons who represent the practising community. A recently published trial of laparoscopic vs. open hernia repair[18] has been sharply criticised for its findings of poor outcomes in the laparoscopic arm, widely attributed to the selection of VA sites and surgeons who, by implication, are inferior in skill level to those in the community.[19] In fact, this trial has demonstrated *effectiveness* for practising surgeons who only occasionally perform laparoscopic herniorrhaphy; in the hands of these non-specialised surgeons, results were less good until the long learning curve had been accomplished.

To be able to compare emerging surgical procedures with other surgical, medical, or watchful waiting strategies in a clinical trial, it is essential that the procedures be standardised in all respects when an emerging technique is under evaluation. Everything, from the incision, materials, extent of the procedure, and postoperative care, must be the same from site to site and surgeon to surgeon. Standardisation to this extent has been a difficult barrier to recruitment of participating surgeons, all of whom believe that their particular technique has special advantages; they are reluctant to accept the rules. If standardisation is ignored, however, analysis of the results of the trial becomes impossible. Site visits, announced or even unannounced, to all centres by the principal investigator or other leaders of the trial, are performed to assure that all procedures are being followed.

When comparing a well-established technique with an alternative, however, it can be argued that a high degree of standardisation will limit the ability to generalise the results and best practices of the surgeon should be allowed so that the results of the evaluation would be representative of usual surgical care, as was done in the VA trial of transurethral resection of the prostate (TURP) vs. watchful waiting for treatment of moderately symptomatic benign prostatic hypertrophy (BPH).[20] In the case of many surgical procedures, video tapes can be reviewed to maintain quality control. Operative reports are also made available for review.

RECRUITMENT OF SUBJECTS

Recruitment of subjects for the trial is the responsibility of each site and surgeon. While enthusiasm often aids in recruitment for trials comparing two or more surgical procedures, enthusiasm is less when trials compare an operation to watchful waiting or to another non-surgical strategy such as medical therapy. In the latter examples, recruitment strategies must be developed that respect the patient's rights and dignity, while meeting recruitment targets. It is acceptable, under IRB rules, to offer payment to subjects for expenses incurred.[21] It is also permitted to advertise in public media for subjects, and offer payment. In a trial comparing watchful waiting to operation for inguinal hernia,[22] newspaper and radio advertising proved highly successful in creating interest in volunteers for the trial. Although many patients recruited for screening through these public methods may not be eligible for the trial, sufficient numbers who are eligible are included.

In trials of medical vs. surgical therapy, a planned escape clause from the medical therapy could aid recruitment. For example, in a proposed trial of medical vs. surgical therapy for morbid obesity, after an ample time for the medical therapy to be evaluated has elapsed, crossover to the surgical arm is permitted. Recruitment for trials comparing watchful waiting to operation is especially difficult, and may require advertising and additional incentives to patients for their participation. The surgeons involved must be highly motivated to maintain target recruitment goals. In our experience with two trials of hernia management, recruitment goals for the trial comparing two surgical procedures were easily accomplished. The second trial of watchful waiting and operation required advertising to the public, continual professional support for the nurse–coordinators involved in screening the patients and keeping them in the trial, frequent contacts with the surgeon participants, and a no-cost extension of the trial by six months in order to meet the needed sample size. Keeping the subjects in the trial until the planned follow-up visits were completed required hard work by the nurse–coordinators, who maintained close contact with the subjects. As promised, the results of these trials were disclosed to the subjects soon after publication.

WHY RANDOMISE?

Some would argue that a trial of a surgical intervention should consider patient preference in the assignment of treatment. It may not be feasible or ethical to evaluate certain surgical techniques in a randomised setting and this approach may be the best available. When randomisation is possible, however, it should be done. Randomisation eliminates systematic biases in the selection of treatment, allowing the evaluation of the new surgical intervention in comparison with the standard of care to be conducted on a level playing field.

Any aspects of study conduct following the randomisation which may differentially affect the treatment groups, apart from the inherent differences in the treatment strategies being employed, should be avoided as much as possible. Thus, follow-up visit schedules and the evaluations conducted at these visits should be identical in the groups being compared. Masking of treatment assignment, when possible, should be accomplished. In most surgical trials, however, this is difficult and many trials attempt to blind the evaluation of the patient's outcomes through independent evaluators, who were not involved in the surgical procedure or the patient's follow-up care, and are blinded to the treatment assignment.

It is also important to maintain the integrity of the randomised groups throughout the study. Thus, study designs should consider the potential impact on the patient's willingness to remain in the study for the planned duration. Studies with substantial drop-out rates cannot rely on the benefits of randomisation to certify the validity of the treatment comparisons. Many studies have established that patients withdrawn early from the trial differ from those who remain in the study. In drug trials, often those who perceive a lack of benefit from their assigned treatment, or experience adverse effects, withdraw early. Thus, the patients for whom the treatment is least successful are eliminated from the analysis of the data. In surgical trials, while poor responders tend to withdraw as they do in drug trials, other patients often drop out because the operation has 'fixed' their problem and they have less motivation to return for follow-up visits, and, therefore, the best cases may be excluded from the analysis. Thus, study design strategies that intentionally drop poorly responding patients or increase the likelihood of certain types of patients withdrawing early introduce bias and diminish the randomised nature of the treatment groups. Because of this, an analysis of results based on the comparison of treatment strategies, i.e. intention to treat, rather than actual treatment received, has highest validity.

Patients who deviate from the assigned treatment protocol, especially those who cross over to the other treatment, can make interpretation of study results difficult, especially if the crossover rates are high. In these cases, an analysis of the data which distinguish those assigned to each treatment and remained on that treatment from those who switched to the other treatment can provide useful information. Although this should never be considered the primary analysis, such analyses can be used to identify factors which influence the decision to change treatments. In turn, subgroups of patients for which a certain treatment may be optimal can be identified. One must be cautious in interpreting this information, however. Generally, these would be considered exploratory analyses which would need validation in another trial.

PRECISION

While deviations from the randomisation can introduce bias, measurement imprecision reduces the power of the study, thus diminishing the ability to detect treatment differences. In planning a surgical trial, it is useful to identify the potential sources of imprecision and plan strategies to minimise these. Such sources include the surgical procedure, the person performing the surgical procedure, the person responsible for follow-up care, instruments used to provide measurement of patient responses, and the person obtaining the follow-up measurements.

OUTCOME MEASURES

As in any clinical trial, the primary outcome measures for surgery trials are specific, applicable to all enrolled in the trial, clear, clinically meaningful, assessable in an unbiased way, achievable and feasible, and timely. A surgeon may, however, consider a procedure to be clinically successful (e.g. the hernia is repaired), while the patient may find the recovery period to be more difficult than he or she anticipated and the procedure has left him or her with unexpected chronic problems, such as pain or fatigue. Thus, it is important to assess patient-centred outcomes in addition to more traditional measures of safety and efficacy.

CLINICAL OUTCOMES

In many surgery trials, an outcome measure of death rates is an attractive choice – it is certainly specific, clinically meaningful, and quantified without controversy. It may, however, not be feasible or achievable within the time constraints of the trial (e.g. survival after procedures for cancer). In most surgery trials, however, death rates are not appropriate outcome measures. Examples of commonly used outcome measures are lengths of hospital stay, rates of complications or amputations, measurement of range of motion, enumeration of events such as arrhythmias, objective endoscopic criteria, and recurrence of disease process. These measures are useful when comparing one operation with another, or medical and surgical therapies. For example, in the trial comparing two types of hernia operation the primary outcome measure was recurrence of the hernia, a clear and readily assessed occurrence.

PATIENT-CENTRED OUTCOMES

Comparing an operative and non-operative treatment with single-outcome measures is more difficult. For example, to compare watchful waiting and operation for inguinal hernia, pain-limiting usual activities and physical function, as measured by the Physical Component Score of the RAND Medical Outcomes Study 36-item short-form health survey (SF-36),[23,24] were used. These measures could be applied to both cohorts of subjects, they were made hernia-specific by including items developed by patient and surgeon focus groups, each could be objectively assessed using standard tools, and they were clinically meaningful to the patient and to society.

Patient-centred, patient-reported outcomes have only recently been introduced into surgical trial design.[25,26] Health-related quality of life, measurements of pain as felt and reported by the patient (in contrast to measures where the numbers of pain pills consumed are counted), and the ability to perform activities of daily living, are now recognised as important in evaluating the results of surgical treatment. To be most valuable, however, the standard measurement tools such as the SF 36 should be amplified by disease-specific items. This approach was used in the two hernia management trials cited above. In the trial comparing one surgical technique with another, and in the trial comparing watchful waiting with operation, patient-centred outcomes were emphasised. The watchful waiting vs. operation trial used patient-centred outcomes as the primary outcomes, and hernia-specific items were used in both trials as secondary outcome measures of pain and ability to perform specified activities. The patient-centred terms for both of these trials were developed with focus groups of patients who had experienced one or the other procedure, and with their surgeons. These tools proved very useful in analysing the differences between the groups, in both the operation and non-operative cohorts.

COST AND COST-EFFECTIVENESS

Determination of cost-effectiveness[27] adds an important dimension to the principal questions of a surgical trial: (1) is the new procedure safe; (2) is it effective; (3) is it as good or better than existing technology; (4) is it cost-effective? Cost-effectiveness, which compares costs with quality-adjusted life years (QALY), has been measured in both of the hernia trials cited above. Of interest in the comparison of the two operative techniques was the finding that the laparoscopic technique was much more expensive to perform than the open procedure, but at two years, the impact on QALY was similar, and was superior in a certain class of hernia repairs. To study cost-effectiveness, the trial must include sequential collection of patient-centred data for which utility scores, which are generally based on patient preferences for different states of health, are used.[28]

THE ROLE OF A DSMB IN A SURGICAL CLINICAL TRIAL

Important in a surgical trial is the constitution of an independent board of experts to evaluate the safety and the integrity of the trial on a periodic basis. A Data and Safety Monitoring Board (DSMB) serves to review the data as they are collected during the trial and to determine primarily if the interventions are safe for the volunteers, and secondarily if there are inconsistencies in the data meriting concern and investigation. In one of the hernia trials, for example, the DSMB detected an excessive rate of surgical complications and unrealistically high recruitment at one of the sites, prompting an audit that led to deletion of this site and all data coming from it. The DSMB also reviews the analyses to determine if endpoints have been reached, so that a trial in which the data show effectiveness or failure of the new intervention can be stopped before the planned time, in order to make the most effective treatment available to all of the randomised volunteers.

PUBLICATION AND REGISTRATION OF SURGICAL CLINICAL TRIALS

Ethical demands for protection of volunteers in any trial require that the trial be designed to meet its endpoints, that the results of the trial be published regardless of the outcome, and that information about the trial, the opportunity to enrol as a volunteer, and its results are made readily available to the public. These goals are the focus of the clinicaltrials.gov initiative housed at the National Library of Medicine. Through this voluntary online registry, all parties to surgical care have access to trials from initiation through completion and publication. Editors of all major medical journals (International Committee of Medical Journal Editors) require, as of 13 September 2005, registration of a clinical trial and assignment of a trial number through the registry, before a manuscript describing the trial can be published.

CONCLUSION

Randomised clinical trials in surgery are urgently needed if the best interventions are to be provided to patients. Evidence of safety and effectiveness, rather than opinion or enthusiasm for a procedure, should be the determination of a treatment strategy. Especially when evidence is coupled with the surgeon's judgement of the needs of the particular patient and the surgeon's expertise, the patient's needs will be best served.

REFERENCES

1. *Lancet* (2005) **366**: cover – 122.
2. A prospective analysis of 1518 laparoscopic cholecystectomies. The Southern Surgeons Club. *New Engl J Med* (1991) **324**(16): 1073–78.
3. Archer SB, Brown DW, Smith CD, Branum GD, Hunter JG. Bile duct injury during laparoscopic cholecystectomy: results of a national survey. *Ann Surg* (2001) **234**(4): 549–58.
4. Bellows CF, Berger DH, Crass RA. Management of gallstones. *Am Fam Physician* (2005) **72**(4): 637–42.
5. Gholson CF, Sittig K, McDonald JC. Recent advances in the management of gallstones. *Am J Med Sci* (1994) **307**(4): 293–304.
6. Gallstones and laparoscopic cholecystectomy. *NIH Consens Statement* (1992) **10**(3): 1–28.
7. Pollock AV. Surgical evaluation at the crossroads. *Br J Surg* (1993) **80**: 964–66.
8. Wennberg JE. An apple a day? *New Engl J Med* (1994) **331**: 815 (letter).
9. van Hout BA, Al MJ, Gordon GS, Rutten FFH. Costs, effects, and c/e-ratios alongside a clinical trial. *Health Econ* (1994) **3**: 309–19.
10. Briggs A, Gray A. Using cost effectiveness information. *BMJ* (2000) **320**: 246.
11. Detsky AS, Naglie IG. A clinician's guide to cost-effectiveness analysis. *Ann Intern Med* (1990) **113**: 147–54.
12. The CONSORT statement. http://www.consort-statement.org/Statement/revisedstatement.htm. Accessed 18 November 2005).

13. Brennan TA, Rothman DJ, Blank L, Blumenthal D, Chimonas SC, Cohen JJ *et al*. Health industry practices that create conflicts of interest: a policy proposal for academic medical centers. *JAMA* (2006) **295**(4): 429–33.
14. Neumayer LA, Gawande AA, Wang J, Giobbie-Hurder A, Itani K, Fitzgibbons RJ Jr *et al*. Proficiency of surgeons in inguinal hernia repair: effect of experience and age. *Ann Surg* (2005) **242**: 344–8; discussion 348–52.
15. Majeed AW, Troy G, Nicholl JP *et al*. Randomized, prospective, single blind comparison of laparoscopic versus small-incision cholecystectomy. *Lancet* (1996) **347**: 989–94.
16. Moseley JB, O'Malley K, Petersen NJ, Menke TJ, Brody BA, Kuykendall DH *et al*. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *New Engl J Med* (2002) **347**(2): 81–8.
17. Freed CR, Greene PE, Breeze RE, Tsai WY, DuMouchel W, Kao R *et al*. Transplantation of embryonic dopamine neurons for severe Parkinson's disease. *New Engl J Med* (2001) **344**(10): 710–19.
18. Neumayer L, Giobbie-Hurder A, Jonasson O, Fitzgibbons R Jr, Dunlop D, Gibbs J *et al*. Open mesh versus laparoscopic mesh repair of inguinal hernia. *New Engl J Med* (2004) **350**(18): 1819–27.
19. Schwaitzberg SD, Jones DB, Grunwaldt L, Rattner DW. Laparoscopic hernia in the light of the Veterans Affairs Cooperative Study 456: more rigorous studies are needed. *Surg Endosc* (2005) **19**(9): 1288–9.
20. Wasson JH, Reda DJ, Brukewitz RG, Elinson J, Keller AM, Henderson WG *et al*. A comparison of transurethral surgery with watchful waiting for moderate symptoms of benign prostatic hyperplasia. *New Eng J Med* (1995) **332**: 75–9.
21. US Food and Drug Administration. Guidance for Institutional Review Boards and Clinical Investigators. 1998 Update: Payment to Research Subjects. http://www.fda.gov/oc/ohrt/irbs/toc4.html#payment. Accessed 31 January 2006.
22. Fitzgibbons RJ Jr, Giobbie-Hurder A, Gibbs JO, Dunlop DD, Reda DJ, McCarthy M Jr *et al*. Watchful waiting vs repair of inguinal hernia in minimally symptomatic men: a randomized clinical trial. *JAMA* (2006) **295**(3): 285–92.
23. Ware JE Jr, Sherbourne C. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Med Care* (1992) **30**: 473–483.
24. Ware JE Jr., Kosinski M, Bayliss MS, *et al*. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. *Med Care* (1995) **33**(Suppl): AS264–279.
25. McCarthy M Jr, Chang CH, Pickard AS, Hurder A, Price DD, Jonasson O *et al*. Visual analog scales for assessing surgical pain. *J Am Coll Surg* (2005) **201**: 246–50.
26. McCarthy M Jr, Jonasson O, Chang CH, Pickard AS, Hurder A, Gibbs JO *et al*. Assessment of patient functional status after surgery. *J Am Coll Surg* (2005) **201**: 171–9.
27. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds, *Cost Effectiveness in Health and Medicine*. New York: Oxford University Press (1996).
28. Nichol MB, Sengupta N, Globe DR. Evaluating quality-adjusted life years: estimation of the health utility index (HUI2) from the SF-36. *Med Decis Making* (2001) **21**: 105–12.

30

Plastic Surgery – Reconstructive Surgery

ERIK ANG[1] AND WOFFLES WU[2]
[1]Suite 1208, Mount Elizabeth Medical Centre, Singapore
[2]Suite 09-02, Camden Medical Centre, Singapore

INTRODUCTION

WHAT IS PLASTIC SURGERY?

Plastic surgery is derived from the Greek word *plastikos*, which means ‘to mould’. A plastic surgeon in essence moulds tissue, replaces what is lost, enhances what is deficient by shifting tissue both living and artificial. This explains the basic characteristic of plastic surgical procedures, which employs meticulous surgical techniques to mould and tailor the various tissues of the body either to fill a defect created by injuries or tumours, or to enhance the appearance of a certain part of the body.

Plastic surgery is all about wound healing, gentle tissue handling and restoration of form and function. Plastic surgeons are comfortable operating in virtually any part of the body and are trained to deal with a variety of surgical problems. These include reconstruction and restoration of congenital or developmental defects, defects arising from trauma, infection, or surgical extirpation of cancers. Basic plastic surgical techniques include gentle atraumatic tissue handling, wound closure with minimal tension, use of local and distant flaps, autogenous grafts and artificial implants, use of lasers to rejuvenate tissue or ablate lesions.

The Key activities in plastic surgery are:

- Wound healing
- Wound and defect coverage
- Shifting or rearrangement of tissue.

MAJOR SUBDIVISIONS IN PLASTIC SURGERY

The field of plastic surgery is vast.[1–3] Major subdivisions include management of burns,[4–8] reconstructive surgery,[9–13] craniomaxillofacial surgery[14–16] and aesthetic surgery (Table 30.1). Many clinical problems in each subdivision are dealt with by multi-disciplinary teams comprising plastic surgeons and a variety of other specialists, both clinical and para-clinical.[17,18] Many fields overlap with other surgical specialities.

Burns

The management of burn patients, especially those with major burn injuries, is a prime

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

Table 30.1. Major subdivisions in plastic surgery

Subspecialty	Common procedures
Burns	• Acute burn management • Post-burn reconstruction • Hypertrophic scar and keloids management
Craniomaxillofacial deformities/trauma and congenital	• Facial trauma correction • Craniomaxillofacial deformity correction • Cleft lip and palate correction • Congenital hand, facial and urogenital abnormality correction
Reconstructive surgery for cancer, trauma and infection	• Breast reconstruction following mastectomy • Head and neck reconstruction following cancer-clearing surgery • Upper limb and lower limb trauma • Post-infection reconstruction of any part of the body • Removal and reconstruction of skin cancers
Aesthetic surgery	• Facial rejuvenation – face lifts (rhytidectomy), fillers, Botolinium toxin therapy, lasers • Eyelid surgery (blepharoplasty) • Body contouring – liposculpture, tummy-tuck (abdominoplasty), buttock lift, arm lift, thigh lift • Breast enlargement/reduction/lift (augmentation mammaplasty, reduction mammaplasty, mastopexy) • Laser surgery – removal of moles, tattoos, resurfacing, rejuvenation

example of a condition requiring many specialists working in concert with each other. Depending on the geographical locality of practice, plastic surgeons will be involved in the acute care, resuscitation and early excision and grafting of burn wounds, and also the post-burn reconstruction procedures. In the Burn Centre at the Singapore General Hospital, administrative and clinical control is through the Department of Plastic Surgery. Anaesthetists are intimately involved in the critical care and airway management aspects in addition to the administration of general anaesthesia during surgical procedures that are carried out within the Burn Centre. Other specialties with involvement in burn management include infectious disease physicians, renal physicians, endocrinologists and general surgeons. Nursing and paraclinical services like physiotherapists and occupational therapists contribute greatly to the acute care and rehabilitation of the burn patients.

Clinical trials in burns of necessity reflect the multidisciplinary nature. Major areas of study include the following areas:

- Wound management
- Acute care and immune modulation
- Infection control
- Post-burn contracture reconstruction.

Craniomaxillofacial and Congenital

Deformities in the craniomaxillofacial region can be due to either congenital and developmental conditions, which manifest in infancy or childhood, or trauma, which occur mainly in adults.

Congenital and developmental conditions Apart from congenital deformities of the extremities, cleft lip and palate is one of the more common congenital deformities that occur. A patient with a cleft lip and palate, depending on the severity and involvement of the particular structures, will often be faced with a few years of surgery and rehabilitation. In infancy, closure of the cleft lip and palate is performed. As the child grows, he or she may require orthodontic treatment and in teenage years may need surgery for speech improvement and for orthognathic purposes. When the patient is older, he or she may require soft tissue correction such as lip revision and nose revision. A variety of techniques and variations of techniques are available to repair these deformities and there is no one vastly superior method over another method as such.

Less common congenital deformities involving the craniomaxillofacial skeleton require multiple surgeries as well. For example, in the case of craniosynostoses, early cranial vault remodeling surgery may be required to relieve intracranial pressure and to allow the rapidly growing infant brain to expand comfortably.

Most of the work in craniomaxillofacial surgery involves osteotomies, moving parts of the facial skeleton and fixation with implants. The type of implants that are used in the fixation is an example of some of the issues in this field of plastic surgery; that is, whether titanium plates or biodegradable plates are used.

Trauma Prior to modern-day craniofacial principles, patients with craniofacial trauma were treated by wiring or by conservative management. Current craniofacial principles of early internal reduction and operative fixation of facial fractures are well established. Clinical issues that concern plastic surgeons involved in facial trauma include choosing the type of implants for fixation, whether the implants are biodegradable or not, and the placement of the implants.

As with burn management, craniomaxillofacial cases are often managed by various different specialities, including plastic surgeons, oral surgeons and maxillofacial surgeons, either singly or in cooperation. In the case of congenital problems like cleft lip and palate patients, neonatologists, paediatricians, geneticists, orthodontists and speech therapists may be involved. Complex craniofacial problems may require neurosurgical input in addition to the other specialties.

Reconstruction

Reconstructive surgery in a nutshell is all about restoring tissue and structures to as near normal or pre-morbidity as possible. Proper tissue handling and wound care coupled with tension-free closure are essential to ensure proper wound closure and healing. When primary closure is not possible because the surgical wound is too large or the wound edges are infected or irradiated precluding proper wound healing, techniques like skin grafts and flaps are necessary. Reconstructive surgery transfers tissue from one part of the body to another part of the body. In oncology patients, this tissue is transferred to replace that which had been removed as part of the cancer-clearing operation. This tissue can be skin, muscle, fascia, bone, nerve, or a combination.

A graft is tissue that does not have a definable vascular pedicle. It depends on the recipient bed for its nourishment and survival. A flap is tissue that has a definable vascular pedicle which has to be transferred together with the tissue to the recipient site in order for it to survive. A flap can be either pedicled, where it is transferred to the donor site still attached to its vascular pedicle, or a free flap, where the flap tissue is detached from its original site and transferred to a distant recipient site, and the vascular pedicle is anastomosed using microsurgical techniques.

Muscle flaps like the rectus abdominis and latissimus dorsi flaps are needed to obliterate dead space, clear infection and protect major vessels. Skin flaps like the radial forearm and the anterior lateral thigh flaps are needed for post-burn contractures or for intra-oral lining. Bone flaps like the fibula are used to reconstruct bony defects: for example, after a portion of the mandible has been removed as part of a cancer-clearing operation.

Plastic surgeons reconstruct defects arising from any part of the integument and are sometimes involved in reconstruction of upper gastrointestinal and urogenital continuity following surgical ablation by oncology surgeons. Common reconstructive procedures include breast reconstruction following mastectomy, and reconstruction in the head and neck region with free flaps. In addition, a large part of a plastic surgeon's work is in the resection of skin cancers, especially in the head and neck region, followed by coverage either primarily or with local flaps or grafts.

RECONSTRUCTIVE AND AESTHETIC SURGERY

Plastic surgeons start off with general surgery training, learning basic surgical principles. Issues include tissue handling, anatomy, physiology, pathology and intensive care. The next stage of the training involves induction into a recognised plastic surgical training programme. The initial rotations usually include training in all forms of reconstruction. In our country, and in many institutions around the world, plastic surgeons look after and manage the burns unit. Burn care takes up a substantial portion of institutional practice. As the plastic surgeon progresses in his or her career, he or she may choose to specialise further in a particular field of interest. Some choose to leave for private practice and concentrate wholly on aesthetic surgery.

This section on plastic surgery is structured in such a way as to reflect this practical issue of the life-cycle of the plastic surgeon in practice. We have chosen to divide our two chapters into one on reconstructive surgery and the other on aesthetic surgery, according to the accepted definition adopted by the American Medical Association[19] which states:

> *Cosmetic* surgery is performed to reshape *normal* structures of the body in order to improve the patient's appearance and self-esteem. Whereas, *reconstructive* surgery is performed on *abnormal* structures of the body, caused by congenital defects, developmental abnormalities, trauma, infection, tumours or disease. Reconstructive surgery is generally performed to improve function, but may also be done to approximate a normal appearance.

The present chapter deals with the basic principles and spectrum of plastic surgery followed by clinical trials in plastic and reconstructive surgery, and the following chapter deals with aesthetic surgery. We have chosen this separation because of the often different and diverse issues concerned in conducting clinical trials associated with each type of surgery.

However, one must add that often there is a grey area between what is reconstructive and what is aesthetic. For example, in a repair of a cleft lip, on one hand it is reconstructing a defect, and on the other hand the lip repair needs to be done in an aesthetically pleasing manner so that the patient is able to function normally in a psychosocial setting. In an 'aesthetic' case of, say, a facelift, the plastic surgeon seeks to improve the aesthetic appearance by re-suspending the skin or deeper structures, whatever the technique he or she may choose, and this is done with principles taken from reconstructive surgery following craniofacial trauma.

SURVEY OF CLINICAL TRIALS IN PLASTIC AND RECONSTRUCTIVE SURGERY

A survey of existing clinical trials in plastic and reconstructive surgery published in the literature was undertaken to obtain an idea of the issues addressed by these trials. An internet-based PubMed Medline search was carried out with the search words 'plastic surgery' and 'reconstructive surgery'. The search parameters included papers between 1980 and 2004, with clinical studies conducted amongst human subjects published in the English language. This survey excluded case reports, case series, review articles, and tips and techniques.

Plastic surgery overlaps with many other subspecialties, so we have confined our search words to plastic and reconstructive surgery, and have not included other words like 'craniofacial surgery', 'cleft surgery', 'burn management' and site-specific reconstruction as these would be beyond the scope of our discussion.

The field of burn management is so vast, encompassing such issues like acute burn care, wound management, intensive care and burn reconstruction, and burn victims are cared for by many different clinical specialties, that our survey excluded the search words 'burn' and 'burn injury', but papers dealing with post-burn reconstruction, which would have come under the search words of 'reconstructive surgery' or 'plastic surgery', were included. In addition, papers that dealt primarily with anaesthetic issues amongst plastic and reconstructive surgical patients were excluded from our survey.

We grouped the papers that we found roughly into the following categories based on the title of the paper and the abstract:

- General plastic surgery
- Reconstructive surgery
- Aesthetic surgery.

GENERAL PLASTIC SURGERY

Two main topics that arose from this survey of clinical trials in general plastic surgery concern wound healing and closure, and infection control (Tables 30.2 and 30.3).

Wound Healing and Closure

Papers reporting on non-randomised trials on wound healing and closure looked at a variety of issues including comparing various types of suture materials,[20,21] and skin expansion devices.[22,23] One study compared coated polyglactin with irradiated polyglactin in ophthalmic plastic surgery work in 32 patients and concluded that the suture material was safe and effective to use.[21] Novel methods of skin closure were also investigated.[24,25]

As with the non-randomised trials section, there were a number of randomised controlled trials in the various aspects of wound closure and healing. One study looked at the role of epidermal growth factor in accelerating wound healing. This study by Brown *et al.*[26] looked at paired donor sites in 12 patients. Another study by Jeschke *et al.*[27] looked at a relatively new reconstructive techniques for the closure of large wounds. This paper investigated the utility of Integra with fibrin glue and negative pressure as compared with conventional management in post-injury wounds. The study was done on 12 patients. Nevertheless, the authors had an interesting result that showed that the new technique resulted in shorter hospital stay and earlier wound coverage.

The debate continues with regards to the superiority of one type of suture material over another, or suturing versus tissue adhesives in the closure of wounds, and several randomised trials were done to try to answer this question. A study of 20 patients compared the use of nylon with polydioxanone in the correction of rectus diastasis[28] and concluded that there was no significant difference between the two groups when measured for width of diastasis post-repair by CT scan.

A larger study comprising 111 patients looked at the aesthetic result and patient satisfaction of skin closure with tissue adhesive versus suturing and concluded that adhesive use yielded better results at one year follow-up. This study was well designed and carried out. It had strict enrolment and exclusion criteria, ensuring that the study groups were evenly matched.[29]

Infection Control

One of the difficulties in conducting clinical trials to determine the most effective way of infection control in plastic surgical patients is that the incidence of post-operative wound infection and bacteriology in plastic surgery depends on whether the operative field is clean, contaminated or dirty, and varies from centre to centre, from country to country and from season to season. Thus clinical trials that compare the use of prophylactic antibiotics versus no antibiotics in plastic surgery procedures may have limited application to other centres and institutions.

For example, a large randomised study of 1400 consecutive patients going for plastic surgical procedures over a six-year period in a Turkish centre concluded that there was no significant difference between those who had sulbactam–ampicillin antibiotic prophylaxis versus those who did not, for all the groups of patients that were studied.[30] A double-blinded randomised study of 339 patients over a nine month period in another centre in Norway showed that there was a significant reduction in wound infection

Table 30.2. Non-randomised clinical trials in general plastic surgery

Authors	Title of paper	Reference
Research in plastic surgery		
Velanovich V, Robson MC, Heggers JP, Smith DJ Jr, Koss N.	Statistical analysis and study design in plastic and reconstructive surgical research.	Plast Reconstr Surg (1987) **80**(2): 308–13.
Wound healing and closure		
Morgan WP, Harding KG, Hughes LE.	A comparison of skin grafting and healing by granulation, following axillary excision for hidradenitis suppurativa.	*Ann R Coll Surg Engl* (1983) **65**(4): 235–6.
Davenport M, Daly J, Harvey I, Griffiths RW.	The bolus tie-over 'pressure' dressing in the management of full thickness skin grafts. Is it necessary?	*Br J Plast Surg* (1988) **41**(1): 28–32.
Bang RL, Mustafa MD.	Comparative study of skin wound closure with polybutester (Novafil) and polypropylene	*J R Coll Surg Edinb* (1989) **34**(4): 205–7.
Keng TM, Bucknall TE.	A clinical trial of tissue adhesive (histoacryl) in skin closure of groin wounds.	*Med J Malaysia* (1989) **44**(2): 122–8.
Blomqvist G, Steenfos H.	A new partly external device for extension of skin before excision of skin defects.	*Scand J Plast Reconstr Surg Hand Surg* (1993) **27**(3): 179–82.
Cruz-Korchin NI.	Effectiveness of silicone sheets in the prevention of hypertrophic breast scars.	*Ann Plast Surg* (1996) **37**(4): 345–8.
Pizzorno R, Bonini F, Donelli A, Stubinski R, Medica M, Carmignani G.	Hyperbaric oxygen therapy in the treatment of Fournier's disease in 11 male patients.	*J Urol* (1997) **158**(3 Pt 1): 837–40.
Eaglstein WH, Alvarez OM, Auletta M, Leffel D, Rogers GS, Zitelli JA *et al*	Acute excisional wounds treated with a tissue-engineered skin (Apligraf).	*Dermatol Surg* (1999) **25**(3): 195–201.
Yamaguchi Y, Kubo T, Tarutani M, Sano S, Asada H, Kakibuchi M *et al*	Epithelial-mesenchymal interactions in wounds: treatment of palmoplantar wounds by nonpalmoplantar pure epidermal sheet grafts	*Arch Dermatol* (2001) **137**(5): 621–8.
Khouri RK, Schlenz I, Murphy BJ, Baker TJ.	Nonsurgical breast enlargement using an external soft-tissue expansion system	*Plast Reconstr Surg* (2000) **105**(7): 2500–12; discussion 513–14.
Talbot AW, Meadows AE, Tyers AG, Shah-Desai S.	Use of 7/0 Vicryl (coated polyglactin 910) and 7/0 Vicryl-rapide (irradiated polyglactin 910) in skin closure in ophthalmic plastic surgery.	*Orbit* (2002) **21**(1): 11–8.
Ferraro GA, Corcione A, Nicoletti GF, Brongo S, Ciccarelli F, D'Andrea F.	The use of recombinant human erythropoietin stimulating factor in plastic surgery.	*Aesthetic Plast Surg* (2004) **28**(3): 174–6.
Infection control		
Exner K, Lang E, Borsche A, Lemperle G.	Efficacy, tolerability and pharmacokinetics of teicoplanin in patients undergoing breast surgery.	*Eur J Surg Suppl* (1992) **567**: 33–8.
Marin-Bertolin S, Gonzalez-Martinez R, Gimenez CN, Marquina Vila P, Amorrortu-Velayos J.	Does double gloving protect surgical staff from skin contamination during plastic surgery?	*Plast Reconstr Surg* (1997) **99**(4): 956–60.

Table 30.3. Randomised controlled clinical trials in general plastic surgery

Authors	Title of paper	Reference
Outcome analysis and issues		
Talmor M, Hydo LJ, Shaikh N, Gayle LB, Hoffman LA, Barie PS.	Clinical features and outcome of patients admitted to the intensive care unit after plastic surgical procedures: implications for cost reduction and quality of care.	*Ann Plast Surg* (1997) **39**(1) 74–9.
Armstrong AP, Cole AA, Page RE.	Informed consent: are we doing enough?	*Br J Plast Surg* (1997) **50**(8): 637–40.
Kokoska MS, Currens JW, Hollenbeak CS, Thomas JR, Stack BC Jr.	Digital vs 35-mm photography. To convert or not to convert?	*Arch Facial Plast Surg* (1999) **1**(4): 276–81.
Marcus JR, Few JW, Chao JD, Fine NA, Mustoe TA.	The prevention of emesis in plastic surgery: a randomized, prospective study.	*Plast Reconstr Surg* (2002) **109**(7): 2487–94.
Wound healing and closure		
Brown GL, Nanney LB, Griffen J, Cramer AB, Yancey JM, Curtsinger LJ 3rd *et al*.	Enhancement of wound healing by topical treatment with epidermal growth factor.	*New Engl J Med* (1989) **321**(2): 76–9.
Davenport M, Daly J, Harvey I, Griffiths RW.	The bolus tie-over 'pressure' dressing in the management of full thickness skin grafts. Is it necessary?	*Br J Plast Surg* (1988) **41**(1): 28–32.
Michie DD, Hugill JV.	Influence of occlusive and impregnated gauze dressings on incisional healing: a prospective, randomized, controlled study.	*Ann Plast Surg* (1994) **32**(1): 57–64.
Jansen DA, Gailliot RV Jr, Galli RA, Escobar JR, Kind G, Parry SW.	An evaluation of fascial staples (a new technique) in wide fascial plication during reconstructive abdominoplasty.	*Ann Plast Surg* (1996) **36**(2): 171–5.
Toriumi DM, O'Grady K, Desai D, Bagal A.	Use of octyl-2-cyanoacrylate for skin closure in facial plastic surgery.	*Plast Reconstr Surg* (1998) **102**(6) 2209–19.
Niessen FB, Spauwen PH, Robinson PH, Fidler V, Kon M.	The use of silicone occlusive sheeting (Sil-K) and silicone occlusive gel (Epiderm) in the prevention of hypertrophic scar formation.	*Plast Reconstr Surg* (1998) **102**(6): 1962–72.
Rossmann JA, Rees TD.	A comparative evaluation of hemostatic agents in the management of soft tissue graft donor site bleeding.	*J Periodontol* (1999) **70**(11): 1369–75.
Nahas FX, Augusto SM, Ghelfond C.	Nylon versus polydioxanone in the correction of rectus diastasis.	*Plast Reconstr Surg* (2001) **107**(3) 700–6.
Ozturan O, Miman MC, Aktas D, Oncel S.	Butylcyanoacrylate tissue adhesive for columellar incision closure.	*J Laryngol Otol* (2001) **115**(7): 535–40.
van Schie CH, Whalley A, Armstrong DG, Vileikyte L, Boulton AJ.	The effect of silicone injections in the diabetic foot on peak plantar pressure and plantar tissue thickness: a 2-year follow-up.	*Arch Phys Med Rehabil* (2002) **83**(7): 919–23.
Ginandes C, Brooks P, Sando W, Jones C, Aker J.	Can medical hypnosis accelerate post-surgical wound healing? Results of a clinical trial.	*Am J Clin Hypn* (2003) **45**(4): 333–51.
Abidia A, Laden G, Kuhan G, Johnson BF, Wilkinson AR, Renwick PM *et al*.	The role of hyperbaric oxygen therapy in ischaemic diabetic lower extremity ulcers: a double-blind randomised-controlled trial.	*Eur J Vasc Endovasc Surg* (2003) **25**(6): 513–8.

(*continued overleaf*)

Table 30.3. (*continued*)

Authors	Title of paper	Reference
Jeschke MG, Rose C, Angele P, Fuchtmeier B, Nerlich MN, Bolder U.	Development of new reconstructive techniques: use of Integra in combination with fibrin glue and negative-pressure therapy for reconstruction of acute and chronic wounds.	*Plast Reconstr Surg* (2004) **113**(2): 525–30.
Infection control		
Schiotz HA, Malme PA, Tanbo TG.	Urinary tract infections and asymptomatic bacteriuria after vaginal plastic surgery. A comparison of suprapubic and transurethral catheters.	*Acta Obstet Gynecol Scand* (1989) **68**(5): 453–5.
Amland PF, Andenaes K, Samdal F, Lingaas E, Sandsmark M, Abyholm F, Giercksky KE.	A prospective, double-blind, placebo-controlled trial of a single dose of azithromycin on postoperative wound infections in plastic surgery.	*Plast Reconstr Surg* (1995) **96**(6): 1378–83.
Andenaes K, Amland PF, Lingaas E, Abyholm F, Samdal F, Giercksky KE.	A prospective, randomized surveillance study of postoperative wound infections after plastic surgery: a study of incidence and surveillance methods.	*Plast Reconstr Surg* (1995) **96**(4): 948–56.
Andenaes K, Lingaas E, Amland PF, Giercksky KE, Abyholm F.	Preoperative bacterial colonization and its influence on postoperative wound infections in plastic surgery.	*J Hosp Infect* (1996) **34**(4): 291–9.
Baran CN, Sensoz O, Ulusoy MG.	Prophylactic antibiotics in plastic and reconstructive surgery.	*Plast Reconstr Surg* (1999) **103**(6): 1561–6.
Berger RS, Pappert AS, Van Zile PS, Cetnarowski WE.	A newly formulated topical triple-antibiotic ointment minimizes scarring.	*Cutis* (2000) **65**(6): 401–4. Erratum in: *Cutis* (2000) **66**(5): 382.
Huether MJ, Griego RD, Brodland DG, Zitelli JA.	Clindamycin for intraincisional antibiotic prophylaxis in dermatologic surgery.	*Arch Dermatol* (2002) **138**(9): 1145–8.

in breast surgery and flap surgery but not secondary cleft surgery when the patients were given azithromycin compared with those who were not.[31] Though the former cited study had a large number of patients, and the latter-cited study had a good study design, their findings may not be universally applicable as bacterial flora and sensitivities change with time and geographical location.

RECONSTRUCTIVE PLASTIC SURGERY

There are a number of papers reporting on clinical trials concerning reconstruction (Tables 30.4 and 30.5). A fair number of these trials concerned principles of reconstruction and breast reconstruction. Some clinical trials on burn and craniomaxillofacial reconstruction were identified. There were papers reporting on non-randomised clinical trials in head and neck reconstruction and limb soft tissue reconstruction, but none of these trials were randomised prospective ones.

Principles of Reconstruction

Non-randomised trials concerning principles of reconstruction dealt with a variety of topics, ranging from sequelae of flap transfer like muscle diameter and sensory recovery with re-innervation, to the use of a novel method to cover donor sites.[32]

This short survey of clinical trials in plastic and reconstructive surgery is by no means exhaustive as the search words were just ‘plastic surgery’ and ‘reconstructive surgery’, but it aims to point

Table 30.4. Non-randomised clinical trials in reconstructive plastic surgery

Authors	Title of paper	Reference
Principles of reconstruction		
Satoh K, Shigehara T.	Clinical trial of a prefabricated secondary hypogastric flap pedicled on the deep inferior epigastric vessel with or without a tissue expander in three patients.	*Plast Reconstr Surg* (1995) **96**(4): 905–11.
Acosta AE.	Clinical parameters of tumescent anesthesia in skin cancer reconstructive surgery. A review of 86 patients.	*Arch Dermatol* (1997) **133**(4): 451–4.
Kauhanen MS, Salmi AM, von Boguslawsky EK, Leivo IV, Asko-Seljavaara SL.	Muscle fiber diameter and muscle type distribution following free microvascular muscle transfers: a prospective study.	*Microsurgery* (1998) **18**(2): 137–44.
Schultes G, Gaggl A, Karcher H.	Reconstruction of accessory nerve defects with vascularized long thoracic vs. non-vascularized thoracodorsal nerve.	*J Reconstr Microsurg* (1999) **15**(4): 265–70; discussion 270–1.
Agarwal R, Agarwal S, Chandra R.	The lateral pectoral flap.	*J Hand Surg [Br]* (1999) **24**(5): 542–6.
Schultes G, Karcher H, Gaggl A.	Histologic and clinical results of reinnervation of the latissimus dorsi transfer with the thoracodorsal nerve.	*J Reconstr Microsurg* (1999) **15**(8): 567–72.
Netscher D, Armenta AH, Meade RA, Alford EL.	Sensory recovery of innervated and non-innervated radial forearm free flaps: functional implications.	*J Reconstr Microsurg* (2000) **16**(3): 179–85.
Sinha UK, Shih C, Chang K, Rice DH.	Use of AlloDerm for coverage of radial forearm free flap donor site.	*Laryngoscope* (2002) **112**(2): 230–4.
Akan M, Yildirim S, Misirlioglu A, Ulusoy G, Akoz T, Avci G.	An alternative method to minimize pain in the split-thickness skin graft donor site.	*Plast Reconstr Surg* (2003) **111**(7): 2243–9.
Burn reconstruction		
Kalaja E.	Acute excision or exposure treatment? Secondary reconstructions and functional results.	*Scand J Plast Reconstr Surg* (1984) **18**(1): 95–9.
Soejima K, Nozaki M, Sasaki K, Takeuchi M, Negishi N.	Reconstruction of burn deformity using artificial dermis combined with thin split-skin grafting.	*Burns* (1997) **23**(6): 501–4.
van Zuijlen PP, van Trier AJ, Vloemans JF, Groenevelt F, Kreis RW, Middelkoop E.	Graft survival and effectiveness of dermal substitution in burns and reconstructive surgery in a one-stage grafting model.	*Plast Reconstr Surg* (2000) **106**(3): 615–23.
Jang YC, Kwon OK, Lee JW, Oh SJ.	The optimal management of pediatric steam burn from electric rice-cooker: STSG or FTSG?	*J Burn Care Rehabil* (2001) **22**(1): 15–20.
Celikoz B, Deveci M, Duman H, Nsanci M.	Reconstruction of facial defects and burn scars using large size freehand full-thickness skin graft from lateral thoracic region.	*Burns* (2001) **27**(2): 174–8.
van Zuijlen PP, Vloemans JF, van Trier AJ, Suijker MH, van Unen E, Groenevelt F *et al*.	Dermal substitution in acute burns and reconstructive surgery: a subjective and objective long-term follow-up.	*Plast Reconstr Surg* (2001) **108**(7): 1938–46
van Zuijlen PP, Lamme EN, van Galen MJ, van Marle J, Kreis RW, Middelkoop E.	Long-term results of a clinical trial on dermal substitution. A light microscopy and Fourier analysis based evaluation.	*Burns* (2002) **28**(2): 151–60.

(*continued overleaf*)

Table 30.4. (*continued*)

Authors	Title of paper	Reference
Breast reconstruction		
May JW Jr, Bucky LP, Sohoni S, Ehrlich HP.	Smooth versus textured expander implants: a double-blind study of capsule quality and discomfort in simultaneous bilateral breast reconstruction patients.	*Ann Plast Surg* (1994) **32**(3): 225–32; discussion 232–3.
Kroll SS, Miller MJ, Schusterman MA, Reece GP, Singletary SE, Ames F.	Rationale for elective contralateral mastectomy with immediate bilateral reconstruction.	*Ann Surg Oncol* (1994) **1**(6): 457–61.
Modena S, Benassuti C, Marchiori L, Mainente M, Zanza A, Perus G *et al*.	Mastectomy and immediate breast reconstruction: oncological considerations and evaluation of two different methods relating to 88 cases	*Eur J Surg Oncol* (1995) **21**(1): 36–41.
Williams JK, Bostwick J 3rd, Bried JT, Mackay G, Landry J, Benton J.	TRAM flap breast reconstruction after radiation treatment.	*Ann Surg* (1995) **221**(6): 756–64; discussion 764–6. Review.
Evans GR, Schusterman MA, Kroll SS, Miller MJ, Reece GP, Robb GL, Ainslie N.	Reconstruction and the radiated breast: is there a role for implants?	*Plast Reconstr Surg* (1995) **96**(5): 1111–15; discussion, 1116–18.
Suominen S, Asko-Seljavaara S, von Smitten K, Ahovuo J, Sainio P, Alaranta H.	Sequelae in the abdominal wall after pedicled or free TRAM flap surgery.	*Ann Plast Surg* (1996) **36**(6): 629–36.
Ribuffo D, Muratori L, Antoniadou K, Fanini F, Martelli E, Marini M *et al*.	A hemodynamic approach to clinical results in the TRAM flap after selective delay.	*Plast Reconstr Surg* (1997) **99**(6): 1706–14.
Blondeel N, Vanderstraeten GG, Monstrey SJ, Van Landuyt K, Tonnard P, Lysens R *et al*.	The donor site morbidity of free DIEP flaps and free TRAM flaps for breast reconstruction.	*Br J Plast Surg* (1997) **50**(5): 322–30.
Gabka CJ, Baumeister RG, Maiwald G.	Advancements of breast conserving therapy by onco-plastic surgery in the management of breast cancer.	*Anticancer Res* (1998) **18**(3C): 2219–24.
Mahdi S, Jones T, Nicklin S, McGeorge DD.	Expandable anatomical implants in breast reconstructions: a prospective study.	*Br J Plast Surg* (1998) **51**(6): 425–30.
Blondeel PN, Demuynck M, Mete D, Monstrey SJ, Van Landuyt K, Matton G, Vanderstraeten GG.	Sensory nerve repair in perforator flaps for autologous breast reconstruction: sensational or senseless?	*Br J Plast Surg* (1999) **52**(1): 37–44.
Blondeel PN.	One hundred free DIEP flap breast reconstructions: a personal experience	*Br J Plast Surg* (1999) **52**(2): 104–11.
Raposio E, Santi PL.	Topical application of DMSO as an adjunct to tissue expansion for breast reconstruction.	*Br J Plast Surg* (1999) **52**(3): 194–7.
Coombs NJ, Royle GT.	How to draw the skin ellipse for a mastectomy.	*Ann R Coll Surg Engl* (1999) **81**(4): 248–50.
Caffo O, Cazzolli D, Scalet A, Zani B, Ambrosini G, Amichetti M *et al*.	Concurrent adjuvant chemotherapy and immediate breast reconstruction with skin expanders after mastectomy for breast cancer.	*Breast Cancer Res Treat* (2000) **60**(3): 267–75.
Contant CM, Menke-Pluijmers MB, Seynaeve C, Meijers-Heijboer EJ, Klijn JG, Verhoog LC *et al*.	Clinical experience of prophylactic mastectomy followed by immediate breast reconstruction in women at hereditary risk of breast cancer (HB(O)C) or a proven BRCA1 and BRCA2 germ-line mutation	*Eur J Surg Oncol* (2002) **28**(6): 627–32.

Table 30.4. (*continued*)

Authors	Title of paper	Reference
Hauben DJ, Shulman O, Levi Y, Sulkes J, Amir A, Silfen R.	Use of the SpaceMaker balloon in sternal wound closure: comparison with other techniques	*Plast Reconstr Surg* (2001) **108**(6): 1582–8; discussion 1589–90.
Hayes AJ, Jenkins MP, Sandhu SS, Baum M.	Subpectoral breast reconstruction using the biodimensional system.	*Ann R Coll Surg Engl* (1997) **79**(5): 355–60.
Kitamura K, Ishida M, Inoue H, Kinoshita J, Hashizume M, Sugimachi K.	Early results of an endoscope-assisted subcutaneous mastectomy and reconstruction for breast cancer	*Surgery* (2002) **131**(1 Suppl): S324–9.
Ichioka S, Nakatsuka T, Ohura N, Sato Y, Harii K.	Clinical use of amrinone (a selective phosphodiesterase III inhibitor) in reconstructive surgery	*Plast Reconstr Surg* (2001) **108**(7): 1931–7.
Baroody M, Tameo MN, Dabb RW.	Efficacy of the pain pump catheter in immediate autologous breast reconstruction.	*Plast Reconstr Surg* (2004) **114**(4): 895–8; discussion 899–900.
Layeeque R, Hochberg J, Siegel E, Kunkel K, Kepple J, Henry-Tillman RS *et al.*	Botulinum toxin infiltration for pain control after mastectomy and expander reconstruction.	*Ann Surg* (2004) **240**(4): 608–13; discussion 613–14.
Lee JW, Chang TW.	Extended latissimus dorsi musculocutaneous flap for breast reconstruction: experience in Oriental patients	*Br J Plast Surg* (1999) **52**(5): 365–72.
Lee SJ, Lim J, Tan WT, Baliarsing A, Iau PT, Tan LK, Lim TC.	Changes in the local morphology of the rectus abdominis muscle following the DIEP flap: an ultrasonographic study.	*Br J Plast Surg* (2004) **57**(5): 398–405.
Head and neck and facial reconstruction		
McKinney P, Pandya S.	Use of pubic fat as a graft for eyelid defects.	*Aesthetic Plast Surg* (1994) **18**(4): 383–5.
Inigo F, Chapa P, Jimenez Y, Arroyo O.	Surgical treatment of lagophthalmos in facial palsy: ear cartilage graft for elongating the levator palpebrae muscle.	*Br J Plast Surg* (1996) **49**(7): 452–6.
Roumanas ED, Markowitz BL, Lorant JA, Calcaterra TC, Jones NF, Beumer J 3rd.	Reconstructed mandibular defects: fibula free flaps and osseointegrated implants.	*Plast Reconstr Surg* (1997) **99**(2): 356–65.
Wei FC, Lutz BS, Chen HC, Tsai MH, Lin PY.	Free transverse colon transplantation for functional reconstruction of intra-oral lining: a clinical and histologic study.	*Plast Reconstr Surg* (1998) **102**(7): 2346–51.
Matsumoto K, Nakanishi H, Urano Y, Kubo Y, Nagae H.	Lower eyelid reconstruction with a cheek flap supported by fascia lata.	*Plast Reconstr Surg* (1999) **103**(6): 1650–4.
Kovacs AF.	The fate of osseointegrated implants in patients following oral cancer surgery and mandibular reconstruction.	*Head Neck* (2000) **22**(2): 111–19.
Munoz Guerra MF, Gias LN, Rodriguez Campo FJ, Diaz Gonzalez FJ.	Vascularized free fibular flap for mandibular reconstruction: a report of 26 cases.	*J Oral Maxillofac Surg* (2001) **59**(2): 140–4.
Wada T, Okamoto K, Nakanishi Y, Nakano H, Iwagami Y, Morita N.	Myofascial flap without skin for intra-oral reconstruction. 2: Clinical studies.	*Int J Clin Oncol* (2001) **6**(3): 143–8.
Lee JH, Kim MJ, Choi WS, Yoon PY, Ahn KM, Myung H *et al.*	Concomitant reconstruction of mandibular basal and alveolar bone with a free fibular flap.	*Int J Oral Maxillofac Surg* (2004) **33**(2): 150–6.

(*continued overleaf*)

Table 30.4. (*continued*)

Authors	Title of paper	Reference
Gok A, Erkutlu I, Alptekin M, Kanlikama M.	Three-layer reconstruction with fascia lata and vascularized pericranium for anterior skull base defects.	*Acta Neurochir (Wien)* (2004) **146**(1): 53–6; discussion 56–7.
Wong TY, Chung CH, Huang JS, Chen HA.	The inverted temporalis muscle flap for intraoral reconstruction: its rationale and the results of its application.	*J Oral Maxillofac Surg* (2004) **62**(6): 667–75.
Craniomomaxillofacial reconstruction		
Karesh JW.	Polytetrafluoroethylene as a graft material in ophthalmic plastic and reconstructive surgery. An experimental and clinical study.	*Ophthal Plast Reconstr Surg* (1987) **3**(3): 179–85.
Cheney ML, Gliklich RE.	The use of calvarial bone in nasal reconstruction.	*Arch Otolaryngol Head Neck Surg* (1995) **121**(6): 643–8.
Childress CS, Newlands SD.	Utilization of panoramic radiographs to evaluate short-term complications of mandibular fracture repair.	*Laryngoscope* (1999) **109**(8): 1269–72.
Friedman CD, Costantino PD, Synderman CH, Chow LC, Takagi S.	Reconstruction of the frontal sinus and frontofacial skeleton with hydroxyapatite cement.	*Arch Facial Plast Surg* (2000) **2**(2): 124–9.
Choung PH, Kim SG.	The coronoid process for paranasal augmentation in the correction of midfacial concavity.	*Oral Surg Oral Med Oral Pathol Oral Radiol Endod* (2001) **91**(1): 28–33.
Staffel JG.	Optimizing treatment of nasal fractures.	*Laryngoscope* (2002) **112**(10): 1709–19.
Wong GB, Burvin R, Mulliken JB.	Resorbable internal splint: an adjunct to primary correction of unilateral cleft lip-nasal deformity.	*Plast Reconstr Surg* (2002) **110**(2): 385–91.
Schopper C, Moser D, Sabbas A, Lagogiannis G, Spassova E, Konig F *et al.*	The fluorohydroxyapatite (FHA) FRIOS Algipore is a suitable biomaterial for the reconstruction of severely atrophic human maxillae.	*Clin Oral Implants Res* (2003) **14**(6): 743–9.
Chiarini L, Figurelli S, Pollastri G, Torcia E, Ferrari F, Albanese M, Nocini PF.	Cranioplasty using acrylic material: a new technical procedure.	*J Craniomaxillofac Surg* (2004) **32**(1): 5–9.
Iannetti G, Cascone P, Saltarel A, Ettaro G.	Le Fort I in cleft patients: 20 years' experience.	*J Craniofac Surg* (2004) **15**(4): 662–9.
Trunk reconstruction		
Nakajima H, Chang H.	A new method of reconstruction for pectus excavatum that preserves blood supply and costal cartilage.	*Plast Reconstr Surg* (1999) **103**(6): 1661–6.
Lardinois D, Muller M, Furrer M, Banic A, Gugger M, Krueger T, Ris HB.	Functional assessment of chest wall integrity after methylmethacrylate reconstruction.	*Ann Thorac Surg* (2000) **69**(3): 919–23.
Daphan C, Tekelioglu MH, Sayilgan C.	Limberg flap repair for pilonidal sinus disease.	*Dis Colon Rectum* (2004) **47**(2): 233–7.

Table 30.4. (*continued*)

Authors	Title of paper	Reference
Limb reconstruction		
Paolini A, Ruggieri M, Leone Sossi FL, Paolini G, Dal Pra G, Scuderi N.	Pectus excavatum in adults: destructive surgery or simple correction of an aesthetic defect?	*Riv Eur Sci Med Farmacol* (1996) **18**(1): 11–17.
Hertel R, Lambert SM, Muller S, Ballmer FT, Ganz R.	On the timing of soft-tissue reconstruction for open fractures of the lower leg.	*Arch Orthop Trauma Surg* (1999) **119**(1–2): 7–12.
Fayomi O, Patel JV, Percival N.	Soft tissue cover for the exposed knee prosthesis.	*Int Orthop* (1999) **23**(1): 51–2.
Lutz BS, Wei FC, Machens HG, Rhode U, Berger A.	Indications and limitations of angiography before free-flap transplantation to the distal lower leg after trauma: prospective study in 36 patients.	*J Reconstr Microsurg* (2000) **16**(3): 187–91; discussion 192.
Meyer C, Hartmann B, Horas U, Kilian O, Heiss C, Schnettler R.	Reconstruction of the lower leg with the sural artery flap.	*Langenbecks Arch Surg* (2002) **387**(7–8): 320–5.
Minami A, Kato H, Suenaga N, Iwasaki N.	Distally-based free vascularized tissue grafts in the lower leg.	*J Reconstr Microsurg* (1999) **15**(7): 495–9.
Penington AJ, Mallucci P.	Closure of elective skin defects in the leg with a fasciocutaneous V-Y island flap.	*Br J Plast Surg* (1999) **52**(6): 458–61.

out the various issues that are involved in conducting clinical trials in this speciality.

The issues are:

- The subject is vast, with considerable overlap with other specialities, and literally covers the entire body.
- Many conditions are complex and require multidisciplinary teams.
- The endpoint is often subjective, related to quality of life issues and aesthetic outcome, unlike trials on pharmaceuticals where there are measurable endpoints like control of blood pressure for antihypertensives, disease-free interval for chemotherapeutic agents, or lowering of blood cholesterol for antilipid medication.
- The patient numbers are small. Many studies have modest numbers of patients accrued.

ROLE OF CLINICAL TRIALS IN PLASTIC SURGERY

Randomised controlled clinical trials are touted to provide the highest level of scientific-based evidence for clinical practice in many branches of medicine and surgery. However, in plastic and reconstructive surgery, there are other forms of clinical studies that are just as important, if not more, when shaping clinical practice. Plastic surgeons are very much like artists and tailors. Often there are multiple solutions to a particular problem or defect. Different patients with different expectations and life styles demand different solutions.[33]

How, then, does a new technique in reconstructive surgery find its way into clinical practice? Initial work is done in the laboratory, followed by animal and cadaveric studies before a technique is translated into clinical practice.[34,35] The vasculature and anatomy of a particular flap are studied in cadavers. This is followed by identifying suitable patients and then employing the new flap on these patients. These are than written up as case reports, clinical series or reviews. Centres with large numbers of suitable patients will then popularise the method. An example is the radial forearm which was first used by the Chinese surgeon Song.[36] This

Table 30.5. Randomised controlled clinical trials in reconstructive plastic surgery

Authors	Title of Paper	Reference
Principles of reconstruction		
Lundeberg T, Kjartansson J, Samuelsson U.	Effect of electrical nerve stimulation on healing of ischaemic skin flaps.	*Lancet* (1988) **2**(8613): 712–14.
Blomqvist L, Malm M, Berg A, Svelander L, Kleinau S.	The inflammatory reaction in elective flap surgery.	*Plast Reconstr Surg* (1998) **101**(6): 1524–8.
Blomqvist L, Rojdmark JS, Malm M.	Serum creatine kinase in fasciocutaneous and musculocutaneous flap surgery.	*Ann Plast Surg* (1997) **39**(5): 532–5.
Jenkins M, Alexander JW, MacMillan BG, Waymack JP, Kopcha R.	Failure of topical steroids and vitamin E to reduce postoperative scar formation following reconstructive surgery.	*J Burn Care Rehabil* (1986) **7**(4): 309–12.
Khouri RK, Sherman R, Buncke HJ Jr, Feller AM, Hovius S, Benes CO *et al*.	A phase II trial of intraluminal irrigation with recombinant human tissue factor pathway inhibitor to prevent thrombosis in free flap surgery.	*Plast Reconstr Surg* (2001) **107**(2): 408–15; discussion 416–18.
Mohmand MH, Sterne GD, Gowar JP.	Home inflation of tissue expanders: a safe and reliable alternative.	*Br J Plast Surg* (2001) **54**(7): 610–4.
Burn reconstruction		
Alexander JW, MacMillan BG, Law EJ, Krummel R.	Prophylactic antibiotics as an adjunct for skin grafting in clean reconstructive surgery following burn injury.	*J Trauma* (1982) **22**(8): 687–90.
Michie DD, Hugill JV.	Influence of occlusive and impregnated gauze dressings on incisional healing: a prospective, randomized, controlled study.	*Ann Plast Surg* (1994) **32**(1):57–64.
Rubegni P, De Aloe G, Mazzatenta C, Cattarini L, Fimiani M.	Clinical evaluation of the trophic effect of polydeoxyribonucleotide (PDRN) in patients undergoing skin explants. A pilot study.	*Curr Med Res Opin* (2001) **17**(2): 128–31.
Pannier M, Martinot V, Castede JC, Guitard J, Robert M, Le Touze A *et al*.	Efficacy and tolerance of Algosteril (calcium alginate) versus Jelonet (paraffin gauze) in the treatment of scalp graft donor sites in children. Results of a randomized study	*Ann Chir Plast Esthet* (2002) **47**(4): 285–90.
Breast reconstruction		
Sinow JD, Cunningham BL.	Intraluminal lidocaine for analgesia after tissue expansion: a double-blind prospective trial in breast reconstruction.	*Ann Plast Surg* (1992) **28**(4): 320–5.
Wickman M, Johansson O, Forslind B.	Dimensions of capsular collagen fibrils: image analysis of rapid compared with slow tissue expansion for breast reconstruction.	*Scand J Plast Reconstr Surg Hand Surg* (1992) **26**(3): 281–5.
Foo IT, Coleman DJ, Holmes JD, Palmer JH, Sharpe DT.	Delay between expansion and expander/implant exchange in breast reconstruction – a prospective study.	*Br J Plast Surg* (1992) **45**(4): 279–83.
Wickman M, Olenius M, Malm M, Jurell G, Serup J.	Alterations in skin properties during rapid and slow tissue expansion for breast reconstruction.	*Plast Reconstr Surg* (1992) **90**(6): 945–50.
Wickman M.	Comparison between rapid and slow tissue expansion in breast reconstruction.	*Plast Reconstr Surg* (1993) **91**(4): 663–70; discussion 671–2.

Table 30.5. (*continued*)

Authors	Title of Paper	Reference
Thuesen B, Siim E, Christensen L, Schroder M.	Capsular contracture after breast reconstruction with the tissue expansion technique. A comparison of smooth and textured silicone breast prostheses.	*Scand J Plast Reconstr Surg Hand Surg* (1995) **29**(1): 9–13.
Wickman M.	Rapid versus slow tissue expansion for breast reconstruction: a three-year follow-up.	*Plast Reconstr Surg* (1995) **95**(4): 712–18.
Brandberg Y, Malm M, Rutqvist LE, Jonsson E, Blomqvist L.	A prospective randomised study (named SVEA) of three methods of delayed breast reconstruction. Study design, patients preoperative problems and expectations.	*Scand J Plast Reconstr Surg Hand Surg* (1999) **33**(2): 209–16.
Tuominen HP, Svartling NE, Tikkanen IT, Asko-Seljavaara S.	The effect of felodipine on endothelin-1 levels, peripheral vasoconstriction and flap survival during microvascular breast reconstruction.	*Br J Plast Surg* (1997) **50**(8): 624–31.
Curran D, van Dongen JP, Aaronson NK, Kiebert G, Fentiman IS, Mignolet F, Bartelink H.	Quality of life of early-stage breast cancer patients treated with radical mastectomy or breast-conserving procedures: results of EORTC Trial 10801. The European Organization for Research and Treatment of Cancer (EORTC), Breast Cancer Co-operative Group (BCCG).	*Eur J Cancer* (1998) **34**(3): 307–14.
Gerber B, Krause A, Reimer T, Muller H, Friese K.	Breast reconstruction with latissimus dorsi flap: improved aesthetic results after transection of its humeral insertion.	*Plast Reconstr Surg* (1999) **103**(7): 1876–81.
Brandberg Y, Malm M, Blomqvist L.	A prospective and randomized study, ‘SVEA,’ comparing effects of three methods for delayed breast reconstruction on quality of life, patient-defined problem areas of life, and cosmetic result.	*Plast Reconstr Surg* (2000) **105**(1): 66–74; discussion 75–6.
Brorson H.	Liposuction gives complete reduction of chronic large arm lymphedema after breast cancer.	*Acta Oncol* (2000) **39**(3): 407–20.
Johansen J, Overgaard J, Rose C, Engelholm SA, Gadeberg CC, Kjaer M *et al*., Danish Breast Cancer Cooperative Group (DBCG) and the DBCG Radiotherapy Committee.	Cosmetic outcome and breast morbidity in breast-conserving treatment – results from the Danish DBCG-82TM national randomized trial in breast cancer.	*Acta Oncol* (2002) **41**(4): 369–80.
Moran SL, Nava G, Behnam AB, Serletti JM, Behnam AH.	An outcome analysis comparing the thoracodorsal and internal mammary vessels as recipient sites for microvascular breast reconstruction: a prospective study of 100 patients.	*Plast Reconstr Surg* (2003) **111**(6): 1876–82.
Futter CM, Weiler-Mithoff E, Hagen S, Van de Sijpe K, Coorevits PL, Litherland JC *et al*.	Do pre-operative abdominal exercises prevent post-operative donor site complications for women undergoing DIEP flap breast reconstruction? A two-centre, prospective randomised controlled trial.	*Br J Plast Surg* (2003) **56**(7): 674–83.

(*continued overleaf*)

Table 30.5. (*continued*)

Authors	Title of Paper	Reference
Di Benedetto G, Aquinati A, Santoli M, Bertani A.	Which is the best position for the remote injection dome using the adjustable expander/prosthesis in breast reconstruction? A comparative study.	*Plast Reconstr Surg* (2004) **113**(6): 1629–33.
Trunk reconstruction		
Wingate GF, Lewis VL Jr, Green D, Wiedrich TA, Koenig WJ.	Desmopressin decreases operative blood loss in spinal cord injury patients having flap reconstruction of pelvic pressure sores	*Plast Reconstr Surg* (1992) **89**(2): 279–82.
Benson JT, Lucente V, McClellan E.	Vaginal versus abdominal reconstructive surgery for the treatment of pelvic support defects: a prospective randomized study with long-term outcome evaluation.	*Am J Obstet Gynecol* (1996) **175**(6): 1418–21; discussion 1421–2.
Fernandez Lobato R, Garcia Septiem J, Ortega Deballon P, Martin Lucas FJ, Ruiz de Adana JC, Limones Esteban M.	Tissucol application in dermolipectomy and incisional hernia repair.	*Int Surg* (2001) **86**(4): 240–5.
Craniomaxillofacial reconstruction		
Munro IR, Boyd JB, Wainwright DJ.	Effect of steroids in maxillofacial surgery.	*Ann Plast Surg* (1986) **17**(5): 440–4.
Hotz G, Novotny-Lenhard J, Kinzig M, Soergel F.	Single-dose antibiotic prophylaxis in maxillofacial surgery.	*Chemotherapy* (1994) **40**(1): 65–9.
Flood TR, McManners J, el-Attar A, Moos KF.	Randomized prospective study of the influence of steroids on postoperative eye-opening after exploration of the orbital floor.	*Br J Oral Maxillofac Surg* (1999) **37**(4): 312–5.
Kane AA, Lo LJ, Yen BD, Chen YR, Noordhoff MS.	The effect of hamulus fracture on the outcome of palatoplasty: a preliminary report of a prospective, alternating study.	*Cleft Palate Craniofac J* (2000) **37**(5): 506–11.
Ysunza A, Pamplona MC, Mendoza M, Molina F, Martinez P, Garcia-Velasco M, Prada N.	Surgical treatment of submucous cleft palate: a comparative trial of two modalities for palatal closure.	*Plast Reconstr Surg* (2001) **107**(1):9–14.
Dietz A, Ziegler CM, Dacho A, Althof F, Conradt C, Kolling G *et al.*	Effectiveness of a new perforated 0.15 mm poly-p-dioxanon-foil versus titanium-dynamic mesh in reconstruction of the orbital floor.	*J Craniomaxillofac Surg* (2001) **29**(2): 82–8.
Ysunza A, Pamplona C, Ramirez E, Molina F, Mendoza M, Silva A.	Velopharyngeal surgery: a prospective randomized study of pharyngeal flaps and sphincter pharyngoplasties.	*Plast Reconstr Surg* (2002) **110**(6): 1401–7.
D'Errico CC, Munro HM, Buchman SR, Wagner D, Muraszko KM.	Efficacy of aprotinin in children undergoing craniofacial surgery.	*J Neurosurg* (2003) **99**(2): 287–90.

flap was then popularised by Soutar for intraoral reconstruction.[37] By the time Evans wrote a review in 1994, the radial forearm was an accepted gold standard for selected head and neck reconstruction.[9]

The fibula flap was described and popularised for mandible reconstruction by Hidalgo[38] and Wei.[39] Since then, many centres have employed the fibula as the accepted standard of care in mandible reconstruction.[40,41] It is not practical nor ethical to compare a pedicled pectoralis major reconstruction with a fibula free flap reconstruction for mandible reconstruction because the results of the fibula free flap are far superior and it is regarded as the standard of care. It is practical, however, to compare the fibula bone flap donor site with the iliac crest bone flap donor site, as Shpitzer *et al.*[42] have done. What a clinical trial might be is perhaps in the refinement of the technique; for example, the use of strong mandible reconstruction plates in fixing the fibula, or use of miniplates. Another aspect that can be studied is perhaps comparing insetting and shaping the flap pre- or post-microsurgical anastomoses and looking at flap survival, operative time, wound healing time, etc.

Another example is how the TRAM and DIEAP flap became popularised for breast reconstruction. In the early years of breast reconstruction, a variety of donor sites were used, including free dermal/fat grafts,[43–45] buttock fat[46] and the opposite breast.[47] In the 1970s, the latissimus dorsi muscle flap, in combination with a silicone prosthesis, became a popular method of choice.[48,49] The rectus abdominis myocutaneous free flap was first used for breast reconstruction in 1979[50] and then popularised for breast reconstruction by Hartrampf[51] in 1982. Later, detailed anatomical studies of the flap were done by Moon.[52] The development of microsurgical techniques brought further flexibility and refinements to breast reconstruction techniques.[53] Soon many centres around the world were utilising the TRAM flap for breast reconstruction.[54–57]

The next major development in breast reconstruction was the use of free perforator flaps. First, detailed experimental studies on the deep inferior epigastric artery (DIEA) perforator flap were done,[58] followed by its application to breast reconstruction.[59–61]

THE ROLE OF CLINICAL TRIALS

Comparative clinical trials thus do not play a major role in the development of new flaps or surgical techniques in reconstructive work, but usually at the refinement stage. In addition, clinical trials do play a role in evaluating various aspects of reconstructive plastic surgery work such as the evaluation of sutures, adhesives, implants, wound dressing material, antibiotics, anaesthetic agents, and chemotherapeutic agents for steroid use in craniofacial surgery. Examples of clinical trials that try to provide refinements of a previously established technique include the following:

- Subciliary versus transconjunctival approach for exposure during craniofacial trauma fixation or lower blepharoplasty procedures to remove eyebags.
- TRAM flap versus DIEAP flap for breast reconstruction.
- Innervated free flap versus non-innervated free flap for head and neck reconstruction.

CONDUCTING A CLINICAL TRIAL IN PLASTIC SURGERY

We need to bear in mind the role of clinical trials in plastic surgery and understand that much of the role of new clinical trials at the moment is in refining pre-existing surgical techniques, be it improving flap survival, reducing morbidity and complications, shortening operative time and hospital stay, and containing costs. But where there is a clinical problem to be answered, the gold standard is a prospective randomised controlled trial, double-blinded if possible.

Conducting a clinical trial in reconstructive plastic surgery is similar to conducting a clinical trial in general in many areas. Essential elements of conducting a clinical trial include the following:

1. Background work and hypothesis.
2. Minimising confounding factors.
3. Determining endpoints.
4. Ethics approval and funding.
5. The study design.
6. Managing results.

In reconstructive surgery, we identify two elements that may vary from clinical trials in general.

MINIMISING CONFOUNDING FACTORS

Minimising confounding factors is of particular relevance to conducting a clinical trial in reconstructive plastic surgery. Some measures that can be taken include:

- Choice of conditions with a uniform defect
- Adequate patient numbers within a limited age group
- Single surgeon
- Single observer.

An example can be given in the area of breast reconstruction, whether it is in comparing implant reconstruction with autogenous reconstruction, or comparing one flap with another. The incidence of breast cancer is relatively high and this would enable adequate numbers of patients to be accrued for a clinical trial, even when allowance is made for for exclusion of certain patients outside of the predetermined age range. The (mastectomy) defect created is relatively uniform, and in unilateral cases there is an opposite breast to compare with when assessing aesthetic results of a reconstruction.

An ideal situation is to have a single surgeon perform all the procedures. Limitations of this approach might include the lack of patient numbers (especially if the surgeon is practising within a competitive situation with other plastic surgeons providing similar reconstructive services), and operator bias, where the surgeon may have a pre-existing preference for one technique over another.[62] A more practical way would be for the lead surgeon in a high-volume institution to engage its team of surgeons, ensuring that the technical aspects of the procedure are as uniform as possible.

DETERMINING ENDPOINTS

The endpoints need to be as quantifiable as possible. In reconstructive plastic surgery, we would be interested in the form (aesthetic appearance) and the function of the reconstruction. Some authors record observations of the aesthetic appearance from three sources – the surgeon performing the procedure, the patient and an independent observer. Usual quantifiable parameters include the use of the visual analogue scale[29] and the Derriford Appearance Scale.[63] Other quantifiable endpoints include quality-of-life scales.[64,65]

A randomised controlled study comparing alternative therapy with conventional methods in the management of burn injuries used the following quantifiable end-points: time to healing of 75% of initial body surface area, visual analogue pain score for analgesic effect, wound colonisation and infection by bacteria, and hospital cost.[8,66]

REFERENCES

1. Georgiade GS, Riefkohl R, Levin LS (eds). *Georgiade Plastic, Maxillofacial and Reconstructive Surgery*, 3rd edn. Baltimore, MD: Williams & Wilkins (1997).
2. Grotting JC. *Reoperative Aesthetic and Reconstructive Plastic Surgery*. St Louis: Quality Medical Publishing (1995).
3. Huemer GM, Bauer T, Gurunluoglu R, Sakho C, Oehlbauer M, Dunst KM. Analysis of publications in three plastic surgery journals for the year 2002. *Plast Reconstr Surg* (2004) **114**: 1147.
4. Engrav LH, Heimbach DM, Reus JL, Harnar TJ, Marvin JA. Early excision and grafting vs. nonoperative treatment of burns of indeterminate depth: a randomised prospective study. *J Trauma* (1983) **23**: 1001–4.
5. Jackson DM. The evolution of burn treatment in the last 50 years. *Burns* (1991) **17**: 329–34.
6. Nguyen TT, Gilpin DA, Meyer NA, Herndon DN. Current treatment of severely burned patients. *Ann Surg* (1996) **223**: 14–24.

7. Yamamoto Y, Yokoyama T, Minakawa H, Sugihara T. Use of the expanded skin flap in esthetic reconstruction of postburn deformity. *J Burn Care Rehabil* (1996) **17**(5): 397–401.
8. Ang ESW, Lee ST, Gan CSG, See PGJ, Chan YH, Ng LH, Machin D. Evaluating the role of alternative therapy in burn wound management: randomized trial comparing moist exposed burn ointment with conventional methods in the management of patients with second-degree burns. *Medscape Gen Med* (2001) **3**(1).
9. Evans GR, Schusterman MA, Kroll SS, Miller MJ, Reece GP, Robb GL *et al*. The radial forearm free flap for head and neck reconstruction: a review. *Am J Surg* (1994) **168**(5): 446–50.
10. Urken ML, Weinberg H, Buchbinder D, Moscoso JF, Lawson W, Catalano PJ, Biller HF. Microvascular free flaps in head and neck reconstruction – report of 200 cases and review of complications. *Arch Otolaryngol Head Neck Surg* (1994) **120**: 633–40.
11. Spear SL, Little JW, Lippman ME, Wood WC (eds). *Surgery of the Breast – Principles and Art*. Philadelphia, PA: Lippincott-Raven (1998).
12. Ang ESW, Tan KC. Free flap reconstruction in head and neck oncology. *Asian J Surg* (2002) **25**(1): 49–53.
13. Ang ESW, Black C, Irish J, Brown DH, Gullane PG, O'Sullivan B, Neligan PC. Reconstructive options in the treatment of osteoradionecrosis of the craniomaxillofacial skeleton. *Br J Plast Surg* (2003) **56**: 92–9.
14. Wornom IL 3rd, Neifeld JP, Mehrhof AI Jr, Young HF, Chinchilli VM. Closure of craniofacial defects after cancer resection. *Am J Surg* (1991) **162**(4): 408–11.
15. Chen CT, Chen YR, Tung TC, Lai JP, Rohrich RJ. Endoscopically assisted reconstruction of orbital medial wall fractures. *Plast Reconstr Surg* (1999) **103**(2): 714–20.
16. Cheney ML, Gliklich RE. The use of calvarial bone in nasal reconstruction. *Arch Otolaryngol Head Neck Surg* (1995) **121**(6): 643–8.
17. Jones NF, Johnson JT, Shestak KC, Myers EN, Swartz WM. Microsurgical reconstruction of the head and neck: interdisciplinary collaboration between head and neck surgeons and plastic surgeons in 305 cases. *Ann Plast Surg* (1996) **36**: 37–43.
18. Harrison LB, Sessions RB, Hong WK (eds). *Head and Neck Cancer – A Multidisciplinary Approach*. Philadelphia, PA: Lippincott-Raven (1999).
19. American Medical Association, House of Delegates, Chicago. June, 1989.
20. Bang RL, Mustafa MD. Comparative study of skin wound closure with polybutester (Novafil) and polypropylene. *J R Coll Surg Edinb* (1989) **34**(4): 205–7.
21. Talbot AW, Meadows AE, Tyers AG, Shah-Desai S. Use of 7/0 Vicryl (coated polyglactin 910) and 7/0 Vicryl-rapide (irradiated polyglactin 910) in skin closure in ophthalmic plastic surgery. *Orbit* (2002) **21**(1): 1–8.
22. Blomqvist G, Steenfos H. A new partly external device for extension of skin before excision of skin defects. *Scand J Plast Reconstr Surg Hand Surg* (1993) **27**(3): 179–82.
23. Khouri RK, Schlenz I, Murphy BJ, Baker TJ. Nonsurgical breast enlargement using an external soft-tissue expansion system. *Plast Reconstr Surg* (2000) **105**(7): 2500–12; discussion 2513–14.
24. Eaglestein WH, Alvarez OM, Auletta M, Leffel D, Rogers GS, Zitelli JA *et al*. Acute excisional wounds treated with a tissue-engineered skin (Apligraf). *Dermatol Surg* (1999) **25**(3): 195–201.
25. Yamaguchi Y, Kubo T, Tarutani M, Sano S, Asada H, Kakibuchi M. *et al*. Epithelial-mesenchymal interactions in wounds: treatment of palmoplantar wounds by nonpalmoplantar pure epidermal sheet grafts. *Arch Dermatol* (2001) **137**(5): 621–8.
26. Brown GL, Nanney LB, Griffen J, Cramer AB, Yancey JM, Curtsinger LJ 3rd *et al*. Enhancement of wound healing by topical treatment with epidermal growth factor. *New Engl J Med* (1989) **321**(2): 76–9.
27. Jeschke MG, Rose C, Angele P, Fuchtmeier B, Nerlich MN, Bolder U. Development of new reconstructive techniques: use of Integra in combination with fibrin glue and negative-pressure therapy for reconstruction of acute and chronic wounds. *Plast Reconstr Surg* (2004) **113**(2): 525–30.
28. Nahas FX, Augusto SM, Ghelfond C. Nylon versus polydioxanone in the correction of rectus diastasis. *Plast Reconstr Surg* (2001) **107**(3): 700–6.
29. Toriumi DM, O'Grady K, Desai D, Bagal A. Use of octyl-2-cyanoacrylate for skin closure in facial plastic surgery. *Plast Reconstr Surg* (1998) **102**(6): 2209–19.
30. Baran CN, Sensoz O, Ulusoy MG. Prophylactic antibiotics in plastic and reconstructive surgery. *Plast Reconstr Surg* (1999) **103**(6): 1561–6.
31. Amland PF, Andenaes K, Samdal F, Lingaas E, Sandsmark M, Abyholm F, Giercksky KE. A prospective, double-blind, placebo-controlled trial of a single dose of azithromycin on postoperative wound infections in plastic surgery. *Plast Reconstr Surg* (1995) **96**(6): 378–83.
32. Sinha UK, Shih C, Chang K, Rice DH. Use of AlloDerm for coverage of radial forearm free flap donor site. *Laryngoscope* (2002) **112**(2): 230–4.

33. Fogarty BJ, Brown AP, Miller R, Khan K. TRAM flap versus nonautologous breast reconstruction: what do patients really think? *Plast Reconstr Surg* (2004) **113**: 1146–52.
34. Mathes SJ, Nahai F. Classification of the vascular anatomy of muscles: experimental and clinical correlation. *Plast Reconstr Surg* (1981) **67**: 177.
35. Cormack GC, Lamberty BGH. A classification of fasciocutaneous flaps according to their patterns of vascularisation. *Br J Plast Surg* (1984) **37**: 80–7.
36. Song R, Song Y, Yu Y, Song Y. The upper arm free flap. *Clin Plast Surg* (1982) **9**: 27.
37. Soutar DS, McGregor IA. The radial forearm flap in intraoral reconstruction: the experience of 60 consecutive cases. *Plast Reconstr Surg* (1986) **78**: 1–8.
38. Hidalgo DA. Fibula free flap: a new method of mandible reconstruction. *Plast Reconstr Surg* (1989) **84**: 71.
39. Wei FC, Seah CS, Tsai YC, Liu SJ, Tsai MS. Fibula osteoseptocutaneous flap for reconstruction of composite mandibular defect. *Plast Reconstr Surg* (1994) **93**: 294–303.
40. Schusterman MA, Harris SW, Raymond AK, Goepfert H. Immediate free flap mandibular reconstruction: significance of adequate surgical margins. *Head Neck* (1993) **15**(3): 204–7.
41. Cordeiro PG, Disa JJ, Hidalgo DA, Hu QY. Reconstruction of the mandible with osseous free flaps: a 10-year experience with 150 consecutive patients. *Plast Reconstr Surg* (1999) **104**(5): 1314–20.
42. Shpitzer T, Neligan PC, Gullane PJ, Boyd JB, Gur E, Rotstein LE *et al*. The free iliac crest and fibula flaps in vascularized oromandibular reconstruction: comparison and long-term evaluation. *Head Neck* (1999) **21**: 639–47.
43. Berson MJ. Dermal fat fascia transplants used in building up the breast. *Surgery* (1994) **15**: 451–6.
44. Bames HO. Augmentation mammoplasty lipotransplant. *Plast Reconstr Surg* (1953) **11**: 404–14.
45. Watson J. Some observation on free fat grafts with reference to their use in mammoplasty. *Br J Plast Surg* (1994) **12**: 263–74.
46. Ortichea MC. Use of buttock to reconstruct the breast. *Br J Plast Surg* (1973) **26**: 304–9.
47. Pontes R. Single stage reconstruction of the missing breast. *Br J Plast Surg* (1973) **26**: 377–80.
48. Oliviari N. The latissimus flap. *Br J Plast Surg* (1976) **29**: 126.
49. Muhlbauer W, Olbrisch R. The LDMF for breast reconstruction. *Chir Plast (Berlin)* (1977) **4**: 27.
50. Holmstrom H. The free abdominoplasty and its use in breast reconstruction. *Scand J Plast Reconstr Surg* (1979) **13**: 423–27.
51. Hartrampf CR, Scheflan N, Black PW. Breast reconstruction with a transverse abdominal island flap. *Plast Reconstr Surg* (1982) **69**: 216.
52. Moon HK, Taylor GI. The vascular anatomy of rectus abdominis musculocutaneous flaps based on the deep superior epigastric system. *Plast Reconstr Surg* (1988) **82**: 815–29.
53. Friedman RJ, Argenta LC, Anderson R. Deep inferior epigastric free flap for breast reconstruction after radical mastectomy. *Plast Reconstr Surg* (1985) **76**: 455–8.
54. Schusterman MA, Kroll SS, Miller MJ, Reece GP, Baldwin BJ, Robb GL *et al*. The free transverse rectus abdominis musculocutaneous flap for breast reconstruction: one center's experience with 211 consecutive cases. *Ann Plast Surg* (1994) **32**: 234–42.
55. Elliott LF, Eskenazi L, Beegle PH Jr, Podres PE, Drazan L. Immediate TRAM flap breast reconstruction: 128 consecutive cases. *Plast Reconstr Surg* (1993) **92**: 217–27.
56. Grotting JC. Immediate breast reconstruction using the free TRAM flap. *Clin Plast Surg* (1994) **21**: 207–21.
57. Shaw WW, Ahn CY. Microvascular free flaps in breast reconstruction. *Clin Plast Surg* (1992) **19**: 917–26.
58. Taylor GI, Corlett RJ, Boyd JB. The versatile DIE flap. *Br J Plast Surg* (1984) **37**: 330–50.
59. Koshima I, Soeda S. Inferior epigastric artery skin flap without rectus abdominis muscle. *Br J Plast Surg* (1989) **42**: 645.
60. Allen RJ, Treece P. Deep inferior epigastric artery perforator flap for breast reconstruction. *Ann Plast Surg* (1994) **32**: 32.
61. Blondeel PN, Boeckx WD. Refinements in free flap breast reconstruction. The free bilateral deep inferior epigastric perforator flap anastomosed to Int. mammary artery. *Br J Plast Surg* (1994) **47**: 495.
62. Chevray PM. Breast reconstruction with superficial inferior epigastric artery flaps: a prospective comparison with TRAM and DIEP flaps. *Plast Reconstr Surg* (2004) **114**: 1077–83.
63. Harris DL, Carr AT. The Derriford Appearance Scale (DAS59): a new psychometric scale for the evaluation of patients with disfigurements and aesthetic problems of appearance. *Br J Plast Surg* (2001) **54**(3): 216–22.
64. Ching S, Thoma A, McCabe RE, Antony MM. Measuring outcomes in aesthetic surgery: a comprehensive review of the literature. *Plast Reconstr Surg* (2003) **112**(7): 1953–4.
65. Curran D, van Dongen JP, Aaronson NK, Kiebert G, Fentiman IS, Mignolet F, Bartelink H. Quality of life of early-stage breast cancer patients treated

with radical mastectomy or breast-conserving procedures: results of EORTC Trial 10801. The European Organization for Research and Treatment of Cancer (EORTC), Breast Cancer Cooperative Group (BCCG). *Eur J Cancer* (1998) **34**(3): 307–14.

66. Ang E, Lee ST, Gan CSG, Chan YH, Cheung YB, Machin D. Pain control in a randomized, controlled, clinical trial comparing moist exposed burn ointment and conventional methods in patients with partial-thickness burns. *J Burn Care Rehabil* (2003) **24**(5): 289–96.

31

Plastic Surgery – Aesthetic Surgery

WOFFLES WU[1] AND ERIK ANG[2]
[1]Suite 09-02, Camden Medical Centre, Singapore
[2]Suite 1208, Mount Elizabeth Medical Centre, Singapore

INTRODUCTION

Aesthetic surgery is the fastest-growing medical specialty today, attracting more recruits than any other specialty, and yet is the most difficult to quantify in terms of outcomes because of its subjective nature. It refers to surgery to enhance, change, augment or reduce different parts of the body in order to improve and enhance the physical appearance. Often by improving the physical appearance, many people improve their psychosocial well-being.[1–4] Plastic surgeons with their extensive training in proper tissue handling and respect for wound healing, and with an astute eye for detail and aesthetics, are well poised and well suited to deal with aesthetic issues.

Aesthetic surgery is a magnet for attention and controversy as it is the current darling of medicine and the tabloid press. Therefore it comes under the microscope of scrutiny and any good or indeed bad outcome has the potential to be blown out of proportion. Originally covering only traditional surgical procedures such as facelifting, blepharoplasties, rhinoplasties, breast augmentation and reduction, liposuction and body contouring, the field has grown exponentially in recent years as its coverage has widened to embrace non-surgical procedures as well. These latter procedures have in turn been fuelled by patient demands for faster results with less downtime away from their social and economic activities, and this is due in part to the culture of instant gratification which we now live in.

These non-invasive or minimally invasive procedures include injection of botulinum toxin, synthetic fillers, barbed threads for facial rejuvenation/lifting(eg APTOS™ Featherlift and the WOFFLES Lift) and numerous skin resurfacing or re-texturising techniques such as intense pulsed light (IPL) therapy, Thermage™ radiofrequency, and an assortment of other lasers, to name a few. Mesotherapy and other modalities for body contouring and fat reduction such as ultrasonic devices have also come under scrutiny

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

and clinical testing to determine their efficacy and safety.

The results are not only subjective but difficult to compare as many studies are retrospective or anecdotal, or they are reports of surgical techniques and not subject to rigid scrutiny. Yet the specialty has to lend itself to rigorous testing and trials so that proper scientific conclusions can be made about the myriad procedures available to practitioners.

ISSUES IN CONDUCTING CLINICAL TRIALS IN AESTHETIC SURGERY

In other branches of medicine, where well-being and cure are concerned, outcomes are easier to quantify. Did the patient get better? Was the patient cured? Did the patient die? How long did the patient survive? Basic questions like these are easily answered. Nevertheless, studies increasingly are looking at quality of life issues in treating cancer, heart or diabetic patients, for example.

In aesthetic surgery we need to not only deal with the surgeon's satisfaction with the result but also to contend with the patients' own perception of the outcome. Cosmesis and normality of result is of paramount importance. There is no point in performing a complicated procedure on a female patient to make her beautiful if the scars are visible or the implants inserted are detectable and indicate the patient is not a natural beauty. People would then know the patient had had an enhancement operation which might therefore be deemed unsuccessful because of its detectability.

In general medicine, these are not important considerations as long as the patient is cured. An abdominal scar is trivial as long as the appendicitis, gallstone or tumour has been removed successfully and the patient is able to resume normal, pain-free living. In aesthetic surgery, however, there is the additional requirement of operating on the patient but making it seem as if there was no operation and the patient was born with those attributes, e.g. breast enhancement.

In the case of a browlift, we have to measure whether there was indeed a lift of the brow, by how much, whether the desired shape was achieved, the longevity of the result, the ease of the procedure, and weigh these against complications such as numbness, visible scarring and detectability, keloids, nerve damage, and so on, before we can evaluate whether it is a safe and worthwhile procedure. We then have to determine the subjective elements of whether the patient and the surgeon were satisfied.

Another dimension is whether we have overcorrected or overdone the browlift, taking the patient out of the realm of normality into the abnormal. In other specialties, the patient is either cured or not, dead or not. It is all or none. But in aesthetic medicine, there is scope for overtreatment, overcorrection, and this brings the patient out of the accepted range of normal beauty into the bizarre.

The patient is the best control. Before and after pictures are important to determine the change in the patient's appearance, but once the patient has been operated on we do not really know how that face would have aged or evolved with time, so we cannot make a true comparison. To determine the success of aesthetic surgical procedures, it would be ideal to operate on one side of the face and compare the outcome with the other side of the face. However, due to the aesthetic nature of the discipline and the patients who seek cosmetic enhancement, this becomes impractical and therefore a stumbling block to true scientific endeavour.

One simple area where 'split-face' treatment can be employed is in the treatment of multiple facial keratoses, a condition due to chronic sun exposure where the face is covered usually in a carpet of little keratoses and sometimes skin tags. Laser removal is the treatment of choice and as it is usually staged due to the downtime involved, both patient and surgeon can agree to treat one side of the face first and come back later to treat the other half. This gives us a direct comparison of one side with the other.

Safety is a key issue in aesthetic surgery where the patients who choose it are not 'sick' in the traditional sense. These patients are essentially normal people seeking cosmetic enhancement or a change in physical appearance due to issues with body image. Complications will mar the final desired outcome and may even make the intended goal impossible to attain. Therefore the margin for error is razor thin. It is a nightmare to have a normal patient undergo aesthetic surgery only to come out scarred or disfigured. Nowhere in medicine is the saying 'if you can't make them better, don't make them worse' more true or pertinent.

In summary, some of the issues in aesthetic surgery that influence the conduct of clinical trials include the following:

- Often the main purpose is in changing the external appearance of a patient. The results are therefore subjective. The patient's expectations and interpretation may differ from the surgeon's. These in turn may differ from the opinions of those who interact with the patient on a regular basis.
- This element of subjectivity in the assessment of patient outcome and results affects the conduct of studies evaluating the efficacy of new technology, in the sense that some surgeons may be pressured by funding considerations from the company providing the new technology to report favourable results.
- Many aesthetic surgeons, including ourselves, consider each procedure on a patient as a work of art – it is quite impossible to conduct a randomised controlled trial comparing an oil painting with a water colour, or a Renoir with a Rembrandt!
- There is a widespread trend towards combination therapy involving minimally invasive and even non-surgical methods to address many aesthetic issues, such as in facial rejuvenation and body contouring. Clinical trials will have to take into account adjunctive procedures like laser resurfacing, use of injectable fillers and botulinum toxin, for example, when comparing results from a sub-periosteal versus a subcutaneous facelift. The results are further blurred when non-medical therapies like massage and spa treatment in combination with body contouring surgery are added to the equation.
- Some of the patients' perception of 'successful' results may be influenced by their total experience in the surgical episode, from the time they initiate a telephone booking to the first time they walk through the clinic door for consultation with the plastic surgeon, before the actual surgery itself.
- Aesthetic patients are healthy and 'well'. Many are in the economically productive age group and are often busy individuals who appreciate minimally invasive procedures with minimal 'downtime' away from their trade or employment. This group of patients may be reluctant to return to the clinic for repeated follow-ups that participating in a clinical trial may necessarily entail.

OBJECTIVE PARAMETERS IN AESTHETIC SURGERY

Nevertheless, in spite of the subjectivity of much of aesthetic surgery, there are certain objective parameters by which results could and should be measured when comparing outcomes of different methods or techniques.

For example, subjective and objective outcomes may be measured in clinical trials on the use of botulinum toxin in the treatment of wrinkles and frown lines. Botulinum toxin is being used extensively in facial rejuvenation for wrinkles and frown lines, repositioning of eyebrows and reduction of bulky lower facial muscles.[5,6] In any trial evaluating this, one may want to evaluate the following treatment outcomes:

- Objective results:
  - facial measurements
  - muscle tension
  - muscle activity
  - histological appearance.
- Subjective results:
  - patient appearance

– patient satisfaction
– observer opinion.

Another area of interest is in breast enhancement surgery. Results from breast augmentation surgery with breast implants depend on a few variables.[7] There are different types of implants available, and different placement planes, and different placement routes. Breast implants have a silicone outer shell, and the filler material can either be saline or silicone. The outer shell can be smooth or textured. The profile of the implant can be round and low, round and high, or tear-drop in shape (known as the anatomical implant). Breast implants can be place in the sub-mammary plane or the sub-pectoral plane. They can be placed via incisions in the infra-mammary crease, nipple-areolar, or axilla. Any clinical trials involving breast implants therefore have to take into account these variables.

Objectively, clinical studies on breast implants might want to look at complication rates like capsular contracture, infection, rupture, explantation, and leakage over a period of time.[8] Breast-feeding ability may also be an issue that might be studied. The aesthetic result may be both objective and subjective. Objectively, the patient might report an increase (or absence of increase) in brassiere size. Measurements of increase in projection of the nipple from the chest wall after a reasonable time frame for post-operative oedema to subside may offer yet another objective parameter. Additionally, pectoralis function can be measured by isometric measurements.[9] The patient's perception of and satisfaction with the breast enhancement surgery may be considered as subjective results.

Therefore, just as in reconstructive plastic surgery, comparative clinical trials in aesthetic surgery do not play a major role at the moment in the development of new surgical techniques, e.g. the use of implants for breast augmentation, or liposuction for body contouring, but in the refinement stage, such as comparing saline with silicone breast implants, or comparing tumescent versus non-tumescent infiltration for liposuction.

APPENDIX

As in the chapter on reconstructive plastic surgery, a survey of existing clinical trials in plastic and reconstructive surgery published in the literature was undertaken to get an idea of the aesthetic surgery issues addressed by these trials. An internet-based PubMed Medline search was carried out with the search words 'plastic surgery' and 'reconstructive surgery'. We did not, however, include the search words 'cosmetic surgery' or 'aesthetic surgery'. The search parameters included papers between 1980 and 2004, with clinical studies conducted amongst human subjects published in the English language. This survey excluded case reports, case series, review articles, and tips and techniques (see Tables 31.1 and 31.2).

FACIAL REJUVENATION

In the 1990s, most facial rejuvenation studies were on the utility of steroids in facial aesthetic surgery, different surgical techniques in performing facelifts or other aesthetic procedures, and evaluating the merits of different treatment protocols like laser resurfacing and chemical peels for rhytids. A number of randomised studies looked at the use of perioperative steroids in reducing the post-operative oedema associated with aesthetic procedures in the face.

At the turn of the century, more papers appeared reporting results of trials on the use of non-invasive rejuvenation techniques, such as the use of lasers, intense pulse light, and radiofrequency, reflecting the development of the new technologies available.

BODY CONTOURING

In the late 1990s, new techniques of liposuction, such as ultrasound guided liposuction, were introduced.

BREAST SURGERY

A number of studies were concerned about comparing textured versus smooth implants for

Table 31.1. Non-randomised clinical trials in aesthetic surgery

Authors	Title	Reference
Facial rejuvenation		
Maimon WN, Schuller DE.	Lidocaine v bupivacaine in facial plastic surgery. A clinical trial.	*Arch Otolaryngol* (1984) **110**(8): 525–8
David LM, Sanders G.	CO_2 laser blepharoplasty: a comparison to cold steel and electrocautery.	*J Dermatol Surg Oncol* (1987) **13**(2): 110–14.
Spear SL, Mausner ME, Kawamoto HK Jr.	Sliding genioplasty as a local anesthetic outpatient procedure: a prospective two-center trial.	*Plast Reconstr Surg* (1987) **80**(1): 55–67.
de la Fuente A, Martin del Yerro JL.	Calibrated nasal tip: review of 100 cases.	*Aesthetic Plast Surg* (1994) **18**(4): 357–61.
Tremolada C, Fissette J, Candiani P.	Anatomical basis for a safe and easier approach to composite rhytidectomy.	*Aesthetic Plast Surg* (1994) **18**(4): 387–91.
Peikert JM, Kaye VN, Zachary CB.	A reevaluation of the effect of occlusion on the trichloroacetic acid peel.	*J Dermatol Surg Oncol* (1994) **20**(10): 660–5.
Ramirez OM.	Endoscopic full facelift.	*Aesthetic Plast Surg* (1994) **18**(4): 363–71.
Reino AJ, Lawson W.	Role of the argon beam coagulator in facial rejuvenation surgery.	*Arch Otolaryngol Head Neck Surg* (1995) **121**(6): 627–33.
Eppley BL, Sadove AM, Holmstrom H, Kahnberg KE.	HTR polymer facial implants: a five-year clinical experience.	*Aesthetic Plast Surg* (1995) **19**(5): 445–50.
Netscher DT, Patrinely JR, Peltier M, Polsen C, Thornby J.	Transconjunctival versus transcutaneous lower eyelid blepharoplasty: a prospective study.	*Plast Reconstr Surg* (1995) **96**(5): 1053–60.
Min YG, Chung JW.	Cartilaginous incisions in septoplasty.	*ORL J Otorhinolaryngol Relat Spec* (1996) **58**(1): 51–4.
Mommaerts MY, Beirne JC, Jacobs WI, Abeloos JS, De Clercq CA, Neyt LF.	Use of fibrin glue in lower blepharoplasties.	*J Craniomaxillofac Surg* (1996) **24**(2): 78–82.
Newman N, Newman A, Moy LS, Babapour R, Harris AG, Moy RL.	Clinical improvement of photoaged skin with 50% glycolic acid. A double-blind vehicle-controlled study.	*Dermatol Surg* (1996) **22**(5): 455–60.
Hwang YJ, Jeon JY, Lee MS.	A simple method of reduction malarplasty.	*Plast Reconstr Surg* (1997) **99**(2): 348–55.
Ellis DA, Tan AK.	Cosmetic upper-facial rejuvenation with botulinum.	*J Otolaryngol* (1997) **26**(2): 92–6.
Lawrence N, Cox SE, Brody HJ.	Treatment of melasma with Jessner's solution versus glycolic acid: a comparison of clinical efficacy and evaluation of the predictive ability of Wood's light examination.	*J Am Acad Dermatol* (1997) **36**(4): 589–93.
Klassen A, Jenkinson C, Fitzpatrick R, Goodacre T.	Measuring quality of life in cosmetic surgery patients with a condition-specific instrument: the Derriford Scale.	*Br J Plast Surg* (1998) **51**(5): 380–4.
Fulton JE.	Simultaneous face lifting and skin resurfacing.	*Plast Reconstr Surg* (1998) **102**(7): 2480–9.
Goldman MP, Fitzpatrick RE, Manuskiatti W.	Laser resurfacing of the neck with the erbium: YAG laser.	*Dermatol Surg* (1999) **25**(3): 164–7; discussion 167–8.

(*continued overleaf*)

Table 31.1. (*continued*)

Authors	Title	Reference
Kim HY, Kang KY.	Epidermal grafts for treatment of stable and progressive vitiligo.	*J Am Acad Dermatol* (1999) **40**(3): 412–7.
Ruiz-Esparza J, Barba Gomez JM.	Long-term effects of one general pass laser resurfacing. A look at dermal tightening and skin quality.	*Dermatol Surg* (1999) **25**(3): 169–73; discussion 174.
Khatri KA, Ross V, Grevelink JM, Magro CM, Anderson RR.	Comparison of erbium:YAG and carbon dioxide lasers in resurfacing of facial rhytides.	*Arch Dermatol* (1999) **135**(4): 391–7.
Cotellessa C, Peris K, Onorati MT, Fargnoli MC, Chimenti S.	The use of chemical peelings in the treatment of different cutaneous hyperpigmentations.	*Dermatol Surg* (1999) **25**(6): 450–4.
Gin I, Chew J, Rau KA, Amos DB, Bridenstine JB.	Treatment of upper lip wrinkles: a comparison of the 950 microsec dwell time carbon dioxide laser to manual tumescent dermabrasion.	*Dermatol Surg* (1999) **25**(6): 468–73; discussion 473–4.
Kelly KM, Nelson JS, Lask GP, Geronemus RG, Bernstein LJ.	Cryogen spray cooling in combination with nonablative laser treatment of facial rhytides.	*Arch Dermatol* (1999) **135**(6): 691–4.
Fisher E, Frodel JL.	Facial suspension with acellular human dermal allograft.	*Arch Facial Plast Surg* (1999) **1**(3): 195–9.
Goldberg DJ.	Non-ablative subsurface remodeling: clinical and histologic evaluation of a 1320-nm Nd:YAG laser.	*J Cutan Laser Ther* (1999) **1**(3): 153–7.
Sumian CC, Pitre FB, Gauthier BE, Levy JL, Bouclier M, Mordon Sr.	A preliminary clinical and histopathological study of laser skin resurfacing using a frequency-doubled Nd:YAG laser after application of Chromofilm.	*J Cutan Laser Ther* (1999) **1**(3): 159–66.
Greene D, Egbert BM, Utley DS, Koch RJ.	The validity of ex vivo laser skin treatment for histological analysis. A prospective controlled study.	*Arch Facial Plast Surg* 1999 **1**(3): 159–64.
Jimenez G, Spencer JM.	Erbium:YAG laser resurfacing of the hands, arms, and neck.	*Dermatol Surg* (1999) **25**(11): 831–4; discussion 834–5.
Koppel RA, Coleman KM, Coleman WP.	The efficacy of EMLA versus ELA-Max for pain relief in medium-depth chemical peeling: a clinical and histopathologic evaluation.	*Dermatol Surg* (2000) **26**(1): 61–4.
Goldberg DJ, Cutler KB.	Nonablative treatment of rhytids with intense pulsed light.	*Lasers Surg Med* (2000) **26**(2): 196–200.
Sclafani AP, Romo T 3rd, Jacono AA, McCormick S, Cocker R, Parker A.	Evaluation of acellular dermal graft in sheet (AlloDerm) and injectable (micronized AlloDerm) forms for soft tissue augmentation. Clinical observations and histological analysis.	*Arch Facial Plast Surg* (2000) **2**(2): 130–6.
Sim RS, Smith JD, Chan AS.	Comparison of the aesthetic facial proportions of southern Chinese and white women.	*Arch Facial Plast Surg* (2000) **2**(2): 113–20.
Trimas SJ, Boudreaux CE, Metz RD.	Carbon dioxide laser abrasion. Is it appropriate for all regions of the face?	*Arch Facial Plast Surg* (2000) **2**(2): 137–40.
Goldberg DJ, Samady JA.	Intense pulsed light and Nd:YAG laser non-ablative treatment of facial rhytids.	*Lasers Surg Med* (2001) **28**(2): 141–4.
Chaushu G, Blinder D, Taicher S, Chaushu S.	The effect of precise reattachment of the mentalis muscle on the soft tissue response to genioplasty.	*J Oral Maxillofac Surg* (2001) **59**(5): 510–6; discussion 517.

Table 31.1. (*continued*)

Authors	Title	Reference
Palaia DA, Rosenberg MH, Bonanno PC.	The use of DDAVP desmopressin reduces the incidence of microhematomas after facioplasty.	*Ann Plast Surg* (2001) **46**(5): 463–6.
Fezza JP, Cartwright M, Mack W, Flaharty P.	The use of aerosolized fibrin glue in face-lift surgery.	*Plast Reconstr Surg* (2002) **110**(2): 658–64; discussion 665–6.
Lupton JR, Williams CM, Alster TS.	Nonablative laser skin resurfacing using a 1540 nm erbium glass laser: a clinical and histologic analysis.	*Dermatol Surg* (2002) **28**(9): 833–5.
Sarkar R, Kaur C, Bhalla M, Kanwar AJ.	The combination of glycolic acid peels with a topical regimen in the treatment of melasma in dark-skinned patients: a comparative study.	*Dermatol Surg* (2002) **28**(9): 828–32; discussion 832.
Tanzi EL, Williams CM, Alster TS.	Treatment of facial rhytides with a nonablative 1,450-nm diode laser: a controlled clinical and histologic study.	*Dermatol Surg* (2003) **29**(2): 124–8.
Chait L, Kadwa A, Potgieter A, Christofides E.	A venturi based suction drainage system used in facelifts.	*Br J Plast Surg* (2003) **56**(2): 150–2.
Jones BM, Grover R.	Reducing complications in cervicofacial rhytidectomy by tumescent infiltration: a comparative trial evaluating 678 consecutive face lifts.	*Plast Reconstr Surg* (2004) **113**(1): 398–403.
Lewis KG, Nahm WK, Schmidt AN, Moy RL.	Lack of difference in the rates of hypopigmentation with 90-microsecond pulsed and longer dwell time carbon-dioxide laser resurfacing.	*J Am Acad Dermatol* (2004) **50**(2): 247–52.
Dahan S, Lagarde JM, Turlier V, Courrech L, Mordon S.	Treatment of neck lines and forehead rhytids with a nonablative 1540-nm Er:glass laser: a controlled clinical study combined with the measurement of the thickness and the mechanical properties of the skin.	*Dermatol Surg* (2004) **30**(6): 872–9; discussion 879–80.
Nahm WK, Su TT, Rotunda AM, Moy RL.	Objective changes in brow position, superior palpebral crease, peak angle of the eyebrow, and jowl surface area after volumetric radiofrequency treatments to half of the face.	*Dermatol Surg* (2004) **30**(6): 922–8; discussion 928.
Body contouring		
Apfelberg DB, Rosenthal S, Hunstad JP, Achauer B, Fodor PB.	Progress report on multicenter study of laser-assisted liposuction.	*Aesthetic Plast Surg* (1994) **18**(3): 259–64.
Matarasso A.	Liposuction as an adjunct to a full abdominoplasty.	*Plast Reconstr Surg* (1995) **95**(5): 829–36.
Coleman Sr.	Long-term survival of fat transplants: controlled demonstrations.	*Aesthetic Plast Surg* (1995) **19**(5): 421–5.
Kuzon WM Jr, Crawford R, Binhammer P, Fielding C, Knowlton R, Levine R.	Effect of electrosurgical technique on wound healing and early complication rate following abdominal dermolipectomy.	*Ann Plast Surg* (1996) **37**(3): 245–50.
Apfelberg DB.	Results of multicenter study of laser-assisted liposuction.	*Clin Plast Surg* (1996) **23**(4): 713–9.

(*continued overleaf*)

Table 31.1. (*continued*)

Authors	Title	Reference
Becker DG, Weinberger MS, Miller PJ, Park SS, Wang TD, Cook TA, Tardy ME Jr, Gross CW.	The liposhaver in facial plastic surgery. A multi-institutional experience.	*Arch Otolaryngol Head Neck Surg* (1996) **122**(11): 1161–7.
Becker DG, Cook TA, Wang TD, Park SS, Kreit JD, Tardy ME Jr, Gross CW.	A 3-year multi-institutional experience with the liposhaver.	*Arch Facial Plast Surg* (1999) **1**(3): 171–6.
Fodor PB, Vogt PA.	Power-assisted lipoplasty (PAL): A clinical pilot study comparing PAL to traditional lipoplasty (TL).	*Aesthetic Plast Surg* (1999) **23**(6): 379–85.
Perry AW, Petti C, Rankin M.	Lidocaine is not necessary in liposuction.	*Plast Reconstr Surg* (1999) **104**(6): 1900–2; discussion 1903–6.
Grippaudo FR, Matarese RM, Macone A, Mazzocchi M, Scuderi N.	Effects of traditional and ultrasonic liposuction on adipose tissue: a biochemical approach.	*Plast Reconstr Surg* (2000) **106**(1): 197–9.
Cardenas-Camarena L, Cardenas A, Fajardo-Barajas D.	Clinical and histopathological analysis of tissue retraction in tumescent liposuction assisted by external ultrasound.	*Ann Plast Surg* (2001) **46**(3): 287–92.
Breast surgery		
Brantner JN, Peterson HD.	The role of vasoconstrictors in control of blood loss in reduction mammaplasty.	*Plast Reconstr Surg* (1985) **75**(3): 339–41.
Hakelius L, Ohlsen L.	A clinical comparison of the tendency to capsular contracture between smooth and textured gel-filled silicone mammary implants.	*Plast Reconstr Surg* (1992) **90**(2): 247–54.
Chajchir A, Benzaquen I, Spagnolo N, Lusicic N.	Endoscopic augmentation mastoplasty.	*Aesthetic Plast Surg* (1994) **18**(4): 377–82.
Serletti JM, Davenport MS, Herrera HR, Caldwell EH.	Efficacy of prophylactic antibiotics in reduction mammoplasty.	*Ann Plast Surg* (1994) **33**(5): 476–80.
Hakelius L, Ohlsen L.	Tendency to capsular contracture around smooth and textured gel-filled silicone mammary implants: a five-year follow-up.	*Plast Reconstr Surg* (1997) **100**(6): 1566–9.
Matarasso A, Wallach SG, Rankin M.	Reevaluating the need for routine drainage in reduction mammaplasty.	*Plast Reconstr Surg* (1998) **102**(6): 1917–21.
Giovanoli P, Meuli-Simmen C, Meyer VE, Frey M.	Which technique for which breast? A prospective study of different techniques of reduction mammaplasty.	*Br J Plast Surg* (1999) **52**(1): 52–9.
Collis N, Mirza S, Stanley PR, Campbell L, Sharpe DT.	Reduction of potential contamination of breast implants by the use of 'nipple shields'.	*Br J Plast Surg* (1999) **52**(6): 445–7.
Luzzati R, Sanna A, Allegranzi B, Nardi S, Berti M, Barisoni D, Concia E.	Pharmacokinetics and tissue penetration of vancomycin in patients undergoing prosthetic mammary surgery.	*J Antimicrob Chemother* (2000) **45**(2): 243–5.
Behmand RA, Tang DH, Smith DJ Jr.	Outcomes in breast reduction surgery.	*Ann Plast Surg* (2000) **45**(6): 575–80.
Chao JD, Memmel HC, Redding JF, Egan L, Odom LC, Casas LA.	Reduction mammaplasty is a functional operation, improving quality of life in symptomatic women: a prospective, single-center breast reduction outcome study.	*Plast Reconstr Surg* (2002) **110**(7): 1644–52; discussion 1653–4.

Table 31.1. (*continued*)

Authors	Title	Reference
Collins ED, Kerrigan CL, Kim M, Lowery JC, Striplin DT, Cunningham B, Wilkins EG.	The effectiveness of surgical and nonsurgical interventions in relieving the symptoms of macromastia.	*Plast Reconstr Surg*. (2002 **109**(5): 1556–66.
Berthe JV, Massaut J, Greuse M, Coessens B, De Mey A.	The vertical mammaplasty: a reappraisal of the technique and its complications.	*Plast Reconstr Surg* (2003) **111**(7): 2192–9; discussion 2200–2.
Wang TD.	Multicenter evaluation of subcutaneous augmentation material implants.	*Arch Facial Plast Surg* (2003) **5**(2): 153–4.

Table 31.2. Randomised controlled clinical trials in aesthetic surgery

Authors	Title	Reference
Facial rejuvenation		
Guyuron B.	Is packing after septorhinoplasty necessary? A randomized study.	*Plast Reconstr Surg* (1989) **84**(1): 41–4; discussion 45–6.
Griffies WS, Kennedy K, Gasser C, Fankhauser C, Taylor R.	Steroids in rhinoplasty.	*Laryngoscope* (1989) **99**(11): 1161–4.
Hevia O, Nemeth AJ, Taylor Jr.	Tretinoin accelerates healing after trichloroacetic acid chemical peel.	*Arch Dermatol* (1991) **127**(5): 678–82.
Hoffmann DF, Cook TA, Quatela VC, Wang TD, Brownrigg PJ, Brummett RE.	Steroids and rhinoplasty. A double-blind study.	*Arch Otolaryngol Head Neck Surg* (1991) **117**(9): 990–3; discussion 994.
Scaccia FJ, Hoffman JA, Stepnick DW.	Upper eyelid blepharoplasty. A technical comparative analysis.	*Arch Otolaryngol Head Neck Surg* (1994) **120**(8): 827–30.
Echavez MI, Mangat DS.	Effects of steroids on mood, edema, and ecchymosis in facial plastic surgery.	*Arch Otolaryngol Head Neck Surg* (1994) **120**(10): 1137–41.
Rapaport DP, Bass LS, Aston SJ.	Influence of steroids on postoperative swelling after facialplasty: a prospective, randomized study.	*Plast Reconstr Surg* (1995) **96**(7): 1547–52.
Humphreys TR, Werth V, Dzubow L, Kligman A.	Treatment of photodamaged skin with trichloroacetic acid and topical tretinoin.	*J Am Acad Dermatol* (1996) **34**(4): 638–44.
Piacquadio D, Dobry M, Hunt S, Andree C, Grove G, Hollenbach KA.	Short contact 70% glycolic acid peels as a treatment for photodamaged skin. A pilot study.	*Dermatol Surg* (1996) **22**(5): 449–52.
Owsley JQ, Weibel TJ, Adams WA.	Does steroid medication reduce facial edema following face lift surgery? A prospective, randomized study of 30 consecutive patients.	*Plast Reconstr Surg* (1996) **98**(1): 1–6.
Nasri S, Newman JP, Goode RL, Koch RJ.	Combined use of superpulsed carbon dioxide laser and cryotherapy for treatment of facial rhytids.	*Arch Otolaryngol Head Neck Surg* (1996) **122**(11): 1169–73. Erratum in: Arch Otolaryngol Head Neck Surg (1997) **123**(1): 46.

(*continued overleaf*)

Table 31.2. (*continued*)

Authors	Title	Reference
Ivy EJ, Lorenc ZP, Aston SJ.	Is there a difference? A prospective study comparing lateral and standard SMAS face lifts with extended SMAS and composite rhytidectomies.	*Plast Reconstr Surg* (1996) **98**(7): 1135–43; discussion 1144–7.
Gilbert SE.	Alar reductions in rhinoplasty.	*Arch Otolaryngol Head Neck Surg* (1996) **122**(7): 781–4.
Camirand A, Doucet J.	A comparison between parallel hairline incisions and perpendicular incisions when performing a face lift.	*Plast Reconstr Surg* (1997) **99**(1): 10–15.
Burns RL, Prevost-Blank PL, Lawry MA, Lawry TB, Faria DT, Fivenson DP.	Glycolic acid peels for postinflammatory hyperpigmentation in black patients. A comparative study.	*Dermatol Surg* (1997) **23**(3): 171–4; discussion 175.
Lim JT, Tham SN.	Glycolic acid peels in the treatment of melasma among Asian women.	*Dermatol Surg* (1997) **23**(3): 177–9.
Ross EV, Grossman MC, Duke D, Grevelink JM.	Long-term results after CO_2 laser skin resurfacing: a comparison of scanned and pulsed systems.	*J Am Acad Dermatol* 1997 **37**(5 Pt 1): 709–18.
Berinstein TH, Bane SM, Cupp CL, DeMarco JK, Hunsaker DH.	Steroid use in rhinoplasty: an objective assessment of postoperative edema.	*Ear Nose Throat J* (1998) **77**(1): 40–3.
Dailey RA, Gray JF, Rubin MG, Hildebrand PL, Swanson NA, Wobig JL, Wilson DJ, Speelman P.	Histopathologic changes of the eyelid skin following trichloroacetic acid chemical peel.	*Ophthal Plast Reconstr Surg* (1998) **14**(1): 9–12.
Marrero GM, Katz BE.	The new fluor-hydroxy pulse peel. A combination of 5-fluorouracil and glycolic acid.	*Dermatol Surg* (1998) **24**(9): 973–8.
Gross EA, Rogers GS.	A side-by-side comparison of carbon dioxide resurfacing lasers for the treatment of rhytides.	*J Am Acad Dermatol* (1998) **39**(4 Pt 1): 547–53.
Duke D, Khatri K, Grevelink JM, Anderson RR.	Comparative clinical trial of 2 carbon dioxide resurfacing lasers with varying pulse durations. 100 microseconds vs 1 millisecond.	*Arch Dermatol* (1998) **134**(10): 1240–6.
Nassif PS, Kokoska MS, Homan S, Cooper MH, Thomas JR	Comparison of subperiosteal vs subgaleal elevation techniques used in forehead lifts.	*Arch Otolaryngol Head Neck Surg* (1998) **124**(11): 1209–15.
Alster TS, Nanni CA, Williams CM.	Comparison of four carbon dioxide resurfacing lasers. A clinical and histopathologic evaluation.	*Dermatol Surg* (1999) **25**(3): 153–8; discussion 159.
Goldman MP, Manuskiatti W.	Combined laser resurfacing with the 950-microsec pulsed CO_2 + Er:YAG lasers.	*Dermatol Surg* (1999) **25**(3): 160–3.
Li YT, Yang KC.	Comparison of the frequency-doubled Q-switched Nd:YAG laser and 35% trichloroacetic acid for the treatment of face lentigines.	*Dermatol Surg* (1999) **25**(3): 202–4.
West TB, Alster TS.	Effect of botulinum toxin type A on movement-associated rhytides following CO_2 laser resurfacing.	*Dermatol Surg* (1999) **25**(4): 259–61.

Table 31.2. (*continued*)

Authors	Title	Reference
Chew J, Gin I, Rau KA, Amos DB, Bridenstine JB.	Treatment of upper lip wrinkles: a comparison of 950 microsec dwell time carbon dioxide laser with unoccluded Baker's phenol chemical peel.	*Dermatol Surg* (1999) **25**(4): 262–6.
Kim SW, Moon SE, Kim JA, Eun HC.	Glycolic acid versus Jessner's solution: which is better for facial acne patients? A randomized prospective clinical trial of split-face model therapy.	*Dermatol Surg* (1999) **25**(4): 270–3.
McDaniel DH, Lord J, Ash K, Newman J.	Combined CO_2/erbium:YAG laser resurfacing of peri-oral rhytides and side-by-side comparison with carbon dioxide laser alone.	*Dermatol Surg* (1999) **25**(4): 285–93.
Alster TS.	Clinical and histologic evaluation of six erbium:YAG lasers for cutaneous resurfacing.	*Lasers Surg Med* (1999) **24**(2): 87–92.
Arambula H, Sierra-Martinez E, Gonzalez-Aguirre NE, Rodriguez-Perez A, Juarez-Aguilar E, Marsch-Moreno M, Kuri-Harcuch W.	Frozen human epidermal allogeneic cultures promote rapid healing of facial dermabrasion wounds.	*Dermatol Surg* (1999) **25**(9): 708–12.
Traikovich SS.	Use of topical ascorbic acid and its effects on photodamaged skin topography.	*Arch Otolaryngol Head Neck Surg* (1999) **125**(10): 1091–8.
Greene D, Koch RJ, Goode RL.	Efficacy of octyl-2-cyanoacrylate tissue glue in blepharoplasty. A prospective controlled study of wound-healing characteristics.	*Arch Facial Plast Surg* (1999) **1**(4): 292–6.
Gladstone HB, Nguyen SL, Williams R, Ottomeyer T, Wortzman M, Jeffers M, Moy RL.	Efficacy of hydroquinone cream (USP 4%) used alone or in combination with salicylic acid peels in improving photodamage on the neck and upper chest.	*Dermatol Surg* (2000) **26**(4): 333–7.
Holmkvist KA, Rogers GS.	Treatment of perioral rhytides: a comparison of dermabrasion and superpulsed carbon dioxide laser.	*Arch Dermatol* (2000) **136**(6): 725–31.
Newman JB, Lord JL, Ash K, McDaniel DH.	Variable pulse erbium:YAG laser skin resurfacing of perioral rhytides and side-by-side comparison with carbon dioxide laser.	*Lasers Surg Med* (2000) **26**(2): 208–14.
Hernandez-Perez E, Khawaja HA, Alvarez TY.	Oral isotretinoin as part of the treatment of cutaneous aging.	*Dermatol Surg* (2000) **26**(7): 649–52.
Stone PA, Lefer LG.	Modified phenol chemical face peels: recognizing the role of application technique.	*Clin Plast Surg* (2001) **28**(1): 13–36. Erratum in: *Clin Plast Surg* (2001) **28**(3): preceding xi.
Carter SR, Seiff SR, Choo PH, Vallabhanath P.	Lower eyelid CO_2 laser rejuvenation: a randomized, prospective clinical study.	*Ophthalmology* (2001) **108**(3): 437–41.
Ross EV, Miller C, Meehan K, McKinlay J, Sajben P, Trafeli JP, Barnette DJ.	One-pass CO_2 versus multiple-pass Er:YAG laser resurfacing in the treatment of rhytides: a comparison side-by-side study of pulsed CO_2 and Er:YAG lasers.	*Dermatol Surg* (2001) **27**(8): 709–15.

(*continued overleaf*)

Table 31.2. (*continued*)

Authors	Title	Reference
Zimbler MS, Holds JB, Kokoska MS, Glaser DA, Prendiville S, Hollenbeak CS, Thomas JR	Effect of botulinum toxin pretreatment on laser resurfacing results: a prospective, randomized, blinded trial.	*Arch Facial Plast Surg* (2001) **3**(3): 165–9.
Bernstein EF, Ferreira M, Anderson D.	A pilot investigation to subjectively measure treatment effect and side-effect profile of non-ablative skin remodeling using a 532 nm, 2 ms pulse-duration laser.	*J Cosmet Laser Ther* (2001) **3**(3): 137–41.
Rostan E, Bowes LE, Iyer S, Fitzpatrick RE.	A double-blind, side-by-side comparison study of low fluence long pulse dye laser to coolant treatment for wrinkling of the cheeks.	*J Cosmet Laser Ther* (2001) **3**(3): 129–36.
de Noronha L, Chin EW, Menini CM, Knopfholz J, Rampazzo JC, Graf R.	Histopathologic and morphometric evaluation of the skin abnormalities induced by erbium:YAG and carbon dioxide lasers in 10 patients.	*Plast Reconstr Surg* (2001) **108**(5): 1380–8.
Powell DM, Chang E, Farrior EH.	Recovery from deep-plane rhytidectomy following unilateral wound treatment with autologous platelet gel: a pilot study.	*Arch Facial Plast Surg* (2001) **3**(4): 245–50.
Oliver DW, Hamilton SA, Figle AA, Wood SH, Lamberty BG.	A prospective, randomized, double-blind trial of the use of fibrin sealant for face lifts.	*Plast Reconstr Surg* (2001) **108**(7): 2101–5; discussion 2106–7.
Baker SS, Hunnewell JM, Muenzler WS, Hunter GJ.	Laser blepharoplasty: diamond laser scalpel compared to the free beam CO_2 laser.	*Dermatol Surg* (2002) **28**(2): 127–31.
Goldman MP, Roberts TL 3rd, Skover G, Lettieri JT, Fitzpatrick RE.	Optimizing wound healing in the face after laser abrasion.	*J Am Acad Dermatol* (2002) **46**(3): 399–407.
Sulimovic L, Licu D, Ledo E, Naeyaert JM, Pigatto P, Tzermias C, Vasquez Doval J, Dupuy P.	Efficacy and safety of a topically applied Avene spring water spray in the healing of facial skin after laser resurfacing.	*Dermatol Surg* (2002) **28**(5): 415–8; discussion 418.
Alam M, Omura NE, Dover JS, Arndt KA.	Glycolic acid peels compared to microdermabrasion: a right-left controlled trial of efficacy and patient satisfaction.	*Dermatol Surg* (2002) **28**(6): 475–9.
Yun PL, Tachihara R, Anderson RR.	Efficacy of erbium:yttrium-aluminum-garnet laser-assisted delivery of topical anesthetic.	*J Am Acad Dermatol* (2002) **47**(4): 542–7.
Alster T.	Laser scar revision: comparison study of 585-nm pulsed dye laser with and without intralesional corticosteroids.	*Dermatol Surg* (2003) **29**(1): 25–9.
Ahn MS, Maas CS, Monhian N.	A novel, conformable, rapidly setting nasal splint material: results of a prospective study.	*Arch Facial Plast Surg* (2003) **5**(2): 189–92.
Poulos E, Taylor C, Solish N.	Effectiveness of dermasanding (manual dermabrasion) on the appearance of surgical scars: a prospective, randomized, blinded study.	*J Am Acad Dermatol* (2003) **48**(6): 897–900.
Paithankar DY, Clifford JM, Saleh BA, Ross EV, Hardaway CA, Barnette D.	Subsurface skin renewal by treatment with a 1450-nm laser in combination with dynamic cooling.	*J Biomed Opt* (2003) **8**(3): 545–51.

Table 31.2. (*continued*)

Authors	Title	Reference
Kargi E, Hosnuter M, Babuccu O, Altunkaya H, Altinyazar C.	Effect of steroids on edema, ecchymosis, and intraoperative bleeding in rhinoplasty.	*Ann Plast Surg* (2003) **51**(6): 570–4.
Gryskiewicz JM, Gryskiewicz KM.	Nasal osteotomies: a clinical comparison of the perforating methods versus the continuous technique.	*Plast Reconstr Surg*. (2004) **113**(5): 1445–56; discussion 1457–8.
Becker FF, Bassichis BA.	Deep-plane face-lift vs superficial musculoaponeurotic system plication face-lift: a comparative study.	*Arch Facial Plast Surg* (2004) **6**(1): 8–13.
Makdessian AS, Ellis DA, Irish JC.	Informed consent in facial plastic surgery: effectiveness of a simple educational intervention.	*Arch Facial Plast Surg* (2004) **6**(1): 26–30.
Nanda S, Grover C, Reddy BS.	Efficacy of hydroquinone (2%) versus tretinoin (0.025%) as adjunct topical agents for chemical peeling in patients of melasma.	*Dermatol Surg* (2004) **30**(3): 385–8; discussion 389.
Bajaj MS, Sastry SS, Ghose S, Betharia SM, Pushker N.	Evaluation of polytetrafluoroethylene suture for frontalis suspension as compared to polybutylate-coated braided polyester.	*Clin Exp Ophthalmol* (2004) **32**(4): 415–19.
Orringer JS, Johnson TM, Kang S, Karimipour DJ, Hammerberg C, Hamilton T, Voorhees JJ, Fisher GJ.	Effect of carbon dioxide laser resurfacing on epidermal p53 immunostaining in photodamaged skin.	*Arch Dermatol* (2004) **140**(9): 1073–7.
Cohen SR, Holmes RE.	Artecoll: a long-lasting injectable wrinkle filler material: report of a controlled, randomized, multicenter clinical trial of 251 subjects.	*Plast Reconstr Surg*. (2004) **114**(4): 964–76; discussion 977–9.
Body contouring		
Courtiss EH, Kanter MA, Kanter WR, Ransil BJ.	The effect of epinephrine on blood loss during suction lipectomy.	*Plast Reconstr Surg* (1991) **88**(5): 801–3.
Lalinde E, Sanz J, Ballesteros A, Elejabeitia J, Mesa F, Bazan A, Paloma V.	Effect of L-ornithine 8-vasopressin on blood loss during liposuction.	*Ann Plast Surg* (1995) **34**(6): 613–18.
Kaplan B, Moy RL.	Comparison of room temperature and warmed local anesthetic solution for tumescent liposuction. A randomized double-blind study.	*Dermatol Surg* (1996) **22**(8): 707–9.
Havoonjian HH, Luftman DB, Menaker GM, Moy RL.	External ultrasonic tumescent liposuction. A preliminary study.	*Dermatol Surg* (1997) **23**(12): 1201–6.
Cook WR Jr.	Utilizing external ultrasonic energy to improve the results of tumescent liposculpture.	*Dermatol Surg* (1997) **23**(12): 1207–11.
Igra H, Satur NM.	Tumescent liposuction versus internal ultrasonic-assisted tumescent liposuction. A side-to-side comparison.	*Dermatol Surg* (1997) **23**(12): 1213–18.
Rubin JP, Bierman C, Rosow CE, Arthur GR, Chang Y, Courtiss EH, May JW Jr.	The tumescent technique: the effect of high tissue pressure and dilute epinephrine on absorption of lidocaine.	*Plast Reconstr Surg* (1999) **103**(3): 990–6; discussion 997–1002.
Man D, Man B, Plosker H.	The influence of permanent magnetic field therapy on wound healing in suction lipectomy patients: a double-blind study.	*Plast Reconstr Surg* (1999) **104**(7): 2261–6; discussion 2267–8.

(*continued overleaf*)

Table 31.2. (*continued*)

Authors	Title	Reference
Lawrence N, Cox SE.	The efficacy of external ultrasound-assisted liposuction: a randomized controlled trial.	*Dermatol Surg* (2000) **26**(4): 329–32.
Gonzalez-Ortiz M, Robles-Cervantes JA, Cardenas-Camarena L, Bustos-Saldana R, Martinez-Abundis E.	The effects of surgically removing subcutaneous fat on the metabolic profile and insulin sensitivity in obese women after large-volume liposuction treatment.	*Horm Metab Res* (2002) **34**(8): 446–9.
Butterwick KJ.	Lipoaugmentation for aging hands: a comparison of the longevity and aesthetic results of centrifuged versus noncentrifuged fat.	*Dermatol Surg* (2002) **28**(11): 987–91.
Omranifard M.	Ultrasonic liposuction versus surgical lipectomy.	*Aesthetic Plast Surg* (2003) **27**(2): 143–5.
Breast Surgery		
Burkhardt BR, Dempsey PD, Schnur PL, Tofield JJ.	Capsular contracture: a prospective study of the effect of local antibacterial agents.	*Plast Reconstr Surg* (1986) **77**(6): 919–32.
Gylbert L, Asplund O, Berggren A, Jurell G, Ransjo U, Ostrup L.	Preoperative antibiotics and capsular contracture in augmentation mammaplasty.	*Plast Reconstr Surg* (1990) **86**(2): 260–7; discussion 268–9.
Coleman DJ, Foo IT, Sharpe DT.	Textured or smooth implants for breast augmentation? A prospective controlled trial.	*Br J Plast Surg* (1991) **44**(6): 444–8.
Burkhardt BR, Demas CP.	The effect of Siltex texturing and povidone-iodine irrigation on capsular contracture around saline-inflatable breast implants.	*Plast Reconstr Surg* (1994) **93**(1): 123–8; discussion 129–30.
Burkhardt BR, Eades E.	The effect of Biocell texturing and povidone-iodine irrigation on capsular contracture around saline-inflatable breast implants.	*Plast Reconstr Surg* (1995) **96**(6): 1317–25.
Asplund O, Gylbert L, Jurell G, Ward C.	Textured or smooth implants for submuscular breast augmentation: a controlled study.	*Plast Reconstr Surg* (1996) **97**(6): 1200–6.
Malata CM, Feldberg L, Coleman DJ, Foo IT, Sharpe DT.	Textured or smooth implants for breast augmentation? Three year follow-up of a prospective randomised controlled trial.	*Br J Plast Surg* (1997) **50**(2): 99–105.
Tarpila E, Ghassemifar R, Fagrell D, Berggren A.	Capsular contracture with textured versus smooth saline-filled implants for breast augmentation: a prospective clinical study.	*Plast Reconstr Surg* (1997) **99**(7): 1934–9.
Liddle AM, Hall AP, Arrowsmith J, Smith G.	Effect of infiltration with ropivacaine on blood loss during reduction mammoplasty.	*Br J Anaesth* (1998) **81**(6): 974–5.
Metaxotos NG, Asplund O, Hayes M.	The efficacy of bupivacaine with adrenaline in reducing pain and bleeding associated with breast reduction: a prospective trial.	*Br J Plast Surg* (1999) **52**(4): 290–3.
Bell RF, Sivertsen A, Mowinkel P, Vindenes H.	A bilateral clinical model for the study of acute and chronic pain after breast-reduction surgery.	*Acta Anaesthesiol Scand* (2001) **45**(5): 576–82.
Nathan B, Singh S.	Postoperative compression after breast augmentation.	*Aesthetic Plast Surg* (2001) **25**(4): 290–1.
Fagrell D, Berggren A, Tarpila E.	Capsular contracture around saline-filled fine textured and smooth mammary implants: a prospective 7.5-year follow-up.	*Plast Reconstr Surg* (2001) **108**(7): 2108–12; discussion 2113.

Table 31.2. (*continued*)

Authors	Title	Reference
Wrye SW, Banducci DR, Mackay D, Graham WP, Hall WW.	Routine drainage is not required in reduction mammaplasty.	*Plast Reconstr Surg* (2003) **111**(1): 113–7.
Cruz-Korchin N, Korchin L.	Vertical versus Wise pattern breast reduction: patient satisfaction, revision rates, and complications.	*Plast Reconstr Surg* (2003) **112**(6): 1573–8; discussion 1579–81.
Platt AJ, Mohan D, Baguley P.	The effect of body mass index and wound irrigation on outcome after bilateral breast reduction.	*Ann Plast Surg* (2003) **51**(6): 552–5.

breast augmentation, especially with regard to capsular contracture rate.

REFERENCES

1. Bradbury E. The psychology of aesthetic plastic surgery. *Aesthetic Plast Surg* (1994) **18**: 301.
2. Pruzinsky T. Psychological factors in cosmetic plastic surgery: recent developments in patient care. *Plast Surg Nurs* (1993) **13**: 64.
3. Goin MK, Burgoyne RW, Goin JM, Staples FR. A prospective psychological study of 50 female face-lift patients. *Plast Reconstr Surg* (1980) **65**: 436.
4. Goin MK, Goin JM. Psychological effects of aesthetic plastic surgery. *Adv Psychosomat Med* **15**: 84, (1986).
5. Rohrich RJ, Janis JE, Fagien S, Stuzin JM. The cosmetic use of botulinum toxin. *Plast Reconstr Surg* (2003) **112**(Suppl.): 177S–187S.
6. Carruthers J, Fagien S, Matarasso SL and the Botox Consensus Group. Consensus recommendations on the use of botulinum toxin type A in facial aesthetics. *Plast Reconstr Surg* (2004) **114**(Suppl.): 1S–22S.
7. Spear SL, Bulan EJ, Venturi ML. Breast augmentation. *Plast Reconstr Surg* (2004) **114**: 73e–81e.
8. Rohrich RJ, Kenkel JM, Adams WP, Beran S, Conner WCH. A prospective analysis of patients undergoing silicone breast implant explantation. *Plast Reconstr Surg* (2000) **105**: 2529–37.
9. Banbury J, Yetman R, Lucas A, Papay F, Graves K, Zins JE. Prospective analysis of the outcome of subpectoral breast augmentation: sensory changes, muscle function, and body image. *Plast Reconstr Surg* (2004) **113**: 701–7.

TRANSPLANTATION

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green

32

Renal Transplantation

VATHSALA ANANTHARAMAN
Renal Medicine, Singapore General Hospital, Singapore

INTRODUCTION

Renal transplantation is the treatment of choice for most patients with end stage renal failure. The ability to reverse the complications of renal failure by implanting a kidney taken from another individual has been hailed as a miracle of modern medicine. Renal transplantation (RTx) extends the lives of recipients and promotes excellent quality of life for patients with renal failure.

However, as every individual's immune system is designed to distinguish self from non-self, kidney transplants are invariably rejected by the transplant recipient as 'non-self' or 'foreign' unless immunosuppressive (IS) drugs are administered. IS drugs, when administered to patients undergoing RTx, lead to suppression of the patient's immune responses and prevent rejection of the renal transplant. By the same mechanisms, however, IS drugs also reduce the transplant patient's ability to fight infection and cancers. As a corollary, patients receiving more potent drugs are at higher risk for these complications Finally, IS drugs are also associated with other pleomorphic side effects that may contribute to other morbidities following RTx. Hence, management of immunosuppression is a major challenge in the field of organ transplantation: inadequate immunosuppression, on the one hand, results in graft failure due to rejection, and, on the other hand, excess immunosuppression results in patient morbidity and mortality due to infections and malignancy or the risks of non-immunosuppressive toxicities of the IS drugs.

There are many approaches to optimising IS therapy for the individual patient. First, identifying patients at higher risk for rejection would allow tailoring of therapy, i.e. permit administration of more potent immunosuppression to those with higher risk. For example, donor–recipient pairs with greater genetic disparity and recipients with greater immune reactivity would be at higher risk for rejection and would thus benefit from more potent immunosuppression. On the other hand, potent immunosuppression would be unnecessary and impose undue risks for the well-matched sibling donor–recipient pair at low risk for rejection. Second, several IS drugs can be used in combination with each other, below their putative 'toxic' threshold, so as to maximise their IS efficacy while minimising their complications. Stratifying an individual patient's risk also allows different patients to receive different combinations of IS drugs. Third, pharmacokinetic (PKA)

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green

monitoring of IS drug levels, so as to achieve target levels within a therapeutic window, is also commonly performed to optimise IS drug doses. Finally, IS therapy can be guided by monitoring of renal function, other laboratory parameters, renal biopsies to detect histological evidence of rejection, fibrosis or even other biomarkers of renal damage.

With the availability of a large number of IS drugs, compounded with the numerous combinations that these drugs can be used in, there are now many options for therapy of the individual patient (Table 32.1). Thus, the decision-making process has become increasingly complex and choice has to be based on evidence, mainly obtained from clinical trials. This chapter will discuss clinical trials in RTx, endpoints in trial design and issues on determination of study population and therapy in the control population. This chapter will also evaluate clinical trials in RTx performed over the last decade, with a focus on the impact of IS drugs on post-transplant

Table 32.1. Immunosuppressive drugs available for use in renal transplantation

Immunosuppressant class	Immunosuppressant drug names	Significant non-immunosuppressive toxicities
Corticosteroids (CS)	Prednisone Methylprednisolone	– Osteoporosis – Hyperglycaemia – Weight gain
Calcineurin Inhibitors (CNI)	Cyclosporine (CYA or Neoral, Sandimmun)[a] Tacrolimus (TAC, FK506 or Prograf)[b]	– Nephrotoxicity – Neurotoxicity – Hypertension – Hyperlipidaemia[a] – Hyperglycaemia[b]
Antimetabolites	Azathioprine (AZA or Imuran)	– Cytopenia – Liver dysfunction – Pancreatitis
	Mycophenolate (MPA) – mofetil (or Cellcept) – sodium (or Myfortic)	– Cytopenia – Gastrointestinal toxicity
Polyclonal antilymphocyte antibodies	Antilymphocyte globulin (ALG) Antithymocyte globulin (ATG)	– Lymphopenia
Anti-CD3 monoclonal antibody	Muromonab-CD3 (OKT3)	– Cytokine release syndrome – Aseptic meningitis
Anti-interleukin 2 receptor monoclonal antibodies (IL2RAb)	Basiliximab (or Simulect) Daclizumab (or Zenapax)	
Inhibitors of mammalian target of rapamycin (MTORi)	Sirolimus (SRL or Rapamune) Everolimus (ERL or Certican)	– Hyperlipidaemia – Cytopenia – Poor wound healing

[a] Complication present more commonly in patients treated with Cyclosporine.

[b] Complication present more commonly in patients treated with Tacrolimus.

outcomes and complications as well as on their clinical use.

MEDICAL BACKGROUND

IS DRUGS

Historically, corticosteroids (CS) and the antimetabolite Azathioprine (AZA) were the principal immunosuppressants used in clinical RTx; by the mid-1980s the calcineurin inhibitor (CNI), cyclosporine (CYA), was approved for use in combination with CS and AZA and polyclonal antilymphocyte antibody preparations and the monoclonal antibody against the lymphocyte surface molecule CD3 (Muromonab-CD3, OKT3) were available for initial immunosuppression or treatment of rejection (Table 32.1). In the last decade, many new IS drugs have been approved for use in clinical RTx: another CNI, tacrolimus (TAC), an improved microemulsion formulation of CYA (Neoral), mycophenolic acid (MPA) analogues, the mammalian target of Rapamycin inhibitor (MTORi), sirolimus (SRL), as well as several new antilymphocyte antibody preparations such as rabbit antithymocyte globulin (ATG) and the anti-interleukin 2 receptor antibodies (IL2RAb), daclizumab and basiliximab. The introduction of these IS drugs has increased the number of options available for immunosuppression.

Each of these classes of immunosuppressant has its unique, pleotrophic, non-IS drug toxicities (Table 32.1) and IS drugs are often used in combination to maximise IS efficacy while minimising toxicities. Usual IS drug regimens consist of a CNI (either CYA or TAC) as monotherapy or in dual or triple combination with or without an antimetabolite (any one) and/or a corticosteroid; CNI-based therapy alone allows for a potential of 16 IS drug regimens. The availability of antibodies for induction therapy and newer combinations using MTORi as base therapy offer additional combinations and add complexity to choosing the correct IS drug regimen for the individual patient.

OUTCOMES IN RTx

Following RTx, the majority of transplants function well for many years; however, a significant proportion may be lost to rejection (acute or chronic), vascular thrombosis, recurrent disease, CNI toxicity or chronic allograft nephropathy, a fibro-sclerotic change which occurs in transplant kidneys likely due to a combination of factors. Though graft loss can occur suddenly, more often, graft loss occurs gradually over time post-RTx; in these cases, renal function gradually deteriorates and when there is inadequate renal function to sustain life (usually with renal function <10% to 15% of normal), the patient with graft loss either returns to dialysis or undergoes retransplantation. RTx patients may also die of various causes such as malignancy, infections and cardiovascular disease; indeed, RTx patients may lose their transplant and die at the same time or soon after.

Hence, there are many measures of the success of RTx as shown in Table 32.2. Graft and patient survivals, the converse of graft loss and patient death, are calculated by actuarial methods

Table 32.2. Measures of successful outcome in renal transplantation

Outcome measure	Incidence/ prevalence
Good graft survival	
1 year	89% to 95%
5 years	66% to 81%
10 years	36% to 58%
Good patient survival	
1 year	95% to 97%
5 years	79% to 90%
10 years	54% to 76%
Low rates of rejection	15% to 39%
Good renal function (serum creatinine <1.5 mg/dL)	57.7% at 1 year
Graft (or patient) half-life ($T_{1/2}$)	9 to 22 years
Reduced risk of complications of:	
Cancer	20% for solid cancers at 20 years
Infections	Varies with organism
Hypertension	35% to 60%
Post-transplant diabetes mellitus	2% to 20%
Hyperlipidaemia	60%
Bone disease	5% to 44%

at specific time points (e.g. 1 year, 5 years, etc.). Though patient death is easily defined and confirmed, patients who die with a functioning transplant are often censored from analysis of graft survival, at time of death, under the premise that their death was due to unrelated factors such as the underlying disease or its co-morbidities. On the other hand, it has been argued that patient death is contributed to by the transplant procedure itself or the IS drugs and is therefore a complication of transplantation; as such, patient deaths with a functioning graft should not be censored in analyses of graft survival, but instead be treated as a graft loss. Graft survivals that are censored for death will appear to be higher than uncensored graft survivals and outcomes can be misconstrued to be better than they actually are if censored graft survivals are reported.

The 'half-life' is yet another commonly reported outcome post-transplant; it is an estimate of the time taken for half the transplants to be lost post-RTx and is calculated by plotting graft or patient survival on a semilogarithmic scale vs. time. However, as 'half-life' reflects projected data and not actual data, their value in reporting outcomes of clinical trials is likely limited. Complications of transplantation, as listed in Table 32.2, are often due to IS drugs and contribute to patient death and graft loss; as such they are important outcome measures. Other outcome parameters that are often relevant in clinical practice are economic and quality of life issues.

Given the wide variations in the patient's risk categories and the type of immunosuppression used, there are wide variations in outcomes and complications post-RTx. As can be seen, the majority of RTx recipients suffer various complications and thus there are many unmet needs in RTx.

CLINICAL TRIALS IN RTx

PHASES OF CLINICAL TRIALS IN RTx

Before an IS drug is subject to clinical trials, it undergoes single dose PKA studies in healthy individuals and renal transplant recipients. Multiple dose PKA, dose searching studies, studies of intra-subject and inter-subject variability, impact of food on IS drug PKA and interactions between IS drugs are other studies performed in RTx patients, often before a full-scale clinical trial is embarked upon. In subsequent Phase II studies, the trial drug is used, generally in a dose-finding study, in combination with other IS drugs, comparing the efficacy and safety of the combination with that of the control population.

Given the potential risk for cytokine release and its side effects following the administration of certain monoclonal antibodies, new guidelines for clinical trials utilising monoclonal antibodies may demand Phase I and II studies in patients rather than healthy subjects.[1]

ENDPOINTS FOR CLINICAL TRIALS IN RTx

In the design of any clinical trial, the study regimen, in comparison with that used in the control population, is hypothesised to improve outcome, be associated with less complications, or alternatively be equivalent. The majority of clinical trials in RTx are of IS drugs and the primary trial endpoints are generally based on the measures of successful post-transplant outcome (Table 32.2) or economic and quality of life issues. Broadly, these outcome measures can be categorised into two groups: (1) those that directly test the efficacy of an intervention and (2) those that test its safety. However, neither efficacy nor safety can be viewed in isolation from each other. For example, an IS drug combination may be so potent and effective that transplant rejection is minimal, but the regimen may yet be so excessively IS, that patients die from either infection or malignancy. Thus, ideal trial endpoints in RTx measure both efficacy and safety parameters or a composite of both.

While the ultimate success following RTx is determined by long-term (e.g. 5 years or 10 years) patient or graft survival, long-term trials are difficult to design and conduct for many reasons. First, waiting for long-term results of IS drugs would prevent their early introduction into the clinical arena. Second, long-term trials would be exceedingly expensive given the long duration of follow-up of a large group of patients.

Hunsicker and Bennett have suggested that if graft or patient survivals at three years were to be used as a trial endpoint, 1500 patients would need to be recruited and followed up for three years, an endeavour not easy to achieve in a rapidly changing IS scene.[2]

As such, most trials in clinical RTx use surrogate endpoints instead of long-term graft and patient survivals. One such surrogate endpoint frequently used is the incidence of acute rejection (AREJ). There are many variations in the diagnosis and reporting of AREJ: clinical, biopsy proven or by protocol biopsy. The severity of AREJ can also be graded clinically: by stratification based on response to CS (steroid-responsive vs. steroid-resistant AREJ) or by histological severity (e.g. BANFF grading).[3,4] Use of standard criteria for diagnosis and grading of AREJ permits comparison across trials.

However, transplant registry data suggest that though the occurrence of a single AREJ episode is associated with reduced short- and long-term graft survivals, this association is not universal.[5] In fact, when the serum creatinine (SCr) at discharge was not significantly impaired, early rejection had little, if any, effect on short- or long-term outcomes. Indeed, from US registry data, the introduction of newer IS drugs into clinical RTx in the United States in the last decade has resulted in a significant reduction in the incidence of AREJ from ~40% to 50% in 1995/1996 to 15% to 17% in 2001/2002.[6] However, one-year graft survival has only improved marginally from 87.9% in 1996 to 89% in 2002, while one-year patient survival has remained virtually unchanged (94.6% in 1996 and 94.5% in 2002).[7,8]

Renal function at six months or one year, as measured by SCr, has also been suggested as a surrogate endpoint.[9] However, it is well established that SCr correlates poorly with renal function in both native kidney disease and following RTx and the search for better surrogate endpoints for clinical trials in RTx continues.[10] The endpoints selected ultimately will determine the size of the study population and the duration of trial follow-up.

SELECTION OF THE SAMPLE POPULATION

As suggested earlier, high-risk patients have worse outcomes than low-risk RTx patients. In clinical trials in RTx, selection of the study population will define the outcomes for the control population and thus determine the number of patients that need to be recruited into the trial. As such, while in the early phases of these clinical trials, low-risk patients are selected as the sample population, high-risk patients are also recruited in the later phases, otherwise there would be little opportunity to determine the efficacy of the IS drug in high-risk recipients.

Ideally, the size of the study population should be predetermined based on the endpoints chosen and their prevalence in the control population. Statistical programs allow the sample size to be calculated for recruitment into clinical trials, given the known incidence/prevalence of the selected endpoint in the control population and the expected incidence/prevalence in the study population. On the other hand, practical considerations such as the numbers of transplant patients who can be recruited within a reasonable time frame will determine the trial size. As such, many investigator-sponsored trials and single centre trials are small in size.

Moreover, transplant trials are expensive and the resources needed to recruit large groups of patients and follow them up require support from pharmaceutical companies. Herein lies one of the dilemmas in the conduct of clinical trials in RTx: while a large clinical trial with adequate power to detect significant differences in trial arms may require sponsorship by a pharmaceutical company, vested interests on the part of the pharmaceutical company may result in bias in the trial design, with design favouring better outcomes in the study group.

TREATMENT IN THE CONTROL GROUP

Ethical principles dictate that the treatment chosen for the control group should not place the control population at higher risk for an event. This is becoming increasingly challenging over the last decade of clinical transplants in RTx.

In 1995, when the first clinical trial of an MPA analogue Mycophenolate Mofetil (MMF) was reported, the control (CYA + placebo) arm had an incidence of AREJ of 56% at six months, an incidence which was acceptable in clinical practice at the time.[11] The study arm that employed MMF had a significantly lower AREJ incidence of 30.3%. If a similar study were to be performed in 2005, the control arm would have to have an AREJ incidence of approximately 15% to keep up with current clinical practice.

However, Curtis and Kaplan have suggested that in transplant trials, the pharmaceutical industry, with the help of investigators, searches for control groups that would do less well.[12] They point out that though some IS drugs, for example IL2RAb, have been declared safe and effective and have become standard therapy at most transplant centres across the United States, these agents were not used either in the control group in trials sponsored by the same pharmaceutical companies. On the other hand, as the use of more potent IS drugs in the control group improves the incidence/prevalence of study endpoints in the control group, outcomes become difficult to improve upon and upcoming trials will necessarily need to recruit larger numbers of patients to remain relevant.

A separate issue is that of placebo control and double blinding in the context of clinical trials in RTx. Given the need for multiple IS drugs to prevent rejection in clinical practice, substitution of one IS drug with a placebo is likely unethical currently as it could lead to higher risk for AREJ in the control population. On the other hand, concealment of allocation is still an important issue as inadequate blinding to study treatment has been suggested to exaggerate treatment efficacy by 30%–40%.[13] The need for monitoring of multiple parameters following IS therapy (e.g. lymphocyte subsets, IS drug PKA) or the need for administration of some IS drugs intravenously have been suggested to hinder blinding in clinical trials in RTx.[14] Nevertheless, blinding can be performed with careful planning, especially for orally administered drugs.

TRIAL REPORTING

The reporting quality of clinical trials in RTx has come under great scrutiny in recent times. Following a literature search of trials on immunosuppression in RTx since 1990, Fritsche *et al.* found 861 publications; of the 63 publications from results of large, randomised, multicentre trials that were reviewed, the overall quality of the studies was poor (JADAD score 2.3).[14] In addition, several other weaknesses were noted: failure to conceal allocation (blinding) in 68.3% of studies, failure to report treated and biopsy-proven AREJ in 54% of studies, failure to report whether graft survivals were censored for death in 27% and short duration of follow-up in 74% of studies (<12 months).[14] The authors concluded that proper design of trials from the outset and adherence to the Consolidated Standards for Reporting Trials (CONSORT) statement at time of reporting would improve the quality of clinical trials in RTx.[15]

Other points of note in the quality of clinical trial reporting in RTx are the use of confusing definitions for endpoints or apparently inappropriate censoring of data. In a study comparing TAC with CYA, the authors reported improved allograft survival at five years for the former.[16] The predefined endpoints in this trial were as follows: 'graft failure', requirement for graft nephrectomy or permanent return to dialysis, graft loss, occurrence of death or graft failure and treatment failure, graft loss or discontinuation of randomised study drug. Though five-year graft and patient survivals between TAC and CYA were comparable on the intent-to-treat analysis, a reanalysis was performed after crossover between drugs due to a rejection episode was redefined also as 'graft failure', yielding their conclusion of better survival in the TAC group.

Likewise, Jurewicz, for the Welsh Transplant Research Group, reported that 'graft survivals' for TAC were significantly higher than for CYA-treated patients (81% vs. 60%, $p = 0.0496$).[17] However, it is apparent from the Kaplan–Meier plot that only graft survival beyond the first year was reported and that all graft losses within the first year were censored. Moreover, there was no explanation as to the reasons or justification for excluding the graft losses or patient deaths.

In a timely manner, the editors of the journal *Transplantation*, together with editors of five other journals in kidney diseases and transplantation, have indicated that the journals would consider research studies of clinical trials for publication only if the trial has been submitted to a free, electronically searchable, clinical trial register.[18] This policy will hopefully prevent the problem of selective reporting in clinical trials, i.e. the failure to report trials with negative results; such discipline will allow the entire body of evidence to be made available for critical review.

CLINICAL TRIALS IN RTx: 1995 TO 2004

A search of the PubMed database using the topics of 'randomised controlled trial' and 'renal transplantation', limited to trials in humans and reported in English for the years 1995 to 2004, revealed a total of 609 publications; of these 326 were trials of IS drugs on graft outcomes while 281 trials were on complications post-RTx. Trials reporting PKA studies and non-randomised studies were excluded. Though not intended as a comprehensive review of all clinical trials published in this period, nor to evaluate their adequacy or quality, this review examines the spectrum of trials and their impact on clinical RTx.

CLINICAL TRIALS ON IS DRUGS

The primary IS drugs studied in these trials have been categorised as shown in Figure 32.1. While 13% of the trials were small, including less than 40 patients, only 33.8% of these trials included at least 200 patients, i.e. with adequate power to detect a change in prevalence of an endpoint from 50% to approximately 30%.

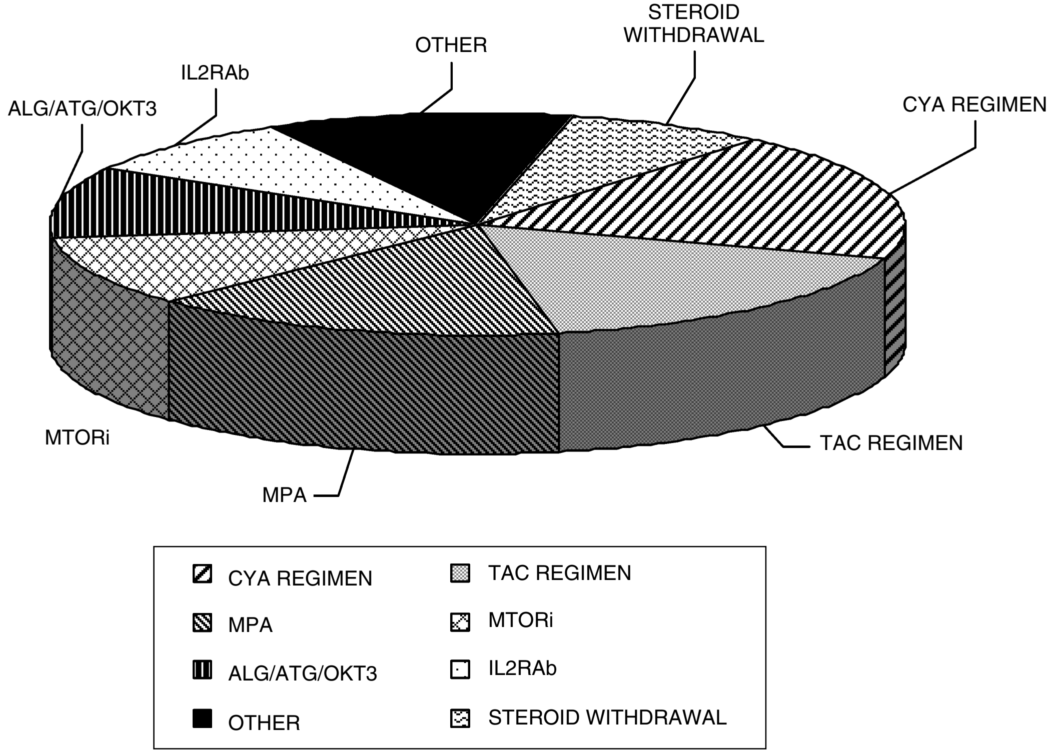


Figure 32.1. Clinical trials on immunosuppressive drugs/regimens in renal transplantation, 1995 to 2004.

The follow-up period for patients, even in the larger trials, was short: 66.7% were for 1 year or less, 29.7% for 1 to 5 years, and only one trial had a follow-up duration of 15 years. Though longer availability of an IS drug should allow for a longer duration of follow-up of outcomes, it appears that once an IS drug has been introduced into clinical use, there is little impetus to publish results on long-term outcomes. For example, despite the availability of the MPA analogue MMF or the IL2RAb since 1995 and 1999 respectively, there is no uncensored data on the long-term outcomes following their use.

Of special concern was the duplication of trial reports, albeit with altered titles in different journals. Of the 111 large (>200 subjects) clinical trials on IS drugs published between 1995 and 2004 that were reviewed for this chapter, only 52 or 46.8% were original; the remaining 53.2% were either duplicate studies or were follow-up studies or *post hoc* analyses at various time points after study initiation.

Figure 32.2 shows the trend for the publications on clinical trials in RTx (as a percentage of trials reported in that year) employing the two CNI, IL2RAb, MPA analogues and MTORi over these years. Despite the lack of trials of high quality and large numbers, the increasing numbers of clinical trials published on MPA (1999), TAC (2000), IL2RAb (2001) and MTORi (2004) have been followed closely by the increasing use of these drugs in clinical RTx over these years in the United States (Table 32.3).[19,20] Clearly, publications on IS drugs from clinical trials have had a significant impact on the clinical use of these IS drugs.

CLINICAL TRIALS ON COMPLICATIONS OF RTx

As suggested earlier, both complications of transplantation per se and complications of the IS drugs used to prevent rejection contribute significantly to post-transplant outcomes (Table 32.2). In contrast to IS drug trials, of which a significant proportion were large trials, only 14.2% of these trials on complications were of more than 200 patients; in fact, 39.9% were with less than 40 patients. It would appear that pharmaceutical companies which sponsor IS drug trials have little interest in initiating trials examining complications post-RTx.

The spectrum of complications covered by the 281 trials published from 1995 to 2004 is shown in Figure 32.3. Cardiovascular disease is the leading cause of death in RTx, contributing to 47% of deaths in some registry series; 19.9% of the published trials addressed issues on dyslipidemia, hypertension or other established risk factors for cardiovascular disease.[21] In contrast, although malignancy and infections are the second and third leading causes of death in RTx, contributing to 15% of deaths each in some series, only 10.7% and 1.8% of the trials in RTx on complications, respectively, addressed these two issues specifically. Even among these studies on complications, only 31% studied the role of an IS drug on these complications, while the remainder studied the role of ancillary therapies in their amelioration.

CONCLUSIONS

Clinical trials are initiated to address an unmet need in a clinical arena. In the field of RTx, there are many unmet needs; however, it is not clear if the clinical trials published in the last decade have addressed their purpose. There remain clearly important issues that are not resolved in the design of trials in RTx: for example, the delineation of appropriate trial endpoints, the optimal duration of follow-up, the appropriateness of censoring of deaths in reporting graft survival and the failure to address important complications post-RTx, to name a few. Moreover, there are additional issues of selective reporting and failure to adhere to quality criteria as enumerated earlier. Indeed, in spite of or because of these deficiencies, IS drugs have emerged from the trial setting to the clinical arena rapidly in the field of RTx. As such, the advent of newer IS drugs has not resulted in better outcomes in one of the most important outcomes in RTx, i.e. fewer patient deaths following RTx. Better design in clinical

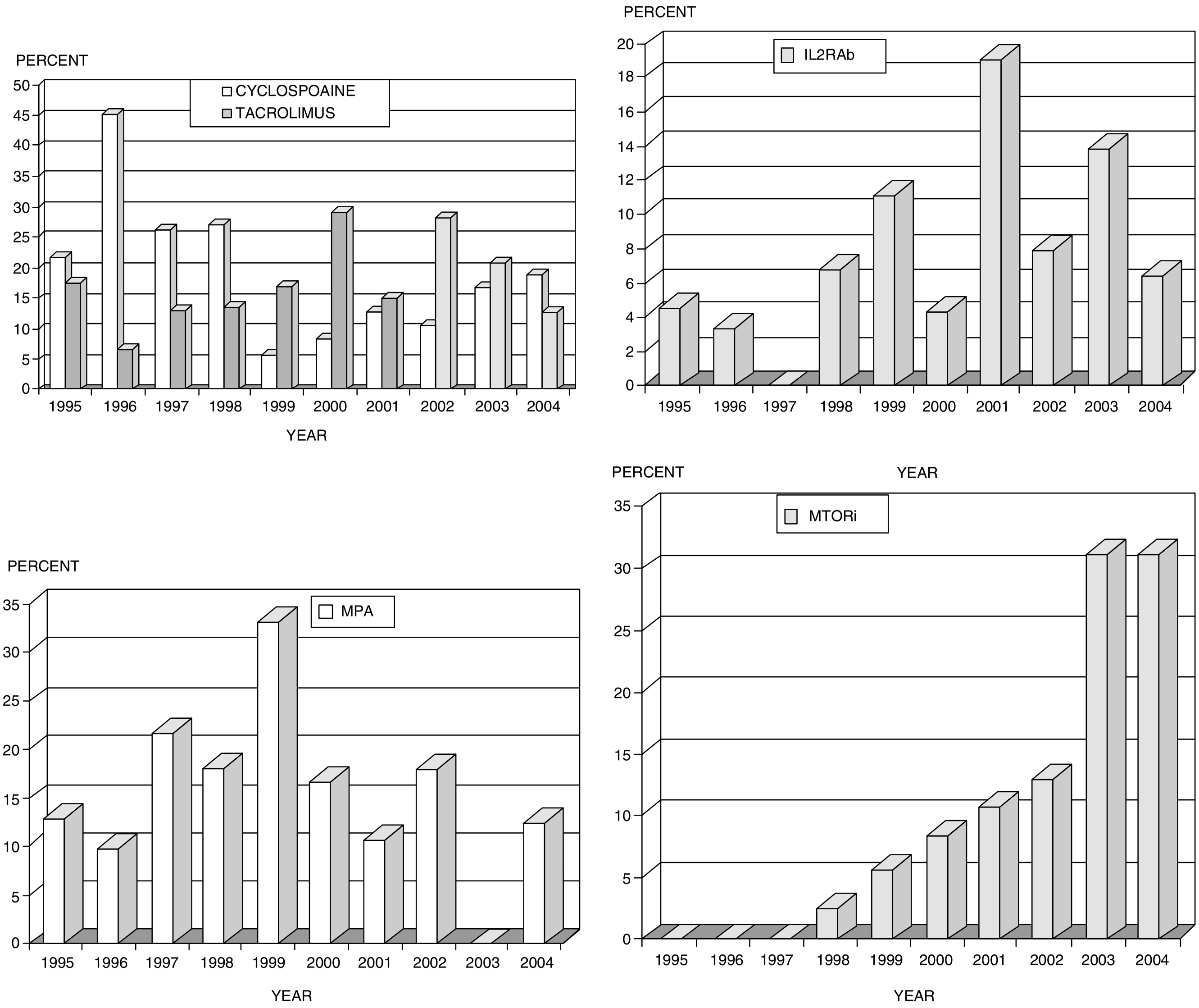


Figure 32.2. Trends of clinical trials in immunosuppressant drugs in renal transplantation, 1995 to 2004.

Table 32.3. Trends in immunosuppressive drug usage in the United States, 1995 to 2003

	1995	1996	1997	1998	1999	2000	2001	2002	2003
IL2RAb use for induction therapy	0.3%	0.8%	0.3%	18%	33.7%	39.7%	39.6%	39.2%	34.2%
Cyclosporine use for maintenance therapy between hospital discharge and 1 year post-RTx	89%	80.8%	74.4%	68.6%	63.3%	53.0%	39.4%	32.3%	–
Tacrolimus use for maintenance therapy between hospital discharge and 1 year post-RTx	14.3%	22.1%	29%	36.9%	42.7%	52.7%	63.6%	69.1%	–
MPA use for maintenance therapy between hospital discharge and 1 year post-RTx	14.6%	43%	72.4%	79.1%	81.7%	79.4%	79.7%	82.3%	–
MTORi use for maintenance therapy between hospital discharge and 1 year post-RTx	0.3%	1.2%	2.3%	3.4%	5.2%	16.7%	21.6%	21.2%	–

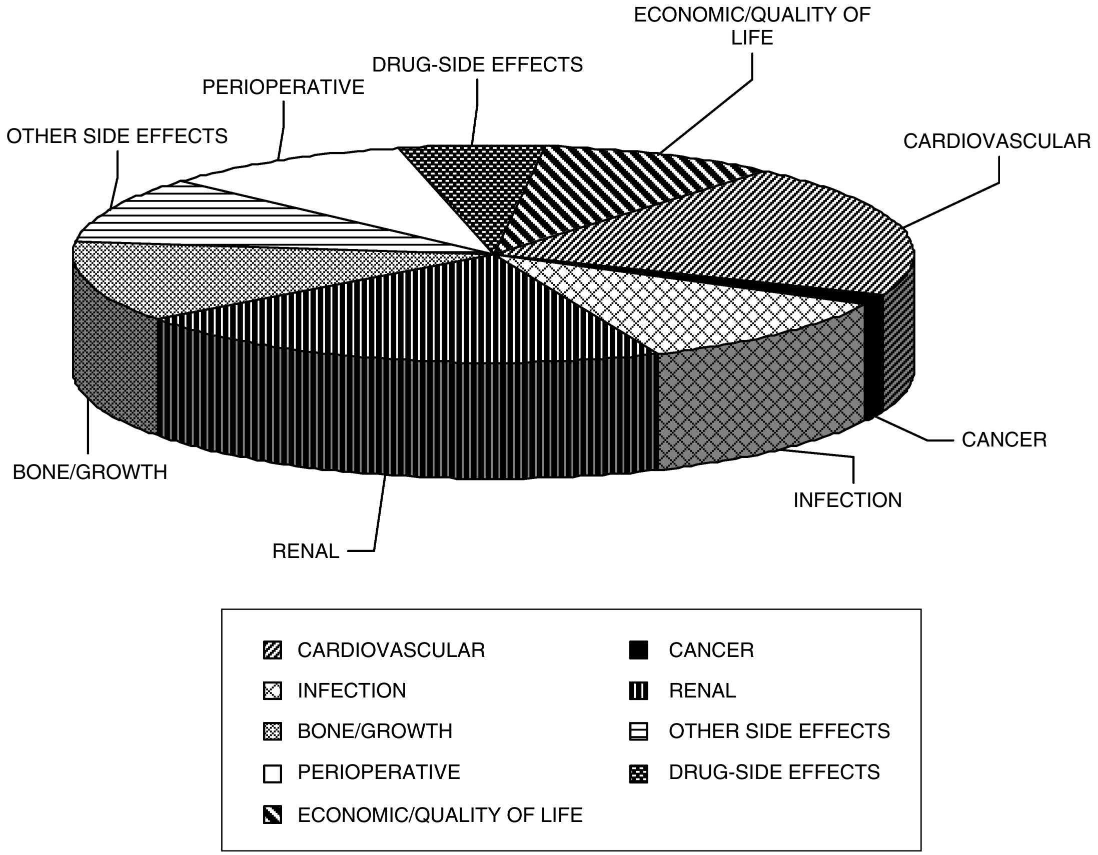


Figure 32.3. Clinical trials on complications in renal transplantation, 1995 to 2004.

trials and adherence to rigorous quality standards in trial reporting in the coming years will hopefully allow the benefits of new IS drugs to be translated into clinical practice.

REFERENCES

1. Turnill A, Baker T. Conducting clinical trials with biological medicines – lessons learnt from the UK TGN1412 trial. Available online at http://www.claytonuz.com/areas_of_law/controller.asp?aolstring=&na=1256
2. Hunsicker LG, Bennett LE. Design of trials of methods to reduce late renal allograft loss: the price of success. *Kidney Int Suppl* (1995) **52**: S120–3.
3. Guttmann RD, Soulillou JP, Moore LW, First MR, Gaber AO, Pouletty P, Schroeder TJ. Proposed consensus for definitions and endpoints for clinical trials of acute kidney transplant rejection. *Am J Kidney Dis* (1998) **31** (6 Suppl 1): S40–6.
4. Racusen LC, Solez K, Colvin RB, Bonsib SM *et al*. The Bannff 97 working classification of renal allograft pathology. *Kidney Int* (1999) **55**(2): 713–23.
5. Cecka JM. The OPTN/UNOS Renal Transplant Registry 2003. *Clin Transpl* (2003): 1–12.
6. OPTN/SRTR 2004 Annual Report. Table 5.6d. Immunosuppression use: antirejection treatment from transplant to one year post transplant, 1993 to 2002. Recipients with kidney transplants. Available online at http://www.ustransplant.org.
7. OPTN/SRTR 2004 Annual Report. Table 1.11b. One year unadjusted graft survival by organ and year of transplant, 1993 to 2002. Available online at http://www.ustransplant.org.
8. OPTN/SRTR 2004 Annual Report. Table 1.12b. One year unadjusted patient survival by organ and year of transplant, 1993 to 2002. Available online at http://www.ustransplant.org.
9. Hariharan S, Kasiske B, Matas A, Cohen A, Harmon W, Rabb H. Surrogate markers for long-term renal allograft survival. *Am J Transplant* (2004) **4**(7): 1179–83.
10. Lachenbruch PA, Rosenberg AS, Bonvini E, Cavaille-Coll MW, Colvin RB. Biomarkers and surrogate endpoints in renal transplantation: present status and considerations for clinical trial design. *Am J Transplant* (2004) **4**(4): 451–7.
11. Placebo-controlled study of mycophenolate mofetil combined with cyclosporin and corticosteroids for prevention of acute rejection. European Mycophenolate Mofetil Cooperative Study Group. *Lancet* (1995) **27**: 345(8961): 1321–5.
12. Curtis JJ, Kaplan B. Transplant immunosuppressive drug trials on trial. *Am J Transplant* (2004) **4**(5): 671–2.
13. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess* (1999) **3**(12): i–iv, 1–98.
14. Fritsche L, Einecke G, Fleiner F, Dragun D, Neumayer HH, Budde K. Reports of large immunosuppression trials in kidney transplantation: room for improvement. *Am J Transplant* (2004) **4**(5): 738–43.
15. Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, Moher D; CONSORT Group. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* (2004) **141**(10): 781–8.
16. Vincenti F, Jensik SC, Filo RS, Miller J, Pirsch J. A long-term comparison of tacrolimus (FK506) and cyclosporine in kidney transplantation: evidence for improved allograft survival at five years. *Transplantation* (2002) **73**(5): 775–82.
17. Jurewicz WA. Tacrolimus versus cyclosporin immunosuppression: long-term outcome in renal transplantation. *Nephrol Dial Transplant* (2003) **18** (Suppl 1): i7–11.
18. Couser W, Drueke T, Halloran P, Kasiske, B, Klahr S, Morris P. A uniform clinical trial registration policy for journals of kidney diseases, dialysis, and transplantation. *Transplantation* (2005) **79**(7): 751.
19. OPTN/SRTR 2004 Annual Report. Table 5.6a. Immunosuppression use for induction, 1994 to 2003. Recipients with kidney transplants. Available online at http://www.ustransplant.org.
20. OPTN/SRTR 2004 Annual Report. Table 5.6c. Immunosuppression use: maintenance between discharge and one year posttransplant, 1993 to 2002. Recipients with kidney transplants. Available online at http://www.ustransplant.org.
21. Briggs JD. Causes of death after renal transplantation. *Nephrol Dial Transplant* (2001) **16**(8): 1545–9.

OTHER PROCEDURES

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

33

Wound Healing

COLIN SONG[1] AND EDWIN CHAN[2]

[1]Department of Plastic, Reconstructive & Aesthetic Surgery, Singapore General Hospital, Singapore

[2]Clinical Trials & Epidemiology Research Unit, SingHealth Health Services, Singapore

INTRODUCTION

Wounds have afflicted people as long as recorded history. It is estimated that chronic wounds currently afflict 0.18% to 1.3% of the US adult population,[1] yet the intransigence of the chronic wound remains an enduring mystery and its successful management the holy grail of wound care. Despite significant advances in the understanding of the basic biology of wound repair, most chronic wounds, especially those on the lower extremity, still fail to heal in a reasonable time. The scientific merits of many wound care products have not been rigorously tested, and in fact many preliminary studies hinted at an efficacy that was later not substantiated in larger trials.[2] Among the reasons for this inconsistency are:

- Poorly designed studies vulnerable to bias.
- Inadequately powered clinical trials such that clinically significant effects could not be shown to be statistically significant.
- Ineffective therapy, e.g. growth factors do not augment healing.
- Ineffective delivery of the novel therapy to the tissues involved in wound repair.
- Poor choice of outcome measures.

ACUTE WOUND HEALING

Wound healing is a complex cascade of events involving cellular and extracellular components resulting in the repair of the injured tissue. The three phases of wound healing are (Figure 33.1):

1. Exudative
   - Inflammation
2. Proliferative
   - Angiogenesis
3. Reparative
   - Epithelization
   - Scar formation.

Following tissue injury, the early inflammatory response is initiated primarily by platelet-derived growth factors and cytokines. These cytokines also orchestrate angiogenesis, endothelial cell migration, deposition of extracellular matrix and macrophage migration to the wound.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

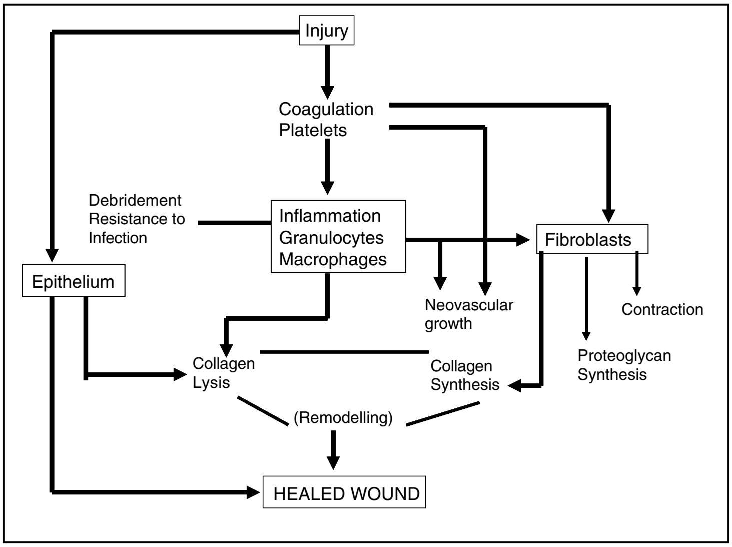


Figure 33.1

The inflammatory phase culminates in the appearance of neutrophils and macrophages which travel along 'guide-rails' of fibrin tracts laid down during the formation of the early fibrin clot. The accumulation of the white cells and activation of macrophages result in the neutralisation and removal of bacteria and tissue debris. This leads to the proliferative phase whereby neovascularisation and the appearance of fibroblasts initiate wound repair by 'filling up' with granulation tissue – a mix of cells, extracellular matrix and collagen. The process of epithelisation occurs under the influence of epidermal growth factor. Epithelial cells lose their intercellular connectivity to adjoining cells, undergo rapid mitosis, and finally close over the wound. However, wound healing is only complete once enough collagen cross-linkages are in place to achieve sufficient functional tensile strength.

CHRONIC WOUNDS

Many factors can disrupt the acute healing process, thus giving rise to the spectrum of problems that manifests itself as a chronic wound – clinically defined as one which has failed to close over or epithelise in a month from time of injury. Chronic wounds are characterized by:

- Deficient acute inflammatory response.
- Inappropriate cellular/cytokine activity – persistence of chronic inflammatory cells, i.e. lymphocytes, macrophages, mast cells, eosinophils.
- Growth factor abundance (as shown by Jude *et al.*[3] for chronic venous leg ulcers) despite which there is failure to heal.
- Increased proteases which break down protein in wounds.[4]
- Lack of vascularity.
- Lack of extracellular matrix and collagen deposition.
- Relatively acellular environment.

The following clinical categories of chronic wounds are frequently encountered:

- Decubitus ulcers
- Diabetic ulcers
- Venous ulcers
- Arterial ulcers.

DECUBITUS ULCERS

The decubitus ulcer and its healing is a complex problem. It is most commonly associated with denervation in the paraplegic patient. In such patients, anatomical regions of the body where soft tissue covers bony prominences are easily subjected to undue pressure. Patients who are incontinent of urine, when dragged across a wet linen surface, suffer shear forces between the cutaneous surface and the linen resulting in the separation of the cutaneous dermis from the deep underlying vascular perforators. This is followed by the development of deep-seated haematomas, which ultimately convert to deep-seated abscesses. The decubitus ulcer is in essence a sinus discharging the contents of the deep-seated abscess. In addition, healing in a denervated wound is delayed because high levels of collagenase enzymes are in action.

DIABETIC ULCERS OF THE LOWER EXTREMITIES

Diabetic patients are at risk of developing chronic wounds for a multitude of reasons. The wounds tend to commonly develop in the leg and feet and usually from trivial trauma.

The problem of chronic diabetic ulcers is fast approaching epidemic proportions in many societies. Multiple factors are involved in the genesis and persistence of chronic diabetic ulcers:

- Poor peripheral vascularity and cardiovascular compromise
- Neuropathy from abnormal neural metabolism and nerve entrapment disorders
- Susceptibility to infections
- Hyperglycaemia and oedema of lower extremities.

VENOUS ULCERS

Undue swelling from venous insufficiency and venous hypertension of the lower extremities predisposes patients with this problem of chronic ulceration. The exact mechanism has not been agreed upon although two hypotheses exist.

First, the 'venous cuff' hypothesis[5] of the pathogenesis of venous ulcers postulates that a 'cuff' of proteinacious deposits in the interstitium derived from accumulated lymphatic fluid around venous capillaries prevents gaseous exchange in the interstitium. The resulting tissue hypoxia and build-up of carbon dioxide impairs the healing of venous ulcers. However, current evidence suggests that fibrin cuffs are not pathognomonic features of these ulcers, do not necessarily impair oxygen diffusion and do not prevent the healing of venous ulcers.[6]

Second, currently the 'white cell hypothesis' holds that microvascular obstruction by thrombosis or leucocyte plugging may be responsible for venous ulceration. Adhering leucocytes degranulate and release potent enzymes and reactive oxygen species that damage capillaries and cause increased microvascular permeability and tissue damage.[7]

The ability of chronic wound fluids to degrade peptide growth factors is associated with increased levels of elastase activity and diminished levels of proteinase inhibitor was demonstrated.[4] Jude *et al.*[3] discovered a growth factor irony in the tissue surrounding the ulcer. The analysis of tissue samples revealed that despite an abundance of growth factors, the cellular healing response was impaired as a result of either target cell receptor blockade or the presence of matrix metallo-proteins in the ulcer.

ARTERIAL ULCERS

Arterial ulcers arise from insufficient vascularity to an area of skin resulting in tissue ischaemia, ulceration and chronic wounds.

The management of chronic wounds remains much of a dilemma. However from a clinical research standpoint, the author adopts a very practical and useful approach which lends itself to avenues for research. Figure 33.2 is the algorithmn for this approach to the management of the four major categories of chronic wounds.

ANIMAL MODELS FOR WOUNDS

Animal wound models may be helpful in detecting healing responses to and potential toxicities

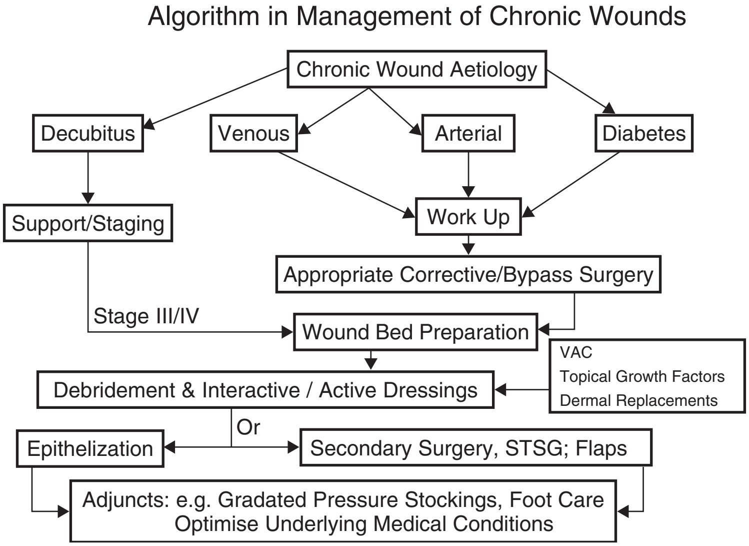

Algorithm in Management of Chronic Wounds
Chronic Wound Aetiology
Decubitus
Venous
Arterial
Diabetes
Support/Staging
Work Up
Appropriate Corrective/Bypass Surgery
Stage III/IV
Wound Bed Preparation
VAC
Topical Growth Factors
Dermal Replacements
Debridement & Interactive / Active Dressings
Or
Epithelization
Secondary Surgery, STSG; Flaps
Adjuncts: e.g. Gradated Pressure Stockings, Foot Care
Optimise Underlying Medical Conditions

Figure 33.2. The management of the four categories of common chronic wounds; adopting the holistic approach and placing the concept of wound bed preparation in perspective. VAC (vacuum-assisted closure), growth factors and dermal replacements are employed primarily as 'tests of vascularity' especially for VAC. This schema for the author serves as a useful methodology for the management of chronic wounds and lends itself to trials in the pursuit of evidence for effective wound care .

of wound care products. The ideal model should exhibit a biological responsiveness to the test agent that mimics that of humans. In practice, they have been poor predictors of efficacy in human clinical trials. Currently there are no ideal animal models for chronic wounds or extensive burns, therefore multiple complementary animal models are used to assess the separate activities of wound healing agents. For example, the process of fibroplasia and stroma formation can be evaluated by subcutaneous injection of some products in rats. Contraction and re-epithelisation can be evaluated by topical application on full-thickness excisional wounds in a pig graft donor site model. Induction of angiogenesis can be evaluated in chick chorioallantoic membrane or rabbit cornea. Breaking strength can be tested in a rat linear incision model. In impaired-healing animal models include infection, necrotising trauma, irradiation, administration of corticosteroids or chemotherapeutic drugs, or drug-induced or genetic diabetes mellitus in mice, rats, hamsters, guinea pigs and young pigs. Induction of wound ischaemia is simplified in the rabbit ear dermal ulcer model since it lacks the vigorous wound contraction seen in rodent models.

CLINICAL TRIALS IN WOUND HEALING

A scientific understanding of the pathophysiology, cellular and molecular basis of wound chronicity is the foundation upon which new therapeutics for wound management should be designed. Ultimately, claims and deductions of efficacy will need to be proven in randomised clinical trials. Long-used traditional, complementary and alternative medicines, ironically, are considered 'new' therapeutics in so far as proof of efficacy and safety has yet to be demonstrated by rigorous scientific standards. The evidence-based medicine paradigm is bringing traditional therapeutics out of the closet into the welcome glare of the scientific spotlight.

CHALLENGES IN CONDUCTING CLINICAL TRIALS ON WOUND HEALING

1. Little information exists on risks of adverse events such as keloids from preclinical animal studies due to basic inadequacies of animal model systems.
2. Large between-subject heterogeneity in healing responses makes it difficult to measure the outcome precisely and probably reflects the effect of differences in unknown prognostic factors amongst subjects. Ideally, comparative studies ought to be conducted with the subject serving as his or her own control.
3. Accurate assessment of the progress of wound healing and closure is difficult. Most clinical studies will conventionally report the rate of surface closure of a wound. This can readily be done with the conventional tracing method or linear dimensional measurements. However, we know that surface closure is not reliably correlated with complete closure. Being a three-dimensional entity, a wound also has depth and assessing this is far from elucidated. Possibly the most accurate way is by volumetric displacement of 'filler' material. More accurate and precise measurements can be achieved using non-invasive imaging methods such as ultrasound.
4. Perhaps *Falanga's*[8] wound bed preparation concept and scoring method for chronic wounds hold promise as a more comprehensive assessment tool in judging the wound closure efficacy of dressings on particular wounds. Wound bed preparation (as distinguished from debridement) can be defined as the global management of the wound to accelerate endogenous healing or to facilitate the effectiveness of other therapeutic measures. Chronic wounds are recognised to have a 'necrotic burden' and a 'cellular burden'. The necrotic burden consists of both necrotic tissue and exudate. The exudate in chronic wounds has been shown to inhibit the proliferation and function of key resident cells and to contain proteases that break down extracellular matrix proteins. A heavy bacterial presence results in a 'biofilm' that isolates the chronic wound from the body's own healing process. The cellular burden refers to those resident cells from the intensely inflammatory response in, for instance, venous ulcers, with resulting heavy exudate production that interferes with healing. The resident cells in chronic wounds, such as fibroblasts and keratinocytes, may become desensitised to certain healing signals such as growth factors.

On this basis Falanga therefore advocated use of the wound bed appearance and amount of wound exudate as more objective assessment of progress. Ultimately, therefore, Falanga used rate of healing as a prognostic factor and claimed it to be 80% sensitive and specific. Any regimen adopted in the treatment of wounds not attaining a rate of healing of 0.08 cm/week would then be considered in need of revision.

Measuring Claims of Effectiveness

The US Food and Drug Administration Center for Drug Evaluation and Research has published a *Guidance for Industry* on trials conducted to assess efficacy in chronic cutaneous ulcer and burn wounds (http://www.fda.gov/cder/guidance/3226dft.htm). These guidelines apply to claims of effectiveness in the two broad categories of: (1) improved wound healing, and (2) improved wound care other than healing.

Claims Related to Improved Wound Healing

Complete wound closure The FDA defines complete closure as skin closure without drainage or dressing requirements. Complete wound closure for chronic, non-healing wounds is considered the most clinically meaningful of the claims related to improve wound healing.

Generally, studies to support such a claim would be designed to measure incidence of complete wound closure in the treatment versus the control groups by a specified time (landmark analysis). Efficacy *success* would be defined as a statistically significantly greater proportion of patients assigned to the 'test' product achieving closure compared with the proportion in the control arm. The prespecified time for endpoint measurement should be based on the natural history of the disease process and the expected response to standard care. The guide adds, 'the linical benefit of wound closure that lasts for a very brief time is at best, highly limited. In general, trials should be designed such that subjects remain on study and continue to be evaluated at least 3 months following complete closure'. The purpose of this follow-up period is to measure the durability of the effect relative to standard care. For some products, the durability of closure is also important for distinguishing wound healing from transient wound coverage. Measurement of partial healing, if prospectively defined, may be supportive of efficacy, but should not be used as primary evidence of clinical efficacy per se. However, as noted below, *partial healing which facilitates subsequent surgical closure* can be an acceptable, clinically relevant outcome to report.

Accelerated wound closure A claim of *accelerated closure* reflects a clinically meaningful decrease in time until complete closure occurs. It may be reported as a median time to complete closure or more commonly as a relative rate of closure using the methods of survival analysis (see the analysis discussion). An accurate and standardised definition of the event needs to be used. Including the time dimension of healing in the efficacy measure provides clinically relevant information and a way of differentiating between products that would otherwise have similar incidences of complete closure within a reasonable time. Another approach is to measure wound size progressively (repeated measures) over time in study subjects and use the methods of longitudinal analysis.

Accelerated healing claims for burns should distinguish between partial thickness burns, full-thickness burns, or donor site wounds. *Accelerated closure* of the donor site produced during harvest of autologous grafts is a claim for which it is especially important to prespecify the clinical benefit expected because these partial thickness wounds heal well in 2–3 weeks with standard care regimens. For example, a product that accelerated healing of donor sites by only one or two days might provide clinical benefit if it could be safely used in extensively burned patients requiring repeated reharvesting of donor sites. If time to reharvest is used as the primary efficacy outcome to support this type of claim, careful attention to masking is important to prevent bias, since reharvest is generally undertaken before the donor site reaches 100% re-epithelisation. Accelerated healing claims based on study of donor sites cannot be generalised to burns and chronic cutaneous ulcers because burns and ulcers do not share the same clinical characteristics of uniform, partial thickness donor sites. However, for systemically administered test products, healing of both the donor sites and the ulcer or burn is an important safety outcome. For example, a product that accelerates the healing of donor sites should not worsen graft take.

Facilitation of surgical closure The FDA does not consider partial healing per se to be an appropriate claim for wound healing agents because the clinical benefit of statistically significant decreases in wound size has not been established. However, agents that heal wounds to the point that surgical closure is more feasible, safer or more effective may lead to the claim of *facilitates surgical closure*. Studies should be designed to measure the incidence of complete wound closure following application of the surgical graft. The durability of surgical wound closure should be assessed over time to ensure adequate quality. Timely excision and grafting have greatly reduced morbidity and mortality in patients with full thickness burns. The clinical benefits of engraftment in burn injury include reduced wound sepsis rates, improved haemodynamic status, and decreased requirement for donor site harvest. Since engraftment rates are high with good standard care, studies of surgical closure of burn wounds may take large numbers of patients to detect a difference between the test product and standard care. It is important to evaluate healing outcomes such as durability, functionality and cosmetic appearance, including scarring.

Improved quality of healing Trials for *improved cosmesis* claims should demonstrate a significant effect on outcomes such as scarring, the contour and feel of the healed skin, or normalisation of skin markings or pigmentation. The appropriateness of an improved cosmesis claim depends on the type and location of the wound. For example, normalisation of skin markings or pigmentation would clearly benefit patients who require grafting of full thickness burns on the face, whereas this outcome would be a less convincing measure of benefit for patients with plantar ulcers. In choosing endpoints to support improved cosmesis claims, it is important to consider whether a reliable assessment tool exists, or can be developed.

Claims of efficacy for scar management are assessed based on, for instance, the *Vancouver Scar Scale* and the like. Products that reduce scarring may also improve function, e.g. range of motion. Standardisation across treatment arms of the use of concomitant therapies, such as pressure garments and rehabilitative therapies (e.g. passive range-of-motion exercises), is important for unconfounded assessment of this outcome.

Claims Related to Improved Wound Care (other than healing)

Products intended for wound management may provide additional important patient benefit apart from affecting the incidence or duration of closure. However, it is then important to also demonstrate that such products do not actually significantly impede healing. Thus, wound healing should be evaluated as a safety outcome for all products with a wound care claim.

Wound infection control Infected wounds do not heal, and the primary efficacy outcome for topical anti-infective wound products can be either *healing* or *control of infection*. Both outcomes should be assessed, and reasonable concordance would be expected. Products for treatment or prophylaxis of infection in serious wounds (e.g. burns, diabetic foot ulcers) should have a well-established and appropriate spectrum of activity.

Debridement It is generally accepted that necrotic tissue inhibits healing by interfering with tissue repair and promoting microbial growth. Thorough debridement of wounds is therefore considered standard care essential to healing. *Partial debridement* is not an acceptable endpoint because the clinical benefit of partial debridement is unclear, and methods for measuring extent of debridement have not been validated. Although there is debate about the optimal design of trials to assess the efficacy of debriding agents, a reasonable endpoint for a debridement claim might be *thorough* removal of necrotic tissue (e.g. produces a wound bed suitable for grafting). Other clinically relevant endpoints, such as pain or blood loss during or immediately following debridement, could provide supportive evidence for clinical benefit when the primary efficacy endpoint is debridement equivalent to that produced by standard mechanical/surgical procedures. For burn wounds, *timeliness* of thorough debridement is an especially important consideration.

Wound pain control Studies of topical products that reduce wound site pain should distinguish between chronic wound pain and acute pain associated with wound care procedures. Appropriate instruments to measure pain should be prospectively defined and properly validated.

Other wound care claims Serious wounds may negatively affect many aspects of patients' lives. Clinically significant improvement in certain activities of daily living not already captured by any of the previously described outcome measures (e.g. decreased drainage when experienced by the patient as an important improvement in ability to function) might support a labelling claim if demonstrated with a validated instrument.

SPECIFIC WOUND CARE CLINICAL TRIAL CONSIDERATIONS

This section discusses specific points of study aim, subjects, intervention, outcome and analysis pertinent to wound trials. It is not intended as exhaustive guidance on wound trial design.

STUDY AIMS

Absorption Studies

For topical drugs, biological and combination products, Phase I evaluations should include quantitation of absorption through the wound. Systemic bioavailability of topically applied products is generally assessed using standard pharmacokinetic measurements with serial serum sampling. Systemic uptake is influenced by wound factors such as size and vascularity, as well as product characteristics such as molecular weight, chemical composition and the presence of excipients. In the case of growth factors, even though relatively little (<1%) absorption typically occurs from chronic ulcer sites, these amounts might still be clinically significant because some growth factors are active *in vitro* at nanogram concentrations. For this reason, it is important to perform sensitive assays against serum background. For products that are absorbed from the wound bed, the systemic dose delivered depends on several factors: the concentration of the active ingredient, the total body surface area treated, the volume applied, frequency of application and duration of contact with the wound. Safety and pharmacokinetic studies for topical wound products should usually be conducted in patients with the indication sought, since absorption through the intact skin of a healthy volunteer would not predict absorption through a wounded surface.

Irritancy or Sensitisation

When preclinical studies or clinical experience suggest that a topical product might induce clinically significant dermatitis, testing for irritancy or sensitivity in healthy volunteers is recommended prior to trials in patients, since superimposed dermatitis is deleterious to wounds. The need for routine testing of the final formulation depends on the product, and sponsors are encouraged to discuss dermal toxicity testing with the clinical team before initiating the studies.

Toxicity Studies

The design of non-clinical toxicology studies for wound products should reflect, as much as possible, the intended clinical use of the product with respect to route of administration, dosing regimen and duration of exposure. It is important to assess any exaggerated pharmacological responses and potential toxicities of wound products. Administration of the wound product at multiples higher than the intended therapeutic dose (determined from wound models) can help provide an estimate of the therapeutic index (toxic dose/effective dose) to aid in the selection of the initial clinical starting dose. Vehicle controls should be employed where appropriate; to evaluate any adverse affects of product formulation components.

STUDY POPULATION

A clearly articulated study objective will guide the choice of the study patient population to sample for inclusion in clinical trials. Typically the sample will be a serial consecutive set of subjects meeting the inclusion criteria who have given informed consent. Investigators need to consider the generalisability (external validity) of the results derived from such a sample.

Chronic Cutaneous Ulcers

The three major categories of chronic cutaneous ulcers are diabetic ulcers venous stasis ulcers and pressure ulcers. In general, separate trials should be conducted for each type of chronic ulcer because they have very different aetiologies and potentially different responses to therapy.

Variability can be reduced by specifying enrolment criteria that exclude conditions known to impede healing. For example, specifying a range for ulcer size will avoid ulcers that would be expected to close rapidly with little intervention (e.g. <1 cm^2), and ulcers that would be less likely to close during a trial (e.g. >50 cm^2). However, if demonstration of efficacy is limited to ulcers of a specific size, and there is a requirement to extrapolate to smaller or larger ulcers, the labelled indication should be similarly qualified.

Burns

The population for burn trials is usually defined by the extent and depth of the burn injury. For most burn wound studies, it is important to determine the depth of target wounds, since this determines the standard of care and the expected time to healing. Important characteristics of the burn wound are its cause (thermal, chemical, electrical), anatomic location, depth (full or partial thickness), duration (left untreated), and extent (% total body surface area). Patient characteristics that affect burn wound healing include age, nutritional status, underlying medical conditions and the presence of concomitant injury (e.g. head trauma, inhalation injury, bone fractures). Patients with serious burns commonly receive multiple concomitant treatments, making it sometimes difficult to detect a treatment effect. For this reason, it is advisable to enrol patients with the least serious burns that still permit assessment of the product's claimed benefit. However, it may also be important to assess the effects of the study treatment used in conjunction with commonly used concomitant therapies. When patients with full-thickness burns are studied, donor sites for autografts are sometimes selected as the target wound to be assessed. As noted earlier, although the patient population is one and the same, demonstrating the safety and efficacy of a product for a donor site wound does not support the safety and efficacy of the product for burn wounds, because burn wounds differ in clinically significant ways from the surgical wounds inflicted at donor sites.

STUDY INTERVENTION – STANDARD CARE

Standard care in the context of this discussion refers to supportive wound care in a clinical trial other than the experimental product. Good standard care procedures in a wound trial are a prerequisite for assessing safety and efficacy of a product. Since varying standard care procedures can confound the outcome of a clinical trial, it is critical that all participating centres agree to use the same procedures. If standard care procedures are not uniform, it is important that the variations be noted and the sample size and collected data be adequate to assess the impact of supportive wound care variations on treatment response. Several professional groups have initiated development of care guidelines for ulcers and burns. Although the FDA does not require adherence to any specific guidance, the basic guiding principle is that standard care regimens in wound trials should optimise conditions for healing and be prospectively defined in the protocol. The rationale for the standard care chosen should be included in the protocol, and the study plan should be of sufficient detail for consistent application in all study centres. It is important to specify in the case report form (CRF), at each visit, the type of ulcer or burn care actually delivered (e.g. extent of debridement, use of concomitant medications). For outpatients, the CRF should also capture compliance with standard care measures, such as wound dressing, off-loading (pressure relief) and dietary intake. In some cases it may be important to assess the effect of experimental treatment across common variations of standard care procedures.

Standard Care for Chronic Cutaneous Ulcers

Basic considerations in choosing standard care procedures for chronic cutaneous ulcer trials include the following:

- Removal of necrotic or infected tissue
- Off-loading of pressure and diabetic foot ulcers
- Compression therapy for venous stasis ulcers
- Establishment of adequate circulation for arterial ulcers
- Maintenance of a moist wound environment
- Infection control
- Nutritional support, including blood glucose control for diabetic ulcer patients
- Bowel and bladder care for patients with pressure ulcers at risk for contamination.

Debridement The presence of necrotic tissue, sinus tracts, exudation or transudation, and infection of soft and hard tissues can interfere with ulcer healing. Appropriate debridement procedures for the indicated ulcer should be specifically defined in the protocol. To avoid bias and confounding of treatment effect, ulcer debridement should precede evaluation of ulcer extent and infection. Enzymatic debriding agents, like other concomitant topical products, can confound results in wound product trials and generally should be avoided. The need for additional debridement, performed after study treatment has started, may indicate product-induced wound deterioration. As such it should be documented and included in the analysis of product safety and efficacy. Discontinuation might be indicated in early trials where little is known about product safety, but not in later trials, where standard debridement procedures may be indicated to optimise patient care (e.g. ongoing removal of callus as part of standard care for diabetic ulcers).

Off-loading/compression Relief of pressure is critical for chronic ulcers. Pressure is the principal cause of decubitus ulcers and off-loading is often difficult to standardise because equipment (e.g. type of bed) may not be available at all sites, and compliance with study procedures is labour intensive (e.g. turning). For diabetic foot ulcers, off-loading options (e.g. casting) must be weighed against the need to apply study treatments and monitor outcome. Similar considerations are important in choosing compression methods for venous stasis ulcers. Every attempt should be made to define a regimen that can be uniformly applied across sites and to document deviations.

Maintenance of a moist wound environment Maintenance of a moist wound environment is generally accepted standard care for all chronic cutaneous ulcers. In choosing test dosing regimens, it is helpful to consider limitations imposed by various standard care dressings. In cases where there is a sound rationale for the expected benefit of a test product, but its use is not compatible with established standard care dressings, alterations in standard care can usually be safely implemented by including adequate discontinuation rules.

Infection control Absence of frank infection is critical. For this reason, wound products whose action is not anti-infective are usually tested in patients with uninfected target ulcers (noting the distinction between colonisation and frank infection of an ulcer). Acceptable ulcers for enrolment can often be achieved during a run-in period with thorough debridement and other good standard care procedures. A high incidence of true infection (as opposed to colonisation) is present at baseline for diabetic foot ulcers. It may not always be necessary to exclude infected diabetic foot ulcers if the infection does not involve underlying structures and is responding to standard systemic antimicrobials. In such cases, it is especially important that the protocol clearly delineate adequate rules for patient discontinuation due to wound deterioration on-study. As for all discontinued patients, safety assessment should continue throughout the trial and these patients should be included in an intention-to-treat analysis.

If an ulcer becomes infected during a study for a topical wound product, and the investigator prescribes topical antimicrobial treatment, it is recommended that the patient be discontinued from study treatment. Use of concomitant topical medication is discouraged in trials for topical products. Systemic antimicrobial therapy for target wound infection may become necessary during the treatment period of the study. Whether or not study treatment should be discontinued in this situation should be planned prospectively. For example, discontinuation might be indicated in early trials, when little is known about product safety and where infection may signal test product-induced deterioration of the wound, but possibly not in later trials where such therapy would be considered standard care (e.g. systemic antimicrobial therapy for diabetic ulcers).

Wound cleansing Agents used for wound cleansing should be prespecified and bland (e.g.

normal saline) because some cleansers retard healing, or can cause irritation and sensitisation.

Nutritional support Caloric intake and metabolic status should be documented if the product is known to have metabolic effects (e.g. anabolic steroids). For products not known to have metabolic effects, these data are still useful if the inclusion criteria include patients significantly above or below ideal body weight (e.g. cachectic patients with pressure ulcers). Maintenance of normoglycaemia is an important factor for patients with diabetic ulcers.

Standard Care for Burns

Standard care for serious burns includes careful attention to the following areas:

- Haemodynamic resuscitation
- Management of co-morbidities
- Timely burn debridement and/or excision
- Wound closure
- Infection control
- Pain control
- Nutritional support
- Rehabilitation, including passive range of motion when burns overlie joints.

Because large burn centres tend to have well-established but unique standard care regimens, analysis of data in multicentre burn trials may require stratification by centre. Since standard care procedures have profound effects on clinical outcome, every effort should be made to reach agreement among site investigators and to document actual care delivered.

OUTCOME ASSESSMENT/QUANTIFICATION

The tools to assess endpoints for a clinical trial should be both prespecified and standardised across clinical sites. For example, if photographs are to be used for measurement and documentation, the lighting and type of camera should be specified. Scoring systems for wounds can be used at baseline to determine eligibility for study, as well as for periodic wound assessment during the study. The use of validated assessment systems is recommended (e.g. Wagner, International Association of Enterostomal Therapists). Proposals for novel assessment systems should include validation data. Regardless of the methodology, the following variables should be noted in all clinical trials of wound care products.

Ulcer Classification

The type of chronic ulcer (venous stasis, diabetic, pressure, arterial insufficiency) can usually be determined from the patient's history and a physical examination. Objective Confirmatory diagnosis can include Doppler sonography to quantify venous or arterial insufficiency, transcutaneous oxygen tension (t_cpO_2) measurements, ankle/brachial index, filament testing to quantify sensory neuropathy, measurement of laboratory markers for diabetes mellitus, and histopathology of ulcer biopsies to detect neoplastic, immune-mediated, or primary infectious disease causes.

Wound Size

Quantitative measurements of wound size are routinely used to assess initial wound size before and after debridement, as well as progress towards closure. For ulcers that tend to be superficial, such as venous stasis ulcers, the area of the wound opening should be measured. This can be accomplished by tracing the wound perimeter or by measuring maximal width and length. For ulcers that extend deeply into tissue, volume or surface area should be measured when feasible. The extent of tissue undermining and sinus tracts is an important part of the evaluation. In the case of diabetic ulcers, qualitative assessment by probing the maximal depth is a frequently used method. For other ulcers, such as pressure ulcers, moulds can be used to provide precise measurement of volume and/or surface area. Alternatively, semiquantitative measurements can be achieved using the maximal width/length/depth and shape coefficient. For acute burns, it is important to determine as well as possible the depth of target burn wounds for the study, as this parameter affects both the choice of standard of care regimen and the expected time to healing. The distinction between partial, full-thickness and indeterminate wounds is currently based on clinical judgement. Clinical parameters include appearance of the tissue, sensation, and bleeding upon debridement. Validated test methods for determining burn depth do not exist currently, but biopsy and Doppler measurement of blood flow are sometimes used. Wound depth heterogeneity is often an impediment to quantitative measures, and burn depth extension in the first 24 to 48 hours following injury frequently necessitates reassessment of wound severity and treatment. Initial clinical assessment of full-thickness wounds should be confirmed by comparison with the total body surface area ultimately grafted. When the target wound is an autograft donor site, the protocol should clearly delineate the method for harvest, and the size, thickness and anatomic location of the donor site.

Wound Imaging

Standardised photographic and wound imaging procedures should be used to document the wound appearance at each clinic visit and to corroborate the measurements captured in the case report form.

Infection

Infection should be assessed clinically by symptoms and signs that include purulent drainage, erythema, warmth, exudation, odour, pain, fever and leucocytosis. Fever, pain and leucocytosis may be absent, however, especially in patients with diabetic foot ulcers. Quantitative and qualitative culture of a viable tissue biopsy can be used at baseline to help determine if the wound is infected or merely colonised and to guide appropriate antimicrobial therapy. This method is generally preferred to quantitative and/or qualitative culture of swab specimens.[9]

Safety and Immune Reactions

It should be realised that wound care products may actually impede healing.

Deterioration of target wounds can manifest as erythema, pain, discharge, infection, tissue necrosis, requirement for repeat debridement or other surgical intervention (i.e. amputation), and/or increase in ulcer size. Undesirable alterations of soft tissues, ligaments, periosteum, or joint capsules underlying deep wounds should also be evaluated. For biological products and some drugs, immunogenicity is generally addressed by measuring antibody titres prior to and after the treatment. Further immunologic characterisation may be recommended, since the development of an immune response can render the product inactive (neutralising antibodies), and/or induce acute or chronic immune reactions (e.g. anaphylaxis, contact sensitisation, autoimmune disease).

STUDY DESIGN CONSIDERATIONS

Randomisation and Stratification

Randomisation is particularly important to reduce bias in trials for wound indications because standard care wound management procedures and baseline wound characteristics have a profound effect on outcome. Because some degree of variation in these factors across patients and sites is unavoidable, stratification by study centre is recommended to ensure balance between the arms and the option to report centre-specific results. In some cases, it may be appropriate to prospectively stratify randomisation by other important covariates, such as wound size or duration, but the total number of variables used for stratification will be limited in practice by the need to ensure adequate sample sizes in each strata.

Comparator Arms

A vehicle control arm is recommended for most wound product studies, with identical standard care procedures included in both the vehicle and investigational product arms. To evaluate the safety and effect of the vehicle, a study arm treated with standard care alone is recommended in Phase II for topical wound products, if the safety of the vehicle has not been previously demonstrated. Within-patient control designs have been used in trials of topical products intended for serious burns, in an attempt to minimise the heterogeneity amongst patients. However, this approach compromises the evaluation of systemic toxicity since subjects are receiving multiple treatments, thus necessitating additional controls or studies to collect adequate safety data.

Masking

In general, masking (blinding) of patients and investigators to the treatment received will reduce bias and should be employed when feasible. Early studies of topical wound products often require an arm that receives only standard care, in addition to an arm receiving vehicle, to establish whether the vehicle has an effect on healing. Often the standard care-only arm cannot be masked. In other cases, especially for some devices, it is impractical or unethical to implement a dummy treatment that mimics the test product and allows masking. In all cases, blinded assessment by an independent third-party evaluator should be implemented.

Trial Stopping Rules

Because the patient populations in burn and chronic ulcer trials often have a high background incidence of serious adverse events, it is recommended that a data safety monitoring group be appointed for blinded trials when the known or suspected risk is significant, and/or the study population is critically ill (e.g. seriously burned patients). Subjects who are discontinued from study treatment should remain in the study for safety assessment and efficacy analysis.

STATISTICAL CONSIDERATIONS SPECIFIC FOR WOUND PRODUCT TRIALS

This section addresses issues that present special statistical considerations for wound product trials.[10]

Analysis

The analysis plan should be prespecified in the protocol, and all point estimates should be accompanied by an estimate of precision, typically a 95% confidence interval. When full closure (yes or no) within a defined time period is the primary outcome variable, the simplest group summary is the *incidence (risk) of closure*, and the between-group comparative (treatment effect) statistic is the relative incidence (risk) of full closure within the defined period. Reporting results as risks (and relative risks) has the advantage of ease of interpretability compared with reporting odds and rates. When analytical adjustment is desired for potential confounders then multiple logistic regression modelling is a convenient approach; however, the magnitude of treatment effect will now be expressed in terms of the odds ratio, and prognostic predictions of risks for groups defined by particular values of model covariates will be given as odds. Regardless, the direct calculation of incidence risk or odds assumes that the follow-up is 'closed' – all study subjects either experience the event (full wound closure) or are followed up for the same time. When this is not the case, i.e. the follow-up time of subjects who do not experience the event are unequal because of 'censoring' by competing events and losses-to-follow-up, it is then necessary to measure also the time to event (*time to closure*) or censoring for each subject. A simple and direct estimate of wound closure rates (person-time) in each group and relative rates may then be made. If it is not appropriate to assume a constant rate, then life table and Kaplan–Meier methods may be used, and if covariate adjustments are further required, the Cox proportional hazards model is a readily available option. Besides meeting the challenge of unequal follow-up times, the advantage of these 'survival analysis' approaches is the estimation of period-specific wound closure risks for treatment groups as well as comparative rates (hazard rate ratio) between groups. For factors that have been randomised within strata, it may be possible to do separate stratified analyses (subgroup analyses by stratum levels) to investigate effect modification (interaction). In most wound trials, the centre or investigator should almost always be considered as a factor in the analysis, due to variations in standard of care. If wound healing is measured by repeated measurements on a continuous scale, then the rate of healing can be evaluated by simple summary measures analysis, growth curve models and a gamut of increasingly sophisticated methods for longitudinal clustered data.

Missing Values

The significance of missing outcome data is their impact on the analysis data set in terms of the actual number of subjects that end up being analysed and effect on the balance of potential confounders within treatment groups. Missing values can cause bias in the estimation if the reasons are associated with treatment outcome. Even if there is no differential reason, exclusion of subjects because of missing data reduces the power of the study. If the outcome is binary, then subjects with missing outcomes can be 'assigned' outcomes to produce a worst-case vs. best–case sensitivity analysis. When a substantial portion of values is missing, concerns arise about the quality of the trial execution. The best approach is to anticipate problem areas and implement preventive measures and prespecify analytical approaches.

Data Transformation and Covariate Analyses

Stratified randomisation should balance the arms for the one or two most important covariates. Regression modelling can be employed to adjust for further variables that might affect the outcome. These covariates should be prespecified, and the analyses should also be prespecified to avoid concerns about interpretability of significance tests. When analysing covariates, experience suggests that it is generally wasteful of information to transform continuous variables into dichotomous variables (e.g. baseline ulcer size ≥ 5 cm^2, duration of the ulcer >1 year). The covariate should be used as a continuous variable unless there are gains in ease of interpretability by categorising. Exploratory analyses may examine subgroups defined by various cut points, but when a particular cut point is deemed to be important in guiding the use of the product (e.g. ulcers greater than 10 cm do not respond), this cut point should be prospectively identified and studied in a confirmatory clinical trial.

Despite great care in designing a trial and in the enrolment of study subjects the results can still leave us with a great deal to ponder over. In a Phase II randomised controlled study of the effects of transforming growth factor (TGF) $\beta 2$ on wound healing in diabetic foot ulcers, Robson *et al.*[9] designed the study to establish the safety and effective dose of transforming growth factor β 2 that would improve healing of chronic foot ulcers in diabetic patients. The design was a double-blind, placebo-controlled, multicentre trial. Randomisation was into five groups:

Group 1	Group 2	Group 3	Group 4	Group 5
Standard care	Placebo topical collagen sponge	TGF $\beta 2$ sponge 0.05 μg/cm^2	TGF $\beta 2$ sponge 0.5 μg/cm^2	TGF $\beta 2$ sponge 5.0 μg/cm^2 TGF $\beta 2$
$N = 24$	$N = 22$	$N = 43$	$N = 44$	$N = 44$

Standard care consisted of sharp debridement, coverage with non-adherent dressing, and weight off-loading of the affected foot.

All the groups were subject to this standard care as well. Outcome measures were divisible into:

Primary – complete closure of the wound and percentage wound area reduction at or before 21 weeks.

Secondary – time to wound closure and durability of wound closure.

Closure rates were:

Group 1	Group 2	Group 3	Group 4	Group 5
71%	32%	58%	57%	61%

Median time to wound closure compared with placebo sponge was significantly reduced in the TGF $\beta 2$ 5.0 $\mu g/cm^2$ group. Durability of wound closure was, however, comparable in all five groups. This study supports the null hypothesis yet it is noted that the placebo sponge group did worst of all the groups, implying some detrimental effect of the sponge. The authors dismiss this by drawing on a platelet-derived growth factor trial which also returned a closure rate of 33%. The better closure rate for the standard care group was explained as an anomaly of the treatment allotment as a result of the small sample size.

SUMMARY OF CLINICAL TRIALS REGISTERED IN THE COCHRANE LIBRARY AS SYSTEMATIC REVIEWS

Objective	No. of RCTs	Odds ratio and 95% CI	Conclusions
1. Prophylactic antibiotics in mammalian bites reduce wound infection[10]	8	0.1 (0.01 to 0.86)	Confirmatory research required
2. Compression hosiery or bandaging prevents recurrence of venous ulcers and is there an optimum pressure to this effect[11]	0 1	0.82 (0.61 to 1.12)	No trials with comparison to control group without compression reported No compression associated with recurrence of ulcer. One trial shows higher compression hosiery more effective
3. Effectiveness and cost-effectiveness of compression bandaging in treatment of venous ulcers[12]	22	N/A	Compression more effective than no compression. Elastic compression more effective. No advantage with four-layered bandage over high-compression systems. Insufficient data to conclude on cost
4. Assess the evidence for the effectiveness of debridement treatment for diabetic foot ulcers[13]	5 3 of hydrogel therapy 1 of surgery 1 of larval	AR 0.23 (0.1 to 0.36)	Inconclusive No evidence to suggest hydrogel increases healing rate of diabetic foot ulcers

Objective	No. of RCTs	Odds ratio and 95% CI	Conclusions
5. Effectiveness of dressings and topical agents for surgical wounds healing by secondary intention[14]	13	Wound healing (1) −25.6 days (to −2.12 days) Pain (4) Patient Satisfaction (3) Costs (2) Length of Hospital stay (4) −30.1 days (−49.82 to −10.38)	No direct comparisons Gauze painful Gauze less satisfied Gauze less cost but more nursing time no difference
6. Effectiveness of electromagnetic therapy in the treatment of pressure sores[15]	2	Not provided	No evidence of benefit based on two studies
7. Effectiveness of electromagnetic therapy in the treatment of venous leg ulcers[16]	3	Not provided	Difference not statistically significant. No current evidence available for EMT
8. Assess benefits and harm of adjunctive hyperbaric oxygen therapy for treatment of chronic ulcers of the lower limb[17]	5 Diabetes amputation reduction (4) Healing (1) Venous ulcer (1) Arterial and pressure ulcers (O)	RR 0.31 (0.13 to 0.71) NNT 3 to 11 RR 2.3 (1.1 to 4.7, $p = 0.03$) WMD 0.33 (0.19 to 0.47% $p = 0.00001$)	HBOT of benefit Drop-outs did not alter result Significant benefit HBOT needs further evaluation
9. Intermittent pneumatic compression for treating venous leg ulcers[18]	4 1 small trial 45 subjects 2 trials 75 subjects 1 trial 16 subjects	RR 11.4 (1.6 to 82) No benefit No difference	Further trials required
10. Effectiveness of low-level laser in the treatment of venous leg ulcers[19]	4 2 vs. sham 1 three-arm study 1 laser vs. UV light	No difference Combination laser + IR light healed more ulcers No difference	No evidence of benefit given four RCTs
11. Effectiveness of enteral and parenteral nutrition on the treatment of pressure ulcers[20]	8 4 on prevention 4 on treatment	1 larger study showed reduced no. of ulcers Heterogeneous trials inappropriate for meta-analysis	8 trials small and poor methodological quality
12. Effectiveness of oral zinc in healing arterial and venous leg ulcers[21]	6	N/A	Trials are small. May be of benefit in venous ulcer patients with lower zinc levels assayed as baseline

Objective	No. of RCTs	Odds ratio and 95% CI	Conclusions
13. Effectiveness of patient education on the prevention of foot ulcers with diabetes mellitus[22]	9 4 intense education with brief interventions 1 in high-risk patients for reduction of ulcer incidence and amputation rate 1 study found patient education as part of complex intervention had reduction	0.28 (0.13 to 0.59) 0.32 (0.14 to 0.71) 0.41 (0.16 to 1.00)	Most studies of poor methodological quality
14. Effects of pentoxifylline for treating ulcers compared with placebo or other therapies in the presence or absence of compression[23]	9 (572 pt) 8 vs. placebo Pentox + compress adverse effects with pentox (vs. defibrotide)	1.41 (1.19 to −1.66) 1.3 (1.10 to 1.54) 1.25 (0.87 to 1.80)	Pentoxifylline appears to be effective adjunct to compression No cost-effectiveness data
15. Effectiveness of pressure-relieving interventions in the prevention and treatment of diabetic foot ulcers[24]	4	N/A	Limited evidence of effectiveness of orthotics for prevention and of therapeutic shoes and of total contact casts
16. Effect of skin grafts for treating venous leg ulcers[25]	9 (579 pt)	N/A	Trials generally of poor quality. Evidence that bilayer artificial skin used in conjunction with compression bandaging increases chance of healing compared with compression and simple dressings. Further research needed
17. Effectiveness of support surfaces and pressure-relieving devices for	41	N/A	Foam alternatives to std mattresses can reduce incidence. Merits of low-pressure devices

Objective	No. of RCTs	Odds ratio and 95% CI	Conclusions
prevention of pressure ulcers[26]			and alternates unclear. Pressure-relieving overlays in operating room effective
18. The effectiveness of therapeutic touch for healing acute wounds[27]	4	N/A	Two studies ($n = 68$) showed significant effect of TT. Remaining two studies favoured the control arm. Author concluded on insufficient evidence
19. Effectiveness of the use of therapeutic ultrasound in the treatment of pressure sores[28]	3	N/A	Two RCTs compared ultrasound vs. sham showed no significant difference. Third RCT comparing US/UV combination with laser found increase in healing rates for US/UV but no statistical significance
20. Effectiveness of therapeutic ultrasound in the treatment of venous leg ulcers[29]	7	N/A	Four RCTs compared US with sham – no difference. Three RCTs compared US with standard treatment – no difference. Author – studies show possible benefit with ultrasound – caution on interpretation of results
21. Effectiveness of dressings, local anaesthesia or topical anaesthesia for pain relief in venous leg ulceration[30]	6	20.6 (29.11 to 12.19)	EMLA cream provides effective pain relief for ulcer debridement. No trials addressing treatment of persistent pain
22. Effectiveness of topical negative pressure in treating chronic wounds[31]	2 ($n = 34$)	N/A	Trial 1 showed wound volume reduction in 6 weeks. Trial 2 showed reduction in no. of

Objective	No. of RCTs	Odds ratio and 95% CI	Conclusions
			days to healing and reduction in wound surface area at 2 weeks in favour of TNP. Author – weak evidence
23. Assess the effects of water compared with other solutions for wound cleansing[32]	3	For chronic wounds odds of infection cleansing with tap water vs N saline 0.16 (0.01 to 2.96) For acute wounds, lower infection rates, 0.52 (0.28 to 0.96)	Author – evidence limited. Tap water in acute wounds reduces infection rate. Quality of tap water is the issue

The Cochrane Reviews represent the accumulation of randomised controlled trials in issues relative to wound healing. Needless to say, *point estimates* and *95% confidence intervals* leading to Forest plots can become meaningful only if the original source trials and specifically randomised controlled trials are *well designed* and *methodically* sound. Trials 11, 13 and 16 were deemed trials with questionable methodology and quality and obviously raise *ethical issues*. In the literature there are many cautionary notes on the necessity for well-designed and executed clinical trials, particularly in wound care, that can make an enormous difference to management of problem wounds, especially since there are a myriad of newer technologies which claim efficacy. This is not so much a damper on innovativeness but the need to have evidence that a newer therapy, albeit expensive, does have unequivocal benefit to the patient, which would go a long way to justifying adoption.

REFERENCES

1. Margolis DJ, Bilker W, Santanna J, Baumgarten M, Venous leg ulcers: incidence and prevalence in the elderly. *J Am Acad Dermatol* (2002) **46**: 381–6.
2. Margolis DJ, Guest Editorial: The swings and roundabouts of randomized controlled studies in wound healing. *Lower Extremity Wounds* (2004) **3**(1): 4–6.
3. Jude EB, Blakytny R, Bulmer J, Boulton AJ, Ferguson MW. Transforming growth factor-beta 1, 2, 3 and receptor type I and II in diabetic foot ulcers. *Diabet Med* (2002) **19**(6): 440–7.
4. Yager DR, Chen SM, Ward S, Olutoye OO, Diegelmann RF, and Cohen IK. The ability of chronic wound fluids to degrade peptide growth factors is associated with increased levels of elastase activity and diminished levels of proteinase inhibitors. *Wound Repair Regen* (1997) **5**: 23–32.
5. Burnard KG, Whimster I, Naidoo A *et al.* Pericapillary fibrin in the ulcer bearing skin of the leg: the cause of lipodermatosclerosis and venous ulceration. *BMJ* (1982) **285**: 1071–72.
6. Falanga V, Kirsner R, Katz MH *et al.* Precapillary fibrin cuff in venous ulceration. Persistence with treatment and during ulcer healing. *J Dermatol Surg Oncol* (1992) **18**: 409–414.
7. Falanga V, Eaglstein WH. The ‘trap’ hypothesis of venous ulceration. *Lancet* (1993) **341**: 1006–8
8. Falanga V. Clarification for wound bed preparation and stimulation of chronic wounds. *Wound Repair Regen* (2000) **8**(5): 347–52.
9. Robson MC, Steed DL, McPherson JM, Pratt BM. Effect of transforming growth factor $\beta 2$ on wound healing in diabetic foot ulcers, *J Appl Res* (2002) **2**(2).
10. Medeiros I, Saconato H. Antibiotic prophylaxis for mammalian bites. *The Cochrane Database of Systematic Reviews* (2001) Issue 2. Art. No. CD001738. DOI: 10.1002/14651858.CD001738.
11. Nelson EA, Bell-Syer SEM, Cullum NA. Compression for preventing recurrence of venous

ulcers. *The Cochrane Database of Systematic Reviews* (2000) Issue 4. Art. No. CD002303. DOI: 10.1002/14651858.CD002303.

12. Cullum N, Nelson EA, Fletcher AW, Sheldon TA. Compression for venous leg ulcers. *The Cochrane Database of Systematic Reviews* (2001) Issue 2. Art. No. CD000265. DOI: 10.1002/14651858.CD000265.
13. Smith J. Debridement of diabetic foot ulcers. *The Cochrane Database of Systematic Reviews* (2002) Issue 4. Art. No. CD003556. DOI: 10.1002/14651858.CD003556.
14. Vermeulen H, Ubbink D, Goossens A, de Vos R, Legemate D. Dressings and topical agents for surgical wounds healing by secondary intention. *The Cochrane Database of Systematic Reviews* (2004) Issue 1. Art. No. CD003554.pub2. DOI: 10.1002/14651858.CD003554.pub2.
15. Flemming K, Cullum N. Electromagnetic therapy for treating pressure sores. *The Cochrane Database of Systematic Reviews* (2001) Issue 1. Art. No. CD002930. DOI:10.1002/14651858.CD002930.
16. Flemming K, Cullum N. Electromagnetic therapy for treating venous leg ulcers. *The Cochrane Database of Systematic Reviews* (2001) Issue 1. Art. No. CD002933. DOI:10.1002/14651858.CD002933.
17. Kranke P, Bennett M, Roeckl-Wiedmann I, Debus S. Hyperbaric oxygen therapy for chronic wounds. *The Cochrane Database of Systematic Reviews* (2004) Issue 1. Art. No. CD004123.pub2. DOI: 10.1002/14651858.CD004123.pub2.
18. Mani R, Vowden K, Nelson EA. Intermittent pneumatic compression for treating venous leg ulcers. *The Cochrane Database of Systematic Reviews* (2001) Issue 4. Art. No. CD001899. DOI: 10.1002/14651858.CD001899.
19. Flemming K, Cullum N. Laser therapy for venous leg ulcers. *The Cochrane Database of Systematic Reviews* (1999) Issue 1. Art. No. CD001182. DOI: 10.1002/14651858.CD001182.
20. Langer G, Schloemer G, Knerr A, Kuss O, Behrens J. Nutritional interventions for preventing and treating pressure ulcers. *The Cochrane Database of Systematic Reviews* (2003) Issue 4. Art. No. CD003216. DOI: 10.1002/14651858.CD003216.
21. Wilkinson EAJ, Hawke C. Oral zinc for arterial and venous leg ulcers. *The Cochrane Database of Systematic Reviews* (1998) Issue 4. Art. No. CD001273. DOI: 10.1002/14651858.CD001273.
22. Valk GD, Kriegsman DMW, Assendelft WJJ. Patient education for preventing diabetic foot ulceration. *The Cochrane Database of Systematic Reviews* (2001) Issue 4. Art. No. CD001488.pub2. DOI: 10.1002/14651858.CD001488.pub2.
23. Jull AB, Waters J, Arroll B. Pentoxifylline for treating venous leg ulcers. *The Cochrane Database of Systematic Reviews* (2002) Issue 1. Art. No. CD001733. DOI: 10.1002/14651858.CD001733.
24. Spencer S. Pressure relieving interventions for preventing and treating diabetic foot ulcers. *The Cochrane Database of Systematic Reviews* (2000) Issue 3. Art. No. CD002302. DOI: 10.1002/14651858.CD002302.
25. Jones JE, Nelson EA. Skin grafting for venous leg ulcers. *The Cochrane Database of Systematic Reviews* (2000) Issue 2. Art. No. CD001737.pub2. DOI: 10.1002/14651858.CD001737.pub2.
26. Cullum N, McInnes E, Bell-Syer SEM, Legood R. Support surfaces for pressure ulcer prevention. *The Cochrane Database of Systematic Reviews* (2004) Issue 3. Art. No. CD001735.pub2. DOI: 10.1002/14651858.CD001735.pub2.
27. O'Mathuna DP, Ashford RL. Therapeutic touch for healing acute wounds. *The Cochrane Database of Systematic Reviews* (2003) Issue 4. Art. No. CD002766. DOI: 10.1002/14651858.CD002766.
28. Flemming K, Cullum N. Therapeutic ultrasound for pressure sores. *The Cochrane Database of Systematic Reviews* (2000) Issue 4. Art. No. CD001275. DOI: 10.1002/14651858.CD001275.
29. Flemming K, Cullum N. Therapeutic ultrasound for venous leg ulcers. *The Cochrane Database of Systematic Reviews* (2000) Issue 4. Art. No. CD001180. DOI: 10.1002/14651858.CD001180.
30. Briggs M, Nelson EA. Topical agents or dressings for pain in venous leg ulcers. *The Cochrane Database of Systematic Reviews* (2003) Issue 1. Art. No. CD001177. DOI: 10.1002/14651858.CD001177.
31. Evans D, Land L. Topical negative pressure for treating chronic wounds. *The Cochrane Database of Systematic Reviews* (2001) Issue 1. Art. No. CD001898. DOI: 10.1002/14651858.CD001898.
32. Fernandez R, Griffiths R, Ussia C. Water for wound cleansing. *The Cochrane Database of Systematic Reviews* (2002) Issue 4. Art. No. CD003861. DOI: 10.1002/14651858.CD003861.

34

Palliative Care

ROBERT S. KROUSE[1] AND DAVID CASARETT[2]
[1]Surgical Care Line, Southern Arizona Veterans Affairs Health Care System, Tucson, Arizona, USA
[2]University of Pennsylvania, Philadelphia, Pennsylvania, USA

INTRODUCTION

Palliative care is an interdisciplinary team approach to care with a focus on comfort and quality of life (QOL) rather than prolongation or 'cure' for patients and their loved ones.[1] The goal of good palliative care is to relieve suffering and to improve QOL. However, it is apparent that access to palliative care is inconsistent, and standards to guide palliative care have not been established clearly. At least in part, these deficiencies exist because of a lack of solid evidence on which to base clinical decisions.[2] Therefore, there is an urgent need for research that can provide evidence to define the standard of care and to increase access to quality care.

Recent years have seen a dramatic increase in palliative care research, defined broadly as activities that are designed to contribute to generalisable knowledge[3] about end-of-life care. This growth has created a heterogeneous field that encompasses both qualitative and quantitative techniques, and descriptive as well as interventional study designs.[4] Although the past 10 years have seen impressive growth in all of these areas, this rate of growth appears to be particularly rapid for interventional research, including controlled trials of pain medications,[5,6] interventional procedures for pain,[7] and other non-pharmacological interventions to improve a variety of aspects of end-of-life care.

OVERVIEW OF PALLIATIVE CARE PROBLEMS

There are multiple examples of problems that can affect the QOL for a patient facing the end of life (EOL). These can be categorised based on symptom or system in the body that is affected. Major symptoms include pain, dyspnea, anorexia and depression. Related to body systems, one can imagine a potential symptom related to each body system. Neurological problems include fatigue, headache and other pain syndromes, and delirium. Pulmonary complications include dyspnea, fatigue and immobility. Cardiac symptoms include shortness of breath, fatigue and pain. Gastrointestinal problems include obstructions, diarrhoea, nausea, vomiting and anorexia. Musculoskeletal complications include fractures, functional loss and pain. Epidermal problems mainly

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

focus on wound problems, but also can include poor cosmesis and pain. Complications related to the haematologic system include infection and fatigue. Urologic problems include ureteral obstructions, bleeding and pain. It is this compendium of problems that palliative care research focuses on, rather than increase in survival time or cure.

INTERDISCIPLINARY NATURE OF CARE FOR PATIENTS FACING THE EOL

There are multiple potential treatments for palliative care problems. These can include medical symptom management, surgical interventions, radiation therapy options, endoscopic opportunities, chemotherapy, as well as many innovative and complementary alternatives. In addition, palliative care teams must include nursing, social work, chaplain services, along with other potential specialties that could impact on a patient's QOL. Therefore, it is in this setting that one must consider palliative research. One must consider comparisons of widely variant therapies to understand the best treatment approaches. In addition, research teams must include investigators from other specialties familiar with a variety of treatment options. Importantly, palliative care research must be imaginative in both options to compare and research methods to better understand the research questions at hand.

IMPORTANCE OF RESEARCH FOR PATIENTS FACING THE EOL

Approximately 2.5 million people die each year in the United States. Most will die with chronic disease, although many deaths are the result of acute illness and injury. Death is a normal process of life, and therefore must be a focus of research to ensure the best therapies possible are offered to patients. Palliative care is fraught with anecdote and opinion. There is a clear lack of evidence related to the best treatments for palliative care problems. Therefore, we may not be delivering the best care for a large number of patients in the United States each year.

There are many potential reasons for the lack of palliative care research. Many of them are related to ethical aspects of this research. There are also innate barriers, such as a lack of trained researchers and challenges of subject recruitment. For instance, many hospice facilities are more willing to increase their involvement in palliative care research, although many institutions may not be willing to enroll patients at this time.[8] Finally, there has been a historical lack of funding for such research, although this may be improving with time, especially with new sources of funding.

It is important to do research for patients facing the EOL, as it is unethical not to do this kind of research. There is frequently the belief that no such research is morally justifiable in this patient population.[10] Although others find the arguments to this conclusion unacceptable,[2] this still might be a prevailing belief among many practitioners. It is imperative that the research and clinical community not bias themselves to palliative care research protocols. In fact, it can be argued, based on the Declaration of Helsinki and the generally accepted ethical code of practice in clinical research, that *not* offering patients at the end stage of life the opportunity to take part in clinical research is unethical.[9] One approach that has been suggested to help with this problem is to identify support personnel to confidentially counsel researchers undergoing psychological stress related to research trials.[11]

TYPES OF RESEARCH PROBLEMS

There are many clinical palliative care questions that need to be studied. Most notably, it is essential to define treatment plans that offer patients the greatest possible QOL. Treatment plans may include medical management, surgical and other invasive interventions, radiation therapy, chemotherapy, and complementary or alternative approaches. Other clinical research needs in palliative care include decision making and preferences, communication skills, education, and issues related to utilisation of services.

Particular attention must be devoted to patients of different cultural backgrounds, as culture fundamentally shapes the way people make meaning out of illness, suffering and dying, and therefore also influences how they make use of medical services at the EOL.[12] Issues of resource utilisation will necessitate study of ethnic and racial issues, as well as issues related to socio-economic and educational background. These issues are particularly important because only a minority of hospitals have palliative care services to address the needs of patients and their families in the dying process.[13] In addition, issues related to poor prognostication,[14] and how this affects utilisation and care need to be studied. Research related to education needs to be at all levels, including patient, physician and other health care workers. Gaps in knowledge affect the entire range of care issues; there needs to be a research focus on these issues.

HISTORICAL PERSPECTIVE OF TRIALS: ARE PALLIATIVE TRIALS 'PALLIATIVE'?

Historically, palliative care treatment studies have typically been retrospective.[15] When prospective studies have been attempted, they have generally enrolled small numbers of patients, and therefore limited conclusions can be drawn. In addition, palliative care trials have only rarely included QOL measures as a primary endpoint. This omission has been clearly outlined in the surgical literature, where it has been noted that documented outcome measures are predominantly physiologic response (69%), survival (64%), and morbidity and mortality (61%).[15] QOL is only noted in 17%, and pain control only in 12%, of articles in the surgical literature.

In addition, a similar absence of QOL data detracts from the usefulness of medical oncology studies. One reason for a lack of attention to QOL in oncology is that once a treatment is initiated, patients frequently have unrealistic goals, and feel that the intent of therapy is 'curative' or to improve their survival.[16] Chemotherapy may offer a QOL improvement in some patients, although this must be considered along with costs and treatment-related morbidity.[16] One example of a chemotherapeutic intervention whose only role is seen to be palliation is Gemcitabine for unresectable or metastatic pancreatic cancer.[17–23] In larger trials with this drug, the primary outcome measure was 'clinical benefit', which was a combination of Karnofsky status, pain medication consumption and pain intensity score. Weight change was also considered a secondary measure.[21] These outcome measures have become the standard for other Gemcitabine studies for pancreatic cancer. While the availability of a single composite measure is useful, it is not clear that this composite measure offers the best possible picture of treatment outcomes. In the literature, though, it is all too common for the outcomes such as response rate and survival to be stressed more than clinical benefit, and therefore clinicians may be discussing the benefit to patients when there may be little or no palliative gain. This may then be misleading and possibly detrimental to patients and their families.

Related to randomised clinical trials (RCT) in palliative medicine, there are few examples of large studies. One of the best examples of attempts at RCTs in the palliative literature is related to the problem of bile duct obstruction seen with periampullary cancers. There are at least five randomised prospective trials comparing stenting with surgical bypass. One of the trials utilised transhepatic endoluminal stenting,[24] while the other studies utilised endoscopic stenting.[25–27] Surgical bypass procedures were variable, as necessitated in the operating room. The studies range in number (from 48 to 202 patients). These studies basically have shown that stenting is as efficacious as operative approaches for biliary obstruction. They also showed a higher recurrence rate for stenting, but due to the fact that the initial stent placement has fewer complications than a surgical procedure, initial stenting has become a standard of care for this problem. These studies come to relatively similar conclusions and make it possible to answer a simple question related to a common palliative care problem. One must note, though, that the outcome measures concentrated on technical success, length of stay, complications, recurrence and survival, and only one study attempted a crude QOL analysis.[26] Validated QOL measures of patient and family satisfaction were not utilised. While these other measures are important and can be used as surrogates for QOL, they do not fully help in understanding the complexity of patients with advanced cancers and the impact of treatments on their QOL. These studies still were able to allow significant change in patient treatment for those facing the EOL.

ETHICAL DILEMMAS AND BARRIERS TO PALLIATIVE CARE RESEARCH

Despite the valuable knowledge that has been produced by palliative research, and the promise of future important advances, its progress has been delayed by a persistent uncertainty about the ethics of these studies. Indeed, there have been concerns raised from several quarters about whether patients near the EOL should ever be asked to participate in any form of research.[8,10] Others have objected to this extreme position.[2] Nevertheless, many providers, Institutional Review Boards (IRBs), ethics committees, study sections and even investigators remain uncertain about the ethical limits of research involving dying patients.

These concerns have considerable intuitive appeal, and must be taken seriously. Indeed, it would be unfortunate if the progress of palliative care research were slowed by the sorts of ethical scandals that have threatened other fields of research that involve vulnerable populations, such as those with mental illness.[29] However, strict oversight and tight limits on palliative care research have the potential to do equal damage to a growing field. Therefore, in order to avoid potential scandals, without excessive regulation and oversight, it will be important that palliative care investigators and clinicians consider these concerns in a fair and balanced way.

There are six ethical aspects of palliative care research that investigators and clinicians should consider in designing and conducting palliative care research. These include: (1) whether a study is research or quality improvement; (2) the study's potential benefits to future patients; (3) the study's potential benefits to subjects; (4) the study's risks to subjects; (5) subjects' decision-making capacity; and (6) the voluntariness of subjects' choices to participate in research. Each of these is discussed, as well as opportunities to enhance the ethics of palliative care research in each of these ways.

BENEFITS TO FUTURE PATIENTS: A STUDY'S VALIDITY AND VALUE

Palliative care research is designed to produce knowledge that will advance understanding of EOL care. Implicit in this goal is the expectation that this knowledge will eventually improve care for future patients. Therefore, the first ethical aspect of palliative care research that deserves consideration is its potential benefits for future patients. These benefits to others can be described in terms of validity and value.

Validity

First, all studies must be valid. That is, they must use techniques of design and data analysis that peer reviewers can agree are appropriate. In addition, all studies must be designed to produce knowledge that is generalisable. Indeed, generalisability is the cornerstone of the Common Rule's definition of research: 'a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge'.[3] These requirements collectively describe a study's validity.[30] Validity is a threshold requirement for all research, because it is unethical to expose human subjects to risks in studies that peer reviewers agree cannot adequately answer a research question.[31] Therefore, at a minimum, investigators routinely consider a study's validity.

Value

Above this threshold of validity, palliative care studies may offer more or less importance or 'value'. Broadly, value can be defined as the likelihood that a study's results will improve the health and well-being of future patients.[32] Like validity, value is an important measure of a study design's scientific quality, but it is also a measure of its ethical quality. Value is an essential aspect of a study's ethical design because a central goal of research is to produce knowledge that will ultimately be 'important',[3,33] 'fruitful',[34] or 'valuable'.[35] In fact, one reason that subjects participate in clinical research is to produce knowledge that will benefit others. Because subjects are willing to accept risks and burdens of research at least in part in order to benefit others, investigators have accepted an ethical responsibility to maximise the probability that a study will be able to do so. Therefore, in addition to widely accepted scientific arguments for valuable research, there are compelling ethical arguments as well.

Maximising Validity and Value in Palliative Care Research

Space does not permit a comprehensive overview of ways in which a palliative care study's validity and value can be assessed and improved. Indeed, such a discussion moves quickly beyond ethics and into the technical language of study design and health measurement. Nevertheless, several broad recommendations are possible.

First, a study's sample size should be adequate to answer the research question that is posed. Problems of underpowered studies, and particularly clinical trials, are both widespread and well described.[36] But issues of power and sample size are particularly relevant to pain and symptom research, in which random variation can be quite large.[37] To minimise these problems, it may be useful to establish consortia or collaborative groups that can participate in multi-centre studies. Such arrangements have been highly effective in promoting research on rare disorders, and may be applicable as well to palliative care research, in which investigators are limited and available patients are often sparse.

Second, palliative care investigators can enhance the ethical quality of a study by taking reasonable steps to increase the generalisability of its results. These steps might include sample size calculations that permit subgroup analysis of groups of patients that have typically not been the focus of investigation, such as patients with non-cancer diagnoses, or elderly patients. The generalisability of a study's results might also be enhanced by recruiting subjects outside academic medical settings, because preliminary evidence suggests that these patients, and their needs for care, may be different than those who receive care in academic settings.

In addition, palliative care investigators can enhance the generalisability, and therefore the value, of their research by making reasonable efforts to include patients who are receiving care at home, and particularly those who are enrolled in a home hospice programme. Substantial barriers may make it difficult to include these patients in research. Nevertheless, few data exist to guide the management of home care patients near the EOL, and palliative care investigators can enhance the value of their research by including this population whenever possible.

Of course, all of these improvements in generalisibility come at a substantial cost. For instance, studies that recruit subjects from several different settings require more elaborate designs for recruitment and follow-up. In addition, investigators who include plans for subgroup analysis in their sample size calculations face rapidly escalating sample size requirements and costs. Nevertheless, steps like these offer an important way to enhance a palliative care study's value, and therefore its ethical quality. Therefore, it will also be important that funding agencies understand the ethical importance of generalisibility, and that generalisibility comes with a financial cost.

BENEFITS TO SUBJECTS

Palliative care investigators can also enhance the ethical rigour of a study by maximising the benefits that it will offer to subjects. Broadly, these benefits can be considered under two categories: benefits to subjects during the study and benefits from the data that are collected. Each of these is discussed below.

Benefits to Subjects During the Study

Investigators may have several opportunities to maximise potential benefits of research to the subjects who participate. Perhaps the first, at least in an interventional study, is in their choice of an intervention. Ideally, a new intervention to be studied should have a reasonable chance of success. More important, though, if it is to offer subjects a significant potential benefit, an intervention should offer the possibility of a meaningful improvement over other interventions that are available to subjects outside the study. For instance, a pain management algorithm that is expected to reduce cancer pain[38] would only offer potential benefits if it is qualitatively or quantitatively different than those that constitute the usual standard of care. On the other hand, a comparison of two medications that are commercially available, such as topical fentanyl and sustained release morphine would not offer subjects any potential benefit compared to regular medical care. This is true even if the study's results offer considerable clinical value.[39]

The potential benefits of a study can also be enhanced by choosing an active control design, rather than a placebo.[39,40] If a placebo is used, a study's potential benefits can also be improved by altering the standard 1:1 randomisation scheme in a placebo-controlled trial in a way that increases subjects' chances of receiving an active agent.[6] The potential benefits of a placebo-controlled trial can also be enhanced by using a crossover design, so that all subjects are offered potential benefits, if the medication's pharmacokinetic profile makes it possible to avoid carryover effects.

These suggestions should be tempered by two caveats. First, the potential benefits of research are never certain. If they were, a randomised trial would not be ethically acceptable. That is, a legitimate argument for the uncertainty that justifies a clinical trial, or equipoise, could not be made.[41] However, investigators generally design studies of interventions for which there is at least some evidence of effectiveness. Therefore, even though these potential benefits are not certain, they are more or less likely, and this assessment of likelihood should be considered in the design of pain research.

Second, palliative care studies need not always offer potential benefits. Indeed, many, and perhaps most, will not. Nevertheless, when a study does offer potential benefits, investigators may consider enhancing a study's potential benefits in these ways. The importance of doing so is particularly great if other aspects of a study raise ethical concerns, which might be the case if subjects' decision-making capacity is limited, or if the study's risks are substantial.

Benefits from Data Collected During a Study

Although the opportunities to enhance potential benefits described above apply largely to studies involving interventions, another opportunity applies equally well, if not better, to research that is descriptive. A common ethical issue in the design of palliative care research, and particularly descriptive research, is the possibility that data gathered may contribute to a subject's care. For instance, data gathered during a descriptive study may identify pain that is inadequately treated,[42–45] dissatisfaction with pain management,[46–50] or related clinical problems like depression.[51,52]

In anticipation of instances like these, investigators can design standard operating procedures that help to ensure that valuable clinical information is made available to the subjects and their clinicians. At the least, these procedures should include data about the presence of unrecognised and untreated symptoms, and concurrent disorders like depression. This is arguably an ethical obligation of symptom-oriented research. Moreover, these procedures offer a significant opportunity for investigators to enhance the potential benefits of pain research.

Benefits to Subjects After a Study Has Ended

Investigators can also enhance the potential benefits for subjects after a study has ended. These sorts of post-study benefits are not usually included in assessments of a study's balance of risks and benefits. They are also components of a study's value, because these benefits generally come from the knowledge that the study produced. Nevertheless, subjects may benefit from the knowledge to be gained from a study if the study's results are applied to their care. Investigators have numerous opportunities to ensure that these results are translated into subjects' care and, by doing so, can enhance the study's potential benefits to subjects.

For instance, subjects in palliative care research can benefit after a study if they learn from the study's aggregate results. This might be the case if a study comparing two pain medications found that one resulted in fewer side effects overall.[39] Subjects in the study would benefit from these data because this knowledge should allow them to make a more informed choice among available medications. Subjects might also benefit from results that are specific to them. For instance, if a subject receives two medications in a blinded crossover trial, and prefers one to the other, the subject would be better able to choose between these medications in future clinical situations, armed with the results of a blinded comparison of the two.[53–55]

Finally, investigators can increase the likelihood that subjects have continued access to medications that are studied. If medications are not available, either due to high cost or because the medication has not yet received regulatory approval, subjects will not benefit (immediately) from the study's results. Thus by arranging reduced rate programmes or open label extension phases, investigators can increase a study's potential benefits for subjects by helping to ensure that subjects will benefit from the study's results.

This benefit may be particularly important in palliative care research, because mortality rates in some studies are very high. This means that subjects may not live long enough to see a study medication's approval for clinical use, or to see a study's results published and translated into improved care. For this reason, it is especially important that investigators consider mechanisms by which results can be applied to the care of research subjects in a timely fashion.

MINIMISING RISKS AND BURDENS

Investigators can also enhance a study's ethical soundness by taking steps to minimise a study's risks and burdens. Although the distinction between risks and burdens is not always clear, a rough heuristic is useful. In general, a risk can be considered as the probability of an adverse medical event or undesirable outcome. Risks might include side effects of a medication, or increased pain during a study. The term 'burden' can be used to describe those unpleasant features of participation in a study that are more certain, and which are better thought of as inconveniences. Additional clinic visits, time spent filling out questionnaires, or time spent waiting in clinic might be described as burdens.

Identifying Risks and Burdens

Attention to the ethical design of pain research, and to the minimisation of research risks and burdens, requires a clear agreement about how they should be defined. The criteria by which study risks and burdens are identified and evaluated uses the concept of incremental or 'demarcated' risks imposed by participation in a study.[56] The application of this standard to interventional pain research would mean that investigators designing a trial to compare the effectiveness of two opioids[39] need not go to great lengths to justify the risks of the opioids being evaluated, if subjects in the trial would have received similar medications, with similar risks, off protocol. Of course, the risks of any medication in a clinical trial should be disclosed in the informed consent process.[3] Nevertheless, investigators are not under the obligation to minimise or justify these risks as they would be if, for instance, the same medications were being given to patients with mild pain, who would not receive them as part of standard care.

Minimising Risks: The Choice of Control

Perhaps one of the most contentious and emotional questions in palliative care research,[57,58] and indeed in research generally,[59–61] is whether a placebo or sham control arm is ethically appropriate. The ongoing debates about the scientific merit of these controls, and the competing advantages of active control superiority trials, and equivalency trials, are beyond the scope of this discussion. However, several general points can be made about the ethics of placebo- and sham-controlled trials. Each of these designs is discussed below.

Broadly, placebos can be defined as interventions that are 'ineffective or not specifically effective' for the symptom or disorder in question.[62] Increased attention to the ethical issue of placebo controls in recent years has produced a growing consensus that all subjects in a clinical trial should have access to the best available standard of care.[63] Thus in infectious disease research, for instance, all subjects with meningitis would have access to an antimicrobial agent that has proven effective. However, this requirement may be difficult to apply to studies of treatment for pain, other symptoms, or depression, in which the placebo response can be quite substantial. These difficulties are compounded when the symptom being studied is transient, such as incident pain (i.e., a sudden increase in pain).[6]

For these reasons, it may not be practical to prohibit placebos in palliative care research, and a placebo control may be ethically acceptable in several situations. First, placebos are acceptable if subjects receive a placebo in addition to the standard of care. For example, subjects might be randomly assigned to receive an opioid for pain, or an opioid plus an adjuvant agent. Second, a placebo arm is justified if the symptom under study has no effective treatment. For example, the transient nature of incident pain often defies adequate treatment on an as-needed basis, and a placebo control might be justified in a RCT of a novel agent for the treatment of incident pain if there is no effective treatment. Third, a placebo control is justified if subjects have adequate access to breakthrough, or 'rescue' treatment. This may in turn alter a trial's endpoints. For instance, the free use of breakthrough dosing in a trial suggests the possible inclusion of these doses as a study endpoint either directly[64,65] or as part of a composite endpoint.[5,66]

Concrete recommendations about sham procedures are somewhat more elusive, in part because sham procedures themselves are difficult to define. In general, though, sham procedures in palliative care research involve the use of a control procedure such as a nerve block, which is administered in a way that makes it ineffective or no real intervention is done.[7] These procedures create ethical concerns because some subjects, or all subjects, depending on the study's design, are exposed to the risks of the procedure without hope of its benefits.[59] Like placebo controls, though, shams also have a role in research, because the nonspecific therapeutic effects of surgery may be substantial. For instance, Leonard Cobb's research in the 1950s effectively debunked a widely used cardiac procedure that, if it had been widely disseminated, would eventually have put thousands of patients at risk.

Investigators have an opportunity to reduce these concerns substantially in the design of a sham-controlled study. For instance, investigators might conduct these studies in a setting in which the procedure itself (whether sham or real) poses few if any additional, or 'incremental' risks above and beyond usual care. Investigators might insert a sham epidural catheter that would then be used for post-operative analgesia.[67] When this is not possible, investigators can choose a crossover design, in which subjects are assigned to receive either the sham or the real procedure, followed by the other. This design does not decrease the incremental risks of the sham procedure. However, it does ensure that all subjects who bear the risks of the sham procedure also have access to the real procedure's potential benefits. This crossover sham design has been used in other settings,[68] and might be appropriate for pain research when the risks or discomforts of the sham procedure are substantial.

Minimising Burdens

For the most part, opportunities to minimise burdens are readily apparent. For instance, it seems reasonable wherever possible to minimise surveys, interviews and additional study visits.[69]

These are all burdens that investigators routinely consider carefully in designing studies. However, there may be other needs and concerns that may be unique to, or more common in, patients near the EOL.

Although it is intuitively obvious that all research subjects would like to avoid the added time commitment and inconvenience of travel to and from additional appointment, this concern may be especially important to patients near the EOL, for whom long periods of time spent sitting in a car can exacerbate discomfort. Similarly, patients may view surveys and questionnaires not only as time consuming, but also as a drain on their energy. Therefore, investigators who conduct palliative care research may have an added reason to minimise the burdens of extra visits and data collection procedures, and to rely on telephone data collection strategies whenever possible.

Palliative care investigators may also need to consider the burdens that a study creates for friends and family members who often take on substantial burdens as caregivers.[70–72] Although most of the burdens of research participation are borne by the subject, the requirements of time, travel, and perhaps time off work create burdens for others. Patients may be very sensitive to these burdens and, for some patients with chronic pain, burdens to others can be influential in the decision whether or not to enroll in a study. By building flexibility into a study design (e.g. use of brief telephone interviews, multiple options for timing of clinic visits) investigators may be able to reduce the burdens of research participation on others.

Ensuring Decision-Making Capacity

Patients who consent to participate in research should have adequate decision-making capacity, which refers to subjects' ability to understand relevant information, to appreciate the significance of that information, and to reason through to a conclusion that makes sense for them.[73] These concerns parallel concerns in research involving patients with dementia,[74] psychiatric illness,[75,76] and patients in the intensive care setting[77] among others. However, deficits in decision-making capacity may create several additional challenges for palliative care investigators.

First, concern about capacity is reasonable given the prevalence of cognitive impairment at the EOL.[78–81] Cognitive impairment occurs in 10 to 40% of patients in the final months and in up to 85% of patients in the last days of life.[79,80] Cognitive impairment may be difficult to identify in palliative care research because decision-making capacity varies over time,[82] and because impairment may result from the experimental or therapeutic medications themselves, such as opioids, benzodiazepines or corticosteroids.[83,84] Investigators who conduct trials of medications will encounter these challenges even more frequently if trials are designed to evaluate treatments for delirium, for which impairment is an inclusion criterion.[85,86]

Second, the effects of cognitive impairment on comprehension may be complicated by clinical depression, which occurs in between 5% and 25% of patients near the EOL.[51,52,87,88,89] Clinically significant adjustment disorders may be even more common.[51] It is possible that these disorders may impair either comprehension or decision making, or both,[81] but studies have not yet supported this conclusion.

Third, even in the absence of overt cognitive impairment or depression, it is possible that severe symptoms or affective disorders may impair subjects' ability to understand the risks and benefits of research participation. For some studies, particularly clinical trials, the presence of one or more of these intractable symptoms is an inclusion criterion.[90–92] It is possible that severe symptoms may impair comprehension if patients are unable to concentrate on the information offered in the informed consent process.[93]

Finally, these challenges may be compounded in prospective studies that require participation over days or weeks. In these studies, even if patients have the capacity to consent at the time of enrollment, they may not retain that capacity throughout the study. Thus days or weeks after patients give consent to participate, they may be unable to understand changes in their condition clearly enough to withdraw. The result can be a 'Ulysses contract' of sorts, in which research subjects find it easier to enroll than they do to withdraw.[94]

None of these challenges is easily remedied. Indeed, it is obstacles like these that lead some authors to argue that patients near the EOL should not be allowed to enroll in research.[10,27] Nevertheless, palliative care investigators have several concrete opportunities to enhance the ethical quality of palliative care research when decision-making capacity is uncertain.

First, at a minimum, investigators whose research involves patients near the EOL who are likely to lack decision-making capacity might institute brief assessments of understanding. Although this strategy cannot assess decision-making capacity, a few simple questions in either open-ended or multiple choice format provide a brief assessment of understanding.[95–97] In some situations, investigators may wish to assess decision-making capacity more formally using validated instruments.[98]

These sorts of safeguards need not be employed in all studies. Instead, their use should be guided by the prevalence of cognitive impairment in a study population and by the balance of risks and benefits that a study offers. For instance, when palliative care research involves only interviews or behavioural interventions that pose minimal risks, informal capacity assessments are generally sufficient. 'Minimal risks' are defined as those risks that are encountered during a patient's usual care, or in everyday life.[3] When research poses greater than minimal risks, but offers potential benefits, some assessment of understanding may be appropriate. This research includes studies that involve a placebo[6] or invasive interventions such as nerve blocks or epidural catheters. When a study that poses greater than minimal risks does not offer potential benefits, or is conducted in a population in which the prevalence of cognitive impairment is high (e.g. an inpatient hospice unit), a formal evaluation of capacity should be considered. This research includes studies that involve a placebo when an effective agent is available,[5] and some pharmacokinetic/pharmacodynamic studies that require repeated blood samples and prolonged observation, without potential benefits.

If a patient does not have the capacity to give consent, a legally authorised representative may be able to give consent for research. This follows from federal guidelines governing research involving children,[3] and is justified by the argument that surrogate decision makers should be allowed to consent to research, just as they are allowed to consent to medical therapy. However, as with other research that involves patients without capacity to consent, investigators should be aware of applicable state laws that may restrict or even prohibit surrogate consent for research. In addition, investigators in this field should be alert to possible future changes in federal regulations that have been discussed.

If a patient does not have the capacity to consent, but is still able to participate in decisions, investigators should obtain assent from the patient and informed consent from the patient's surrogate. This 'dual consent' ensures that patients are as involved in the decision as possible, yet provides the additional protection of a surrogate's consent.

If a patient has decision-making capacity intermittently, or is expected to lose capacity, investigators may obtain advance consent. This approach has been used in a study of treatment for delirium, in which informed consent was obtained from patients while they had decision-making capacity.[85] Advance consent should be obtained only for specific studies, and should be obtained close to the planned start of research, for instance at the time of hospitalisation or enrollment in a hospice or palliative care programme.

PROTECTING VOLUNTARINESS

Another way that investigators can enhance the ethical soundness of a study's design is to examine ways in which subjects' voluntary participation can be protected. In general terms, a choice is voluntary if it is made without significant controlling influences. At first glance, assurances of voluntariness appear to be an issue of informed consent, and in fact for the most part they are. However, a study's design and plan for subject selection and recruitment may have as great an influence on subjects' freedom to refuse research participation as does the informed consent process. In particular, two features of a study's design are relevant. First, a prospective subject's choice must be made with full knowledge of available alternatives.[3] Second, a subject choice must be made with the understanding that the subject can withdraw at any time.[3] Each of these creates opportunities in a study's design to ensure voluntariness that are discussed below.

Reasonable Alternatives to Participation

First, investigators can make sure that a study recruits subjects from an environment with excellent standards of palliative care. If patients generally receive excellent care, they will be best able to make a free and uncoerced choice about research participation. If, however, patients do not have access to a bare minimum of treatment options and expertise, they may view research participation more favourably, out of desperation.

One solution, albeit a somewhat draconian one, would be to require that palliative care research be conducted only in settings in which patients have access to a full range of services, treatment and expertise. Although this requirement would reduce the potential for research participation out of desperation, it would effectively limit research to a small number of academic centres, with a possible loss of generalisibility. Another more practical option might be to include a lead-in phase when clinical pain research is conducted in settings where the standard of care is poor. A lead-in phase allows an opportunity to optimise palliative care prior to recruitment. This strategy not only has ethical value but scientific value as well because it provides a uniform baseline prior to randomisation.

Opportunities to Withdraw

Investigators can also enhance the ethics of a study's design by ensuring that subjects are able to withdraw at any time. Although a subject's ability to withdraw should be a fundamental aspect of any ethical research,[3] there may be unique barriers to withdrawal from palliative care research. For instance, subjects who withdraw from clinical pain research that involves one or more medications will usually need access to a different medication upon withdrawal. This problem may be straightforward in many cases, but can be very challenging in an interventional study if the investigational medication is an opioid, which requires the subject to get a new prescription and get it filled. Most states have created considerable barriers to opioid prescribing, including triplicate prescriptions, which may make it very difficult for a subject to obtain a new prescription and get it filled in a timely manner. If a subject has medication available, the process may be easier. Nevertheless, considerable challenges of calculating an equianalgesic dose remain. For both of these reasons, investigators can enhance the ethical design of pain research by developing mechanisms to ensure that subjects who drop out continue to receive adequate pain treatment with as little interruption as possible.

PALLIATIVE CARE PROBLEMS AND CLINICAL TRIAL METHODS: ADOPTING STUDY DESIGN TO SETTING AND PROBLEM

STUDY SETTINGS

One major difficulty for palliative care trials is the location that they occur. Frequently, patients are in hospice care near the EOL. Because a hospice is designed to provide care for patients in their homes, recruiting hospice patients for research can be very difficult. These patients are seriously ill, and approximately one-half of hospice patients die within three weeks of hospice enrollment. Furthermore, hospice clinicians who provide home care often become the sole link between patients and the health care system. When these hospice clinicians are reluctant to approach patients to participate in research, the result is often slow recruitment, underpowered trials, and studies that cannot be completed.

Investigators may face similar challenges in a variety of other palliative care settings, including inpatient settings and palliative care units.

STUDY POPULATIONS

There are different considerations related to populations that are to be studied. Problems exist related to minority patients and their exclusion in trials, as well as insufficient study related to differences in important QOL measures for diverse populations. It will be necessary to focus recruitment efforts on underserved populations in order to ensure all populations are studied and results are generalisable for the entire population. Tailored recruitment strategies may be necessary to successfully recruit ethnic and racial minorities.

SAMPLE SIZE

As with all studies, appropriate sample size must be determined to garner useful differences. The most important point of sample size that arises in palliative care in which QOL measures are an important outcome is due to uncertainty about likely effect sizes. Specifically, the minimal clinical relevant difference to estimate sample size is often unknown.[99] There may not be sufficient evidence to define changes that one should expect or consider for palliative care trials.[99] Sample sizes utilising QOL data can be used for sample size estimation, though.[100,101] As with all studies, the adequate sample size must be carefully considered to detect meaningful differences in QOL.

TIMING

Consideration of all endpoint measurements must take into account the timing of evaluations. When QOL is the major endpoint, this takes on critical relevance. Symptoms and overall QOL change over time. QOL data are dependent on the timing of questionnaire administration and data collection. Different treatments, especially if divergent, may also have different trajectories of complications, as well as benefits. These issues must be considered in the initial planning of palliative care studies.

Longitudinal research is likely the most appropriate design for palliative care QOL assessments. This has been recommended to be on a weekly schedule.[102] This is due to the short median survival of patients, dramatic QOL and symptom changes nearest the EOL, and this is the shortest intervention period that is likely to give a clinically significant effect in the management of patients with advanced disease.

MISSING DATA

This issue is so important in palliative care trials that it is worth separate consideration. The missing data problem can be subdivided into missing forms and missing items.[99] It may be reasonable to pilot the data collection system prior to the study to ensure the team can perform the study as anticipated. This can include a debriefing form to better understand what items patients might avoid and why they did so.[99] Missing data may bias data in unpredictable ways.[99] If missing data are very frequent, then the entire study may not be able to answer the questions for which it was designed.

OUTCOME MEASURES

To detect benefit in a palliative care trial, the appropriate outcome measures must be utilised. There are many potential outcome measures available, which are either specific to a symptom or disease, or which are more general (Table 34.1).[103] General QOL instruments can assist in understanding the overall QOL of a patient, as well as indicating basic functional status. Disease-or symptom-specific tools help to understand individual palliative care problems, so one can potentially understand the effects of treatments. Using QOL instruments, one can also attempt to quantify symptoms. One issue may be that instruments are not necessarily validated for patients with advanced disease. Therefore, in addition to using a validated questionnaire,

Table 34.1. Taxonomy of QOL Instruments[104]

Generic instruments	Health profiles Preference-based measures
Specific instruments	Disease-specific (e.g. diabetes) Population-specific (e.g. frail elderly) Function-specific (e.g. sexual functioning) Condition- or problem-specific (e.g. pain)

one must also specifically target content valid instruments. This requires extensive knowledge of QOL issues that are relevant to the trial. One example of a QOL problem in the palliative setting is cognitive failure. There are at least 10 instruments that have been used in this setting, most commonly the Mini-Mental State Exam (MMSE).[105] In addition to knowing which instrument to use, one must consider who will administer the exam. For example, related to cognitive failure, the MMSE and Bedside Confusion Scale can be given by non-clinicians, which may be an advantage in many studies.

One must also understand that as disease progresses, a patient's ability to provide data often diminishes. Therefore, proxy respondents may be necessary. Proxies may be very helpful in providing objective data, but are likely to be less useful for subjective outcomes such as pain and depression.[106] The nature of the relationship is also important, as well as other cultural factors. Also, other defined measures can be utilised as surrogate markers for QOL. For example, time out of the hospital, ability to eat food, or need for a nasogastric tube may all be used as measures for clinical benefit for malignant bowel obstruction. But without fully understanding the needs and goals of the patient, the wrong outcome measure could be used. Therefore, in the setting of palliative care, it may be reasonable to consider individualising measures for each patient. In theory, one might develop a flexible composite outcome measure that indicates whether a patient's goals were met. This approach has the advantage of being highly sensitive to subjective endpoints and can accurately determine whether a particular patient benefited from an intervention. In a sense, this approach is similar to global ratings of change used in other research fields. The problem, however, is that each patient in a trial may have a somewhat different composite measure.

QUALITATIVE RESEARCH

Qualitative medical research utilises personal interview[107–109] and focus group interview[109,110] of patients and families to better understand patients' experiences and opinions. These methods include a rich variety of data collection, analysis procedures, and standards for validity and reliability.[111] These types of studies help create an understanding about what disease means and how it can change throughout its course. Ideas that may not have been considered by the research team, or common themes, can then be further explored to better understand causality and potential interventions for patients facing the EOL. More generally, mixed method research that includes both quantitative and qualitative methods offers a promising way to supplement a study's main outcomes with qualitative data.

OBSERVATIONAL STUDIES

Observational data fall into the categories of descriptive studies and analytic or comparative studies. Descriptive studies include case reports, case series and cross-sectional studies. These are simply reports based on description of a disease process according to patient characteristics. Comparative studies attempt to compare treatments, against either retrospective controls or prospective groups of subjects. These include case–control and cohort studies. Observational studies are useful in generating hypotheses, baseline data for randomised studies, and provide the rationale for sample size estimation; they may also be of great use in understanding the natural history of disease progression. In fact, a major problem in the design of prospective palliative care studies is a dearth of prospective observational data to appropriately understand the natural course of disease, especially as it related to important outcome measures.

While there may be an abundance of retrospective trials published, they may give little insight into what occurs throughout the course of a palliative care problem and how things change from the patient's perspective. Good attempts can be made with case-controlled studies to indicate best treatments. One example is with the common palliative care problem of gastric outlet obstruction. There are three basic treatment approaches available. Historically, an open surgical bypass was the standard of care. Recently, minimally invasive laparoscopic as swell as endoscopic stenting approaches are becoming available in many institutions. A recent case–control study matched these three treatment approaches over a 10-year period of time.[112] There were 16 patients in the open surgical and endoscopic approaches, and 14 patients in the laparoscopic arm. The study found that patients undergoing an endoscopic procedure fared better in QOL surrogate markers such as complications, time to eating, and length of stay after procedure. This type of study does allow practitioners to make treatment-related decisions. While prospective data may allow for a richer database with alternative patient–centred outcomes, this type of study does add to the literature, and in a very difficult population may obviate the need for some RCTs. In fact, as RCTs in this population of patients may be quite difficult, prospective observational studies are an important future method of research for patients facing the EOL. Through this method, one may be able to deduce the better treatment approach, but more importantly one can better understand the natural history of disease and what this means to the patient and family in a longitudinal fashion.

PHASE I STUDIES

Studies that focus on toxicity in a population facing the EOL have great limitations. While it is rare to have responses on a Phase I trial, the patient does have at least some potential to improve survival or even possibly cure. While QOL is not the major goal of treatment, these outcomes may become the primary outcome for future studies, as with Gemcitabine. Medications whose sole purpose is QOL, such as anti-emetics, need to undergo the same dosing and toxicity studies, but it is difficult to initiate them in patients who are facing the EOL, and are more likely to be tested in a healthier population. While outcomes may then be assumed to be useful for sicker patients, this may not always hold true.

PHASE II STUDIES

Phase II studies focus on safety and efficacy, which are imperative in the palliative setting. In fact, when the focus is on QOL and not survival, ensuring that treatments have as little a chance as possible to lead to a poorer QOL is especially important. Unfortunately, Phase II trials still frequently focus on survival instead of other more relevant outcomes. It is reasonable to report response rates and survival data, as this is information that is important to patients, families and practitioners.

One example of a treatment that has been extensively studied in the Phase II palliative setting is Gemcitabine, with or without some form of combinational therapy, for pancreatic adenocarcinoma.[16–19] As a rule, these studies report response rates and survival data, but they also make attempts to examine other QOL-related measures. For example, one study examined bi-weekly Gemcitabine in 43 patients in this setting, and reported response rate (21%), time to progression (5.3 months), median survival (8.8 months), and probability of surviving beyond 12 months (26.3%).[26] Of the 43 patients, 36 did not have symptoms. The study notes a symptom response rate of 44%, using three different outcome measures: pain, Karnofsky performance status, and weight. This is noted to be the common set of outcome measures for these trials. Unfortunately, as with many studies in this population, the authors did not control well for the multiple other supportive treatments that were surely being utilised, such as appetite stimulants, anti-emetics, alterations in pain medications, or other interventions. Pain was measured based on improvement of pain using a visual analogue scale or a decrease in analgesic medications. Karnofsky score was utilised as a surrogate for QOL, which may not be relevant to many patients, especially based on their mental status. There was no clear primary outcome measure to display benefit, but an attempt to add the three outcomes chosen to come to an overall clinical benefit. Still, this study was a laudable attempt to examine a true palliative benefit of chemotherapy.

PHASE III STUDIES

It is imperative to initiate RCTs for patients facing the EOL. As long as the RCT is the standard by which effectiveness is judged, the field whose interventions have not been proven by this test is at risk of being relegated to second-class status in the medical hierarchy.[113] As noted, there are many obstacles to overcome to initiate Phase III studies for palliative care problems. The problem must be common enough to get enough patients to enroll to adequately power the study. Even if the problem is relatively common, it still may be too rare to get sufficient numbers of patients, especially with the innate difficulties in patient accrual. Therefore, these studies must be accomplished through cooperative groups. While cooperative groups have historically not embraced such trials, this must change. Palliative care problems must have equipoise in treatment options in the minds of research teams, as well as potential referral sources. There should be sufficient background data to establish meaningful outcome measures. In all, while there are many hurdles to overcome, and there are relatively few well-controlled Phase III studies in the palliative care literature, researchers must still strive for this type of study as a standard to ensure the best treatment options are available to their patients.

FUTURE DIRECTIONS

Clearly, the future for palliative care trials mandates the full compendium of research to ensure the best treatments are available. There is great hope that research for palliative problems will expand in the future, as there is more of a national and international focus on patients facing the EOL.

REFERENCES

1. Krouse RS, Rosenfeld KE, Grant M, *et al.* Palliative Care Research: Issues and Opportunities. *Cancer Epidemiol Biomarkers Prev* (2004) **13(3)**: 337–9.
2. Mount B, Cohen R, MacDonald N, Bruera E, Dudgeon D. Ethical issues in palliative care research revisited. *Palliat Med* (1995) **9**: 165–70.
3. Department of Health and Human Services. Protection of Human Subjects. Title 45 Part 46: Revised. *Code of Federal Regulation* (1991).
4. Corner J. Is there a research paradigm for palliative care? *Palliat Med* (1996) **10**: 201–8.
5. Dhaliwal HS, Sloan P, Arkinstall WW *et al.* Randomized evaluation of controlled-release codeine and placebo in chronic cancer pain. *J Pain Symptom Manage* (1995) **10**(8): 612–23.
6. Farrar JT, Cleary J, Rauck R, Busch M, Nordbrock E. Oral transmucosal fentanyl citrate: randomized, double-blinded, placebo-controlled trial for treatment of breakthrough pain in cancer patients. *J Natl Cancer Inst* (1998) **90**(8): 611–16.
7. Polati E, Finco G, Gottin L, Bassi C, Pederzoli P, Ischia S. Prospective randomized double-blind trial of neurolytic coeliac plexus block in patients with pancreatic cancer. *Br J Surg* (1998) **85**: 199–201.
8. Casarett DJ, Karlawish J, Hirschman KB. Are hospices ready to participate in palliative care research? Results of a nationwide survey. *J Palliat Med* (2002) **5**(3): 397–406.
9. Kaasa S, De Conno F. Palliative care research. *Eur J Cancer* (2001) **37**: S153–9.
10. Annas GJ. *Some choice: law, medicine, and the market*. New York: Oxford University Press (1998).
11. Clark D, Ingleton C, Seymour J. Support and supervision in palliative care research. *Palliat Med* (2000) **14**: 441–6.
12. Kagawa-Singer M. Negotiating cross-cultural issues at the end of life: 'You got to go where he lives'. *JAMA* (2001) **286**(23): 2993–3001.

13. Billings JA, Pantilat S. Survey of palliative care programs in United States teaching hospitals. *J Palliat Med* (2001) **4**(3): 309–14.
14. Lamont EB, Christakis NA. Prognostic disclosure to patients with cancer near the end of life. *Ann of Intern Med* (2001) **134**(12): 1096–105.
15. Miner TJ, Jaques DP, Tavaf-Motamen H, Shriver CD. Decision making on surgical palliation based on patient outcome data. *Am J Surg* (1999) **177**: 150–4.
16. Doyle C, Crump M, Pintilie M, Oza AM. Does palliative chemotherapy palliate? Evaluation of expectations, outcomes, and costs in women receiving chemotherapy for advanced ovarian cancer. *J Clin Oncol* (2001) **19**: 1266–74.
17. Ulrich-Pur H, Kornek G.V., Raderer M, Haider K, Kwasny W, Depisch D, Greul R, Schneeweiss B, Krauss G, Funovics J, Scheithauer W. A phase II trial of biweekly high dose gemcitabine for patients with metastatic pancreatic adenocarcinoma. *Cancer* (2000) **88**(11): 2505–11.
18. Feliu J, Mel R, Borrega P, Lopez Gomez L, Escudero P, Dorta J, Castro J, Vazquez-Esteves SE, Bolanos M, Espinosa E, Gonzalez Baron M. Phase II study of a fixed dose-rate infusion of gemcitabine associated with uracil/tegafur in advanced carcinoma of the pancreas. *Ann Oncol* (2002) **13**: 1756–62.
19. Rothenberg ML, Moore MJ, Cripps MC *et al*. A phase II trial of gemcitabine in patients with 5-FU-refractory pancreas cancer. *Ann Oncol* (1996) **7**: 347–53.
20. Matano E, Tagliaferri P, Libroia A, Damiano V, Fabbrocini A, De Lorenzo S, Bianco AR. Gencitabine combined with continuous infusion 5-fluorouracil in advanced and symptomatic pancreatic cancer: a clinical benefit-oriented phase II study. *Br J Cancer* (2000) **82**(11): 1772–5.
21. Burris HA III, Moore MJ, Andersen J, Green MR, Rothenberg ML, Modiano MR, Cripps MC, Portenoy RK, Storniolo AM, Tarassoff P, Nelson R, Dorr FA, Stephens CD, Von Hoff DD. Improvements in survival and clinical benefit with Gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. *J Clin Oncol* (1997) **15**(6): 2403–13.
22. Colucci G, Giuliani F, Gebbia V, Biglietto M, Rabitti P, Uomo G, Cigolari S, Testa A, Maiello E, Lopez M. Gemcitabine alone or with Cisplatin for the treatment of patients with locally advanced and/or metastatic pancreatic carcinoma. *Cancer* (2002) **94**: 902–10.
23. Smith D, Gallagher N. A phase II/III study comparing intravenous ZD9331 with gemcitabine in patients with pancreatic cancer. *Eur J Cancer* (2003) **39**: 1377–83.
24. Bornman PC, Harries-Jones EP, Tobias R, Van stiegmann G, Terblanche J. Prospective controlled trial of transhepatic biliary endoprosthesis versus bypass surgery for incurable carcinoma of head of pancreas. *Lancet* (1986) **1**(8472): 69–71.
25. Smith AC, Dowsett JF, Russell RCG, Hatfield ARW, Cotton PB. Randomised trial of endoscopic stenting versus surgical bypass in malignant low bileduct obstruction. *Lancet* (1994) **344**: 1655–60.
26. Andersen JR, Sørensen SM, Kruse A, Rokkjær M, Matzen P. Randomised trial of endoscopic endoprosthesis versus operative bypass in malignant obstructive jaundice. *Gut* (1989) **30**: 1132–5.
27. de Raeve L. Ethical issues in palliative care research. *Palliat Med* (1994) **8**: 298–305.
28. Shepherd HA, Royle G, Ross APR, Diba A, Arthur M, Colin-Jones D. Endoscopic biliary endoprosthesis in the palliation of malignant obstruction of the distal common bile duct: a randomized trial. *Br J Surg* (1988) **75**: 1166–8.
29. Hilts PJ. VA hospital is told to halt all research. *New York Times.* March 25 (1999).
30. Freedman B. Equipoise and the ethics of clinical research. *New Engl J Med* (1987) **317**: 141–5.
31. Rutstein DR. The ethical design of human experiments. In: Freund PA, ed., *Experimentation with Human Subjects*. New York: George Braziller (1970) 383–401.
32. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? *JAMA* (2000) **283**: 2701–11.
33. Brody BA. World Medical Association, Declaration of Helsinki. *The Ethics of Biomedical Research. An International Perspective*. New York, Oxford University Press (1998).
34. The Nuremberg Code. Reprinted in: Brody B. *The Ethics of Biomedical Research. An International Perspective*. New York: Oxford University Press (1947) 213.
35. Freedman B. Placebo-controlled trials and the logic of clinical purpose. *IRB: A review of human subjects research* (1990) **12**: 1–6.
36. Meinert CL. *Clinical Trials. Design, conduct, and analysis*. Oxford: Oxford University Press (1986).
37. Moore RA, Gavaghan D, Tramer MR, Collins SL, McQuay HJ. Size is everything–large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. *Pain* (1998) **78**: 209–16.
38. Du Pen SL, Du Pen AR, Polissar N *et al*. Implementing guidelines for cancer pain management: results of a randomized controlled clinical trial. *J Clin Oncol* (1999) **17**(1): 361–70.
39. Ahmedzai S, Brooks D. Transdermal fentanyl versus sustained-release oral morphine in cancer

pain: preference, efficacy, and quality of life. The TTS-Fentanyl Comparative Trial Group. *J Pain Symptom Manage* (1997) **13**(5): 254–61.

40. Wong JO, Chiu GL, Tsao CJ *et al*. Comparison of oral controlled-release morphine with transdermal fentanyl in terminal cancer pain [published erratum appears in *Acta Anaesthesiol Sin* (1997) **35**(3): 191]. *Acta Anaesthesiol Sin* (1997) **35**(1): 25–32.
41. Freedman B. Scientific value and validity as ethical requirements for research: a proposed explication. *IRB* (1987) **9**: 7–10.
42. Ingham J, Seidman A, Yao TJ, Lepore J, Portenoy R. An exploratory study of frequent pain measurement in a cancer clinical trial. *Qual Life Res* (1996) **5**(5): 503–7.
43. Twycross R, Harcourt J, Bergl S. A survey of pain in patients with advanced cancer. *J Pain Symptom Manage* (1996) **12**(5): 273–82.
44. Bernabei R, Gambassi G, Lapane K *et al*. Management of pain in elderly patients with cancer. *JAMA* (1998) **279**: 1877–82.
45. Parmelee PA, Smith B, Katz IR. Pain complaints and cognitive status among elderly institution residents. *J Am Geriatr Soc* (1993) **41**: 517–22.
46. McCracken LM, Klock PA, Mingay DJ *et al*. Assessment of satisfaction with treatment for chronic pain. *J Pain Symptom Manage* (1997) **14**(5): 292–9.
47. Ward SE, Gordon DB. Patient satisfaction and pain severity as outcomes in pain management: a longitudinal view of one setting's experience. *J Pain Symptom Manage* (1996) **11**(4): 242–51.
48. Desbiens NA, Wu AW, Broste SK *et al*. Pain and satisfaction with pain control in seriously ill hospitalized adults: findings from the SUPPORT research investigations. For the SUPPORT investigators. Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment. *Crit Care Med* (1996) **24**(12): 1953–61.
49. Payne R, Mathias SD, Pasta DJ, Wanke LA, Williams R, Mahmoud R. Quality of life and cancer pain: satisfaction and side effects with transdermal fentanyl versus oral morphine. *J Clin Oncol* (1998) **16**(4): 1588–93.
50. Pellino TA, Ward SE. Perceived control mediates the relationship between pain severity and patient satisfaction. *J Pain Symptom Manage* (1998) **15**: 110–16.
51. Derogatis LR, Morrow GR, Fetting J *et al*. The prevalence of psychiatric disorders among cancer patients. *JAMA* (1983) **249**: 751–7.
52. Kathol RG, Mutgi A, Williams J, Clamon G, Noyes R. Diagnosis of depression in cancer patients according to four sets of criteria. *Am J Psychiat* (1990) **147**: 1021–4.
53. Heiskanen T, Kalso E. Controlled-release oxycodone and morphine in cancer related pain. *Pain* (1997) **73**(1): 37–45.
54. Deschamps M, Band PR, Hislop G, Rusthoven J, Iscoe N, Warr D. The evaluation of analgesic effects in cancer patients as exemplified by a double-blind, crossover study of immediate-release versus controlled-release morphine. *J Pain Symptom Manage* (1992) **7**: 384–92.
55. Yalcin S, Gullu IH, Tekuzman G *et al*. A comparison of two nonsteroidal antiinflammatory drugs (diflunisal versus dipyrone) in the treatment of moderate to severe cancer pain: a randomized crossover study. *Am J Clin Oncol* (1998) **21**(2): 185–8.
56. Freedman B, Fuks A, Weijer C. Demarcating research and treatment: a systematic approach for the analysis of the ethics of clinical research. *Clin Res* (1992) **40**: 653–60.
57. Kirkham SR, Abel J. Placebo-controlled trials in palliative care: the argument against. *Palliat Med* (1997) **11**(6): 489–92.
58. Hardy JR. Placebo-controlled trials in palliative care: the argument for. *Palliat Med* (1997) **11**(5): 415–18.
59. Macklin R. The ethical problems with sham surgery in clinical research. *New Engl J Med* (1999) **341**: 992–6.
60. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *New Engl J Med* (1994) **331**: 394–8.
61. Temple RT, Ellenberg SS. Placebo-controlled trials and active control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. *Ann Intern Med* (2000) **133**: 455–63.
62. Shapiro AK, Shapiro E. The placebo: is it much ado about nothing? In: Harrington A, ed., *The Placebo Effect*. Cambridge, MA: Harvard University Press (1998).
63. *World Medical Association International Code of Medical Ethics; amended by the 35th World Medical Assembly*, Venice (2000).
64. Broomhead A, Kerr R, Tester W *et al*. Comparison of a once-a-day sustained-release morphine formulation with standard oral morphine treatment for cancer pain. *J Pain Symptom Manage* (1997) **14**(2): 63–73.
65. Maxon HRd, Schroder LE, Hertzberg VS *et al*. Rhenium-186(Sn)HEDP for treatment of painful osseous metastases: results of a double-blind crossover comparison with placebo. *J Nucl Med* (1991) **32**(10): 1877–81.
66. Silverman DG, O'Connor TZ, Brull SJ. Integrated assessment of pain scores and rescue morphine use during studies of analgesic efficacy. *Anesth Analg* (1993) **77**: 168–70.

67. Haak van der Lely F, Burm AG, van Kleef JW *et al*. The effect of epidural administration of alfentanil on intra-operative intravenous alfentanil requirements during nitrous oxide-oxygen-alfentanil anaesthesia for lower abdominal surgery. *Anaesthesia* (1994) **49**(12): 1034–8.
68. Hahn AF, Bolton CF, Pillay N *et al*. Plasma-exchange therapy in chronic inflammatory demyelinating polyneuropathy. A double-blind, sham-controlled, cross-over study. *Brain* (1996) **119**(Pt 4): 1055–66.
69. Bruera E. Ethical issues in palliative care research. *J Palliat Care* (1994) **10**: 7–9.
70. Family Caregiving: Agenda for Action, Improving Services and Support for America's Family Caregivers. National Health Council, Washington, DC (1999).
71. Steele RG, Fitch MI. Needs of family caregivers of patients receiving home hospice care for cancer. *Oncol Nurs Forum* (1996) **23**: 823–28.
72. Emanuel EJ, Fairclough DL, Slutsman J, Alpert H, Baldwin D, Emanuel LL. Assistance from family members, friends, paid care givers, and volunteers in the care of terminally ill patients. *New Engl J Med* (1999) **341**: 956–63.
73. Grisso T, Appelbaum PS. *Assessing Competence to Consent to Treatment*. New York: Oxford University Press (1998).
74. Marson DC, Schmitt FA, Ingram KK, Harrell LE. Determining the competency of Alzheimer patients to consent to treatment and research. *Alzheimer Dis Assoc Disord* (1994) **8**(Suppl 4): 5–18.
75. Elliott C. Caring about risks: are severely depressed patients competent to consent to research? *Arch Gen Psychol* (1997) **54**: 113–16.
76. Benson P, Roth L, Winslade WJ. Informed consent in psychiatric research: preliminary findings from an ongoing investigation. *Soc Sci Med* (1985) **20**: 1331–41.
77. Lemaire F, Blanch L, Cohen SL *et al*. Informed consent for research purposes in intensive care patients in Europe–part II. An official statement of the European Society of Intensive Care Medicine. Working Group on Ethics. *Intensive Care Med* (1997) **23**(4): 435–9.
78. Weinrich S, Sarna L. Delirium in the older patient with cancer. *Cancer* (1994) **74**: 2079–91.
79. Breitbart W, Bruera E, Chochinov H, Lynch M. Neuropsychiatric syndromes and psychological symptoms in patients with advanced cancer. *J Pain Symptom Manage* (1995) **10**(2): 131–41.
80. Pereira J, Hanson J, Bruera E. The frequency and clinical course of cognitive impairment in patients with terminal cancer. *Cancer* (1997) **79**(4): 835–42.
81. Massie MJ, Holland J, Glass E. *et al*. Delirium in terminally ill cancer patients. *Am J Psychiat* (1983) **140**: 1048–50.
82. Bruera E, Franco JJ, Maltoni M, Watanabe S, Suarez-Almazor M. Changing pattern of agitated impaired mental status in patients with advanced cancer: association with cognitive monitoring, hydration, and opioid rotation. *J Pain Symptom Manage* (1995) **10**(4): 287–91.
83. Bruera E, MacMillan K, Hanson J *et al*. The cognitive effects of the administration of narcotics. *Pain* (1989) **39**: 13–16.
84. Stiefel FC, Breitbart W, Holland JC. Corticosteroids in cancer: neuropsychiatric complications. *Cancer Invest* (1989) **7**: 479–91.
85. Breitbart W, Marotta R, Platt MM *et al*. A double-blind trial of haloperidol, chlorpromazine, and lorazepam in the treatment of delirium in hospitalized AIDS patients. *Am J Psychiat* (1996) **153**(2): 231–7.
86. Chong SF, Bretscher ME, Mailliard JA *et al*. Pilot study evaluating local anesthetics administered systemically for treatment of pain in patients with advanced cancer. *J Pain Symptom Manage* (1997) **13**(2): 112–17.
87. Brown J H, Henteleff P, Barakat S *et al*. Is it normal for terminally ill patients to desire death? *Am J Psychiat* (1986) **143**: 208–11.
88. Morris JN, Suissa S, Sherwood S *et al*. (1986). Last days: a study of the quality of life of terminally ill cancer patients. *J Chron Dis* **39**: 47–62.
89. Massie MJ, Holland JC. Depression and the cancer patient. *J Clin Psychiat* (1990) **51**: 12–17.
90. Eisenach JC, DuPen S, Dubois M, Miguel R, Allin D. Epidural clonidine analgesia for intractable cancer pain. The Epidural Clonidine Study Group. *Pain* (1995) **61**(3): 391–9.
91. Pappas GD, Lazorthes Y, Bes JC *et al*. Relief of intractable cancer pain by human chromaffin cell transplants: experience at two medical centers. *Neurol Res* (1997) **19**(1): 71–7.
92. Plancarte R, de Leon-Casasola OA, El-Helaly M *et al*. Neurolytic superior hypogastric plexus block for chronic pelvic pain associated with cancer. *Reg Anesth* (1997) **22**(6): 562–8.
93. Kristjanson LJ, Hanson EJ, Balneaves L. Research in Palliative Care populations: ethical issues. *J Palliat Care* (1994) **10**(3): 1010–15.
94. Dresser R. Bound to treatment: the Ulysses contract. *Hastings Cent Rep* (1984) **14**: 13–16.
95. Miller CK, O'Donnell DC, Searight HR, Barbarash RA. The Deaconess Informed Consent Comprehension Test: an assessment tool for clinical research subjects. *Pharmacotherapy* (1996) **16**(5): 872–8.

96. Penman DT, JHolland JC, Bahna GF *et al*. Informed consent for investigational chemotherapy: patients' and physicians' perceptions. *J Clin Onco* (1984) **2**: 849–855.
97. Tankanow RM, Sweet BV, Weiskopf JA. Patients' perceived understanding of informed consent in investigational drug studies. *Am J Hosp Pharm* (1992) **49**: 633–5.
98. Grisso T, Appelbaum PS. The MacArthur Treatment Competence Study III. *Law Hum Behav* (1995) **19**: 149–74.
99. Groenvold M. Methodological issues in the assessment of health-related quality of life in palliative care trials. *Acta Anaesthesiol Scand* (1999) **43**: 948–53.
100. Julious SA, George S, Campbell MJ. Sample sizes for studies using the short form 36 (SF-36). *J Epidemiol Community Health* (1995) **49**: 642–4.
101. Julious SA, George S, Machin D, Stephens RJ. Sample sizes for randomized trials measuring quality of life in cancer patients. *Qual Life Res* (1997) **6**: 109–17.
102. Tang ST, McCorkle R. Appropriate time frames for data collection in quality of life research among cancer patients at the end of life. *Qual Life Res* (2002) **11**: 145–55.
103. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. *Pharmacoeconomics* (2000) **1**: 13–35.
104. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* (1993) **118**: 622–9.
105. Hjermstad MJ, Loge JH, Kaasa S. Methods for assessment of cognitive failure and delirium in palliative care patients: implications for practice and research. *Palliat Med* (2004) **18**: 494–506.
106. McPherson CJ, Addington-Hall JM. Judging the quality of care at the end of life: can proxies provide reliable information? *Soc Sci Med* (2003) **56**: 95–109.
107. Bauman LJ, Adair EG. The use of ethnographic interviewing to inform questionnaire construction. *Health Educ Q* (1992) **19**: 9–23.
108. McCracken G. *The Long Interview*. Thousand Oaks, CA: Sage (1988).
109. Merton RK, Fiske M, Kendall PL. *The Focused Interview: A Manual of Problems and Procedures*, 2nd edn. New York: Free Press (1990).
110. Coreil J. Group interview methods in community health research. *Med Anthropol* (1995) **16**: 193–210.
111. Miles MB, Huberman AM. *Qualitative Data Analysis*. Thousand Oaks, CA: Sage (1994).
112. Mittal A, Windsor J, Woodfield J, Casey P. Lane M. Matched study of three methods for palliation of malignant pyloroduodenal obstruction. *Br J Surg* (2004) **91**: 205–9.
113. Casarett D. 'Randomize the first patient': old advice for a new field. *AAHPM Bull* (2002) **2**(Spring): 4–5.

35

Complementary Medicine

PING-CHUNG LEUNG
Institute of Chinese Medicine, The Chinese University of Hong Kong, Hong Kong

INTRODUCTION

Medicine has been an art of healing. Although there is no complete account of its history of development in the prehistorical and extremely primitive days, it must be closely related to the very ancient people's eating habits and their observations of animal behaviour. Ancient people fallen sick must prefer light meals with plenty of drinks. The latter might mean fruits and plant-related products, which are the forerunners of medicinal herbs.

Ancient people lived with animals, keeping them either as domestic friends or as meat providers. Animal instincts and behaviour lent the ancient people much wisdom of healing. As dogs ate up special grasses and leaves when they fell sick, followed by vomiting or diarrhoea, sometimes bringing out special unwanted ingested food or worms, the ancient people noted the special grasses and leaves. When they desired to clear their guts under difficult circumstances, they recalled those grasses and leaves and hence imitated the animals, hoping to achieve the same remedy. In this way, the primitive art of healing started.

What followed must have been more and more observations on more and more grasses and leaves which became considered as "herbs". Taking herbs as a means to remove symptoms and ailments is, therefore, the standard early stage of the healing art in human history.[1,2] The valuable observations and experiences were kept until today.

All primitive tribal populations today still use herb treatment, as the standard popular method of healing. The practice does not rule out trial uses of new herbs and their combinations, but mostly depends on past experience and documentations. These early clinical trials were not the result of imagination but initiated after observations on the anecdotal effects of different herbs.[3,4]

Traditionally there was no real need for large-scale clinical trials for complementary medicine. The need came only when scientific healers became interested in complementary medicine and started making use of herbs and other methods in their attempts to supplement modern medicine. They wanted to know whether, by utilising the same logic of analysis commonly practised in modern medicine, they could prove that herbal treatment constituted a logical substitution or supplement to scientific medicine.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green

This chapter explores the promises and fallacies of clinical trials in herbal medicine and acupuncture, and identifies the similarities and difficulties, the developments and limitations.

TYPES OF COMPLEMENTARY MEDICINE

The current mainstream of medicine is scientific variety. Other forms of health care outside the mainstream fall into the category of complementary or alternative medicine.

If one uses history as the criterion of identification and considers ancient medicine equivalent to complementary medicine, one sees four main systems of ancient healing. They are: Chinese, Indian, Ayurveda, Greek and Egyptian. Geographically, the four systems are separated and yet nearby areas do have similarities. China and India certainly did communicate, so did Greece and Egypt. China probably also obtained information from Greece, i.e. Europe, later in history through the 'silk-route'.[5]

The four different systems have two main unique features. The Greek and Egyptian systems concentrated on the use of single herbs, while the Chinese and Indian systems used multiple combinations. Combined formulae are most commonly prescribed in Chinese herbal medicine.

After thousands of years, the four ancient systems of medicine still survive well. Greek medicine in Europe has established itself as a homeopathic healing art, while the other three systems enjoy persistent but varying popularity.

In the modern sense, alternative/complementary medicine includes not only the herbal streams, but any other form of medicine that is unrelated to the modern scientific stream. When the American Medical Association did a survey in the United States aiming at the revelation of the popularity and users of alternative medicine, 17 modalities were targeted (Table 35.1).[6]

Table 35.1. Modalities of alternative medicine

1. Relaxation techniques
2. Herbal medicine
3. Massage
4. Chiropractice
5. Spiritual healing by others
6. Megavitamins
7. Self-help group
8. Imagery
9. Commercial diet
10. Folk remedies
11. Life-style diet
12. Energy healing
13. Homeopathy
14. Hypnosis
15. Biofeedback
16. Acupuncture
17. Self-prayer

Of these varieties, the one that commanded the highest popularity was acupuncture as a form of pain control.

The author cannot possibly be knowledgeable about all the varieties of complementary medicine and would not be able to discuss everything on their clinical trials. Rather, he would prefer to concentrate on the two varieties that he is familiar with, namely herbal medicine and acupuncture. Daily the discussion, examples of clinical trials will be presented, based on his own personal interests and experience.

FUNDAMENTAL CONSIDERATIONS FOR CLINICAL TRIALS ON CHINESE MEDICINE

How should clinical trials of Chinese medicine be conducted? Are there differences between such trials and others designed for modern medicine?

We have explained earlier that, originally, complementary medicine and its practitioners did not demand clinical trials. However, clinical trials are indicated for modern scientists because once the efficacy is proven, an alternative methodology of treatment can be endorsed.[7]

If modern medicine were not totally successful, there would be a real need for supplementing with alternative medicine. Generally speaking, the success of modern medicine is well known in most areas. It is therefore necessary to look to complementary medicine only in those areas where the scientific mainstream encounter deficiencies.

WHERE ARE THE DEFICIENCIES?

The deficient areas lie where modern medicine, in spite of recent advances, fails to get good solutions.

Modern medicine has developed from the logic of modern science which follows the deductive approach. The problem is first thoroughly understood by identifying the cause. The cause can then be removed by working out an effective means. In the situation of a disease, when the cause is simple and straightforward, removing it is easy. On the other hand, when the cause is complicated, not well understood or multiple, removal becomes difficult or impossible. Examples of simple disease inducing causes include straightforward infections and hormonal deficiencies. The former is easily tackled with an efficient antibiotic while the latter could be treated with hormonal replacement.

When the causative agent is not thoroughly known, e.g. viral infections, treatment becomes difficult.

When the cause is complicated, e.g. in allergic conditions, treatment does not guarantee effective results.

When the cause is complicated, e.g. involving many factors like physiological, social and psychological aspects, modern scientific medicine becomes obviously deficient or incapable.[8–10]

Therefore the deficient areas in modern medicine that deserve contributions from complementary medicine include a number of specific areas (Table 35.2).

Table 35.2. Specific areas in modern medicine that could benefit from alternative medicine

1. Allergic conditions
2. Autoimmune diseases
3. Cancers
4. Chronic pain
5. Chronic derangements
6. Degenerative diseases
7. Nerve damage
8. Viral infections
9. Other areas where modern conventional therapy fails.

INDICATIONS AND PHILOSOPHY OF APPLYING COMPLEMENTARY MEDICINE

Success of current medical treatment is judged by its effectiveness and the statistical chances of obtaining good results. Modern medicine has developed as direct corrective measures. Hence, when it is effective, the probability of repeatedly arriving at good results is very high. Unless it is not available, there is therefore no reason why modern medicine should not be endorsed as the primary mainline treatment.

Although there are still confident herbal practitioners who believe and declare that whatever modern scientific practitioners can do, they can substitute with other herbal remedies, the number who remain committed is getting less and less. Indeed, today, most herbal practitioners accept the role of functioning as supplementary or alternative healers in a combined effort of cure and care.[11]

In this context, complementary herbal treatment is seldom used as the only healing modality. Instead, it is often given as an adjuvant treatment, either together with the mainline or after completion of the mainline treatment. Users of herbal preparations, moreover, frequently look to a tonic supportive supplement, rather than a curative drug.

HOW DOES HERBAL MEDICINE REALLY WORK?

Traditionally the system of herbal medicine was built on the rich experience of herb users or herbalists, accumulated over more than 2000 years in China since the early Chinese culture. For some reason, while basic medical sciences, e.g. anatomy and physiology, developed gradually in European territories around the Renaissance period, Chinese healers never felt the need to explore the basic medical sciences. Without a sound knowledge of anatomy and physiology, i.e. biological structure and function of the human body, it would not be possible to explore abnormal structures and functions, i.e. pathology. Without understanding the pathology, it would not be possible to develop direct means of removing the pathology. Herbal practitioners therefore, try to heal, not by direct confrontation with the pathological problem, but by indirectly supporting individuals to overcome their own difficulties.[12,13]

HOW DO INDIVIDUALS OVERCOME THEIR OWN PATHOLOGICAL PROBLEMS?

First, by surviving harmful disturbances imposed by the pathological processes. Second, by supporting the unaffected organs and systems so that their proper functions can be maintained. Third, by preventing future pathological mishaps while the current problem is being solved.

The herbal practitioner has the means to suppress those symptoms that are manifestations of the pathology. Suppression of symptoms like cough, diarrhoea or dyspnoea helps the sick individual to survive.

While waiting for the pathological damage to heal naturally, the unaffected organs and systems need to be supported to maintain their efficient function, which in turn will support the overall function and metabolic harmony of the living individual.

Prevention in the modern biological sense frequently refers to an immunological mechanism through which the individual becomes more resistant to future attacks of similar pathological nature.

The main focus of disease management for Chinese medicine is often the control of adverse symptoms. The ultimate goal is maintaining the well-being of the biological system. Aetiological considerations are therefore not directed towards the actual cause of the disease (of which the herbal expert has no idea) but a general conceptual state of the biological balance of the human bodily functions. The ancient healers correlated this conceptual state with the Taoist philosophy and imagined that bodily function was kept in a balanced state between the Yin and Yang (i.e. negative and positive). Any loss of balance led to ailment and disease.

The aim of treatment is therefore to restore the balance. The Yin and Yang include other contrasting opposing forces like cool and heat, superficial and deep, emptiness and solid. The causes of imbalance could be traced to a lack of balance of any pair of opposing forces. In the actual treatment, therefore, all efforts are spent on maintaining balance, by a supplement of the deficient force, or a decrement of the excessive one.

Since the pathological causes of the symptomatology are unclear to the herbal expert, he or she would need to observe the changes of symptoms and adjust the day-to-day protocol accordingly. This approach differs very much from conventional modern medicine which successfully identifies a pathological cause of disease, chooses a method of cure with a good chance of success, then administers it with all effort and persists with the commitment, until total removal of the pathology is achieved.

While the aetiology, epidemiology and natural course of a disease affect the design of clinical trials for modern medicine, it is now clear that in Chinese medicine, there is little analogy of aetiological and epidemiological considerations. The course of events in a disease, for a herbal expert, is the appearance of the symptoms: the loss of biological well-being due to the lack of balance between the vital forces. The aim of treatment is the re-establishment of balance; once balance is re-established, either naturally or through herbal intervention, well-being will be re-established. Treatment consists of a dynamic application of symptomatic relief with the goal of re-establishing the balance.[14]

Clinical trials for Chinese medicine or herbal medicine, therefore, could follow the line of thought for scientific planning on data collection and subsequent data meta-analysis. However, the pre-treatment data would be confined mainly to symptomatology. Other parameters, though carrying little weight for the herbal expert, could still be included for more scientific knowledge in clinical trials.

GENERAL CONSIDERATIONS FOR CLINICAL TRIALS ON HERBS

In the modern scientific world, only up to date methodology should be adopted. The set of common methodology for conducting clinical trials in modern medicine has been logical, useful and has made wonderful contributions to the clinical testing of new drugs and new methods of clinical treatment. The proper analysis of data and the use of statistics have revealed the trustworthiness of certain accumulated experience, as well as, the fallacies of some well-accepted and widely practised methods.[15]

The common methodology of random selection, blinding and placebo control, followed by statistical analysis, should be adopted. In the design of the trial, good clinical practice should be the aim. However, due to the nature of herbs, which have different origins and many different species, it is not uncommon to encounter situations where the basic principles cannot be strictly kept.[16] Until the day of good agricultural practice (GAP) arrives with maturity, herb supply remains uncertain.

THE OLD APPROACHES

Herbal experts fervently respect case reports and anecdotal reports, particularly when results appear promising. Of course the reason behind this is that they do not make use of statistics. Moreover, they believe that treatment results are different with different patients. Once good results are known to be possible, the expert can try to achieve equally good or even better results by wisely manipulating the varieties of treatment.

In this chapter, we do not endorse this traditional approach. We want to apply modern assessment tools for a better understanding of herbal or Chinese medicine treatment we do not want to degrade the value of anecdotal observations in ancient Chinese medicine. After all, the development of this system of healing depended solely on anecdotal analysis.

Good clinical practice insists that the prescribed drug for the clinical trial should be thoroughly known and uniform. However, using herbal preparations for clinical trials faces the difficulties of thorough technical knowledge and uniformity.

Pharmaceutical trials demand that details be known about the chemistry, the mode of action and metabolic pathways before clinical tests are conducted. What is the chemistry of specific herbs? What are the pathways of action and metabolic degradations? Are there adverse effects in the process of metabolism? A lot of work has been done in the past 50 years on the basic understanding and yet not much has come out. Each and every herb contains so much complicated chemistry that many years of research might not yield much fruit. Actually, at least 400 herbs are popular and possess records of therapeutic action and impressive efficacy. To demand thorough knowledge of just this popular selection of herbs is just not practical, not to mention the less commonly used extra 1–2000 varieties.[17]

Uniformity is another difficult area. Strictly speaking, since herbs are agricultural products, uniformity should start with the sites of agricultural production. The sites of production have different weather, different soil contents and the ways of planting are also different. At the moment, there may be over 50% of popular Chinese herbs produced on special farms in China. However, these farms are scattered over different provinces, which have widely different climates and soil environments. Good agricultural practice demands that environmental and nurturing procedures be uniformly ensured. Procedures include soil care, watering, fertilisers, pest prevention harvests, and storage. When such procedures are not uniform and there are no means to ensure common practice, good agricultural practice is not possible.

Not only is there a lack of uniformity in the mode of herb production, but different species of the same herb are found or planted in different regions and provinces. These different species have different chemical contents. Herbal experts have extensive experience and knowledge of some special correlations between the effectiveness of particular herbs and their sites of production. Some commonly used herbs are even labelled jointly with the best sites of production. With the development of molecular biology, coupled with modern means of assessment for active ingredients within a chemical product, species-specific criteria can be identified, using the DNA ‘finger-printing’ technique. Uniformity today should therefore include screening using both chemical and molecular ‘finger-printing’ techniques.

When we consider the other 50% of herbs that are only available from the wilderness, i.e. around mountains, highlands or swamps, and cannot be grown on agricultural farms, the insistence on product uniformity becomes even more difficult.

Putting together what we have discussed so far, to insist strictly on good clinical practice in clinical trials for herbal medicine is largely impossible. We have to accept a compromise. Indeed, in the past 50 years, many attempts have been made in comprehensive analytical studies of herbal preparations. The intention was: submit the herb to processes of extraction, analyse the important ingredients, then try to work out the chemical equation which could account for the clinical effects.

Extraction eliminates the useless and concentrates the effective components, which not only cuts down the volume of herbs used, but also intensifies the biological action. Knowing the actual effective ingredients and working out the chemical formulae would be ideal for modernisation of herbal preparations with the aim of converting the preparations into proper pharmaceuticals.

However, in spite of the efforts and resources put into herbal extractions and chemical analyses in the past 50 years, successful examples have not been impressive. The results of such efforts certainly do not match the resources put in.[18]

This unsatisfactory outcome has initiated a new approach. Instead of following the scientific pathway already taken by pharmaceuticals, which has shown more difficulties than promise, a more practical line has been endorsed. Since most, if not all, of the herbs have been used for hundreds of years, there should be a sufficient amount of reliability on the safety and efficacy of most of the herbs. The safety and efficacy are already well documented, but their practical utilisation in specific clinical circumstances needs to be further established. The traditional use of herbs had been focused on symptomatic control. Nowadays, the aim of clinical management is directed towards the curing of disease. We need to acquire an updated understanding of the effectiveness of the herbal preparations on different diseases. That is why we cannot be satisfied with records on efficacy alone but should start a series of clinical trials to further prove the efficacy of herbs.[19]

The National Institutes of Health of the United States have openly endorsed the approach of accepting traditional methods of healing as basically safe measures and then submitting them to proper clinical trials.[20] The recognition of acupuncture as a practical effective means of pain control started in 1998.[21] The subsequent formation of a special section devoted to research on complementary/alternative treatment followed. The National Centre for Complementary, Alternative Treatment (NCCAM) was properly formed and given a substantially large budget.

Clinical trials to be discussed within this chapter follow the efficacy-driven principle. They are planned strictly according to the principles set out under the modern methodology of clinical trials aiming at the production of objective evidences for the effectiveness of the methods used. It is, however, understood that product uniformity and quality cannot be absolutely guaranteed and that although GMP (Good Manufacturing Process) can be assured, GCP (Good Clinical Practice) cannot be absolutely ensured because of the lack of guarantee for any herbal preparation.

In our discussion full reference will be given to what is being recommended in China, which undoubtedly harbours most activities in Chinese medicine.

HERBAL DRUGS IN CHINA

In 1999 the National Bureau on Drug Control defined new drugs as ‘a manufactured product for medical treatment that is produced for the first time or an old product reproduced with different formulation and different indications’.

New drugs are divided into five categories

I. Group 1
   Artificial derivatives from Chinese herbs
   Newly discovered Chinese herbs and derivatives

Extracts from Chinese herbs and derivatives
Extracts from decoctions of herbs.

II. Group 2
Herbal injections
Herbal preparations processed inside animals
Extracts from complex decoctions.

III. Group 3
New preparations of decoctions
Combined herbal and chemical preparation
Imported herbal preparations.

IV. Group 4
Converted formulary
Cultivated herbal and domestic animal preparations.

V. Group 5
Herbal preparations with extended uses.

STAGES RECOMMENDED FOR HERBAL RESEARCH

The usual four stages are recommended:

Phase I Study of the general acceptance of the human being after consumption of the herbal preparation.

Normally Phase I refers to toxicity study. The code of practice given under 'Code of Practice for the Scrutiny of New Drugs' in China, however, recommends that the general well-being of the individual after consumption be observed.[22] The logic of skipping toxicity tests is probably based on an assumption that Chinese herbal preparations have been used safely for centuries, therefore a special toxic screening is not necessary. The author has strong reservations about this attitude and would recommend that toxicity clearance should remain the first phase of clinical trials.

Phase II Study of the safety and efficacy while working out the effective dosage.

Phase III Expand on the Phase II study, collecting more reliable confirmation on safety and efficacy.

Phase IV Further study of the safety and efficacy after the new drug is put on the market. More observations on adverse effects are expected.

It has been pointed out that, as far as herbal medicine is concerned, it is not unusual to find that correlation does not exist between laboratory research and clinical trials. When studies on the pharmacology, pharmacodynamics and pharmacokinetics are carried out after the clinical trials, positive values, in support of the clinical observations, might not be impressive.

The possible explanation of this observation may lie in the fact that the clinical consumption of herbal preparations involves multiple, complex, *in vivo* biological interactions, whereas laboratory tests consist of only simple unidirectional biological interactions. A multidirectional approach to the design of biological investigations should therefore be adopted.

The reverse could also be observed, i.e. impressive biological activities are not well matched with clinical trial observations. Clinicians facing the challenge of this dilemma should review the methodology adopted for the clinical trials and consider some reorganisation.

HOW DO CONCEPTS OF TRADITIONAL HEALING AFFECT CLINICAL TRIALS ON CHINESE MEDICINE?

Earlier in this chapter, the author mentioned the unique concepts in Chinese medicine, which are different from modern scientific medicine. The application of modern concepts in the area of clinical trials leads to an inevitable sacrifice of some of the fundamental principles of Chinese medicine practice. Experienced herbal experts, therefore, might not like to participate.

The following list includes the important concepts in Chinese medicine practice being sacrificed:

1. *Symptom and syndrome identification principle*. Following this principle, the herbal expert adjusts details of the treatment according to observations of the day-by-day changes in symptomatology. Different drugs may then be used for the same symptoms or the same drug used for different symptoms. Proper clinical trials can only use a uniform choice of treatment modality. This violates the symptom and syndrome identification principle.

2. *Holistic approach.* Chinese Medicine emphasises holistic care and holistic response, whereas clinical trials prefer objective, specific data as endpoints. The inclusion of specific data in herbal research probably does not invite objection from the herbal expert, as long as general data like different aspects of well-being, i.e. quality of life, are included. However, a highly specific endpoint does not have a strong Chinese medicine appeal.
3. *Response to pathological processes.* Chinese Medicine emphasises the response of healthy organs to disease. The ability of the healthy organs to respond to pathological changes ensures that the individual would be able to better resist adversities. Modern clinical trials aim mostly at diseased organs or specific pathological processes.
4. *Old system of clinical observation.* Herbal experts utilise a system of clinical observations which might be considered today as obsolete and over-subjective. This system of clinical signs includes tongue observation, pulse detection and a collection of subjective feelings.[23] Modern clinical trials insist on objective data that could be monitored. We therefore have to develop means to objectively assess the subjective signs in the tongue and the pulse or sacrifice the old system of observations. Herbal experts might not appreciate either choice.
5. *Strong tradition.* Herbal experts have genuine confidence in anecdotal observations and the experience of single patients. Insisting on the need to investigate collective observations and condemning single case experience would not be welcomed by herbal experts. This conceptual difference directly affects the participation and cooperation of traditional and modern experts.

While thoroughly recognising the unique nature of Chinese medicine and having pointed out the lack of harmony between the old tradition and modern science, one may realize that the current compromise adopted in China is to insist on a modern scientific approach as far as possible. Hence in standard textbooks in China, the following are advocated,[24] as standard instructions for clinical trials:

1. Use the principles of randomisation, blinding and repetition.
2. Adopt good protocols for clinical trials.
3. Avoid bias at all cost.
4. Eliminate chance factors.
5. Establish new standards of clinical assessment.
6. Establish unique outcome studies.
7. Establish unique quality of life assessments.
8. Insist on using modern statistics.

ADVERSE EFFECTS

Historically, great herbal masters in China in the ancient days did produce records of adverse effects and toxic problems with some herbs. As early as the Han dynasty (second century), documents were produced on herbs that needed to be utilised with great care.[25] This tradition was followed closely in subsequent centuries.[26]

More reports became available on methods and means with which toxicities and adverse effects could be reduced.[27] These included preparation techniques and special combinations of herbal choices.

In spite of good past experience, the prevalent belief is that Chinese medicine herbs are safe. On the other hand, more and more reports appeared on adverse effects and toxicities, and non-users of herbs tend to exaggerate the negative reports.

It must be pointed out that when new preparations come on to the market, the innovative processes of extraction and/or production might have produced or initiated new possibilities of adverse affects or toxicity. This experience is already well recorded in a number of modernised preparations, particularly those for injection.[28] Among the adverse effects, allergic reactions are commonest.

Todate, standard instructions on clinical trials for Chinese medicine define adverse drug reaction in exactly the same way as modern scientific clinical trials, and explanations on the reactions have been identically identified.[29]

Categories of adverse reactions include the following:

1. *Reactions to herbs*. Reactions are defined as harmful and unexpected effects while the standard dosages are used in certain drug trials. It is especially pointed out that for Chinese medicine, the harmful reactions could be due to the chemical nature of the herb or a poor choice of indication. These reactions do not include allergic responses.
2. *Dosage-related adverse effects*. Using an unnecessarily high dose could induce excessive effects, side effects or even toxic effects. Secondary effects like electrolyte imbalance might also be observed.
3. *Dosage-unrelated adverse effects*. These adverse effects could be the result of unfavourable preparation, contaminants in the herbs, sensitivity of the consumer, allergic reactions or specific inductive effects of the herb.
4. *Drugs interactions*. Classically, records are available in old Chinese Medicine literature on combined effects of herbs, their facilitatory and antagonistic effects. Nowadays, not only are drug interactions between herbs important, but possible interactions between herbs and commonly used pharmaceutical preparations are becoming issues of great concern since users of herbal preparations are increasing. In the area of anaesthesia, drug interactions between herbs and modern medicine could induce life-threatening reactions. Table 35.3 illustrates some studies currently done on this issue.[30]
5. *Delayed adverse effects*. Adverse effects of delayed nature include induction to cancer formation, foetal abnormalities and even blockage of bacterial sensitivities.
6. *Drug dependence*. There might be suspicions that herbal preparations might lead to drug dependence. Apart from a few opium-related herbs, Chinese herbs, in fact, are well known to be non-addictive because of their gross lack of specificities.

From the above account it might appear obvious that adverse effects in clinical trials using Chinese medicine in fact follow closely the experience encountered in other drug trials.

As far as the grading of adverse effects is concerned, it would be appropriate to categorise the effects as mild, moderate and severe.

With regard to the overall assessment of adverse effects, a convenient recommendation for Chinese Medicine trials is that of Naranjo.[31]

Naranjo's system of grading adverse drug reactions (ADRs) according to fact-finding results is shown in Table 35.4

The overall assessment is:

ADR confirmed	≥ 9
ADR likely	5–8
ADR possible	1–4
ADR unlikely	≤ 0

Detection and recording of adverse effects should bear different emphases at different phases of the trial, e.g. the Phase I trial aims at detection of adverse effects in relation to dosage, Phase II and III collect details, whereas Phase IV is concerned mainly with marketed drugs.

Whatever the situation, detection of adverse effects should include both clinical observations and laboratory data, and detection should be followed with follow-up observations. The summation of observations should be thoroughly analysed so that a comprehensive explanation of the adverse effects may be eventually worked out.

REPORTING OF ADVERSE EFFECTS

It is currently required in China that adverse effects should be reported to the relevant monitoring body as soon as possible. Once a drug is marketed, adverse effects should be continuously reported to the National Control Bureau, within the first five years.

Adverse effects detected at the post-market Phase IV might be particularly important for Chinese medicine trials. Since herbal preparations do not have clear, definite information about the effective contents of the herbs, bias and chance might be more likely than other trials on simpler

Table 35.3. A number of commonly-used medicinal herbs and their known interactions with some commonly-used drugs

Herb	Drug	Interaction	Mechanism
Radix Salviae Miltiorrhizae (Danshen)	Warfarin	Increased INR Prolonged PT/PTT	Danshen decreases elimination of Warfarin in rats
Radix Angelicae Sinensis (Danggui)	Warfarin	Increased INR and widespread bruising	Danggui contains coumarins
Ginseng (Radix Ginseng)	Alcohol	Increased alcohol clearance	Ginseng decreases the activity of alcohol dehydrogenase and aldehyde dehydrogenase in mice
Garlic	Warfarin	Increased INR	Post-operative bleeding and spontaneous spinal epidural haemorrhage
Herbal ephedrae (Ma Huang)	Pargyline, Isoniazid, Furazolidone	Headache, nausea, vomiting, bellyache, blood pressure increase	Pargyline, Isoniazid and Furazolidone interfere with the inactivation of noradrenalin and dopamine; ephedrine in herbal ephedrine can promote the release of noradrenalin and dopamine
Ginkgo Biloba	Aspirin	Spontaneous hyphema	Ginkgolides are potent inhibitors of PAF
Cornu cervi pantotrichum Fructus crataegi Radix polygoni multiflori	Adrenomimetic Levodopa Opium	Strengthens the effect of increasing blood pressure Increased blood pressure and heart rate Central excitation	Natural MAOIs in Cornu cervi pantotrichum, Fructus crataegi, and Radix polygoni multiflori inhibit the metabolism of adrenomimetic, levodopa and opium
Bitter melon	Chlorpropamide	Decreased urea glucose	Bitter melon decreases the concentration of blood glucose
Liquorice	Oral contraceptives	Hypertension, oedema, hypokalaemia	Oral contraceptive may increase sensitivity to glycyrrhizin acid
St John's wort	Warfarin Cyclosporin	Decreased INR Decreased concentration in serum	Decreases the activity of Warfarin
Radix Isatidis (Banlangen)	Trimethoprin (TMP)	Significantly increases anti-inflammation effect	
Liu Shen pill	Digoxin	Frequent ventricular premature beat	
Tamarind	Aspirin	Increases the bioavailability of aspirin	
Yohimbine	Tricyclic antidepressants	Hypertension	

Note: ACE: Gangiotension-converting enzyme
INR: international normalised ratio
PT: prothrombin time
PTT: partial thromboplastin time
PAF: platelet-activating factor
AUC: Garea under the concentration/time curve
MAOIs: monoamine oxidase inhibitors.

Table 35.4. Naranjo's system of grading adverse drug reactions according to fact-finding results

		Yes	No	Not clear
1.	Are there decisive records about the ADR?	+1	0	0
2.	Are the ADRs found after consumption of other drugs?	+2	−1	0
3.	Are the ADRs improved after consumption of antidote?	+1	0	0
4.	Are there repeated ADRs or repeated administration?	+2	−1	0
5.	Are there other predisposing factors?	−1	+2	0
6.	Are there ADRs after placebo?	−1	+1	0
7.	Has the blood level of drug giving ADRs been investigated?	+1	0	0
8.	Do the ADRs correlate to dosages?	+1	0	0
9.	Is there past history?	+1	0	0
10.	Is there objective proof	+1	0	0

drugs at the early phases. The large trial population during Phase IV gives a better chance of elimination of bias and allows a better opportunity of objective detection of adverse effects. During the Phase IV trial, the following aspects deserve particular attention:

1. Actual danger level of the adverse effects. the degree of danger of course depends on the incidence of occurrence. The requirement for treatment and the financial implications are also important.
2. More thorough studies at Phase IV should be considered according to epidemiological principles. Randomised controlled trials should be insisted on. Cohort studies might be convenient and useful, but need to have markedly obvious differences between series of comparisons before results could be instructive. Case reports might still be useful but might function as special warning signals to call for more serious studies.[32]

QUALITY OF LIFE

While clinical trials aim at a thorough scientific understanding of the effectiveness of specific forms of treatment, endpoints of measurement are set to give objective standards of evaluation. Primary endpoints are unique, focused, specific criteria which indicate the situation of the target against which the trial is directed. Changes of primary endpoints illustrate the efficacy directly. Secondary endpoints are supplementary criteria created to support observations on changes and efficacy. Secondary endpoints become more important when predictably, primary endpoints do not give clear-cut, impressive results. Secondary endpoints become more important when primary endpoints are expected to change slowly and are particularly important when chronic problems are being faced.

Since Chinese medicine, under most circumstances, does not operate via a direct, confrontational route but rather acts indirectly to support the healthy organs and helps to maintain vitality and prevent functional deterioration, critical and detailed assessment of the secondary endpoints is therefore of utmost importance.

Quality of life (QoL) is an important aspect in the assessment of care given to the chronically ill. QoL often measures the competency of the care and the ethical standard of the society in mental disorders and other disorders that demonstrate strong social orientations. Not infrequently, when technical endpoints are used as results of clinical trials, a reasonable outcome is observed, and yet patients might not be satisfied with their QoL. QoL is therefore multifocal: it differs between people in developed and developing areas; it also differs in different cultural circles.[33] Different countries and regions therefore try to develop their own data to be included within their own studies of QoL.[34] Meanwhile global, generally acceptable QoL charts are also being planned, examined and validated.[35]

Before an acceptable general data chart is ready, one has to accept the achievements already revealed in different fields. Generally speaking, QoL data sheets take in information about physiological well-being, psychological well-being, social well-being and the individual's subjective feeling on the treatment received and the rehabilitation underway. Different specialties and special areas of concern have created charts of their own and all these are valuable information when equivalent studies come up. Usually they are adopted right away or after validation. Hence there are instruments already developed for children and the elderly, and different medical specialties and subspecialities likewise have created their own charts. Just to mention a few, special QoL charts are available for the mentally ill, cardiovascular diseases, rheumatological disease respiratory problems gynaecological problems, and special infections.[36] QoL charts for Chinese medicine studies need to be developed.

WHAT ARE THE RECOMMENDATIONS FOR CLINICAL TRIALS OF CHINESE MEDICINE?

The simplest thing to do is to refer to the available charts in whichever clinical trial is being conducted and think about amendments to make them even more suitable.

ARE THERE UNIQUE FEATURES THAT NEED TO BE OBSERVED?

There are features related to health which are derived from the philosophy of Chinese medicine ever since its initial development. Chinese people at all walks of life are influenced by this philosophy without being aware of it, at all stages of their life. The belief that health depends on harmony between contrasting forces prompts the individual to feel either 'hot or cold', 'light or heavy' 'sick inside or sick outside'. After treatment, the feeling might remain, might reverse or might become balanced. No wonder the feeling is subjective, but in any clinical trial including the data of QoL, can one ignore the outcome of the philosophical guideline traditionally respected as being crucial for the whole system of the healing art?

Henceforth, it is obvious that QoL studies are particularly important for clinical trials of Chinese medicine and research should be done on special inclusions of data which are unique for it.

ARE CLINICAL TRIALS USING COMPLEMENTARY MEDICINE DIFFERENT FROM OTHER TRIALS?

Clinical trials are a scientific practice with established rules and regulations so that no matter which medical specialty we are addressing, the same basic approach needs to be followed. Complementary medicine, however, does not enjoy a history of scientific research and its system of trust is based on users' experience and practitioners' wisdom and honesty. While, in the modern world, both users and practitioners demand more objective evidence, the quickest way is to transfer the whole methodology system established for evidence-based medicine to testify to the validity of complementary medicine.

However, the basic philosophy of healing influences and determines the expectations of healing, the assessment of healing and evaluation of the final outcome. Modern medicine is built on a deductive, specific-target-orientated, confrontational, problem-solving philosophy. It is different from the non-specific, harmonising, holistic approach of complementary medicine. Expectedly, applying the same methodology to assess clinical effects cannot be perfect. Nevertheless, complementary medicine is only at its early stage of evaluation; applying a common methodology is convenient. With the accumulation of more experience, the methodology could be modified to better suit holistic medicine.

PHASING

The specific Phase I–IV trials are standard procedures created to assess the effectiveness of new drugs at different stages of maturity so that toxicity comes ahead of dosage and dosage ahead of efficacy, and even after marketing it is necessary to assess the drugs further.

Clinical trials for complementary medicine could follow the same logic. Checking a compound (mixture or extract) for safety, without knowing the exact chemical equation, is a difficult job. Therefore State 1 is going to be lengthy and expensive, although absolute accuracy cannot be expected until the biologically active chemical equation is identified. To be able to skip the Phase I trial for complementary medicine would be a significant shortcut to its research on clinical efficacy.

If one accepts the logic that single herbs or combinations of many herbs in classical formulae are not likely be unsafe for consumption, because thousands of people have taken them in the past and yet no serious adverse effects have been recorded, one must agree that the expensive and lengthy Stage 1 trial could be skipped. Indeed, after many years of gross scepticism of alternative/complementary medicine, NIH the adopted a practical approach to the research on Chinese medicine, by allowing Phase I to be skipped, when classic, conventional items or formulae are used, and when sufficient literature supports the assumption of safety.

Clinical trials on complementary medicine could therefore start with Phase II, provided that the herbs and mixtures used are common, frequently used ones, and that sufficient literature is available for reference, verifying that toxicity is not a problem.

When clinical trial starts with Phase II, efficacy is explored together with dosage. Classics in herbal medicine give recommendations on the choice of herbs and mixtures, and dosages are clearly given. Phase II therefore starts with the classically recommended dosage which is to be carefully verified, and Phase II study therefore actually intrudes into Phase III, where efficacy is the priority. Very often, after the completion of the initial trial, when efficacy is established, dosage verification needs serious reconsideration.

The history of clinical research on complementary medicine is so short that commercialisation after Phase III study is uncommon. The Institute of Chinese Medicine at the Chinese University of Hong Kong did have some experience of putting commercial herbal products in clinical trials. But such experience was not with Phase IV studies, because Phases II and III had not been completed. When regulations become more mature, more clear-cut phasing of clinical trials can be expected.

TRIAL DESIGN

Basic trial designs closely follow evidence-based medical practice. Randomisation and double blinding with placebo control form the backbone of clinical research.

PLACEBO

Placebo is a special problem. In affluent communities, the standard means of healing is modern medicine. Most of the time, patients turn to complementary/alternative medicine when modern conventional treatment has failed. Under such special circumstances, such a patient would demand a positive trial: making sure that the alternative method of healing could be positively tried. Refusal against placebo is therefore expected.

While most placebos are made with flour, sugar and oil, good experience with modern pharmaceuticals might not be transferable to alternative medicine. Thus an externally polished placebo tablet or capsule might look perfect, but it would lack the herbal smell. It is equally difficult to make identical placebo preparations with the same taste.

In communities where complementary medicine is commonly practiced, the suggestion of randomisation and that placebo might be given might also influence the registered candidates to secretly take their own alternative medicine in addition to the trial.

CROSSOVER AND PATIENT SATISFACTION

Facing difficulties in randomisation and placebo control, the arrangement of crossover becomes an important practical issue to consider. The assurance that, eventually, the herbal preparation under trial will be provided has the practical effect of gaining genuine support for randomisation and placebo control. A wash-out period is usually necessary before the crossover to the opposite treatment regime.

Very often, complementary/alternative medicine is used to treat chronic illnesses. Long-term treatment is required in the clinical trial. Review of clinical results, therefore, is slow and lengthy. After completion of the clinical trial, the registered clients would need to know what is required after the trial. Should they stop abruptly and just stop treatment? Should they shift to the other treatment? Should they continue with the alternative treatment? If they need to continue with the trial preparation, where and how do they get it? All these questions need to be answered at the start of the clinical trial, and the trial preparations need to be in place long before completion of the trial.

NATIONAL REGULATIONS THAT AFFECT THE DESIGN OF THE CLINICAL TRIAL

China has a regulation that herbal preparations put under trial need to be totally classical, i.e. exactly similar to Chinese medicine classics, or that the preparation is a commercial product on the market. Placebo control trials are not recommended. Instead, trials comparing the effects of a herbal preparation with standard treatment models, e.g. pharmaceutical products or herbal preparations in market circulation, are standard practice.

With this background, it is no surprise that most of the clinical trials done in China belong to simple parallel group studies. With this background, it is also obvious that ethical approvals for clinical trials using alternative medicine in China would be directed along different orientations, compared with conventional approaches.

EXAMPLES OF CLINICAL TRIALS ON CHINESE MEDICINE

To provide more information about clinical trials in Chinese medicine being done in Hong Kong, the following paragraphs are devoted to summaries of such trials.

Synopsis I

Name of Study Medication: Phyllanthus SP. Compound

Title of Study: A Prospective Randomised, Double-Blind, Placebo-Controlled, Parallel Study to Evaluate the Effect of Phyllanthus SP. Compound (葉下珠) in the Treatment of Chronic Hepatitus B Virus Infection.

Study Centre: Single-centre

Objective:

Primary

- To evaluate the efficacy of normalisation of liver enzyme, seroconversion of HbeAg and disappearance of HBV DNA in serum.
- To evaluate the safety of Phyllanthus SP. Compound in patients with hepatitis B.

Secondary

- Proportion of patients with end-of-treatment HbeAg aeroconversion (HbeAg to anti-Hbe, normalisation of ALT and disappearance of HBV DNA at the end of treatment).
- Proportion of patients with HbeAg to anti-Hbe.
- Proportion of patients with sustained normalisation of ALT.
- Proportion of patients with undetectable HBV DNA.

Design: A single-centre, prospective randomised, double-blind, placebo-controlled, parallel study. Patients will be randomised to one to the four treatment groups and treated for duration of 6 months.

Study Population: A minimum of 85 hepatitis B patients will be enrolled, 25 subjects per treatment group, 10 subjects in control group, total 4 groups.

Definition of Endpoints:

- The primary safety endpoint is tolerability.
- Tolerability failure is defined as a permanent discontinuation of Phyllanthus PLUS as the result of an adverse event.
- The primary endpoint is a reduction in HBV DNA level from the baseline.
- The secondary endpoint is HbeAg negative, anti-Hbe positive and a decrease in ALT level from baseline.

Study Regimen: Subjects will be randomly and alternatively assigned to receive Phyllanthus PLUS or placebo for 6 months prospective parallel study.

Duration of Treatment: 6 months

Statistical Methods: Efficacy: Summary statistics for the change of HBV DNA, HbsAG, HbeAg and ALT from baseline will generated and provided for each treatment group.

Safety:

- The incidence of adverse events and laboratory toxicity will be summarised by treatment group and severity. Change from baseline in vital signs will be summarised by treatment group.

Synopsis II

Name of Study TCM: Danggui Buxue Tang (當歸補血湯)

Title of Study: A Randomised, Double-Blind, Comparison Study of the Effect of Danggui Buxue Tang (當歸補血湯) with Oestradiol on Menopausal Symptoms and Quality of life in Hong Kong Chinese Women.

Study Centre: Single-Centre

Objective:

Primary

- To compare the effects of Danggui Buxue Tang (當歸補血湯) with Oestradiol on menopausal symptoms of hot flushes and sweating.
- To evaluate the safety of Danggui Buxue Tang (當歸補血湯) in patients with menopausal symptoms.

Secondary

- To evaluate the quality of life of the patients with menopausal symptoms.

Design: A single-centre, randomised, double-blind and comparison study. Subjects will be randomised to one to the two treatment groups and treated for duration of 6 months and follow-up of 18 months.

Study Population: A minimum of 100 patients with menopausal symptoms will be enrolled, 50 subjects per treatment group.

Definition of Endpoints:

- The primary safety endpoint is tolerability. Tolerability failure is defined as a permanent discontinuation of Danggui Buxue Tang (當歸補血湯) as the result of an adverse event.
- The primary efficacy endpoint is the change in severity and frequency of hot flushes and night sweats.
- The secondary efficacy endpoint is the changes in score for the domains measured in the Menopause Specific Quality of Life Questionnaire.

Study Regimen: Subjects will be randomly and alternatively assigned to receive Danggui Buxue Tang (當歸補血湯) or placebo for 6 months.

Duration of Treatment: 6 months' treatment period and 18 month follow-up.

Synopsis III

Name of Study TCM: Danggui Buxue Tang

Title of Study: A Randomised Comparison Study of the Effect of Danggui Buxue Tang with Tranexamic Acid on Dysfunctional Uterine Bleeding and Quality of Life in Hong Kong Chinese Women.

Study Centre: Single-centre

Objective:

Primary

- To compare the effects of Danggui Buxue Tang (當歸補血湯) with tranexamic acid on menstrual blood loss per month.
- To compare the patient's satisfaction between using Danggui Buxue Tang (當歸補血湯) and tranexamic acid.
- To evaluate the safety of Danggui Buxue Tang (當歸補血湯) in patients with dysfunctional uterine bleeding.

Secondary

- To evaluate the improvement of anaemia.
- To evaluate the status of iron deficiency.
- To evaluate the unwanted side effects.

Design: A single-centre, randomised comparison study. Subjects will be randomised to one of the two treatment groups and treated for duration of 6 months and follow-up of 24 months.

Study Population: A minimum of 125 patients with dysfunctional uterine bleeding will be enrolled, 63 subjects in Danggui Buxue Tang (當歸補血湯) group and 62 subjects in tranexamic acid group.

Definition of Endpoints:

- The primary safety endpoint is tolerability. Tolerability failure is defined as a permanent discontinuation of Danggui Buxue Tang (當歸補血湯) as the result of an adverse event.
- The primary efficacy endpoint is change in frequency and severity of menstrual bleeding.
- The secondary efficacy endpoint is improving anaemia and iron deficiency.

Study Regimen: Subjects will be randomly and alternatively assigned to receive Danggui Buxue Tang (當歸補血湯) or tranexamic acid for 6 months' treatment and 24 months of follow-up.

Duration of Treatment: 6 months' treatment and 24 months of follow-up.

Synopsis IV

Name of Study TCM: Formula A (托毒生肌顆粒) and Formula B (味地黃顆粒劑)

Title of Study: A Randomised, Double-blind, Placebo-controlled Study on the Clinical Effects of Integrated Western Medicine and Traditional Chinese Medicine for Diabetic Foot Ulcer.

Study Centre: Multi-centre

Objective:

Primary

- To evaluate the wound healing effect of Formula A (托毒生肌顆粒) and Formula B (味地黃顆粒劑) in patients with diabetic foot ulcer.

- To evaluate the safety of Formula A (托毒生肌顆粒) and Formula B (味地黃顆粒劑) in patients with diabetic foot ulcer.

Secondary

- To evaluate the effect of control of the local infection.

Design: A multi-centre, randomised, double-blind, placebo-controlled study. Subjects will be randomised to one of the two treatment groups and treated for duration of 6 months.

Study Population: A minimum of 80 diabetic foot ulcer patients will be enrolled, 40 subjects per treatment group.

Definition of Endpoints:

- The primary safety endpoint is tolerability. Tolerability failure is defined as a permanent discontinuation of Formula A (托毒生肌顆粒) and Formula B (味地黃顆粒劑) as the result of an adverse event.
- The primary efficacy endpoint is diabetic foot ulcer healing and to avoid leg amputation.
- The secondary efficacy endpoint is the control of local infection.

Study Regimen: Subjects will be randomly and alternatively assigned to receive Formula A (托毒生肌顆粒) and Formula B (味地黃顆粒劑) or placebo in a 6 months' prospective parallel study.

Duration of Treatment: 6 months

Synopsis V

Name of Study TCM: Relieve Wheezing Tablet (培補鈉氣丸)

Title of Study: A Randomised, Doubled-Blind, Placebo-Controlled Parallel Study of the Effect of Relieve Wheezing Tablet (培補鈉氣丸) in the Treatment of Childhood Asthma

Study Centre: Single-centre

Objective:

Primary

- To evaluate the medication score, including daily use of inhaled steroids.
- To evaluate the symptom score, including cough at daytime and night-time, wheeze/chest tightness at daytime and nighttime, degree of shortness of breath on exertion.

Secondary

- To evaluate the spirometry lung function result.
- To evaluate the breakthrough attacks requiring medical attention from A & E doctors, family physicians of hospitalisation.
- To evaluate the degree of skin allergy.
- To evaluate the changes in peripheral blood and Eosinophilic Cationic Protein (ECP).

Design: A single-centre, randomised, double-blind and placebo-controlled, parallel study. Subjects will be randomised to one of the two treatment groups and treated for duration of 6 months.

Study Population: A minimum of 80 patients with moderate to severe perennial asthma will be enrolled, 40 subjects per treatment group.

Definition of Endpoints:

- The primary safety endpoint is tolerability. Tolerability failure is defined as

a permanent discontinuation of Relieve Wheezing Tablet (培補鈉氣丸) as the result of an adverse event.
- The primary efficacy endpoint is a change in improving the symptoms of asthmatic children and use of inhaled steroids.
- The secondary efficacy endpoint is improvement of lung function.

Study Regimen: Subjects will be randomly and alternatively assigned to receive Relieve Wheezing Tablet (培補鈉氣丸) or placebo for 6 months.

Synopsis VI

Name of Study Medication: Danshen and Radix Puerariae Compound

Title of Study: A Prospective Randomised, Double-Blind, Placebo-Controlled, Parallel Study to Evaluate the Effect of a Herbal Preparation with Compound Formula of Danshen (丹蔘) and Radix Puerariae (葛根) as Cardiovascular Tonic in Cardiac Patients.

Study Centre: Single-centre

Objective:

Primary

- To evaluate the safety of Danshen and Radix Puerariae Compound as adjunctive therapy in patients with coronary artery disease.
- To evaluate the efficacy of treatment and secondary prevention of cardiovascular diseases.

Secondary

- To evaluate the lipid and homocysteine-lowering effect of Danshen and Radix Puerariae Compound.

Design: A single-centre, prospective randomised, double-blind, placebo-controlled, parallel study. Patients will be randomised to one of the two treatment groups and receive Danshen and Radix Puerariae Compound or placebo for a duration of 24 weeks.

Study Population: A total of 100 patients with Coronary Artery Disease (CAD) will be enrolled, 50 subjects treated with Danshen and Radix Puerariae Compound and 50 treated with placebo.

Definition of Endpoints:

- The primary safety endpoint is tolerability.
- Tolerability failure is defined as a permanent discontinuation of Danshen and Radix Puerariae Compound as the result of an adverse event.
- The primary endpoint is improving cardiovascular function (endothelial function and carotid intima-medial thickness) from the baseline.
- The secondary endpoint is decrease of plasma lipid and homocysteine levels.

Study Regimen: Subjects will be randomly assigned to receive Danshen and Radix Puerariae Compound (TCM) or placebo for 24 weeks in a prospective parallel study.

Duration of Treatment: 24 weeks.

Duration of project: 30 months

ACUPUNCTURE

Acupuncture is a practical procedure using a special needle to enter special regions of the human body surface by which symptoms suffered by the patient are removed. Like other aspects of Chinese medicine, it aims at symptom control, not at the treatment of a specific disease entity. The most popular use is in the field of pain control.

In 1998, the NIH held a consensus conference on the use of acupuncture for pain control. The conclusion was that acupuncture should be accepted as an effective means of pain control for musculo-skeletal problems and under other specific situations.[21] Since then, interest in the use of acupuncture in the United States grew and many clinical studies were started.

Of course, acupuncture has been used for the control of other symptoms. Examples include nerve damage, allergic conditions like rhinitis, asthmatic attacks, and general feelings of 'unwell', often labelled as 'derangement'.

How are clinical trials on acupuncture being conducted? Could the clinical trials on acupuncture be put online with modern epidemiological requirements? Or would it be even more difficult compared with herbal medicine?

We have, first of all, to look at the procedures involved and the explanations given for the effects produced.

Acupuncture involves the insertion of thin needles, through specific points on the body surface, to varying depth of soft tissue, then allowing the needles to remain for some minutes. While the needles are inside the soft tissue, the acupuncturist may employ regular or occasional rotary movements of the needle. In recent years, acupuncturists have applied direct electrical current stimulation, so as to unify the stimulations, widen the effects and save labour. Acupuncture is an invasive process directly aiming at the removal of symptoms. It is easy to imagine, then, that patients would not agree to participate in a study where they would not be able to enjoy the puncture treatment and function as recipients of 'sham' puncture. Likewise, if there were other placebo punctures which fulfilled the requirement of randomisation and placebo control, very few patents would be willing to participate.[37,38]

However, studies of placebo acupuncture have started and the varieties include the following:

1. Placebo points – entry points are sites outside the acupuncture meridians.
2. Sham puncture – puncture lightly then withdraw
3. Hiding entry points – while entry points are hidden, it might be possible to achieve a real placebo effect. Hiding of entry points may be achieved by puncturing through plastic tubes or soft plastic blocks.
4. Camouflage puncture by which a needle is just taped to the skin.

None of these methods can be thoroughly endorsed as 'placebo' because the requirements for placebo in the strict sense are far from being satisfied since most recipients could differentiate right away whether it is a true or false puncture. The conventional application of acupuncture depends on a subjective feeling of 'numbness' felt within the punctured area. Puncturing without checking this feeling is not considered appropriate. This requirement makes 'placebo' puncture impossible. The use of electrical stimulation is a means to enhance the effects in modern situations where there is insufficient experience on acupoint identification and actual puncturing techniques. It is also considered as a method of modernising acupuncture. When electrical stimulation is used, placebo becomes absolutely impossible because the electrical stimulation is always felt.

Considerations are further complicated by the theories of acupuncture. There are two acceptable theories: the neurological and the humoural. The neurological theory observed that since some of the meridians and most of the acup points are related to the peripheral nerves, stimulation of these points leading to physiological effects could be working via neurological pathways, possibly through proprioceptive receptors.[39] The humoural theory assumes that needle stimulation produces humoural (hormonal) reactions, manifested as the serological appearance of functional factors which possess pain control effects and other regulatory functions.[40]

Whichever theory is at work, it specifies that the tiny area of puncture is producing chain reactions, either directly or indirectly. An apparently harmless, non-productive action on the skin and soft tissue, imitating acupuncture, might trigger off similar effects and would be far from being a placebo.

Therefore standard epidemiological planning for clinical trials in Chinese medicine including acupuncture would be very difficult, if at all possible. Randomisation would not be acceptable to patients, whereas in situations of acupuncture where sham puncture is insisted on, it is both unacceptable to patients and short of the placebo requirements.

Carefully planned cohort studies aiming to compare the effects of different means of pain control and other treatment expectations are therefore the only reasonable means to look objectively at the clinical effects of acupuncture. Many cohort studies of this nature are being done in the study of back pain, neck pain, arthritis of the knee, etc. The effects of puncture were compared with conventional techniques using physiotherapy and other means.

Another common application for acupuncture is on restoration of nerve function. Damaged neurological tissues suffer from a real lack of regeneration. Peripheral damage feasible for repair carries reasonable promise. When cell bodies are involved, either intra-cranially or in the spinal cord, loss of neurological and secondary muscular functions becomes permanent. Acupuncture is widely used under such difficult situations. Although many reports of impressive results are available, it is difficult to appreciate the real value. Scientific data coming from well-planned cohort studies for the observation of functional restorations are still difficult to interpret, since the damage is not uniform and the factors affecting the different aspects of rehabilitation and functional return are multiple and complicated. We are therefore still relying on careful case studies.

However, one can appreciate the obvious limitations. After all, in the field of rehabilitation, experience in the last three decades has already shown that qualified, broad, trustworthy clinical trials are not possible.[41] Although meta-analysis has ruled out absolute scientific justification of all the rehabilitation attempts like the different forms of physiotherapy, massage, bracing, and even invasive techniques like injection and surgery, we can still rely on them because we have to relieve our patients of suffering. We are all aware of the fact that we are not certain which patient is the best candidate to receive which treatment.[42]

CONCLUSION

Complementary medicine does not have a history of modern scientific development. It builds up its knowledge by relying on observations and experience. Now that we are trying to make use of this traditional stream of medicine in a scientific world, we need to explain why it works in our area of concern. Very often, these areas are not well served by scientific medicine. This makes the scientific explanations even more important.

The way to go about giving scientific explanations of healing processes involves the application of methodologies that are well known and accepted by all clinical scientists. The standard way to begin a scientific approach to clinical trials using traditional Chinese medicine would be just an application of the same methodologies. However, this approach is not ideal and would probably remain doubtful in spite of growing enthusiasm. We are barred from a smooth application of the scientific methodology, basically because of the different philosophy behind the traditional Chinese way of healing. Moreover, the lack of knowledge of the exact chemistry for the active component in the herbal remedy when herbal drug trials are being done further jeopardises the validity of the clinical trials carried out.

In spite of the essential difficulties, efficacy-driven trials are still being carried out, utilising the principles of evidence-based medicine. As long as the scientific gap is successfully narrowed, practical use of complementary medicine will become safer, more logical and deserve wider application. At the same time, workers in complementary medicine should compile a unique, relevant system of assessment for the clinical effects.

REFERENCES

1. Bensky D, Gamble A. *Chinese Herbal Medicine*. Materia Medica Seattle: Eastland Press (1993).
2. Tang W, Eisenberg G. *Chinese Drugs of Plant Origin: Chemistry, Pharmacology and use in Traditional and Modern Medicine*. Berlin, New York: Springer (1992).
3. Su WT. *History of Drugs*. Beijing: United Medical University Press (1992).
4. McGuire MB. *Ritual Healing in Suburban America*. New Brunswick, NJ: Rutgers University Press, (1988).
5. Kong YC. *Formulation's from South – West Asia*. Beijing: Chinese University Press (1998).
6. Eisenberg DM, Roger DB. Trends in alternative medicine use in the USA 1990–1997. *JAMA* (1998) **280**(18): 1569–75.
7. Lai SL. *Clinical Trials of Traditional Chinese Materia Medica.* Guangdong: People's Publishing House (2000) Chapter 1.
8. Leung PC. Editorial-seminar series on evidence-based alternative medicine. *Hong Kong Med J* **7**(4), (2001).
9. Kaptchuk TJ, Eisenberg Dm. The persuasive appeal of alternative medicine. *Ann Intern Med* (1998) **129**(12): 1061–1065.
10. Eisenberg DM, Kessler RC, Foster C. Unconventional medicine in the USA – prevalence, costs and patterns of use. *New Eng J Med* (1998) **328**(4): 246–52
11. Fair WR. The role of complementary medicine in urology. *J Urol* (1999) **162**: 411–420.
12. Leung PC. *Traditional Chinese Medicine in China*. Seminar on Health Care in Modern China, Yale–China Association (2001).
13. Leung PC. Development of Traditional Chinese Medicine in Hong Kong and Its Implications for Orthopaedic Surgery. In: *Hong Kong Journal of Orthopaedic Surgery* (2002) Vol. 6 No. 1, pp. 1–5.
14. Ho T. An anthropologist's view on traditional Chinese medicine. In: *A Practical Guide to Chinese Medicine* Singapore: World Scientific (2002).
15. Chalmers I. The Cochrane Collaboration: preparing, maintaining and disseminating systemic reviews of the effects of health care. *Ann NY Acad Sci* (1993) **203**: 166–72.
16. Tyler VE. *Herbs of Choice: The therapeutic use of Phytomedicinals*. New York: Pharmaceutical Product Press (1994).
17. Ernst E. Harmless herbs? A review of the recent literature. *Am J. Med* (1996) **104**: 170–8.
18. Bisset Ng. *Herbal Drugs and Phytopharmaceuticals*. Boca Raton, FL: Medpharm Scientific Publisher (1982).
19. Leung PC. *Clinical trials in Chinese medicine – the efficacy driven approach*. Hospital Authority Convention, Hong Kong (2001).
20. NIH. National Centre for Complementary and Alternative Medicine. Five Year Strategic Plan 2001–2005, NIH Document (2000).
21. NIH. Acupuncture. *NIH Consens Statement* (1997) **15**(5): 1–34.
22. Wang PY. *New Chinese Medicine: research production and registration*. Chinese Medicine Press, China, (1995).
23. Chen KC. *Research in Chinese Medicine*. Beijing: China Health Publisher (1996).
24. Lai SL (ed). *Clinical Trials of Traditional Chinese Materia Medica*, Guangdong: People's Publishing House (2001) Chapters 2, 3, 4.
25. Chang CK (Han Dynasty). *Discussions of Fever*. Classics.
26. Li ZC (Ming Dynasty). *Materia Medica*. Classics.
27. Suen TM (Tang Dynasty). Important prescriptions. Classics.
28. National Burean on Drug Supervision. *Regulations on Clinical Trials Using Drugs and Herbs*. Burean Publication (1999).
29. Suen TY, Wang SF, Wang KL. *Adverse effects of drug treatment*. Beijing: People's Health Press (1998).
30. De Smet P, D'Arcy PF (eds). Drug interactions with herbal and other non-toxic remedies in mechanisms of drug interactions. Berlin: Springer (1996).
31. Sackett DL. *Clinical epidemiology*, 2nd edn; Boston: Little, Brown (1991) 297–9.
32. Uppsala Monitoring Centre, WHO. Safety monitoring of medicinal products: guidelines for setting up and running a pharmaco vigilance centre (2000.)
33. Jaeschke R, Singer J. Measurement of health status. Ascertaining the minimally important difference. *Controlled Clin Trials* (1989) **10**: 407–15.
34. Wyrwick KW, Nienaber NA. Linking clinical relevance and statistical significance in health related quality of life. *Med Care* (1999) **37**: 469–78.
35. Stewart AL, Greenfield, Hayo RD. Functional status and well-being of patients with chronic conditions. *JAMA* (1989) **262**: 907–13.
36. Keininger DL. Why develop IQOD? International Health Related Quality of life Outcome Database. *QoL News*(1999) **Sept–Dec** (23): 2.
37. Park J, White AD, Lee H, Ernst ED. Development of a new sham needle. *Acupunct in Med* (1999) **17**(2) 347–52.

38. Streiberger K, Kleinhenz J. Introducing a placebo needle into acupuncture research. *Lancet* (1998) **352**: 364–5.
39. Godfrey CM, Morgan P. A controlled trial of the theory of acupuncture in musculoskeletal pain. *J Rheumatol* (1978) **5**: 121–4.
40. Lewith GT, Machin D. On the evaluation of the clinical effects of acupuncture. *Pain* (1983) **16**: 111–27.
41. Nachemson E. *Causes and treatment of low back pain*. American Academy of Orthopaedic Surgery Lecture Series (1989).
42. Spitzer WO, Leblanc FE. Scientific approach to the assessment and management of activity-related spinal disorders. *Spine* (1987) **12**(Suppl 1): 51–9.

36

Cluster Randomized Trials in Health Care Research

RANDALL D. CEBUL, NEAL V. DAWSON AND THOMAS E. LOVE
Center for Health Care Research and Policy, Case Western Reserve University School of Medicine at MetroHealth Medical Center, Cleveland, Ohio, USA

INTRODUCTION AND OVERVIEW

THEMES IN HEALTH CARE RESEARCH

The mantra of health care research is ‘access, cost, and quality’. These three themes motivate methods designed to investigate and demonstrably improve the care and outcomes of persons and populations at risk for, or suffering from, medical conditions. To these three themes, many would add a fourth – ‘value’ – which simply relates the quality of health care services to their costs. The relevant disciplinary underpinnings of health care research are broad, including methods from the population sciences (epidemiology, biostatistics and clinical research), social sciences (economics, psychology and sociology), and the decision and information sciences (decision and cost-effectiveness analysis meta-analysis and applied informatics).

In the United States and developed countries worldwide, approaches to health care research increasingly are focusing on fundamental challenges in human behaviour and on potential solutions that can be found by re-engineering the systems of care delivery. Structural problems in health care delivery are highlighted in work that focuses on inadequacies of the health information infrastructure and on the suboptimal organisation of caregivers, leading to waste, errors and concerns about patient safety. Work motivated by the disciplines of economics and psychology variously highlights either ‘carrots’ or ‘sticks’ – incentives that are designed to change the personal or group behaviours of patients and/or their providers, or the activities and processes of care of entire health care systems. As the last century witnessed an epidemiologic shift towards chronic illnesses, health care interventions increasingly are being evaluated for their impact on economic and quality of life measures, and not solely the length of life.

CLUSTER-RANDOMISED TRIALS, AND THEIR RELEVANCE TO HEALTH CARE RESEARCH

Cluster-randomised trials (CRTs) are studies that typically examine the effects of non-therapeutic

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

interventions such as life-style modification, educational programmes, or innovative approaches to the organisation or delivery of health care – approaches that are central to the themes of health care research. CRT interventions are targeted at groups (as examples, patients, families, providers, group practices, hospitals, or even entire communities or countries) and must acknowledge the possibility, and frequently seek to take advantage of, intragroup interaction. Because the alternative and more familiar approach to clinical trials, i.e. person-level allocation, risks contamination (e.g. individuals in the same family being allocated to alternative dietary interventions) or the results of 'learning effects' (e.g. patients of the same physician being allocated to different interventions designed to improve adherence to published standards of care), conventional randomised controlled trials (RCTs) of such interventions may blunt the intended intervention effect and thereby enhance the probability of a Type II error. Trends in the nature and quality of CRTs in health care research are highlighted in the following section of this chapter.

In a conventional RCT, the unit that is randomised (typically a patient) also is the unit of analysis. With a large enough sample size, this process increases the likelihood that both observed and unobserved attributes of the subjects will be distributed similarly across groups, allowing statistical tests to be conducted that assume independence of individuals within and across groups. In contrast, the central design feature of a CRT is that it is a comparative trial in which: (1) the units of assignment to interventions are identifiable groups; and (2) the units of analysis are members of those groups.[1] In non-health-care contexts, CRTs also are referred to as 'group randomisation' or 'community intervention trials', the former label fostering a general understanding of common features *within* (as contrasted to *across*) groups, while the latter label highlights the allocation of experimental interventions at the level of entire communities or political subdivisions.[1,2] Regardless, and in contrast to person-level randomisation, persons *within* clusters (groups, communities) are understood to have more in common with others in the *same* cluster than they have with persons in *other* clusters. We prefer and will use the term *cluster*-randomised trial because of the literal meaning of cluster ('a number of similar things grouped together in association or in physical proximity')[3] and because of its connotation, as in 'birds of a feather flock together'.

Despite often compelling reasons for undertaking a CRT in health care research, the effects of clustering are: (1) to potentially reduce the effective sample size and power of the trial in comparison with an analogous trial using patient-level allocation; (2) to promote consideration of CRT designs that better balance study clusters on important attributes prior to allocating interventions; and (3) to require analysis strategies that account for the extent of clustering. These issues are described in more detail in the section on clustering and its impact. The last section of this chapter addresses selected additional issues in the design and conduct of CRTs in health care research, including standards for their evaluation and the likely impacts of practice-based research networks and electronic health data on the use and quality of future CRTs. While we believe that methods used in the conduct of CRTs will evolve rapidly with these emerging trends, interested readers can take advantage of excellent textbooks that serve as references for key methods.[1,4]

HISTORY AND TRENDS IN HEALTH CARE CRTs

The analytical challenges unique to CRTs have been recognised relatively independently in such disciplines as psychology, educational research and medicine.[5] In medical treatment, it appears that the idea of a CRT may have preceded the concept of the individual randomised trial. Perhaps the earliest example of a type of CRT of medical treatment was considered by Van Helmont in 1648[6] to assess the efficacy of bloodletting for treating fever. He proposed forming two groups by casting lots and then randomly treating only one group with phlebotomy to cure fever. Despite a conceptual ‘head-start’ of several centuries, the development of rigorous designs and analyses for CRTs has lagged behind that of individual randomised trials. This has been attributed, at least partially, to the additional requirements that potential dependencies in responses within clusters place on the research architecture and analytic requirements.[5] Another possible explanation is that effecting changes in life style and systems of care have, until recently, seemed to be intractable goals,[7] while advances associated with disease-specific therapeutics (the subject of many patient-level RCTs) have long had the strong support of a large medical–industrial research complex.[8]

The potential advantages of individual randomisation of subjects in trials of medical treatments was largely unappreciated until the appearance of Bradford Hill’s series of articles on the use of statistical methods in medical research, published in *Lancet* in 1937.[9] Within a decade, the first example of a clinical trial that contained an adequately randomised control group was developed by the Medical Research Council in a study of the use of streptomycin to treat tuberculosis.[10]

Analytical advancements important to the development of modern CRTs came with improvements in cluster sampling from the field of survey research. For example, a variance inflation factor identical to that used in today’s CRTs was developed by Hansen and Hurwitz in 1942.[11] Additional advances came from trials of manoeuvres to inhibit the effects of contagious diseases in populations[12] and to alter the behaviours of members of communities that are susceptible to the development and untoward consequences of chronic illnesses.[13,14] Both disadvantages (potential imbalance of important prognostic factors) and advantages (increasing subject participation, reducing the likelihood of treatment spillover, and convenience of administration) of cluster randomisation were recognised and described.[12] The cited advantages have been recognised more recently as being potentially important to enhancing the effectiveness (in a practice or community setting) of an intervention for which efficacy had been proven in a patient-level RCT.

In the 1980s and 1990s, new research methods were described that combined aspects of both RCTs and CRTs. ‘Firm system trials’ were based on parallel group practices that used ongoing randomisation of patients who received care in the outpatient and/or inpatient services of several Internal Medicine training programmes.[15,16] This movement was the focus of a national conference in 1990 sponsored by the NIH Office of Medical Applications of Research, with proceedings described in a supplement to the journal *Medical Care*.[17] To enhance comparability at several levels of the hierarchy, some programmes randomly allocated not only patients to firms, but also house staff (who turn over nearly completely over a three-year period) and physician–faculty.[18] Firm systems have been described as ‘laboratories’ in which continuous quality improvement, educational and related interventions can be performed on an ongoing basis. Several studies have documented the similarities across practices that can result from ongoing randomisation of patients and house staff.[18]

Despite these advances, improvements in the design and analysis of CRTs have been inconsistently applied, even in grant-funded research reflected in peer-reviewed publications. As summarised in 1995 by Simpson and colleagues,[19] fewer than 20% of CRTs published between 1990 and 1993 adequately addressed sample size and power in their design, and only 57% adequately addressed these issues in their analyses. A more recent systematic review of 152 CRTs in primary care published between 1997 and 2000 reported continued variability in the quality of trial design and analysis.[20] The continuing need for methodologic investigation and the imperative to consistently apply existing knowledge has been increasingly recognised and discussed.[5,20,21] Further increases in the prevalence and methodologic rigour of cluster trials in health care delivery also may be facilitated by the advent of large systems of health care, with multiple practices or hospitals that are linked organisationally by shared electronic medical records.

CLUSTERING AND ITS IMPACT

REASONS FOR CLUSTERING

CRTs must consider the context in which a non-therapeutic intervention occurs, including the possibility that relatively intact groupings ('clusters') exist through natural or explicit selection processes. In CRTs related to public health interventions, clustering also may occur through the influence of covariates at the group level, such as sharing exposure to a common environment, or through the tendency of infectious agents to spread more rapidly among those in close proximity (such as within families, or on a common hospital ward) than among those at greater distance (such as across families or in different hospitals). In the health care delivery setting, without an ongoing random allocation process (such as that described among firm system practices, above), selection may occur for many reasons. To name but a few of these, health-care-related groupings may exist because of socio-economic, educational and social forces, health insurance policy restrictions, geography, and gender/age/race-related preferences of patients and/or providers. To the extent that these selection factors tend to be common across individuals in a given health care or geographic setting, persons within that setting are likely to be more similar to others in that setting than to persons in other settings. Because selection factors may influence the probability of adopting the study intervention and/or achieving the outcome of interest, clustering needs to be acknowledged in the design and analysis of the CRT. As an illustration in the field of public health, a trial to influence smoking cessation across multiple communities would need to consider the possibility that people with respiratory conditions (including those that are smoking related) might seek to reside in regions with lower humidity, or that warmer climates might be disproportionately populated by older people.[20] These characteristics, in turn, may affect the probability that the residents would participate in an educational programme promoted through a media campaign[21] or develop a smoking-related adverse outcome. Similarly, a trial designed to influence children's dietary habits should acknowledge the family context within which most nutritional habits come about (e.g. by randomly allocating the intervention to intact family units) as well as the variations in dietary practices across families of different socio-economic, racial or ethnic backgrounds.

Trials across clinical practices, hospitals or health care systems must similarly acknowledge the likelihood of systematic variation in the characteristics of those patients, providers and systems. At the patient level, patients may generally prefer physicians of the same sex; people in a demographic minority group may preferentially seek care by doctors of similar ethnic or language background; and poor or uninsured patients may cluster in 'safety net' provider organisations. These patient attributes, in turn, may be associated with systematic differences in the likelihood of patients' adopting a particular (life-style, behavioural, or other health-care-related) intervention and/or to their being susceptible to the targeted health outcome. At the provider level, Wennberg and others have amply documented that practice styles of physicians may vary widely and systematically by geographic region, or by age, prior training or specialty focus.[22,23] At the health care 'system' level, systematic variations exist in information infrastructure,[24] the way care delivery is organised, and the types of financial incentives (or disincentives) that influence the delivery of preventive services as well as the ordering of costly tests and treatments. Since these physician- and system-related variations also may influence health outcomes or the likelihood that CRT interventions may be adopted, trials that allocate interventions at the level of the clinical practice or health care system need to account for these variations in selecting a specific trial design, estimating sample sizes and trial power, and conducting analyses.

HOW CLUSTERING IS DESCRIBED AND MEASURED

Figure 36.1 depicts the subjects of a simple CRT in which the intervention is allocated at the

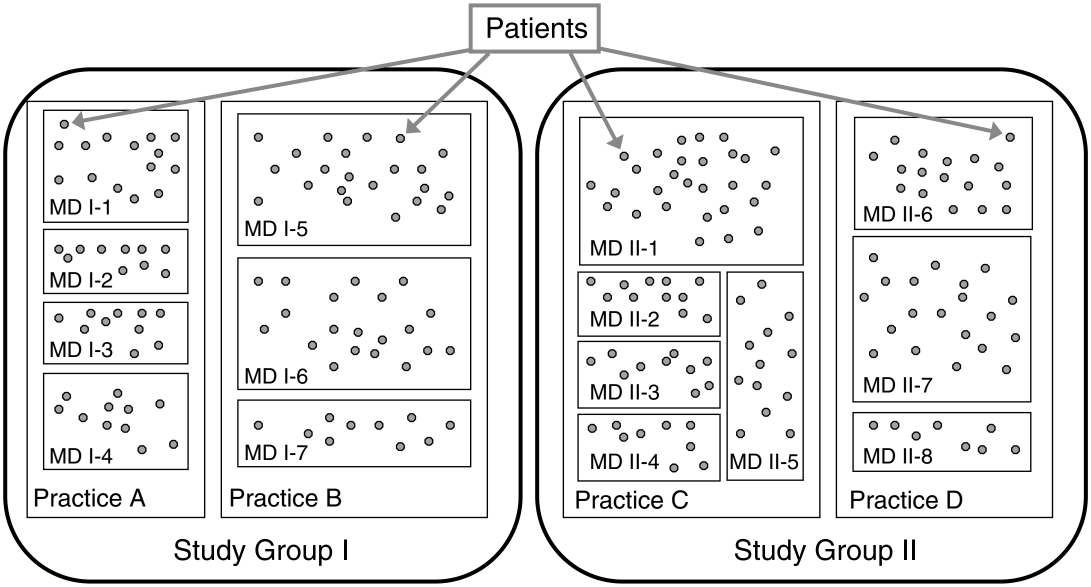


Figure 36.1. A simple four-level cluster-randomised trial.

clinical practice level, with the patient panels of four group practices allocated into two study groups (study group I, composed of patients in practices 'A' and 'B', and study group II, composed of patients in practices 'C' and 'D'). This trial may be described as having four-level clustering, or 'nesting', with patients clustered (or 'nested') within physicians, who in turn are nested within group practices, which are then allocated to study groups. If a coin toss or other method for random allocation is used to assign the intervention to study groups, the central question is the extent to which there are baseline differences across study groups. Relevant baseline differences that must be considered include differences in likelihood of adopting the proposed intervention, and differences in susceptibility to the outcome(s) of interest. The answers to these questions, in turn, must consider both the nature of the intervention and the type of outcome(s) the investigator seeks to affect.

Figure 36.2 represents the pre-intervention values for haemoglobin A1c (a measure of glucose control) among patients of four group practices in a hypothetical trial to improve glycaemic control among diabetic patients, where higher values reflect poorer baseline control. By visual inspection, while it is apparent that each practice's patients exhibit variation in glycaemic control, baseline values for practices C and D are higher (worse) than values for practices A and B.

The fundamental measure that is used to reflect cross-cluster differences is the intracluster (or intraclass) correlation coefficient, or ICC, denoted by the Greek letter ρ For a continuous variable with sample variances between and within clusters of s^2_{Between} and s^2_{Within}, respectively, we define the ICC as $\rho = s^2_{\text{Between}}/(s^2_{\text{Between}} + s^2_{\text{Within}})$. We then interpret ρ for a continuous variable as merely the proportion of overall variance that can be attributed to between-cluster variance. Small values of ρ (approaching zero) imply virtually complete statistical independence among members of a cluster, while larger values of ρ (approaching one) imply greater degrees of statistical dependence, with responses virtually identical to other responses within that cluster and different from responses in the comparison cluster. The estimated ICC is dependent on the specific study design (e.g. completely

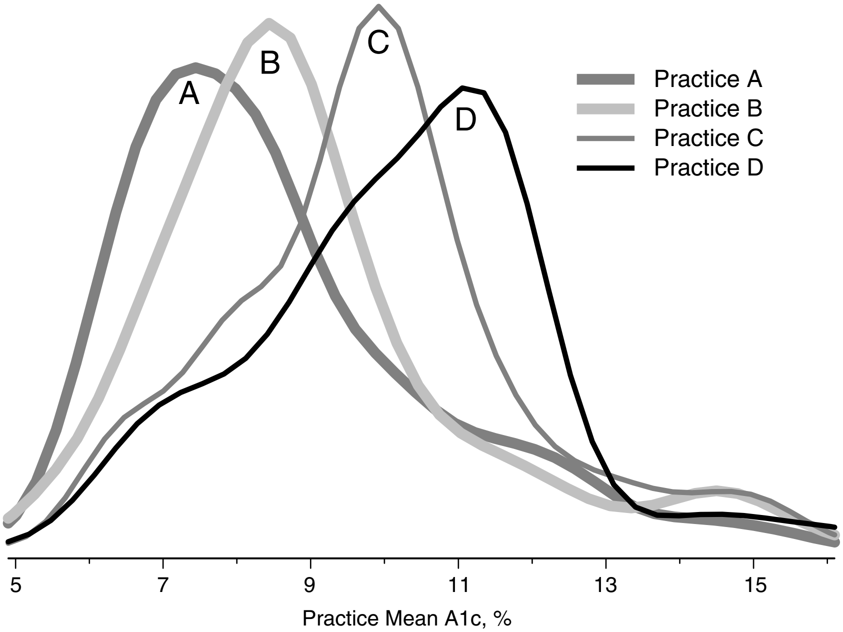


Figure 36.2. Density of pre-intervention haemoglobin A1c values among subjects in four practice panels.

randomised, matched-pair, or stratified),[25] the number of clusters being examined, and the number of subjects participating in the planned trial.[25] Historically, a major challenge to accurate pre-trial estimation of the ICC has been insufficient published data, in terms of small numbers of prior publications on the populations and outcomes of interest, infrequency of published ICCs when trial results are reported, and, frequently, small numbers of clusters in reported trials.[4] In the growing number of health care systems with ready access to electronic medical data on patients and practices, some of these challenges may be addressed by empirical determination of risk factor or outcome rates, and past changes in risk factor or outcome rates over time, in the practices being considered for study.

EFFECTS OF CLUSTERING

The value of the ICC in a CRT affects its required sample size, its power to detect meaningful differences across intervention groups, and the analytic approach used to estimate the intervention's effect size. Compared to CRTs with small ICCs larger values of ICC (ρ) mandate larger sample sizes in order to avoid Type I errors. Alternatively, larger values of ρ reduce the 'effective sample size' in a trial of a fixed number of subjects, reducing the trial's power to detect meaningful between-group differences at the individual level. Although a detailed discussion of sample size and power is beyond the scope of this chapter, the variance of a parameter's mean value in a CRT (such as the variation in haemoglobin A1c in our hypothetical diabetes trial) typically is larger than that which would be expected under the assumption of statistical independence. The factor describing this difference has been called the design effect (DEFF) or the variance inflation factor. The DEFF is mathematically summarised by $1 + (m - 1)\rho$, where m is the number of subjects.[21] The influence of DEFF on 'effective sample size' is given by the simple formula $m/[1 + (m - 1)\rho] = m/\text{DEFF}$. Thus, when $\rho = 0$, DEFF $= 1.0$, and the effective sample size is analogous to the sample size in a person-level RCT and the same as the number of individual subjects under study. Under the condition of complete dependence, on the other hand, in which $\rho = 1.0$, the effective sample size is one per cluster, in which the total information provided by each cluster is no more than what would be provided by any one member of that cluster.

Despite increasing calls for values of ICC to be published routinely in trial reports, an impediment to the planning of CRTs in health care has been the relative paucity of published data for relevant processes and outcomes.[26–29] While exceptions generally have pertained to the publication of ICCs in public health settings (e.g. worksite or epidemiological studies), Murray and Blitstein recently reported almost 1900 ICCs from 21 studies of health promotion and disease prevention.[30] Two additional recent reports are noteworthy. Campbell and colleagues[28] examined ICCs from 21 health care research data sets (mainly from the United Kingdom) and examined hypotheses related to the comparative ICC values for process of care *versus* outcome variables, outcomes of primary *versus* secondary care providers, outcomes with lower *versus* higher base rate prevalences, studies with larger *versus* smaller cluster sizes, and values for 'subjective' *versus* 'objective' measures. Overall, significant differences in magnitude were found among the 220 available ICCs, with values ranging from 0 to 0.415. Higher ICCs were observed for process as compared with outcome variables and for secondary care as compared with primary care outcomes. Interestingly, the effects of prevalence and cluster size were less clear cut, and there was no evidence to suggest that subjective measures (such as those obtained through self-report) had higher ICCs than objective measures. Parker *et al.*[31] recently reported ICC and design effect results for several cardiovascular measures from primary care practices in the CEART trial. For several measures (including weight, total and LDL cholesterol, and glucose), ICC values were <0.02 with corresponding design effects ranging from 1.0 to 2.3. Other measures (including smoking status and body mass index) had intermediate values, while the largest ICCs (0.05–0.12) and design effects (4.4–9.4) were observed for height and diastolic blood pressure.

As mentioned above, in a trial with a fixed number of subjects, larger values for ICC reduce the power to detect clinically meaningful differences across groups. Figure 36.3, illustrates this effect in a trial to improve glycaemic control among 1200 patients clustered in two large group practices, displaying the ICC impact over the range of values reported by Campbell and colleagues.[28] In the figure, we assume that we want to detect an absolute difference in changes across groups in haemoglobin A1c of 0.5% (e.g. if one group (changes) by 1.0%, the second group improves (changes) by 1.5% or more). In addition to assuming a two-tailed 95% confidence level, this simulation assumes that the intervention effect has an associated standard deviation of 2.0. The figure describes three separate design scenarios that examine the effect of enrolling the 600 patients across different numbers of physicians in each group. Holding the total number of patients constant, as might be necessary for trial cost considerations, these designs examine the effect on power of allocating the 600 patients to *fewer* physicians (as displayed on the bottom line for 25 physicians, or about 24 patients per physician, on average) as compared with designs that allocate the patients to more physicians (upper two lines in the figure).

From the figure, we see: (1) that this planned trial has excellent power to detect cross-group differences at low values for ICC (the power for all scenarios exceeds 0.99 at $\rho = 0$, as one might expect for a patient-level RCT); (2) that power declines substantially at values for ρ that are within the published range for some measures (as described by Campbell and colleagues)[28]; and (3) that designs that allocate fewer patients per physician to more physicians (top lines in the figure) are more efficient statistically than those that allocate more patients per physician to fewer physicians. Assuming that we want 90%

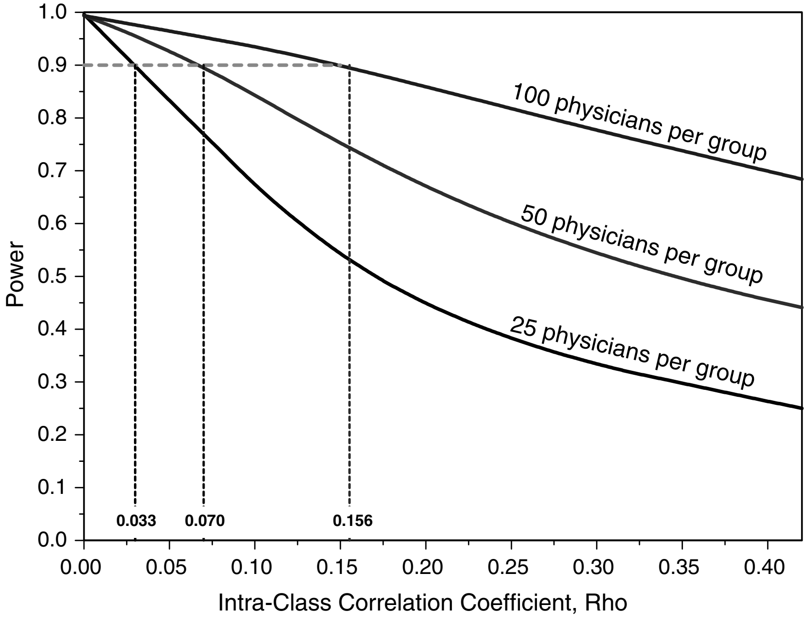


Figure 36.3. Relationship of statistical power to the intraclass correlation coefficient.

power to detect important differences for the measure of interest (in this case, haemoglobin A1c), the dashed lines in the figure represent the maximum estimated ICC that would be acceptable across the three design scenarios. In the most statistically efficient design (top line in the figure), high power can be achieved even at values of ICC as high as 0.16, whereas in the least efficient design, much lower ICCs would be necessary. For example, with the 600 patients allocated across 25 physicians per group, the maximum acceptable ICC value falls to about 0.03. Since from our own work and published reports from others[31] we have determined ICC values for glycaemic control that are less than 0.02, acceptably high power could be achieved in any of the scenarios, and decisions about trial design might be influenced by factors other than the particular strategy for allocating patients across available physicians.

The analysis of a CRT also must account for the effects of clustering. As CRTs represent a subset of research designs variably called nested, hierarchical, multilevel or clustered, analytic approaches to CRTs should account for clustering at one or more levels. Extending our hypothetical diabetes trial, clustering may occur among patients across different physicians; physicians may cluster in practices with other providers of similar (or different) specialties; and practice 'styles' may cluster in health care organisations that have different systems of care delivery or financial incentives. The specific analytic approach, while beyond the scope of this chapter, must take into consideration the specific design of the trial (e.g. completely randomised, matched or stratified), the relevant unit of analysis and intended effect (e.g. on cohorts or sequential cross-sections), and the nature of the primary outcome (e.g. binary, continuous, 'count', time-to-event, or categorical) and its distribution in the study sample. Illustrative methods that account for clustering include generalised estimating equations,[32] multilevel[33] or hierarchical modelling,[34] and robust variance estimation techniques.[35,36] More detailed descriptions of these approaches, as they pertain to CRTs, are described elsewhere.[1,4]

SELECTED CONSIDERATIONS IN PLANNING AND CONDUCTING HEALTH CARE CRTs

Detailed discussions of methodologic issues in the design, conduct and analysis of CRTs are provided in excellent textbooks by Murray[1] and Donner and Klar;[4] in this section, we highlight selected design issues and ethical considerations. Other recent systematic reviews have identified standards for evaluating the methodologic quality of published CRTs.[20,26] Table 36.1 summarises several of these issues, as modified from Eldridge and colleagues.[20] Because CRTs are more complex to design, execute and analyse, the use of a clustered design always should be clearly justified. As described in the previous section, clustering should be accounted for in sample size calculations, in estimating power for important process and outcome measures in a trial of a fixed sample size, and in trial analyses by using appropriate statistical methods. A sufficient number of clusters per intervention group should be identified and enrolled to provide adequate power for the primary analyses. Donner and Klar[37] suggest that at least four clusters per intervention group typically are necessary, although this depends on other considerations as well, and larger numbers may be required, for example, in analyses of cross-sectional designs with binary outcome measures.[38] Pre-randomisation balancing of clusters, through stratification or matching procedures, should generally be undertaken, especially with limited numbers of clusters. Finally, despite efforts to increase the number of clusters and to balance pre-randomisation baseline characteristics, imbalances on important confounders and other covariates may remain that should be accounted for analytically.

DESIGN CONSIDERATIONS

Murray[1] describes four main features that differentiate CRT designs: (1) main effects or factorial; (2) schedule of data collection; (3) cohort or cross-sectional; and (4) presence or absence of pre-randomisation matching or stratification. Each of these features is determined by the nature of the study question(s) as well as cost and other practical considerations. In turn, these design features significantly influence trial sample size and approaches to analysis. Interest in the effect of a single intervention enables the investigator to choose a relatively straightforward main effects design, while important interests in two or more interventions will cause the investigator to select a factorial design. A separate decision pertains to whether the investigator plans to gather data at discrete time intervals or to establish a form of continuous surveillance for process and outcome measures. In general, if the research

Table 36.1. Key characteristics and recommendations in design, analysis and reporting of CRTs (adapted from Eldridge *et al.*)[20]

Characteristic	Recommendation
Complexity in design, conduct, and analysis	Use of a clustered design should always be justified
Clustering – impact on effective sample size and power	Account for clustering in sample size and power calculations
Clustering – impact on inference	Account for clustering in analyses using appropriate statistical methods
Number of clusters	
– impact on pre-randomisation balancing between experimental and control groups	Carry out stratification or justify not doing so
– impact on trial power	Include at least four clusters per intervention group or justify not doing so
Unavoidable baseline imbalances	Account for baseline imbalances in analyses or justify not doing so

question involves changes over time in an entire population, a serially sampled cross-sectional design may be preferred, and the population will be sampled at discrete time intervals, i.e. the investigator will measure each intervention group at multiple points over time, although each member of each group would be measured only once.[1] If, as is frequently the case in health care delivery trials, the primary research question pertains to changes in individual patients over time, the investigator typically will choose the statistically more powerful cohort design, with multiple measures taken over time in the same patients.

Two contemporary trends in health care delivery and research foster more effective balancing of clusters before they are assigned to intervention groups. The increasing prevalence and support for practice-based research networks (PBRNs)[39] is leading to quality improvement research across multiple community-based practices, expanding the accessibility of clusters and cooperating practices for assignment to intervention groups. Simultaneously, electronic medical records with rich clinical data now can readily identify large groups of similar patients and facilitate the identification of groups of practices, before allocating them to interventions, that are better balanced on key prognostic attributes. As summarized by Raab and Butcher,[40,41] balancing may be undertaken through stratification, randomisation within matched pairs of similar groups, by adapting a minimisation procedure developed for individual RCTs, or by selecting balanced allocation from among all possible allocations of clusters to study interventions. Design, analytic and inferential considerations of such 'restricted randomisation' procedures are discussed by Rosenbaum[42] and Braun and Feng.[43]

ADDRESSING ETHICAL ISSUES

From the preceding sections, it should be clear that CRTs exist in a wide spectrum of designs to address a variety of questions and interventions at the group level. At one end of the spectrum, CRTs can resemble the more familiar individual RCT in which individuals must decide whether or not to participate. At the other end, the trial may involve the random allocation of whole practices, communities, or even countries, precluding subjects from individually deciding on participation. The UK Medical Research Council clinical trials series[44] refers to trials with structures that do not allow participation decisions by individuals as Type A trials and those that allow individual participation decisions as Type B trials. For both types of trials, the usual sets of ethical concerns apply and are reflected in internationally adopted guidelines.[45–47] The categories of concerns may be summarised as: (a) the trial must hold the potential to produce findings that may be used to improve the health and/or welfare of humans, (b) trial participants must face a favourable balance of risks and benefits, (c) when in conflict, the interests of participants must prevail over those of society and science, (d) when possible, participants must give voluntary informed consent (alternative safeguards apply when this cannot be done), and (e) the trial research proposal must be reviewed by an independent ethics review committee.

Ethical concerns for Type B trials mimic those for individual RCTs, including individual informed consent. Analogous to the concern for individuals in Type B trials, Type A trials require a mechanism for representing the interests of the cluster.[44] Certain quality improvement studies and administrative trials may require prior notification of subjects.[44,48] Administrative trials include studies that do not intrude on the ability of patients and physicians to decide about individual care as well as studies that examine aspects of the health care operation that may indirectly affect patients but about which patients do not make decisions. An illustrative administrative trial might ask what is the optimal method for overbooking patient visits to minimise unused clinic slots without impeding patient flows through the clinic. In other CRTs, individual consent concerns may be addressed through 'opt-out' strategies.[49]

SUMMARY

The CRT represents a vital approach to addressing important questions concerning the health and health care of individuals grouped loosely as patients seen by providers in clinical practices or other health care settings. Interventions commonly seek to change the behaviour of patients, often by attempting to change the behaviour of their providers or their systems of care delivery through organisational means, educational programmes, or incentives. CRTs are an increasingly important part of the methodologic armamentarium of health care researchers conducting trials of quality improvement, effectiveness and 'implementation' research.

Because the unit of randomisation is the cluster, while the unit of analytic interest is the individual patient, the CRT has unique and complex features pertaining to its design, execution and analysis. Most importantly, the CRT must account for the correlation among individuals within clusters, and the investigator needs to account for this correlation in the trial's design and analysis. Recent methodological developments, along with increasing numbers of practice-based research networks and use of electronic medical records, may foster the design of trials with better a priori balance on important prognostic features.

ACKNOWLEDGEMENTS

This chapter was supported in part by grant #R01 HS15123 of the United States Agency for Healthcare Research and Quality.

REFERENCES

1. Murray DM. *Design and Analysis of Group Randomized Trials.* New York: Oxford University Press (1998).
2. Last JM. *A Dictionary of Epidemiology*, 3rd edn. Oxford: Oxford University Press (1995).
3. Gove PB, Merriam-Webster I. *Webster's Third New International Dictionary of the English Language Unabridged.* Springfield, MA: Merriam-Webster (1986).
4. Donner A, Klar N. *Design and Analysis of Cluster Randomisation Trials in Health Research.* London: Arnold (2000).
5. Donner A, Klar N. *Design and Analysis of Cluster Randomisation Trials in Health Research.* London: Arnold (2000) 16–23.
6. Bloom BS. Controlled studies in measuring the efficacy of medical care: a historical perspective. *Int J Technol Assess Health Care* (1986) **2**(2): 299–310.
7. Eisenberg JM. Physician utilization: the state of research about physicians' practice patterns. *Med Care* (2002) **40**(11): 1016–35.
8. Stossel TP. Regulating academic-industrial research relationships–solving problems or stifling progress? *New Engl J Med* (2005) **353**(10): 1060–5.
9. Editor. Mathematics and medicine. *Lancet* (1937) **1**: 31.
10. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J* (1948) **2**: 769–82.
11. Hansen MH, Hurwitz WN. Relative efficiencies of various sampling units in population inquiries. *J Am Stat Assoc* (1942) **37**: 89–94.
12. Pollock TM. *Trials of prophylactic agents for the control of communicable diseases: a guide to their organization and evaluation*, *Monograph Series* No. 52. Geneva: WHO (1966).
13. Tuomilehto J, Nissinen A, Salonen JT, Kottke TE, Puska P. Community programme for control of hypertension in North Karelia, Finland. *Lancet* (1980) **2**(8200): 900–4.
14. Farquhar JW, Maccoby N, Wood PD, Alexander JK, Breitrose H, Brown BW Jr *et al.* Community education for cardiovascular health. *Lancet* (1977) **1**(8023): 1192–5.
15. Cebul RD. Randomized, controlled trials using the Metro Firm System. *Med Care* (1991) **29**(7 Suppl): JS9–18.
16. Cohen DI, Breslau D, Porter DK, Goldberg HI, Dawson NV, Hershey CO *et al.* The cost implications of academic group practice. A randomised controlled trial. *New Engl J Med* (1986) **314**(24): 1553–7.
17. Research on the delivery of medical care using hospital firms. Proceedings of a workshop. April 30 and May 1, 1990, Bethesda, Maryland. *Med Care* (1991) **29**(7 Suppl): JS1–70.
18. Dawson NV. Organizing the Metro Firm System for research. *Med Care* (1991) **29**(7 Suppl): JS19–25.
19. Simpson JM, Klar N, Donner A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health* (1995) **85**(10): 1378–83.
20. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* (2004) **1**(1): 80–90.

21. Murray DM. *Design and Analysis of Group Randomized Trials*. New York: Oxford University Press (1998) 6–9.
22. Wennberg JE, Freeman JL, Shelton RM, Bubolz TA. Hospital use and mortality among Medicare beneficiaries in Boston and New Haven. *New Engl J Med* (1989) **321**(17): 1168–73.
23. McPherson K, Wennberg JE, Hovind OB, Clifford P. Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway. *New Engl J Med* (1982) **307**(21): 1310–14.
24. Cutler DM, Feldman NE, Horwitz JR. US adoption of computerized physician order entry systems. *Health Aff (Millwood)* (2005) **24**(6): 1654–63.
25. Donner A, Klar N. Sample size estimation for cluster randomization designs. *Design and Analysis of Cluster Randomisation Trials in Health Research*. London: Arnold. (2000), Chapter 5.
26. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess* (1999) **3**(5): iii–92.
27. Smeeth L, Ng ES. Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC Trial of the Assessment and Management of Older People in the Community. *Controlled Clin Trials* (2002) **23**(4): 409–21.
28. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials* (2005) **2**(2): 99–107.
29. Campbell M, Grimshaw J, Steen N. Sample size calculations for cluster randomised trials. Changing Professional Practice in Europe Group (EU BIOMED II Concerted Action). *J Health Serv Res Policy* (2000) **5**(1): 12–16.
30. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval Rev* (2003) **27**(1): 79–103.
31. Parker DR, Evangelou E, Eaton CB. Intraclass correlation coefficients for cluster randomized trials in primary care: the cholesterol education and research trial (CEART). *Contemp Clin Trials* (2005) **26**(2): 260–7.
32. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* (1986) **73**: 13–22.
33. Goldstein H. *Multi-level Statistical Models*, 2nd edn. London: Arnold (1995).
34. Bryk AS, Raudenbush SW. *Hierarchical Linear Models: Application and Data Analysis Methods*. Newbury Park, CA: Sage (1992).
35. Neuhaus JM, Segal MR. Design effects for binary regression models fitted to dependent data. *Stat Med* (1993) **12**(13): 1259–68.
36. Brook RH, Ware JE Jr, Rogers WH, Keeler EB, Davies AR, Donald CA *et al*. Does free care improve adults' health? Results from a randomized controlled trial. *New Engl J Med* (1983); **309**(23): 1426–34.
37. Donner A, Klar N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *Am J Epidemiol* (1994) **140**(3): 279–89.
38. Ukoumunne OC, Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Stat Med* (2001) **20**(3): 417–33.
39. Genel M, Dobs A. Translating clinical research into practice: practice-based research networks–a promising solution. *J Investig Med* (2003) **51**(2): 64–71.
40. Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med* (2001) **20**(3): 351–65.
41. Raab GM, Butcher I. Randomization inference for balanced cluster-randomized trials. *Clin Trials* (2005) **2**(2): 130–40.
42. Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. *Stat Sci* (2002) **17**: 286–304.
43. Braun TM, Feng ZD. Identifying settings when permutation tests have nominal size with paired, binary outcome, group randomized trials. *J Nonparametr Stat* (2003) **15**: 653–63.
44. *Cluster Randomised Trials: Methodological and Ethical Considerations*, Clinical Trials Series. London: Medical Research Council (2002).
45. International Statistical Institute. *Declaration of Professional Ethics*. Voorburg: ISI (1985). http://www.cbs.nl/isi/ethics.html.
46. Council for International Organisations of Medical Sciences. *International Guidelines for Ethical Review of Epidemiological Studies*. Geneva: CIOMS (1991).
47. World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. Amended by the 52nd WMA General Assembly. *Ferney-Voltaire: WMA*. (2000).
48. Goldberg HI. Ethical issues in administrative continuous improvement. Applying the concept of prior notification to the conduct of firm trials. *Med Care* (1990) **28**(9): 822–33.
49. Littenberg B, MacLean CD. Passive consent for clinical research in the age of HIPAA. *J Gen Intern Med* (2006) **21**(3): 207–11.

37

Nursing

NICOLA CRICHTON
Institute of Primary Care and Public Health, Faculty of Health and Social Care,
London South Bank University, London, UK

INTRODUCTION

A widely held view of nursing is that it is the general care of sick people, as opposed to medical attention to their disease. Nightingale[1] in essence says that what nursing attempts to do is 'put the patient in the best condition for nature to act upon him'. The definition of nursing has been and continues to be a topic of wide debate, but a generally accepted definition of nursing, adopted by the International Council of Nurses, is that of Henderson[2] that the unique function of the nurse is 'to assist the individual, sick or well, in performance of those activities contributing to health or its recovery (or to a peaceful death) that he would perform unaided if he had the necessary strength, will or knowledge, and to help him gain independence as rapidly as possible'. This definition stresses the holistic and empowering nature of nursing care and in spite of increasing specialisation and extended roles still provides a viable definition of nursing.

In 2003 the Royal College of Nursing (RCN) launched its Defining Nursing document,[3] the culmination of research and evaluation work incorporating wide consultation with RCN members. The definition is expressed as a core with six defining characteristics. The core states that 'Nursing is the use of clinical judgement in the provision of care to enable people to improve, maintain, or recover health, to cope with health problems, and to achieve the best possible quality of life, whatever their disease or disability, until death' (RCN 2003,[3] p. 3). The six defining characteristics detail the purpose of nursing, the mode of nursing interventions, the specific domain of nursing, the focus of nursing, the value base of nursing and a commitment to partnership. These defining characteristics again stress the holistic and empowering nature of nursing and the partnership working with patients, their relatives and carers and with the multidisciplinary team of health and social care professionals.

In secondary care (hospital care) nurses are the health care professionals who have the greatest day-to-day contact with patients and provide the majority of hands-on care. In primary care the role of nurses has become more prominent with the introduction of Nurse Practitioners, who are taking on the first point of contact role previously the domain of General Practitioners

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
ISBN: 0-470-01014-2

as well as having an increased role together with Practice Nurses and District Nurses in providing monitoring, care, advice and support for those with chronic conditions such as diabetes and asthma. A nurse taking on tasks formerly undertaken by a doctor is described as extending the role of the nurse.[4,5]

The last three decades have seen the introduction of Clinical Nurse Specialists[6] who have referral caseloads from nurses and doctors and are offering specialist care and advice to both patients and their care team. The range of nurse specialists is huge, including, for example, tissue viability, continence, infection control, pain, palliative care, as well as particular condition specialists such as a lymphoma nurse specialist. These positions are described as expanding the role of the nurse: that is, enhancement of existing nursing roles through greater autonomy based on increased depth of nursing knowledge.[7]

Nursing permeates all areas of the health care arena, so potentially does nursing research and clinical trials both of and in nursing. Florence Nightingale began the tradition of research by nurses; her initial research activities focusing on the importance of a health environment in promoting physical and mental well-being of the patient.[1] This work changed the attitudes of both the military and society towards the care of the sick. However, it was not until 1952 that the journal *Nursing Research* first appeared in the United States, this being the first nursing research journal. It was even later, in 1976, that the *Journal of Advanced Nursing*, the UK's first nursing research journal, was started. Thus research carried out by nurses only really began to influence nurse training and patient care about 50 years ago.

Nurses have not conventionally led clinical research programmes. Reasons for this include difficulties with regard to competition for funding; nurse education failing to provide sufficient high-quality training and understanding of research and its conduct; and status within the clinical team, which is frequently medically led. The Briggs Report[8] suggested the need for nursing practice to be based on research and this has been emphasised in subsequent policy documents. A major political victory for nursing research in the United States was the creation of the National Center for Nursing Research in 1985. This, together with centres now established in other countries, has provided some focus for nursing research. Nursing courses in many developed countries are now generally within a university structure rather than in individual schools of nursing and this has placed nurse training within a stronger research culture. This has helped strengthen the training in understanding, using and conducting research.

The recent development of the role of Consultant Nurse,[9] although often difficult to distinguish from the clinical nurse specialist,[10] is perhaps seen as the pinnacle of a clinical nursing career[11] and has emphasised the increasingly expert and advanced nature of nursing knowledge. Such roles also include in their remit professional leadership, training and development, practice development, research and evaluation. Unfortunately there is evidence that some nurse consultants do not have the necessary background and qualifications to prepare them to achieve in all four domains of the role. Woodward *et al.*[12] identify the difficulties that some are having is fulfilling aspects of their role including the research aspect, and in addition many seemed to continue feel inferior or unequal to doctors.

Although it is now more common to find programmes of research being led and conducted by nurses, much of this research uses qualitative methodologies such as phenomenology, ethnography or grounded theory rather than clinical trials. Many would argue that qualitative approaches to research fit better with the holistic and patient-centred philosophy espoused by nursing theories. However, pressure for evidence-based care in all health and social care arenas and the gold standard status of randomised clinical trials for assessing effectiveness of care have increased the number of clinical trials both in nursing and of nursing.

Delivery of patient care is generally through a multidisciplinary team in which the nurse's role, status and level of autonomy with regard to decision making and changing care plans will vary enormously depending on the clinical specialty. Thus most research in nursing has been conducted by multidisciplinary teams that include nurses, but often involving the nurses as the caregivers and data collectors rather than as the principal investigator or leader of the research. Because of the multidisciplinary nature of research involving nurses and the fact that nurses work in almost all clinical specialties, it becomes difficult to define which clinical trials can be considered as trials in nursing. For the purposes of this chapter, examples of clinical trials in nursing will be selected from arenas in which nurses tend to have more autonomy and a greater responsibility for patient care, e.g. pain control, wound care, smoking cessation and infection control.

With the advent of expanded and extended nursing roles there has been considerable interest in demonstrating the effectiveness of nurses in these new roles, in particular to demonstrate that care delivered by nurses is equivalently good or better than the care it replaces, e.g. a clinical trial comparing the effectiveness of care by Nurse Practitioners with that provided by General Practitioners.[13] Such clinical trials can be considered as trials of nursing rather than in nursing but will also be a focus in this chapter.

CLINICAL TRIAL DESIGNS IN NURSING

The trials we are considering here are rarely drug development trials so the taxonomy of Phase I–IV trials used within the pharmaceutical industry[14] is not really relevant. This taxonomy is described in detail in other chapters of this book. The overwhelming majority of trials that are in nursing, of nursing or relevant to nursing would be considered to fall within the class of Phase III trials as they are evaluations of interventions.

In common with many therapeutic application areas, the most widely used trial design is a parallel group design, but other designs such as crossover design, cluster randomised designs, pre-test, post-test design and N-of-1 trials are sometimes used and are worthy of discussion. Almost all trials are pragmatic rather than explanatory; that is, they try to mirror real life[15] and measure the benefit the treatment produces in routine clinical practice rather than attempting to measure treatment benefit under ideal conditions.

PARALLEL GROUP DESIGN

This design involves a comparison between different groups of participants, usually two groups, who are receiving different treatments. Generally this will mean that one group of participants gets the novel intervention (the experimental group) and the other group of participants gets the standard intervention (the control or comparison group). For example, a parallel group design was used by Girou *et al.*[16] to compare the efficacy of handrubbing with alcohol-based solution with conventional handwashing with antiseptic soap, and in this study health care workers were randomised to use either the alcohol-based solution or the standard antiseptic soap.

The two groups will both be running at the same time, hence the term parallel group design: they are parallel in time. As suitable patients agree to take part in the trial they will be allocated to one or other group by a predetermined process, usually by a randomisation process, which will be considered in more detail later.

Lee *et al.*[17] conducted a study to compare the effects of music on preprocedure anxiety in patients undergoing day procedures. Patients were assigned to either the control group (usual own-choice relaxing activity) or the intervention group (listening to their choice of music in reclining chairs) depending on the day of their procedure. This experiment had parallel groups but did not randomise patients to groups; it used systematic allocation, in this case alternate day allocation, in order to avoid unnecessary disturbance in the pre-operative waiting area.

These examples of very different areas of health care in which nurses are often involved indicate the wide applicability of the parallel group design with patients allocated to treatment group by randomisation. Wakefield *et al.*[18] used a parallel group design to investigate whether a motivational interviewing intervention increased successful smoking cessation attempts by patients with cancer compared with a usual care control group. The cancer patients were randomly allocated to either the motivational interviewing intervention group or the usual care group. Roykulcharoen and Good[19] used a similar design in order to investigate the effectiveness of systematic relaxation to relieve post-operative pain. Patients who were undergoing abdominal surgery were randomly allocated to either the relaxation group, who used the relaxation intervention for 15 minutes in bed during recovery following the first ambulation after surgery, or the control group who were asked to lie quietly in bed for 15 minutes after first ambulation. In a study of treatments for head lice Hill *et al.*[20] used a parallel group design. Young people with head lice were randomly allocated either to use the Bug Buster kit or to use over-the-counter pediculicide treatments.

Parallel group designs have also been used to investigate the effectiveness of nursing. Nurse management of patients with minor illnesses was investigated by randomly allocating patients to treatment either by a specially trained nurse or by a general practitioner.[21] In another study to investigate Nurse Practitioner and General Practitioner care for patients requesting same-day consultation in primary care the patients were randomly allocated to be seen either by the Nurse Practitioner or by the General Practitioner.[13] To evaluate the effectiveness of a nurse-led clinic in primary care for secondary prevention of coronary heart disease, eligible patients were randomly allocated to either the intervention group to attend the nurse-led clinics or the control group who received the usual care.[22] To evaluate the effectiveness of trained nurses in pre-operative assessment patients were randomly assigned to be assessed pre-operatively either by an appropriately trained nurse or by a house officer.[23]

In some studies more than two groups will run in parallel. In assessing whether cognitive–behavioural interventions are effective at reducing pain and anxiety among adolescents following major orthopaedic surgery, the patients were randomised to four groups. One group got concrete–objective information only, one group got coping instruction only, one group got both concrete–objective information and coping instruction, and the final group was a control group and got the standard information about the surgery experience.[24]

As parallel group trials are comparing a new intervention with an existing intervention it seems natural to question why we need to run the groups in parallel rather than allocating all new patients to the new treatment and using past patients to provide information about the control intervention. There are several difficulties with the use of historical controls. First, there is an ethical issue that the patients treated in the past did not consent to the trial and have not given permission for information about them to be used in this way. Second, when information was recorded about their illness, treatment and response to treatment no one had thought of the trial, so it is unlikely that all the information needed for the trial will have been recorded, let alone recorded in a systematic way. It may even be difficult to check whether the past patient fulfils eligibility criteria for the trial. Finally, in our ever-changing world there may be conditions apart from the treatment that have changed between the period when the controls were treated and when recruiting to the new treatment began; for example, a change in referral practice or ancillary care may influence the response to treatment. Investigations carried out into the effect of using historical controls have shown that such an approach tends to exaggerate the benefits of the new treatment.[14] Thus it is important that the experimental and control groups are being recruited and treated at the same time.

CROSSOVER DESIGNS

In a crossover design each patient receives all the treatments in some order. Usually the order in which an individual patient receives the treatments is decided at random. In its simplest form such a trial will involve two treatments, the new treatment and the existing treatment. In this design some patients receive the new treatment for a period of time and are then changed (crossed) over to the existing treatment for a period, whilst other patients will receive the existing treatment in the first period and then change to the new treatment for the second period.[25] The main advantage of such a design is that patients are compared with themselves; that is, they act as their own control. This can be a great benefit particularly when dealing with a rather subjective and difficult-to-measure outcome such as pain.

Another advantage of crossover trials is that as each patient is effectively playing two roles, they are in both the intervention group and the comparison group, thus less participants will be required than would be needed in a parallel group study of the same potential to detect significant treatment effects if they are present (this is referred to as power). However, each participant will need to take part for longer than in a parallel group trial as they need to complete all the treatment periods, i.e. at least two periods. As patients are required for longer, the burden for each individual patient is greater and there is more likely to be a problem with drop-out in a crossover design than in an equivalent parallel group design.

Whilst there are attractive advantages of using a crossover design, such designs are only suitable for stable long-term conditions. For example, it would not be possible to use such a design for a new treatment of venous leg ulcers because we know that existing treatments are effective and we would therefore find that the leg ulcer reduced during the first treatment period. Thus the treatment being used in the second period would be applied to a smaller ulcer and this may advantage the treatment in the second period. In other words, the condition would not be equally bad in all treatment periods, so later treatments would be advantaged by having a less severe disease to treat. The only way we could ensure that the leg ulcer was equally severe in the second treatment period would be to have a time of no treatment between the two treatment periods and allow the patient to get worse again, which would clearly be both unprofessional and unethical.

Other disadvantages of the crossover design are the problems of carryover and period effects. Carryover happens if the treatment given in the first period continues to have an effect in the second period and occurs because many treatments, particularly drugs, remain active in the body for some time after the last dose is given. The effect of this carryover may interfere with the effectiveness of the treatment given in the second period, possibly enhancing or possibly suppressing its effect. One way of overcoming this problem is to have a wash-out period between the two treatment periods, which would usually mean a period without treatment; however, it is often unsafe or unethical to do this. For example it would not be safe to leave an asthmatic patient without an inhaler and it would not be ethical to leave an arthritic patient without pain relief just for the sake of the experiment. The problem of period effects is when the effectiveness of the treatment is affected by the place in the sequence that the treatment is given. For example, it might be that the treatment given first is most effective regardless of which treatment that is, because after weeks of suffering the patient suddenly has relief and when the treatments switch the patient only briefly revisits his or her original suffering, so the second period of treatment might not make the same impression of relief. Alternatively, perhaps the patient remembers better the treatment he or she had in the second period so might prefer this, particularly in situations in which there is little to choose between the treatments. If carryover is likely to occur and a wash-out period is not possible, or if period effects are likely to occur, then use of a crossover design is unlikely to be sensible. Senn[26] discusses these difficulties together with more complex problems such as period by treatment interactions, and explores alternative ways of trying to overcome these disadvantages, such as multiple crossovers.

In a study of pain control during dressing changes for burns patients Finn *et al.*[27] used a two-period, two-treatment crossover design to compare patient-controlled intranasal (PCIN) fentanyl with oral morphine. Patients who were not prescribed intravenous analgesia and requiring identical wound care procedures on two consecutive mornings were recruited and randomised to receive either PCIN fentanyl and oral placebo or oral morphine and intranasal placebo on the first day followed by the alternative on the second day. In such patients there is unlikely to be a significant improvement in the wound between one day and the next and the analgesic effect of these two treatments is short acting so will not still be effective at the time of the day 2 wound care. Thus the disadvantages of period and carry-over effects are not anticipated. Pain is difficult to measure and highly personal, with each patient having a different pain tolerance, so to be able to compare treatments within each individual patient is a major advantage.

Berg and Seidler[28] conducted a randomised crossover comparison of adhesively coupled colostomy pouching systems. Although several adhesively coupled two-piece systems are on the market, there have been few controlled trials comparing pouching system effectiveness. In this study under the supervision of ostomy care nurses in six outpatient clinics, clinical performance and patient preferences for two adhesively coupled, closed-end pouching systems were compared during normal use. By using a crossover design it was possible to allow patient preference expressed through measures of comfort, flexibility, wear time and ease of removal to be incorporated into the evaluation of the products.

Crossover trials can be run with more than two treatments. For example, Nikoletti *et al.*[29] compare the effect of using standard care, standard care plus plain ice or standard care plus flavoured ice in the prevention of oral mucositis in patients during three cycles of chemotherapy.

CLUSTER-RANDOMISED DESIGNS

This is a design which incorporates the desirable property of randomising to novel intervention or comparison intervention, but rather than randomly allocating individual participants, whole groups of participants are allocated together. For example, a study might be using participants from several general practices and we randomly allocate some of the general practices to use the novel intervention with all their patients, whilst other general practices are allocated to the control group and use the comparison intervention with all their participants. Thus we are randomising general practice each of which is a cluster of patients and we are evaluating the intervention at the patient level.

In a cluster-randomised design we must take the clustering into account both in deciding the sample size of the study and in the analysis[30,31] because participants within clusters are likely to be more similar to each other than they are to participants in other clusters. Bland[32] reports a large increase in the number of trial reports that use the term ‘cluster random’ and that in 2003 all such reports in the *British Medical Journal* did show awareness of the need to allow for clustering in the analysis, whilst in 1993 and before clustering was ignored in most trials.

Several studies of nursing practice use general practices as clusters. In a study to assess the effect of additional training of Practice Nurses and General Practitioners in patient-centred care on the life-style, psychological and physiological status of patients with newly diagnosed type II diabetes, Kinmouth *et al.*[33] randomly allocated practices to routine care or to routine care plus additional training. The effectiveness of the intervention was evaluated using type II diabetic patients in the control and intervention practices. In this example it would be impossible for a practice to contribute patients to both groups because the staff either have or have not received the additional training.

In a study to determine whether asthma specialist nurses, using a liaison model of care, reduced unscheduled asthma care Griffiths *et al.*[34] randomised general practices to either the intervention group or a control group. In the intervention group the practice had discussions with the specialist nurse about guidelines for managing asthma patients and the patients were reviewed by the specialist nurse for asthma control; telephone or face-to-face support was available if needed. In the control group the specialist nurses visited the practice to discuss the guidelines and patients had their inhaler technique checked; participants otherwise continued usual care. In this example it would not be possible for an individual practice to have randomised patients to the two treatment groups because of the high risk of contamination of the control group, nor would it be logistically possible for the specialist nurses to support all the practices.

A cluster-randomised trial of smoking cessation in pregnant women randomised practices to three groups.[35] In each group midwives in the practice delivered the allocated intervention to the pregnant smokers in the practice. The interventions were based on the transtheoretical (stages of change) model (TTM) compared with standard care. One group of practices delivered standard care, the second group delivered TTM-based self-help manuals, and the final group of practices delivered TTM-based self-help manuals plus sessions with an interactive computer program giving individualised smoking cessation advice. Again there would be high risk of contamination if a practice tried to recruit patients to more than one group, so a cluster-randomised design appears to be the most effective design for this investigation.

Studies of children are often run through schools with a school providing a cluster of children who are more similar to each other than a random sample of children would be. This increased similarity comes about because of their shared school environment: for example, the influence of the school physical activity regime, the school tuck shop policy and contact with the same set of teachers. Hamilton *et al.*[36] determined the impact of a school-based harm minimisation smoking intervention in comparison with traditional abstinence-based approaches. It would not be feasible to assess both programmes within an individual school because children in the different groups would be bound to discuss the interventions, leading to contamination.

PRE-TEST, POST-TEST DESIGNS

All experiments are to some degree pre-test, post-test designs in the sense that information is collected from patients at the start of an experiment before the intervention happens, i.e. at baseline, and then information is collected after the intervention has taken place. However, in a parallel group design or a crossover design both the baseline measurement and the after intervention measurement are made in the same patient. What is really envisaged with a pre-test, post-test design is that information is collected from people or environments in period 1, perhaps considered as a baseline period (pre-test), and then an intervention is made that affects all those that come later, and whilst measurement is made after the intervention (post-test) this will not be in the same people.

This is not the same as the idea of historical controls in that the pre-test period (control period) was not before the experiment had been conceived, so the collection of data from individuals for the purposes of the experiment can be by consent and with the same eligibility criteria and same data collection forms as for the post-test period (intervention period). However, the major difficulty remaining with this design is that other things in the environment and world around the experiment might have changed and that might affect outcome as well as the fact that the intervention has taken place.

As the change in the environment is likely to be unmeasurable why is such a design used, given there is likely to be scepticism about the results? Sometimes this might be the only way of attempting to measure an effect. For example, if there is a change in law or enforced policy then it will be impossible to evaluate the effect of the law change by randomising some to comply with the law and some to not comply. Sometimes it is not so much that the policy is changed, but if, for example, the treatment is that of providing patients with an education package, it might be difficult if the control and intervention groups are running in parallel for nurses to be sure they do not introduce components of the package even for control group patients, thus introducing contamination to the control group. A pre-test, post-test design is one way of preventing such contamination, though a cluster-randomised design would be another alternative for the education package example.

Various studies of handwashing have suggested that more accessible sinks would increase compliance with handwashing protocols; however, it is difficult to test this claim as it is usually impractical and costly to consider relocating the sinks in a unit. The rebuilding and relocation of a tertiary referral centre on the same campus offered Whitby and McLaws[37] an opportunity to investigate whether accessibility of sink location does affect compliance with handwashing protocol. In the new hospital no clinical activity could occur more than 5 metres (usually less) from a sink. Covert observation of the nursing staff on four units for a period of 24 hours spread over three days took place two months before the relocation to the new hospital (pre-test period). The covert observation was repeated in the same units in the new hospital 1 month after and 10 months after the relocation (post-test period). Using two post-test observation periods allowed the researchers to investigate whether any improvement in compliance that occurs soon after the move to the new building was sustained once the novelty of the new building had warn off. Clearly more has changed in the environment than just the sink location, potentially confounding the outcome of this study, but it is difficult to envisage how this issue could be investigated using a more robust parallel group design.

In their study of patient information after ruptured intracranial aneurysm, von Vogelsang *et al.*[38] used a pre-test, post-test design, which they called a quasi-experiment. The participants were divided into two groups, a control group who received only oral information and an intervention group who received written and oral intervention. The control group was recruited first (pre-test) and then the intervention group (post-test). The authors argue that it was necessary to recruit the control group first in order to ensure that the written information in the intervention did not provide contamination and hence bias. In this example alternative designs are possible. One alternative would be a parallel group design in which patients are allocated at random to either the control or intervention treatments; whilst this might risk contamination it does reduce the risk of selection bias. Another alternative is to use several centres and use a cluster-randomised design, which would minimise risk of contamination but relies on comparability between centres. Choosing between these designs is in essence about trading one type of potential bias for another.

N-OF-1 TRIALS

These are experiments that take place within individual patients in an attempt to find the best treatment or best dose for that individual patient. The clinician and patient are, if possible, blinded and outcomes are measured on each treatment. The treatment periods are then repeated, ideally in a randomised sequence until the clinician and patient are convinced the treatments are different and the best treatment for the patient is determined. This is rather like a crossover design within an individual patient. Such trials can be particularly useful when clinicians or patients have concerns about a treatments efficacy or safety profile. Nurses are increasingly responsible for the long-term care of patients with chronic conditions and this type of design provides a systematic way of evaluating treatment options for an individual patient so promotes individualised care.

Nurses are most likely to encounter or have been involved in N-of-1 trials for pain relief. For example, March *et al.*[39] report a series of N-of-1 trials in patients with osteoarthritis comparing a non-steroidal anti-inflammatory drug (NSAID) with paracetamol with three treatment cycles each of two weeks of paracetamol and two weeks of NSAID, the order of the treatments within each cycle being decided randomly. In the conclusion of this study, as well as reaching a clinically useful decision about treatment for 19 of 25 patients which included withdrawing patients from drugs they no longer needed, they report that patients felt they had a better understanding of their arthritis and liked taking part in discussion of their results and future treatment. Wegman *et al.*[40] also conducted a similar series of N-of-1 trials in patients with osteoarthritis comparing NSAIDs with paracetamol, but in this study five treatment cycles were used.

Skeletal muscle cramps of the leg affect many ambulatory elderly people. Woodfield *et al.*[41] report a series of N-of-1 trials of quinine versus placebo for muscle cramp. Following a two-week wash-out period each patient received three four-week treatment blocks of quinine sulphate and matched placebo capsule with an individual randomised crossover design.

Nikles *et al.*[42] examine patient perspectives and experiences of N-of-1 trials. They consider N-of-1 trials for osteoarthritis comparing paracetamol and ibuprofen and also N-of-1 trials for attention-deficit hyperactivity disorder (ADHD) comparing dexamphetamine or methylphenidate and placebo. Participants or their carers were surveyed before and after the trial and reported that their participation had led to an increased knowledge, awareness and understanding of their condition, their body's response to it and its management. This led to a sense of empowerment and control as well as improved individually focused care. Patient empowerment through involvement in making decisions about individual care fits well with nursing philosophy and the concept of patient-centred care.

CLINICAL TRIAL METHODS IN NURSING

RANDOMISATION

The process of allocating participants to treatment group is a key step in avoiding bias. As Doll[43] explains, prior to 1948 a widely used technique for allocating patients to treatment groups was to use alternate allocation. This has a major problem in that the investigator knew which treatment the next patient would receive and this might affect their decision about whether the next patient was suitable for inclusion in the trial. Thus to minimise the risk of bias it is important that the investigator recruiting patients to the trial does not know which group the patients will join, should they consent to take part in the trial. Randomisation is used to allocate patients to treatment groups in order to prevent bias.[44]

For studies that compare two treatments the most widely used method of randomisation is effectively to toss a coin to decide the treatment group once the patient has consented to take part, thus the patient has a 50:50 chance of joining either treatment group. This is known as simple randomisation. In reality a computer program is likely to be used to make the random allocation in advance of the trial starting. Once a patient has consented the investigator will either take the next envelope in the sequence, which will contain the randomised treatment allocation, or call a central office to be informed of the treatment allocated to that patient. Hill *et al.*[20], Wakefield *et al.*[18] and Girou *et al.*[16] all used simple randomisation to allocate participants to treatment groups.

If simple randomisation is used it will usually produce groups of similar but not equal size. This will not create any problem with analysing the experiment, although sample size calculations usually assume groups of equal size, so the power of the experiment may be reduced if the numbers allocated to the different treatment groups are very unequal. If the investigators want to ensure the allocation to treatments groups is equal or close to equal then they might use random permuted blocks of a set size, e.g. random permuted blocks of size 10. If the experiment has two treatment groups and uses random permuted blocks of size 10 then the randomisation is carried out so that for every 10 participants recruited, 5 will be allocated to treatment A and 5 will be allocated to treatment B. Shum *et al.*[26] and Kinley *et al.*[23] both used random permuted blocks of size 4 to allocate patients to be seen by a doctor or nurse.

Simple randomisation should, for large studies, produce treatment groups that will be comparable with regard to participant characteristics that might affect the outcome of treatment. For example, simple randomisation provided Wakefield *et al.*[18] with two treatment groups that were similar (not statistically significantly different) with regard to age, gender mix, marital status, educational level, age they began smoking, number of cigarettes smoked daily, etc., so simple randomisation has resulted in two treatment group that are similar with regard to variables that might be expected to affect the outcome of the intervention. However, simple randomisation is not a guarantee of balance, particularly in small or moderate-sized studies.

When an investigator knows that a participant characteristic is an important predictor of treatment outcome, e.g. severity of disease may have an impact on the likely outcome of treatment with more severely ill patients doing less well, the researcher might want to be more certain that the treatment groups will be balanced with respect to the important variable. There are two commonly used ways of ensuring the treatment groups are balanced with regard to particular identified variables, called stratification and minimisation.

Stratified randomisation involves using a separate randomisation list for each stratum. So for example if the research decides that age (less than 50, 50 and over) and severity of illness (severe, not severe) matter then this would create four strata: one is patients less than 50 who are severely ill, another is patients aged 50 and over who are severely ill, and so on. A randomisation is run separately for each stratum and then when a patient is recruited you identify the strata they belong to and take the next envelope in the stack for that particular stratum. Stratified randomisation was used by Murchie *et al.*[22] who used age, sex and practice as stratifying variables. Kinmouth *et al.*[33] used stratified randomisation to allocate practices to treatment groups in their cluster-randomised study. The difficulty with this approach is that if there are several variables that need to be taken into account, the number of strata increases very quickly and the process becomes unmanageable.

The technique of minimisation is often used instead of stratified randomisation when there are several stratifying variables that need to be taken into account. The investigator determines which factors they want to see equally represented in the two treatment groups. The first patient is randomised to either treatment A or treatment B. Then when each subsequent new patient is recruited their prognostic characteristics are noted and they are allocated to the group, which would minimise any differences in these factors. For example, suppose we wish to take into account three variables: gender, age (as >50 years or ≤ 50 years) and ethnicity (White, Black, Asian). Table 37.1 gives the hypothetical distribution of the characteristics of the first 30 patients entered. If the next patient recruited is female, aged 43 and White, if we were to allocate her to treatment A the imbalance in gender will decrease ($6 + 1$ versus 7), but the imbalance in age will increase ($7 + 1$ versus 5) and the imbalance in ethnicity will increase ($11 + 1$ versus 9). The way we formally decide the appropriate allocation is to sum over the three characteristics for the numbers of participants with the same characteristics as the new recruit. So for

Table 37.1. Hypothetical distribution of the characteristics of the first 30 patients entered

	Treatment group A ($n = 15$)	Treatment group B ($n = 15$)
Gender: Male	9	8
Gender: Female	6	7
Age: ≤50 years	7	5
Age: >50 years	8	10
Ethnicity: White	11	9
Ethnicity: Black	2	3
Ethnicity: Asian	2	3

$$\text{Treatment A: } 6 + 7 + 11 = 24$$

$$\text{Treatment B: } 7 + 5 + 9 = 21$$

Thus the imbalance will be minimised by allocating this new recruit to the treatment with the smallest total, so in this case we should allocate to treatment B.

Pocock,[14] Senn[45] and Altman and Bland[46] all provide further examples of how an individual patient is allocated by minimisation. A slight sophistication of the method is to introduce an element of random allocation by effectively using a loaded dice weighted in favour of the allocation that will minimise the differences. The method of minimisation is used by Roykulcharoen and Good[19] to allocate patients to the treatment groups and is used by both Griffiths *et al.*[34] and Lawrence *et al.*[35] to allocate practices to groups in their cluster-randomised trials.

Randomisation is the best method of removing selection bias, but the process can be compromised if the concealment of the randomisation is inadequate. A common method of concealing the randomisation is to use sealed envelopes. So once the participant is entered into the trial the next envelope in the sequence is opened and the participant receives that treatment regime. However, if the researcher has an opportunity to tamper with the envelopes, perhaps opening and resealing them, or if the envelope is not opaque so the assignment can be seen when held up to the light, then there is potential for bias to be introduced. Shum *et al.*[21] when discussing their methods explain that they concealed the randomisation by using sequentially numbered, non-resealable, opaque envelopes so the allocation would appear to have been adequately concealed in this study. Torgerson and Roberts[47] present examples of inadequately concealed randomisation. Hewitt *et al.*[48] reviewed the adequacy of allocation concealment in 234 trials and found that trials using inadequate concealment tended to show significant differences between treatment groups more often that trials using adequate concealment.

If participants asked to take part in a randomised controlled trial are to give truly informed consent they must understand the concept of randomisation. Featherstone and Donovan[49] investigated patients' perspectives of participation in a randomised controlled trial and found that many found the concept of randomisation difficult and that inaccurate patient information and lay interpretations of common trial terms caused confusion. They recommend that in addition to clear and accurate patient information, the participants may need an opportunity to discuss the purposes of randomisation in order to understand them fully so as to be able to give informed consent.

BLINDING

In most studies there is concern that knowledge of treatment group might have an affect on both the participant's perception of response to treatment and on the investigator's expectation and perception of response. This risk of bias can be eliminated if the participant and investigator are unaware of the treatment the participant is receiving. If both the participant and the investigator who is collecting the outcome information are unaware of the treatment group the participant is in then this is described as a double-blind study. If the participant is aware of the treatment group but the person who is assessing outcome is unaware of the group then this is described as a single-blind study.

In studies of drugs it is often possible to ensure the participant is blind to treatment group, so such studies can usually be double-blind. For example, Finn *et al.*[27] in their crossover study of wound care used a double-dummy method so that the study remained double-blind. One of the treatments was being given intranasally and the other treatment was being given orally, so when patients were receiving active PCIN fentanyl they also received placebo orally and when they received active oral morphine they received PCIN placebo. Thus neither the patient nor the investigator were aware which treatment the patient received on which day.

Sometimes it is not possible to keep the patient blind to treatment. For example, in the investigation of Nikoletti *et al.*[29] of three methods of preventing oral mucositis the patients will be aware of whether they are having ice in addition to standard care and will also know whether they are having plain ice or flavoured ice. However, the oral mucositis was assessed by a nurse; that nurse need not have been involved in the patient treatment and could therefore remain unaware of which treatment the patient received. Thus the study could be single-blind.

With the non-drug interventions more commonly used in trials in and of nursing it may be impossible to arrange for the study to be double-blind, but single-blind may be achievable. For example, in the handwashing study conducted by Girou *et al.*,[16] whilst the participants were aware of which group they were in, the microbiologist examining the culture plates was blind to the hand hygiene method used, so the study was single-blind. In the investigation of specialist nurse intervention to reduce unscheduled asthma care, Griffiths *et al.*[34] used researchers blind to the randomisation status of the general practice to obtain information from medical records and to interview participants, so this study was single-blind. Similarly the study of head lice treatments conducted by Hill *et al.*[20] was single-blind; whilst the participants were aware of the treatment they had used the study nurses who recorded the presence, number and stage of lice were unaware of the treatments used.

In studies such as Shum *et al.*[26] (2000) the patients were aware of whether a nurse or a General Practitioner had treated them. Since the main outcome measures were self-reported patient measures such as general satisfaction and health status, outcome assessment was also not blind to treatment allocation. In studies of pain or anxiety, such as Lee *et al.*[17] and Roykulcharoen and Good[19] in which the patient is not blind to treatment group and the outcome measures are patient-completed assessment tools, then it will be very difficult to have any level of blinding in the study. Generally measures such as pain scales and anxiety scales will be nurse administered and it could be arranged that the nurse is blind to treatment group, so the way the nurse administers the measures will not be affected by knowledge of treatment. However, whether this is enough to claim the study is single-blind is debatable when the patients responding to these questions are well aware of the treatment they received.

SAMPLE SIZE

In all studies consideration needs to be given to the necessary size of the study. A study needs to be large enough to have a good chance of demonstrating a treatment difference exists when indeed there truly is a clinically relevant treatment difference. However, the study should not be unnecessarily large as this would be considered a waste of money, time and patient resource, as well as potentially unethical in the sense of continuing longer than necessary to deny some patients the most effective treatment. Sample size calculations are advocated for clinical trials to guide the size of the trial.

In conducting any study in which use is made of hypothesis testing there is a risk of two types of error, the first concluding that the treatments are different when in fact they are not different effectively and the second concluding that the treatments are not different when in fact one is more effective than the other. How much chance there is of the first type of error is determined by the significance level we decide to set, so usually this is 5%. The power is a measure of how likely we are to produce a statistically significant result for a treatment difference of a given magnitude. In practical terms it indicates the ability to detect a true difference of clinical importance, so we want the power to be high. Usually power should be at least 80% and ideally higher, perhaps 95%. The higher the power, the larger the study will need to be.

In order to calculate the necessary sample size the researcher needs to decide the main outcome variable and how it will be measured, as this will affect the way the sample size is calculated. For example, will the difference between treatment groups be determined by a difference in means or by a difference in proportions? Consideration will also need to be given to the size of difference that would be clinically relevant and some assessment of the level of variability expected. If difference between treatment groups is determined by difference in mean then the size of clinical difference and the variability are often expressed as an effect size, i.e. the mean difference divided by the standard deviation. Machin *et al.*[50] provide a comprehensive book of tables for calculating sample size in different circumstances and a variety of statistical packages are available to carry out these calculations.

Hill *et al.*[20] provide an example of the specification necessary for a sample size calculation when the outcome is a proportion, in this case the proportion recovered. They specify that a clinically relevant difference to detect would be a 30% difference between the groups, anticipating 80% success on one head lice treatment and 50% success on the other. They specify a significance level of 5% and a power of 80%. In fact they recruited more participants than the 49 per group the sample size calculation suggested in order to allow for participant drop-out and those with incomplete follow-up data. It is usually a good idea to make allowance for some participant withdrawal so that the number of patients who have complete follow-up information will be sufficient for the specified study power.

Shum *et al.*[21] provide an example of a sample size calculation for a difference in mean, though rather than stating the mean difference and standard deviation they express this as an effect size of 0.2SD. Specifying a significance level of 5% and a power of 90%, they find they need 530 participants in each group; that is, 530 who complete a consultation with the nurse and 530 who are seen by a general practitioner.

If the study is to use a cluster-randomised design, then the sample size calculation is more complex and the researchers additionally need to estimate the intracluster correlation, which will be required in the calculation. Unfortunately researchers often have little information on which to base this estimate, but it will usually be small. Griffiths *et al.*[34] were able to base their estimate of 0.05 for the intracluster correlation on previous studies. In cluster-randomised trials it is mostly the number of clusters that matter, rather than the number of participants per cluster.

If the study is to investigate whether two treatments are equivalent or check that a new treatment is not inferior to an existing treatment then the study will need to be large, usually much larger than a study whose aim is to demonstrate difference. The large size is needed to ensure that we are retaining the hypothesis of no difference for the correct reason, that there really is no difference, rather than because the study is too small to demonstrate difference. Machin *et al.*[50] also provide tables to deal with this situation. Kinley *et al.*[23] wanted to demonstrate that nurses were not inferior to preregistration house officers in pre-operative assessment; the sample size calculation for this study recommended 1125 patients in each group, so a total of 2250 patients are required for this non-inferiority trial.

A common criticism directed at nursing studies of effectiveness is that they are too small.[51] Lack of evidence of a sample size justification through a sample size calculation appears still to be common in clinical trials published in nursing journals. For example, Roykulcharoen and Good,[19] Lee *et al.*[17] and Nikoletti *et al.*[29] present no evidence of a sample size calculation prior to starting the study. If there is no pre-study sample size calculated it makes interpretation of non-significant findings difficult, because we will be unable to determine whether the lack of significance is because there is no treatment effect or because the study is of low power and hence too small to provide convincing evidence of small treatment effects.

OUTCOME MEASURES

Identifying suitable measures of outcome is a challenge for many nursing clinical trials. In drug trials the treatment will specifically target a specified symptom and changes in that symptom can then be specifically measured. Whilst this will be the case in some nursing trials, in many the intention of the intervention is more diffuse than improvement in a single symptom. Nursing, as discussed in the introduction to this chapter, is about provision of care to enable patients to recover and to help them cope with health problems. These are ill-defined concepts, but to break them down into component parts and measurable features may oversimplify the impact of the intervention. It is very possible that the whole effect will be more than the sum of the measurable components.

Another challenge for nursing trials is that having identified possible outcome features to measure, the instruments for measuring that feature may be considered weak or lacking robust testing. Rarely will the outcome be as simple as a blood test or making a measurement such as blood pressure. For example, in a study of an intervention for pain, the researcher will need to select from the many available pain measurement tools, all of which require a level of patient participation and are effectively self-reported by the patient. In addition a study of a pain treatment may need to measure anxiety, another variable requiring patient self-report, in order to be clear whether the intervention is having a direct impact on pain or perhaps an indirect impact by reducing anxiety.

A further challenge is that the intervention is often something that is applied to the nurses but the effect of that intervention is assessed in the impact on patients, so the measurement of effect is several stages removed from the intervention. Of course, if the intervention is an education package for nurses one possible outcome measure is to test the nurses' knowledge before and after the intervention. However, the intention of the education is usually to improve patient care and improving nurses' knowledge does not necessarily translate into any change in practice. For example, Seers *et al.*[52] report a small clinical trial that attempts to assess the impact of nurses practising evidence-based pain management. In this study wards were randomised either to receive training and support in developing an evidence-based approach to post-operative pain management immediately, or to receive the training later. Thus there were some control wards. The impact of the intervention was assessed by comparing reported levels of patient pain before the training with reported levels of pain after the training and also by considering the variety of drugs that were used before and after the training. The challenges of running a study of such a complex intervention[53] and the confounding factors such as continual staff changes and ward reorganisation make it difficult to detect an effect at the patient level and make Seers *et al.*[52] question whether a randomised controlled trial is necessarily the most robust way of assessing the impact of a complex change when limited environmental control is possible.

ANALYSIS AND REPORTING

In general, in studies of nursing and in nursing, the analysis will be undertaken assuming intention to treat. That is, participants will be included in the analysis in the group to which they have been allocated whether or not the participant complied with that treatment. Carrying out the analysis in this way reflects what happens in the real world, participants do not always complete treatment as intended, and it will not be easy to identify at the outset those who will not comply. Thus an intention-to-treat analysis will give a more realistic estimate of the size of the treatment effect we could expect in clinical practice than analysis that omits participants that were not fully compliant with the allocated treatment.

When reporting clinical trials some nursing journals, in common with many major medical journals, advise authors to follow the reporting guidelines in the CONSORT statement of 2001.[54] Authors should, for example, be providing a flow chart showing the participants' route through the trial so that it is clear to readers how many participants were eligible, how many consented and were randomised, and when participants dropped out and why. Wakefield *et al.*[18] provide a clear flow chart showing how the 3200 patients screened were reduced to only 88 patients who completed the six-month follow-up. Clear patient flow charts also appear in Hill *et al.*,[20] Kinley *et al.*,[23] Shum *et al.*,[21] Kinnersley *et al.*[13] and Murchie *et al.*[22] accounting for participants lost to follow-up and withdrawn. Nursing journals do not seem to be enforcing their recommendation of following the CONSORT guidelines for reporting as none of Roykulcharoen and Good,[19] Lee *et al.*,[17] LaMontagne *et al.*,[24] Nickletti *et al.*[29] show participant flow charts.

CONCLUSION

Use of clinical trials in nursing has lagged behind other areas with regard to both the number of clinical trials conducted and the quality of those trials. Whilst there are areas of nursing care in which other types of research approach are appropriate, there are certainly areas that would benefit from randomised controlled trials, perhaps incorporating some qualitative investigations allowing deeper exploration of the effects of complex interventions, particularly when blinding and outcome assessment are difficult. Oakley (1998)[55] stresses that the randomised controlled trial is seen as primarily associated with medicine, but there is also a long tradition of using this design in experimental sociology from which nursing might learn. The emphasis on evidence-based health care and the status of randomised controlled trials as the highest level of evidence for effectiveness means nurses will need to continue to rise to the challenge of using controlled trials as part of their research base to demonstrate effectiveness of the care they provide.

REFERENCES

1. Nightingale F. *Notes on Nursing: What it is, and what it is not.* London: Harrison and Sons (1860).
2. Henderson V. *The Nature of Nursing*. New York: Macmillan (1966).
3. Royal College of Nursing. *Defining Nursing*, London: Royal College of Nursing, publication code 001 998 (2003).
4. DHSS *The extending role of the clinical nurse*, London: HMSO (1977).
5. Department of Health. *Developing key roles for nurses and midwives: a guide for managers*. London: Department of Health (2002).
6. Woods LP. *The enigma of advanced nursing practice*. Salisbury: Mark Allen (2000).
7. Hunt G, Wainwright P (eds). *Expanding the role of nursing*. Oxford: Blackwell Scientific (1994).
8. Briggs A. *Report of the Committee on Nursing*. London: HMSO (1972).
9. NHS Executive. *Nurse, midwife and health visitor consultants: Establishing posts and making appointments*. Health Service Circular 1999/217, London: National Health Service Executive (1999).
10. Maylor M. Differentiating between a consultant nurse and a clinical nurse specialist. *Br J Nurs* (2005) **14**: 463–8.
11. Redfern L. Clinical pinnacle. *Nurs Stand* (2003) **17**(22): 96.
12. Woodward VA, Webb C, Prowse M. Nurse consultants: their characteristics and achievements. *J Clin Nurs* (2005) **14**: 845–54.
13. Kinnersley P, Anderson E, Parry K, Clement J, Archard L, Turton P, Stainthorpe A, Fraser A, Butler CC, Rogers C. Randomised controlled trial of nurse practitioner versus general practitioner care for patients requesting ‘same day’ consultations in primary care. *Br Med J* (2003) **320**: 1043–8.
14. Pocock S. *Clinical trials: A practical approach*. Chichester: John Wiley & Sons, Ltd (1983).
15. Roland M, Torgerson DJ. Understanding controlled trials: what are pragmatic trials? *Br Med J* (1998) **316**: 285.
16. Girou E, Loyeau S, Legrand P, Oppein F, Brun-Buisson C. Efficacy of handrubbing with alcohol based solution versus standard handwashing with antiseptic soap: randomised clinical trial. *Br Med J* (2002) **325**: 362–5.
17. Lee D, Henderson A, Shum D. The effect of music on preprocedure anxiety in Hong Kong Chinese day patients. *J Clin Nurs* (2004) **13**: 297–303.
18. Wakefield M, Olver I, Whitford H, Rosenfeld E. Motivational interviewing as a smoking cessation intervention for patients with cancer: randomized controlled trial. *Nurs Res* (2004) **53**: 396–405.
19. Roykulcharoen V, Good M. Systematic relaxation to relieve postoperative pain. *J Adv Nurs* (2004) **48**: 140–8.
20. Hill N, Moor G, Cameron MM, Butlin A, Preston S, Williamson MS, Bass C. Single blind, randomised comparative study of the Bug Buster kit and over the counter pediculicide treatments against head lice in the United Kingdom. *Br Med J* (2005) **331**: 384–7.
21. Shum C, Humphreys A, Wheeler D, Cochrane MA, Skoda S, Clement S. Nurse management of patients with minor illnesses in general practice: multicentre, randomised controlled trial. *Br Med J* (2000) **320**: 1038–43.
22. Murchie P, Campbell NC, Ritchie LD, Simpson JA, Thain J. Secondary prevention clinics for coronary heart disease: four year follow up of a randomised controlled trial in primary care. *Br Med J* (2003) **326**: 84–7.
23. Kinley H, Czoski-Murray C, George S, McCabe C, Primrose J, Reilly C, Wood R, Nicolson P, Healy C, Read S, Norman J, Janke E, Alhameed H, Fernandes N, Thomas E. Effectiveness of appropriately trained nurses in preoperative assessment: randomised controlled equivalence/non-inferiority trial. *Br Med J* (2002) **325**: 1323–7.

24. LaMontagne LL, Hepworth JT, Cohen F, Salisbury MH. Cognitive-behavioral intervention effects on adolescents' anxiety and pain following spinal fusion surgery. *Nurs Res* (2003) **52**: 183–90.
25. Sibbald B, Roberts C. Understanding controlled trials: Crossover trials. *Br Med J* (1998) **316**: 1719–20.
26. Senn S. *Cross-over trials in clinical research*. Chichester: John Wiley & Sons, Ltd (1993).
27. Finn J, Wright J, Fong J, Mackenzie E, Wood F, Leslie G, Gelavis A. a randomised crossover trial of patient controlled intranasal fentanyl and oral morphine for procedural wound care in adult patients with burns. *Burns* (2004) **30**: 262–8.
28. Berg K, Seidler H. Randomized crossover comparison of adhesively coupled colostomy pouching systems. *Ostomy Wound Manage* (2005) **51**: 30–2.
29. Nikoletti S, Hyde S, Shaw T, Myers H, Kristjanson. Comparison of plain ice and flavoured ice for preventing oral mucositis associated with the use of 5 fluorouracil. *J Clin Nurs* (2005) **14**: 750–3.
30. Kerry SM, Bland JM. Statistics notes – sample size in cluster randomisation. *Br Med J* (1998) **316**: 549.
31. Kerry SM, Bland JM. Statistics notes – analysis of a trial randomised in clusters. *Br Med J* (1998) **316**: 54.
32. Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodology* (2004) **4**: 21.
33. Kinmouth AL, Woodcock A, Griffin S, Spiegal N, Campbell MJ. Randomised controlled trial of patient centred care of diabetes in general practice: impact on current wellbeing and future disease risk. *Br Med J* (1998) **317**: 1202–8.
34. Griffiths C, Foster G, Barnes N, Eldridge S, Tate H, Begum S, Wiggins M, Dawson C, Livingstone AE, Chambers M, Coats T, Harris R, Feder G. Specialist nurse intervention to reduce unscheduled asthma care in a deprived multiethnic area: the east London randomised controlled trial for high risk asthma (ELECTRA). *Br Med J* (2004) **328**: 144–8.
35. Lawrence T, Aveyard P, Evans O, Cheng KK. A cluster randomised controlled trial of smoking cessation in pregnant women comparing interventions based on the transtheoretical (stages of change) model to standard care. *Tob Control* (2005) **12**: 168–77.
36. Hamilton G, Cross D, Resnicow K, Hall M. A school-based harm minimisation smoking intervention trial: outcome results. *Addiction* (2005) **100**: 689–700.
37. Whitby M, McLaws ML. Handwashing in healthcare workers: accessibility of sink location does not improve compliance. *J Hosp Infect* (2004) **58**: 247–53.
38. Von Vogelsang AC, Wengstrom Y, Forsberg C. Patient information after ruptured intracranial aneurysm. *J Adv Nurs* (2004) **48**: 551–9.
39. March L, Irwig L, Schwarz J, Simpson J, Chock C, Brooks P. N of 1 trials comparing a nonsteroidal anti-inflammatory drug with paracetamol in osteoarthritis. *Br Med J* (1994) **309**: 1041–6.
40. Wegman ACM, van der Windt DAWM, de Haan M, Deville WLJM, Chin A Fo CT, de Vries TPGM. Switching from NSAIDS to paracetamol: a series of n of 1 trials for individual patients with osteoarthritis. *Ann Rheum Dise* (2003) **62**: 1156–61.
41. Woodfield R, Goodyear-Smith F, Arroll B. N-of-1 trials of quinine efficacy in skeletal muscle cramps of the leg. *Br J Gen Prac* (2005) **55**: 181–5.
42. Nikles CJ, Clavarino AM, Del Mar CB. Using n-of-1 trials as a clinical tool to improve prescribing. *Br J Gen Pract* (2005) **55**: 175–80.
43. Doll R. Controlled trials: the 1948 watershed. *Br Med J* (1998) **317**: 1217–20.
44. Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *Br Med J* (1999) **318**: 1209.
45. Senn S. *Statistical issues in drug development*. Chichester: John Wiley & Sons, Ltd (1997).
46. Altman DG, Bland JM. Treatment allocation by minimisation. *Br Med J* (2005) **330**: 843.
47. Torgerson DJ, Roberts C. Understanding controlled trials: Randomisation methods: concealment. *Br Med J* (1999) **319**: 375–6.
48. Hewitt C, Hahn S, Torgerson DJ, Watson J, Bland JM. Adequacy and reporting of allocation concealment: review of recent trials published in four general medical journals. *Br Med J* (2005) **330**: 1057–8.
49. Featherstone K, Donovan JL. Random allocation or allocation at random? Patients' perspectives of participation in a randomised controlled trial. *Br Med J* (1998) **317**: 1177–80.
50. Machin D, Campbell M, Fayers P, Pinol A. *Sample size tables for clinical studies*, 2nd edn. Oxford: Blackwell Science (1997).
51. Thomas LH, Bond S. The effectiveness of nursing: a review. *J Clin Nurs* (1995) **4**: 143–51.
52. Seers K, Crichton N, Carroll D, Richards S, Saunders T. Evidence-based postoperative pain management in nursing: is a randomized-controlled trial the most appropriate design? *J Nurs Manage* (2004) **12**: 183–93.

53. Hawes P, Sheill A, Riley T. Complex interventions: how ‘out of control’ can a randomised control be? *Br Med J* (2004) **328**: 1561–63.

54. Altman DG, Schulz KF, Moher D for the CONSORT group. The revised CONSORT statement for reporting randomised trials: explanation and elaboration. *Ann Intern Med* (2001) **134**: 663–94.

55. Oakley A. Experimentation and social interventions: a forgotten but important history. *Br Med J* (1998) **317**: 1239–42.

SPECIAL POPULATIONS

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

38

Clinical Trials in Paediatrics

JOHAN P.E. KARLBERG
Clinical Trials Centre and Department of Paediatrics,
The University of Hong Kong, Hong Kong SAR

WHY SHOULD WE DO CLINICAL TRIALS IN CHILDREN?

Children are subject to many of the same diseases as adults, and are often treated with the same drugs and biological products. However, many drugs on the market used to treat children are inadequately labelled for use with paediatric patients; and many carry disclaimers stating that safety and effectiveness in paediatric patients have not been established. Information about the safety and effectiveness of treatments for some paediatric age groups is particularly difficult to find. Even today, no treatment is available for many of the thousands of rare and serious diseases that largely affect neonates, infants and children. Most drugs used to treat common diseases in both children and adults have not been investigated in children at all. Over 50% of all drugs prescribed in paediatric practice are either 'unlicensed' or 'off label'.

The paediatric medical community has for decades tried to persuade regulatory authorities and the pharmaceutical industry to test new drugs in the paediatric population in parallel with the adult studies. The motto of the campaign has been 'Children are not simply Small Adults' and its *Guidelines for the Ethical Conduct of Studies to Evaluate*, published in 1995, reported that:

- In 1973, 78% of medications included a disclaimer or lack of dose information for children.
- In 1991, 81% of listed drugs were restricted for certain age groups.
- In 1992, 79% of 19 new molecular entities approved were not labelled for use in children.

As a result of effectively being denied access to well-studied drugs, paediatricians either do not treat children with potentially beneficial medications, or treat them with medications based either on adult studies or anecdotal empirical experience in children. Such non-validated administration of medications may place more children at risk than if the drugs were administered as part of well-designed, controlled clinical trials. There is therefore a moral imperative to formally study drugs in children, so they can enjoy equal access to existing as well as new therapeutic agents.

The US National Institutes of Health (NIH) published regulations in 1999 clearly defining what human studies could be funded by NIH

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

to the exclusion of paediatric subjects. The exclusionary circumstances were:

- Research topic irrelevant to children;
- Laws or regulations barring inclusion of children;
- The knowledge is available for children or will be obtained from another ongoing study;
- The relative rarity of the condition in children;
- The number of children is limited;
- Insufficient data are available in adults to judge potential risk in children.

Not until recently have children been more regularly included in clinical studies to investigate drugs. Considerable differences between the pharmacokinetics and pharmacodynamics of drugs in children and in adults frequently make it impossible to bridge conclusions from data obtained in adults. Children cannot even be considered a homogeneous group, since age groups differ in their absorption, distribution, metabolisation and excretion of drugs and their effect on developing organ systems. The anatomical structure of children's organs differ from adults, causing different pharmacodynamic characteristics observed during childhood.

The lack of paediatric safety information in product labelling exposes paediatric patients to the risk of age-specific adverse reactions unexpected from adult experience. The absence of paediatric testing and labelling may also expose paediatric patients to ineffective treatment through under-dosing, or may deny paediatric patients therapeutic advances. Failure to develop a paediatric formulation of a drug or biological product, where younger paediatric populations cannot take the adult formulation, may also deny paediatric patients access to important new therapies.

Three conclusions can therefore be drawn about paediatric drug studies: studies must be made in different age groups; describing the pharmacokinetics and pharmacodynamics is crucial; and the safety of drugs must be studied to identify potential severe side effects.

REGULATORY ISSUES OF CLINICAL TRIALS IN CHILDREN

Regulatory authorities in the US and Europe have in recent years taken important steps to address the problem of inadequate paediatric testing and inadequate paediatric use information in drug and biological product labelling. But these efforts have, thus far, not substantially increased the number of products entering the marketplace with adequate paediatric labelling. The regulatory authorities have therefore concluded that additional steps are necessary to ensure the safety and effectiveness of drug and biological products for paediatric patients. Manufacturers of new and marketed drugs and biological products must now evaluate the safety and effectiveness of the products in paediatric patients if the product is likely to be used in a substantial number of children, or provide a more meaningful therapeutic benefit to paediatric patients than existing treatments.

Since 2000 in both the US and Europe, pharmaceutical companies have been obliged to include paediatric data in all new drug applications and licence extensions provided that substantial use of the drug in children and a meaningful therapeutic benefit are expected. The strength of this legislation is, however, different in the two regions – and so is the extension of market exclusivity.

In recent years an independent 'Orphan' drug regulation has been in force in the countries of the European Community as well as in the US. This creates incentives for the development of drugs for rare serious diseases, but is unlikely to achieve effective improvement in paediatric drug therapy. The Food and Drug Administration (FDA) Modernization Act established economic incentives for pharmaceutical manufacturers to conduct paediatric studies on drugs for which patent protection or exclusivity is available under the Drug Price Competition and Patent Term Restoration Act or the Orphan Drug Act. These provisions attach six additional months of marketing exclusivity to any existing exclusivity or patent protection of a drug for which the FDA has requested paediatric studies.

However, there is likely to be a consensus during the coming years – at least in the International Conference on Harmonisation (ICH) GCP regions – over requirements for conducting clinical trials on new drugs and other therapies in children. But before this consensus can be reached, a number of points have to be addressed and discussed, underlined by the following two examples.

EXAMPLE 1 – ONGOING DISCUSSIONS OF THE CONSENT PROCESS IN PAEDIATRIC TRIALS

The significant increase in the number of children participating in clinical trials continues to raise ethical and procedural concerns. The FDA addressed this issue in April 2001, calling on institutional review boards to review study protocols that include children and ensure they adopt safeguards to protect young research participants. A group in the US is currently examining the 'best practices' related to research involving children. The study will address:

- Process for obtaining informed consent and assent from children and their parents or legal representatives.
- How well participants in paediatric studies and their guardians understand direct benefits and risks of study involvement.
- Definition of 'minimal risk' related to healthy and ill child study participants.
- Whether regulations and policies should vary for children of different ages (for example, teenagers and infants).
- Appropriateness of payments to children, parents, or legal representatives for participation in research.
- Role of IRBs in monitoring compliance with regulations related to paediatric studies.

EXAMPLE 2 – ONGOING DISCUSSIONS OF THE LEGISLATION OF PAEDIATRIC TRIALS

Based on feedback from a consultation document, the European Commission was expecting to prepare draft legislation on paediatric medicinal development by Autumn 2002. This legislation is considered by many to be pressing, creating the conditions needed to improve medicines for children. Nearly all involved parties in Europe supported a legal and regulatory framework for improving child health, especially regarding the labelling of medicines. The consultation document concluded:

- A robust ethical framework for European paediatric research needs to be created, including guidance for informed consent, ethical review, recruitment of subjects, and safety and oversight.
- A robust paediatric clinical study infrastructure needs to be created in Europe, since as a result of reluctance to perform such studies up to now, there is a serious shortage of trained and experienced people and centres of excellence.
- Greater cooperation should be stimulated between public sector research and private sector research in paediatrics, in the interest of developing a European dimension to improving medicines for children.
- A clear framework should be developed for assembling international data and information regarding paediatric trials and medicines – to ensure that unnecessary trials are not carried out in Europe, and that European paediatricians have the benefit of up to date and comprehensive information regarding medicinal products for their patients, wherever in the world that information has been generated.
- A greater public dialogue is required in Europe regarding the benefits and risks of paediatric research for individual children participating in research, as well as for public health in general.

POST-LAUNCHING THE ICH GUIDELINE E11

The E11 ICH Guideline – Clinical Investigation of Medicinal Products in Paediatric Population – came into operation in January 2001. This guideline has been the basis for the development of important legal documents addressing the conduct of clinical trials in paediatric populations.

For instance, one example is the US '2002 Best Pharmaceuticals for Children Act (BPCA)' that became law on 4th January 2002. This Act establishes an additional mechanism for obtaining information in the paediatric populations for off-patent drugs. It provides a mandate for the FDA and the NIH in the US to collaborate in the study of 'off-patent' and 'on-patent' drugs that industry does not want to study. Another example is the '2003 Paediatric Research Equity Act (PREA)' that became law in the US on 3rd December 2003. This Act requires paediatric studies of certain drugs and biological products; new indications; new dosage forms; new routes of administration; new dosing regimens; and new active ingredients. Table 38.1 indicates the number of Proposed Paediatric Study Requests (PPSRs) submitted to the FDA under those two new Acts, i.e. the so-called 'Paediatric Exclusivity' requests. It is possible in the US to obtain an additional six months of exclusivity if the sponsor submits requested (PPSR) information relating to the use of the drug in the paediatric population. As seen from Table 38.1, the largest number of requests is in the area of metabolic, endocrine, neuropharmacological and cardio-renal drugs.

The ICH E11 Guideline has also had a significant influence on the drafting of the European Commission proposal for a Regulation of the Council and of the Parliament on Medicinal Products for Paediatric Use $^{(2004/0217(COD))}$. This regulation was formally adopted on 13th July 2005 and it addresses the following key elements;

- a six-month extension of a patent/supplementary protection certificate (SPC);
- an increase of exclusive commercial rights of 'orphan drugs', intended to treat rare illnesses, from 10 to 12 years (if invented specifically for children);
- establishment of a Medicines Investigation for the Children of Europe (MICE) Fund, a special EU programme for research into medicines for children;
- establishment of a network of researchers and research centres under the supervision of the European Medicines Agency to avoid duplication of research and tests on children;
- establishment of a Paediatric Committee as the 'cornerstone' of the European paediatric R&D system.

Table 38.1. The number of Proposed Paediatric Study Requests (PPSRs) submitted to the US FDA under 'Paediatric Exclusivity', as of 30th June 2005

Drug product	Requests, *N*
Cardio-renal	39
Neuropharmacological	52
Oncology	22
Medical imaging and radiopharmaceutical	4
Anaesthetic, critical care and addiction	25
Gastrointestinal and coagulation	33
Metabolic and endocrine	62
Anti-infective	7
Anti-viral	25
Dermatologic and dental	23
Anti-inflammatory, analgesic and ophthalmologic	33
Over-the-counter	6
Pulmonary	18
Reproductive and endocrine	11
Special pathogen and immunologic	16
Total	376

With this recent legislation in the US and EU, we will be able to speed up the process to identify more effective and safe drugs for use in children. Already, after a few years of trial experiences in children, we have identified some key elements specifically of importance for drug administration in children such as:

- Pharmacokinetics are variable in children, more so than first anticipated.
- Adverse reactions that are paediatric specific are being defined.
- Trial designs are being modified.
- Ethical issues have to be reassessed from the paediatric perspective.

WHEN INITIATING A PAEDIATRIC PROGRAMME (SUMMARY POINTS OF ICH GCP E11)

The decision to proceed with a paediatric development programme for a certain medicinal product requires consideration of factors such as:

- Prevalence of the condition in the paediatric population;
- Seriousness of the condition;
- Availability and suitability of alternative treatments;
- Unique paediatric indications;
- Unique paediatric-specific endpoints;
- Age ranges of paediatric patients likely to be treated;
- Unique paediatric safety concerns;
- Unique paediatric formulation development.

The most common considerations when discussing the need and timing of a paediatric programme are:

- Most important is the presence of a serious or life-threatening disease for which the medicinal product represents a potentially important advance in therapy. This situation suggests relatively urgent and early initiation of paediatric studies.
- For medicinal products for diseases predominantly or exclusively affecting paediatric patients, the entire development programme will be conducted in the paediatric population, except for initial safety and tolerability data, which will usually be obtained in adults.
- For medicinal products intended to treat serious or life-threatening diseases occurring in both adults and paediatric patients, for which there are currently no (or limited) therapeutic options, there is need for relatively urgent and early initiation of paediatric studies.
- For medicinal products intended to treat other diseases and conditions there is less urgency. Trials would usually begin at later phases of clinical development or, if a safety concern exists, even after a substantial post-marketing period in adults. Testing of these medicinal products in the paediatric population would usually not begin until Phase II or III – since very early initiation of testing in paediatric patients might needlessly expose them to a compound of no benefit.

TYPES OF STUDIES (SUMMARY POINTS OF ICH GCP E11)

Selection of the type of study should be on the same principles as studies planned for adults. However, several considerations are of specific importance for paediatric studies. Some of the most important are:

- When a medicinal product is to be used in the paediatric population for the same indication(s) as those studied and approved in adults, the disease process is similar in adults and paediatric patients, and the outcome is likely to be comparable, extrapolation from adult efficacy can be appropriate. In such cases, pharmacokinetic (PK) studies in all the age ranges of paediatric patients likely to receive the medicinal product, together with safety studies, may provide adequate information.
- When a medicinal product is to be used in younger paediatric patients for the same indication(s) as those studied in older paediatric patients, the disease process is similar, and the outcome is likely to be comparable, extrapolation of efficacy from older to younger paediatric patients may be possible. In such cases, pharmacokinetic studies in the relevant age groups of paediatric patients together with safety studies may be sufficient.
- Many diseases in preterm and term newborn infants are unique or have unique manifestations precluding extrapolation of efficacy from older paediatric patients and call for novel methods of outcome assessment.
- Where the disease course/outcome of therapy in paediatric patients is expected to be similar to adults, but the appropriate blood levels are not clear, it may be possible to use measurements of a pharmacodynamic (PD) effect related to clinical effectiveness. Thus,

a PK/PD approach combined with safety and other relevant studies could avoid the need for clinical efficacy studies.
- When unique indications are being sought for the medicinal product in paediatric patients, or when the disease course and outcome of therapy are likely to be different in adults and paediatric patients, clinical efficacy studies in the paediatric population are needed.

Pharmacokinetics

PK studies generally should be performed to support formulation development and determine PK parameters in different age groups. PK studies in the paediatric population are generally conducted in patients with the disease. Single-dose or steady-state studies are the choice of PK study:

- For medicinal products that exhibit linear pharmacokinetics in adults, single-dose PK studies in the paediatric population may be sufficient.
- When there is a nonlinearity in absorption, distribution and elimination in adults and difference in duration of effect between single and repeated dosing in adults suggests steady-state studies in the paediatric population.

Special considerations should be taken when blood is drawn more than once in paediatric subjects, such as in PK/PD studies. Several approaches can be used to minimise the amount of blood drawn and/or the number of venipunctures:

- Use of sensitive assays;
- Use of laboratories experienced in handling small volumes;
- Using routine clinical blood samples for PK analysis;
- Use of indwelling catheters;
- Use of population pharmacokinetics and sparse sampling.

Efficacy

The principles in study design, statistical considerations and choice of control groups are detailed in other ICH guidelines and apply to paediatric efficacy studies. But there are also certain features unique to paediatric studies.

- Extrapolation of efficacy from studies in adults to paediatric patients, or from older to younger paediatric patients, as mentioned above.
- For efficacy studies it may be important to employ different endpoints for specific age groups.
- Measurement of subjective symptoms requires different assessment instruments for patients of different ages.
- The response to a medicinal product may vary among patients because of the developmental stage of the patient.

Safety

ICH guidelines (E2 and E6) describe adverse event reporting and apply to paediatric studies. But there are certain safety aspects unique to paediatric studies.

- Medicinal products may affect physical and cognitive growth and development, and the adverse event profile may differ in paediatric patients, compared with adults.
- The dynamic processes of growth and development may not manifest an adverse event at once, but at a later stage of growth and maturation.
- Long-term studies or surveillance data may be needed to determine possible effects on skeletal, behavioural, cognitive, sexual and immune maturation and development.
- Post-marketing surveillance may provide important safety and/or efficacy information for the paediatric population.
- Age-appropriate, normal laboratory values and clinical measurements should be used in adverse event reporting.

AGE CLASSIFICATION OF PAEDIATRIC PATIENTS (SUMMARY POINTS OF ICH GCP E11)

Decisions on how to stratify studies and data by age need to take into consideration developmental biology and pharmacology. The identification of which ages to study should be medicinal product-specific and justified.

- *Preterm newborn infants:* Preterm newborn infants have a unique pathophysiology and responses to therapy. The complexity of and ethical considerations involved in studying preterm newborn infants requires a careful protocol development with expert input from neonatologists and neonatal pharmacologists. Only rarely can we extrapolate efficacy from studies in adults or in older paediatric patients to the preterm newborn infant.
- *Term newborn infants (0 to 27 days):* Newborn infants are more mature than preterm newborn infants, but many of the physiologic and pharmacologic principles for preterm infants also apply to them.
- *Infants and toddlers (28 days to 23 months):* This is a period of rapid CNS maturation, immune system development and total body growth. By 1–2 years of age, clearance of many drugs on a mg/kg basis may exceed adult values and then it may be dependent on specific pathways of clearance.
- *Children (2 to 11 years):* Most pathways of drug clearance are exceeding adult values. Changes in clearance of a drug may be dependent on maturation of specific metabolic pathways. The protocols should ascertain assessment of the effect of the medicinal product on growth and development. Recruitment of patients should ensure adequate representation across the age range in this category. Puberty can affect the activity of enzymes that metabolise drugs, and dose requirements for some medicinal products may decrease dramatically.
- *Adolescents (12 to 16–18 years (dependent on region)):* This is a period of sexual maturation and medicinal products may interfere with the actions of sex hormones. Medicinal products and illnesses that delay or accelerate the onset of puberty can have a profound effect and may affect final height. Many diseases are also influenced by the hormonal changes around puberty and hormonal changes may thus influence the results of clinical studies. Non-compliance is a special problem and compliance checks are important.

ETHICAL ISSUES IN PAEDIATRIC STUDIES (SUMMARY POINTS OF ICH GCP E11)

The paediatric population represents a vulnerable subgroup. Therefore, the following special measures are needed to protect the rights of paediatric study participants.

- Participants in clinical studies are expected to benefit from the clinical study, except under special circumstances.
- When protocols involving the paediatric population are reviewed, there should be IRB/IEC members or experts consulted by the IRB/IEC who are knowledgeable in paediatric ethical, clinical and psychosocial issues.
- Paediatric study participants are dependent on their parent(s)/legal guardian to assume responsibility for their participation in clinical studies. Participants of appropriate intellectual maturity should personally sign and date either a separately designed, written assent form or the written informed consent.
- Information that can be obtained in a less vulnerable, consenting population should not be obtained in a more vulnerable population or one in which the patients are unable to provide individual consent.
- Studies in handicapped or institutionalised paediatric populations should be limited to diseases or conditions found principally in these populations, or when it is expected that the disease may alter the effects of a medicinal product.
- To minimise risk in paediatric clinical studies, those conducting the study should be trained and experienced in studying the paediatric population, including the evaluation and management of potential paediatric adverse events.
- In designing studies, every attempt should be made to minimise the number of participants and of procedures, consistent with good study design.

- To ensure that experiences of the study subjects are positive and to minimise discomfort and distress.

ETHICAL CONSIDERATIONS IN PAEDIATRIC STUDIES

Of all the problems surrounding research in children, the one that poses perhaps the most complex question is research ethics. Children are not legally able to provide consent and the extent to which children are able to understand the meaning, risks and potential benefits of participating in clinical trials varies enormously according to age and background. For this reason it may be appropriate to address some points related to the IRB review, including the informed consent process, in paediatric trials more specifically than outlined in the ICH GCP E11 guideline. One document that addresses this topic at more depth is the Review and Award Codes for the NIH Inclusion of Children Policy from 1999. The following partly originates from this document, but also incorporates sources listed at the end of this chapter.

First studies that promise no demonstrable benefits to the child participating in the study or to children in general should not be conducted, irrespective of the minimal nature of the attendant risks. The risks include discomfort, inconvenience, pain, fright, separation from parents or surroundings, effects on growth and development of organs, and size or volume of biological samples.

The proposed research must be of value to children in general and, in most instances, to the individual child subject:

- The research design must take into consideration the unique physiology, psychology and pharmacology of children and their special needs and requirements as research subjects.
- The design should minimise risk while maximising benefits.
- The study design must take into account the racial, ethnic, gender and socioeconomic characteristics of the children and their parents.
- A placebo/observational control group may be acceptable when there is no commonly accepted therapy, or the commonly used therapy is of questionable efficacy, or the commonly used therapy has high frequency of side effects, i.e. larger than the benefits.

PAEDIATRIC INFORMED CONSENT

Children are not legally able to provide consent and the extent to which children are able to understand the meaning, risks and potential benefits of participating in clinical trials varies enormously according to age and background. Children are counted as members of a vulnerable population at risk for exploitation and are given special protection in clinical research. In paediatric trials, just as in adult trials, materials in an understandable language, opportunities to discuss the trial, and freedom to withdraw without penalty must be provided to potential subjects.

Investigators are ultimately held responsible for ensuring adequate informed consent. More than two decades of enquiry into the process of consent have shown that adults are less than adequately informed about risks, benefits and participation in research. The process is even more problematic for research involving individuals with limited abilities in decision-making. The evolving psychological and emotional development of children and adolescents presents challenges to paediatric investigators not encountered when dealing with adult subjects. Unless opposing evidence is identified, capacity to understand and provide informed consent has long been assumed in adults. Results from studies in healthy and sick children suggest that also children have this capacity. Several investigators have evaluated the degree to which minors from school age through adolescence are capable of providing assent. Even very young children demonstrate inquisitiveness about the proposed research. By the age of 9, children can understand purpose, risk and the right to withdraw from the study. Even 7-year-old children can understand the purpose of a study. Such observations support the requirement by most ethics boards that assent be obtained in children aged 7 and older. However, information regarding scientific versus therapeutic study objectives for both research and alternative treatment is not well understood in 7 to 20-year-old subjects. Paediatric subjects can thus provide an informed agreement to participate, but the assent process should be conducted using discussions that encourage questions.

Obtaining Informed Permission – Assent – to Participate

Regulations permit studies involving minimal risk in children, with the provision that permission from parents and assent from subjects are obtained. Research involving greater than minimal risk, but providing potential direct benefit to the child, is also permitted with the same provision. There are some exceptions to the requirement for assent and consent. Assent is not necessary for research expected to directly benefit the child. Assent must be an active affirmation from any child with an intellectual age of 7 years or older. Assent should be obtained from children who are competent to understand; and the purpose, risks and benefits of a study should be explained to them. The following guideline has therefore been proposed:

No greater than minimal risk

- Assent of the child and permission of at least one parent.

Greater than minimal risk and prospect of direct benefit

- Assent of the child and permission of at least one parent.
- Anticipated benefit justifies the risk.
- Anticipated benefit is as least as favourable as alternative approaches.

Greater than minimal risk and no prospect of direct benefit

- Assent of the child and permission of both parents.
- Likely to yield generalisable knowledge about the child’s disorder.

SPECIFIC PROBLEMS OF PAEDIATRIC STUDIES

SUBJECT RECRUITMENT

Insufficient enrolment of children is the most common reason for discontinuing paediatric studies. Creating and expanding networks for paediatric pharmacology studies, such as in the US and Europe, are steps in the right direction to recruit enough subjects. Many reasons for this poor recruitment rate for paediatric studies include:

- Strict inclusion and exclusion criteria;
- Limited size of the paediatric population;
- That each age group has to be considered separately;
- Inconvenience for the parents in having their children participate in a clinical study;
- Fear of making one’s own child available as a ‘guinea pig for research’.
- Doctors are wary of jeopardising the doctor–patient relationship, or losing the trust of parents.

EARLY TESTING

There are no healthy paediatric volunteers. The lack of volunteers for Phase I studies is a special problem and makes the planning of therapeutic studies in children difficult. The requirements for paediatric study designs are for this and other reasons different from studies in adults. Alternative study designs and alternative statistical methods are required.

STUDY MANAGEMENT

To obtain a sufficient number of subjects requires a large number of study centres. Moreover, the cost for each individual step of a paediatric study is usually higher than for studies in adults – both to pharmaceutical companies, as sponsors of the studies, and to the participating doctors. For instance, explaining the nature of a study – to obtain permission from parents and ensure their cooperation during the course of the study – is a very time-consuming process. Explanatory material and information has to be not only prepared for parents, but also adapted for the children. Caring for the children during their visits to the study centre also requires creativity, patience and time.

FINAL COMMENTS

Faced with heavy workloads, paediatricians may often be reluctant to assume what looks like the extra work of clinical trials. But a shortage of investigators is not the only problem that slows paediatric trials. It takes many subjects to satisfy the requirements for an adult drug to be adequately studied in children – and frequently the population of paediatrics with a certain disease does not exist. So not only do studies need to be designed to use small populations efficiently; they also need to be designed with children in mind. Just taking adult protocols, then changing the age in the inclusion criteria and the dose, is not good enough. With a limited number of investigators and a limited number of potential subjects, study design is critical for successful development of new safe and life-saving therapeutic entities for paediatric usage.

FURTHER READING

PUBLICATIONS

- Shirkey H. Therapeutic orphans (editorial comment). *J Pediatr* (1968) **72**: 119–20.
- Rogers LC, Cocchetto DM. Labeling prescription drugs for pediatric use in the United States. *Appl Clin Trials* (1997) **10**: 50–6.
- Tarnowski KJ, Allen DM, Mayhall C, Kelly PA. Readability of pediatric biomedical research informed consent forms. *Pediatries* (1990) **85**: 58–62.
- Susman EJ, Dorn LD, Fletcher JF. Participation in biomedical research: the consent process as viewed by children, adolescents, young adults, and physicians. *J Pediatr* (1992) **121**: 547–52.
- Ethical Conduct of Studies to Evaluate Drugs in Pediatric Populations. Committee on Drug. *Pediatrics* (1995) **95**: 286–94.
- Wechsler J. Science, pediatric studies, and surrogates. *Appl Clin Trials* (1999) **12**: 28–33.
- Nahata MC. Lack of pediatric drug formulations. *Pediatrics* (1999) **104**: 607–9.
- Conroy S, McIntyre J, Choonara I. Unlicensed and off label drug use in neonates. *Arch Dis Child Fetal Neonatal Ed* (1999) **80**: F142–5.
- Conroy S, McIntyre J, Choonara I, Stephenson T. Drug trials in children: problems and the way forward. *Br J Clin Pharmacol* (2000) **49**: 93–7.
- Kauffman RE. Clinical trials in children: problems and pitfalls. *Paediatr Drugs* (2000) **2**: 411–18.
- Simar MR. Pediatric drug development; the International Conference of Harmonization Focus on Clinical Investigation in Children. *Drug Inf J* (2000) **34**: 809–19.

OFFICIAL DOCUMENTS/GUIDELINES

- Review and award codes for the NIH inclusion of children policy March 26, 1999.
- Regulation (EC) No. 141/2000 of the European Parliament and the Council of 16 December 1999 on orphan medicinal products.
- World Medical Association, Declaration of Helsinki (amended October 2000).
- Food and Drug Administration, Department of Health and Human Services, The Pediatric Exclusivity Provision January 2001 Status Report to Congress.
- Food and Drug Administration, Additional Safeguards of Children in Clinical Investigations of FDA-Regulated Products: Interim Rule, Federal Register 66 (79) 20589 (24 April 2001).
- Food and Drug Administration, Update of list of approved drugs for which additional paediatric information may produce health benefits in the paediatric population (May 20, 2001), Docket No. 98N0056ICH Topic Ell, 'Clinical Investigation of Medicinal Products in the Pediatric Population'. CPMP/ICH/2711/99, January 2001, adopted July 2000.
- European Commission, Better Medicines for Children: Proposed regulatory actions on paediatric medicinal products (European Commission, Brussels, Belgium, 2002).
- 2002 Best Pharmaceuticals for Children Act (BPCA) became law in the US on 4 January 2002.
- 2003 Pediatric Research Equity Act (PREA) became law in the US on 3 December 2003.
- REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on medicinal products for paediatric use and amending Regulation (EEC) No. 1768/92, Directive 2001/83/EC and Regulation (EC) No. 726/2004, 29 September 2004 and formally adopted 13 July 2005 (2004/0217 (COD)).

39

Clinical Trials Involving Older People

CAROL JAGGER[1] AND ANTONY J. ARTHUR[2]
[1]Trent Institute for HSR, University of Leicester, Leicester LE1 6TP, UK
[2]School of Nursing, University of Nottingham, Queen's Medical Centre, Nottingham NG7 2UH, UK

INTRODUCTION

As few diseases or conditions present for the first time in later life, there are few treatments prescribed solely to older people. There is also little consensus on which population constitutes the 'elderly' or who can be accurately defined as an 'older person' since ageing is a continuous process from birth until death. However, the prevalence of a large number of physical and mental health problems increases with age and therefore the presence of co-morbid conditions is greater among older age groups. In the UK, those aged 65 years and over make up 18% of the population but they receive nearly half of all prescriptions.[1] By 2010 in most of the developed countries, this age group will form over 15% of the total population, but in certain countries like Japan, they will account for over 20%.

Older people not only are the highest users of drugs but have the highest incidence of side effects.[2] The physiological changes that accompany the ageing process may alter the way in which older people respond to drugs. The ability of the body to handle drugs is particularly impaired among older people in poor health.[3]

Older patients, and those responsible for prescribing their treatment, should be able to expect that these treatments have been tested and investigated among samples that reflect this population. Therefore, for drugs used to treat conditions where increased age is a risk factor, older people should be oversampled in studies designed to test their efficacy. However, the reverse is generally true. Older people are under-represented in trials of treatments for conditions that predominantly affect older people. Clinicians may be unaware of the paucity of older people studied, resulting in the late recognition of serious side effects when drugs tested on predominantly younger adult populations are finally released and prescribed to large numbers of older people. Perhaps the most famous, or infamous, case of this was benoxaprofen, a non-steroidal anti-inflammatory drug marketed as Opren, which was withdrawn after a

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

report of the death of five elderly patients who had taken the drug.[4]

Trials of the efficacy of interventions need to cover the age groups who are affected.[5] In this chapter we explore how older people have often been explicitly, through the use of upper age limits, and implicitly, through other criteria, excluded from participation in clinical trials. The evidence to support the use of different approaches to trial recruitment and retention that maximise participation among older people is examined. The process of informed consent and how it can be ethically applied without creating barriers to the inclusion of older people in treatment studies is discussed. Finally, we make suggestions as to how researchers, when designing clinical trials, may increase the opportunities for older people to take part so that those responsible for the treatment of older people will be able to base their practice on high-quality and relevant evidence.

ELIGIBILITY

EXPLICIT EXCLUSION

There is ample evidence to suggest that older people are often explicitly excluded from clinical trials, usually without justification and often when the treatment being tested is of direct benefit to older people. An analysis of studies reported in four leading journals (*Gut,* the *BMJ*, *The Lancet* and *Thorax)* found 35% (170) excluded older people with no justifiable reason.[6] In a review of study protocols submitted for ethical approval, of the 155 of relevance to older people, over half had an upper age limit that was unjustified, although this exclusion was not commented on by the reviewing committee.[7]

Reviews of published trials that have focused on specific treatments or conditions have identified that a significant proportion of trial evidence is based on studies that included upper age limits. Of 47 published trials of statins, 31 included age-based exclusion criteria,[8] although there was some evidence that North American studies were likely to be more inclusive than those carried out elsewhere. The situation is similar with treatments in the participation in Phase III acute myelogenous leukaemia trials where those potentially eligible for treatment are excluded from clinical trials on the basis of their age.[9]

Although there is evidence that age is the most important predictor of outcome following acute myocardial infarction,[10] a review of studies examining the efficacy of treatments for acute myocardial infarction found that 60% explicitly excluded those aged over 75 years.[11] Age-based exclusion was more common when the treatment under investigation was invasive and in more recent trials suggesting that in terms of older people's participation in clinical trials, the situation is not improving. Similar evidence for this trend over time is provided in a review of clinical trials for Parkinson's disease where subjects older than 75 years were more likely to be excluded in studies published between 1987 and 1996 than those published prior to this period.[12] This review found only 38% of studies included subjects over 75 years of age. With a large and growing proportion of people with Parkinson's disease aged at least 75 years, this suggests that the gap between the characteristics of trial samples and the population in need of treatment is growing.

Operating an upper age limit for trials has often been used to limit the problem of co-morbid conditions and drug interactions that may occur with increasing age. This stems from the belief that most adverse drug reactions in older people are simply a consequence of advancing age. A review of pertinent studies suggests that this may be misguided since the physiological and functional characteristics of the patient, rather than chronological age per se, appear to be the most important in drug interactions.[13]

IMPLICIT EXCLUSION

Even when age limits are not imposed, older patients are often implicitly excluded because of other exclusion criteria or because of investigator, cultural or other biases in enrolment.[14] In a review of acute myocardial infarction trials, comparison of the age distributions of patients in trials with and without age exclusions showed no differences, suggesting that factors other than explicit age restrictions were at play.[11] Moreover, since more women than men survive to older age and in some cases, such as cardiovascular disease, women develop diseases later in life than men, exclusion on the basis of age, either implicitly or explicitly, disadvantages women.[14]

The process of patient selection and recruitment mostly aims to produce a homogeneous study population with the purpose of increasing the statistical power to detect the effects of drugs.[15] The resulting clinical trial, conducted in 'sterile' conditions, bears little resemblance to clinical practice and cannot be extrapolated to the general population. Reports that general practitioners frequently miss depression among older people may be a reflection of the lack of evidence for the most appropriate treatment strategies for this age group. The evidence that exists applies to uncomplicated major depression, which accounts for only 15% of depressed older people seen in primary care.[16] A study of patients receiving treatment for rheumatoid arthritis found that the majority would be excluded according to current entry criteria for trials of rheumatoid arthritis treatments.[17]

Indeed, although tight eligibility criteria may aim to produce very similar participants, interpatient variability is such that a truly homogeneous group of patients is difficult, if not impossible, to identify. Important prognostic variables will be measured at baseline, but even if study participants are the same on these criteria, they will still vary in the course of their disease and on unmeasured prognostic factors.[18] Thus the gain in attempting to study a group of homogeneous patients is outweighed by the loss in generalisability and clinical applicability of the results.

Even when treatment trials are specifically designed for older people, overly stringent exclusion criteria can produce highly skewed and non-representative patient populations. Many trials of treatments for Alzheimer's disease have excluded patients with behavioural problems despite such problems being common with increasing cognitive impairment. Since there is considerable scope for improving such symptoms with drugs that enhance cognition, these trials may well be missing opportunities.[19] It is also argued that the use of depression scales to identify and exclude patients with Alzheimer's disease with concomitant depression is inappropriate due to a lack of data providing evidence that these scales are valid for use with this particular population.[20]

Studying a narrow group of patients also misses the potential to identify subgroups of patients who may respond particularly well to the drug being tested. A trial comparing the efficacy of sertraline and nortriptyline in major depression included patients aged 60 years and over, but a subgroup analysis of the 76 patients aged 70 years and over suggested that treatment with sertraline may confer even greater benefit in this older age group than patients aged 60 years and over.[21] In trials of intervention packages or services rather than drugs, similar tensions exist between maximising the detection of a significant effect of the intervention in a population unencumbered with concurrent illness, and a need to assess effectiveness as close as possible to a viable model of service provision after the trial. The advantages of wide eligibility criteria for entering patients into clinical trials are summarised in Box 39.1.[18]

Box 39.1 Advantages of wide eligibility criteria for entering patients into clinical trials (Yusuf *et al.*[18])

1. Easier screening and recruitment. Large trials are more feasible and economical.
2. Large study sizes reduce random error, providing more reliable overall results.
3. Wider applicability of results. Therefore greater clinical and public health impact.
4. Greater opportunity to test subgroup hypotheses.

RECRUITMENT

We have argued that criteria for clinical trial participation can explicitly exclude older people through the use of upper age limits or do this more implicitly by the requirement to meet other criteria that disproportionately affect older age groups. A challenge to all researchers running a clinical trial is recruitment of sufficient participants within the desired time frame. The way in which recruitment is planned and executed can potentially have a great effect on the success of recruiting motivated participants but particular considerations need to be given to recruiting older participants. The reasons why potential participants agree to take part in research are many and varied and therefore recruitment approaches need to be broad and flexible to maximise the uptake of invitations to participate.

In studies that have followed up eligible participants who choose not to enrol in clinical trials, older age is a consistent predictor of non-participation. When recruitment to a randomised controlled trial of thrombolytic strategies was compared with recruitment to two observational studies with similar eligibility criteria, trial participants were significantly younger.[22] In another study, enrolees were on average three years younger than non-enrollees to a randomised controlled trial of arrhythmia therapy post-myocardial infarction.[23] One report, of recruitment to a trial of weight loss and dietary sodium reduction for older people following withdrawal of anti-hypertensive medications, found a threefold difference in yield between those aged 60 to 64 years and those aged 75 to 80 years.[24]

Although the experience of earlier trials on strategies to maximise recruitment may not be immediately transferable across time and place, they may provide researchers with ideas that can be applied to their own context. Experiences in recruiting older people to trials have been described in the treatment of hypertension with both pharmacological[25] and non-pharmacological interventions[24] and in trials of exercise.[26] Many of the reports simply describe the experience of one or two particular strategies, though mass mailing, media advertising, community-based screening, clinical practice screening, participant referrals and other recruitment methods have been compared in a trial of the efficacy of weight loss and sodium reduction for preventing hypertension in the elderly.[24] This study concluded that mass mailing of a brochure or letter describing the study resulted in the greatest yield in terms of per cent randomised (76%; $N = 737$) though it is less clear whether this applied to all subgroups of the population. Similar results favouring electoral roll mail-out and newspaper advertising were found in a randomised trial of vitamin E in the prevention of cataract and age-related maculopathy.[27] However, the authors of a primary prevention trial of low-dose aspirin found general practice recruitment produced a greater yield than approach via the electoral roll or local community, although the latter was the most cost effective in terms of the cost per participant.[28]

Trials recruiting volunteers rather than clinical populations may result in a population of older people more likely to remain throughout the length of the study, but may not always provide evidence applicable to the general population of older people. Older volunteers tend to be more likely than younger ones to be healthy and living independently, and of particular importance for trials of interventions involving exercise since volunteers may not be the subjects most likely to benefit.[26] However, a combination of volunteers and clinically referred patients in a trial of maintenance therapy for clinical depression resulted in no difference between the two sources of recruits in terms of treatment response although the groups differed in terms of demographics.[29]

Rarely does one single strategy succeed in recruiting adequate numbers of representative patients. It is important therefore that the characteristics of participants are regularly monitored throughout the trial, and compared with the general population, so that, if necessary, specific demographic groups, such as the oldest-old or particular ethnic groups may be targeted. Such mixed-mode recruitment has produced representative samples of high-risk older people for a trial of geriatric evaluation and management.[30] The final sample should aim to be as representative as possible and a list of strategies that could be used if shortfalls occur during recruitment should be developed at the design stage of the trial.

When potential participants who declined to take part in a trial of influenza vaccination were asked why they had chosen to refuse, over half reported being reluctant to take part in a research project while over a quarter objected to the term 'Geriatric Medicine' on the letter of invitation.[31] The authors of this study suggest avoiding direct reference to ageing when recruiting for a study designed for older people where the majority are likely to be in good health. In contrast, an earlier trial attempting to recruit healthy older people with early cognitive impairment found that a symptomatic approach, focusing on early diagnosis and treatment of existing cognitive impairment, was five times more effective than a normative approach, where emphasis was placed on normal cognitive changes associated with ageing.[32]

Older people are less likely to have their own transport than younger age groups and the requirement to travel to take part in a research study can seriously affect the recruitment of older participants. There is evidence that greater distance between study site and own residence is predictive of non-participation in clinical trials.[33] Those conducting tests to screen for trial eligibility and ongoing monitoring during the trial need to consider the potential for conducting tests in participants' homes or set up outreach sites in local community areas to lower the burden, both real and perceived, on participants. Home visits or provision of transport to clinic for assessments allow frail and home-bound older people to participate. Although such a strategy is inevitably more costly, this may be offset by greater yield.[34] There is a suggestion in the literature that recruiting older people from ethnic minority groups may be particularly problematic if there is an expectation to travel to take part in research.[35] Furthermore, research that takes part in the community is likely to be perceived as more trustworthy by the local population and provides an opportunity for word of mouth to be used to publicise the trial.

Family and caregiver involvement in recruitment may be essential if those who are frail, disabled or resident in nursing homes are to be targeted for recruitment. In focus groups, caregivers for people with Alzheimer's disease said they were likely to get involved in research if they could see the potential for the person they care for to benefit from a new treatment and perceive there to be adequate social and emotional support by those running the trial.[36] Conversely, caregivers were reluctant to participate if the benefit was not made explicit or the involvement was seen as particularly burdensome.

For older people with a pre-existing health problem, the perceived potential benefit of their involvement in randomised trials is offset by the possibility of receiving placebo treatment.[20] With the concept of randomisation often difficult to explain in lay terms, it becomes an inevitable barrier to trials based on a randomised design. Data from focus groups of 225 'eligible refusers' to a primary prevention trial of aspirin found an unwillingness to give up choice in favour of random allocation and potentially risk their hold on good health.[37] Clinical trialists need a good understanding of the attractiveness of the intervention or treatment being tested. A trial of antidepressants in the treatment of older people in the community with depression or anxiety failed because only 6 of 54 participants were willing to take the study medication.[38]

The evidence for whether participants take part for altruistic reasons is inconsistent. A survey of clinical trial participants found that 80% believed they were helping medical science.[39] However, the most frequently reported reason for taking part in a randomised clinical trial of behavioural therapy for chronic heart failure was benefit to the participant, although helping others was also cited.[40] For potential participants with pre-existing conditions, there may be less altruistic motivation than among healthy older people.[37] Older people may incorrectly assume that their involvement in research is of less value than that of younger people and decline to take part. This may be a reflection, particularly among those with health problems, of a nihilistic attitude towards treatment,[41] although it may also stem from a fear that they will not be able to comply with the demands of trial involvement.[42]

In practice, the initial approach to potential participants may be via the clinician directly responsible for their care. In many countries this is now becoming the expected route of recruitment and therefore support from clinical staff becomes key to trial success.[43] Where access to a sample is via a gatekeeper, researchers have to ensure that gatekeepers are clearly aware of the value of the research and that this message is not lost when participants are approached. In this sense the concept of 'refusal' becomes more complex than a numerical value representing the proportion of those invited who decline to take part.[44]

The enthusiasm with which a doctor or nurse assumes their role in the recruitment process is likely to be related to the burden of recruitment and participant involvement, and their view of the patient's suitability for the trial which may draw on factors other than eligibility criteria. Although this may negatively affect recruitment by introducing an additional barrier to recruitment,[45] there is evidence that involvement of a service provider in face-to-face contact with a potential participant may increase recruitment.[46] Over-enthusiasm on the part of service providers is possible too, particularly in unblinded trials when one intervention is considered preferable to others.[42]

INFORMED CONSENT

Given the need to use proactive recruitment strategies to ensure sufficient participants in a clinical trial, a potential ethical dilemma occurs when this risks being coercive and violating privacy rights.[47] This fine line needs to be negotiated carefully to ensure that research involving a potentially vulnerable group is both ethical and valid. The provision of informed consent from patients before randomisation is a universal requirement, although legal requirements across countries may differ. As with eligibility and recruitment, the means of gaining informed consent from subjects enrolling should be addressed at the design stage of the trial and the information required for a patient to give informed consent is listed in Box 39.2. A synopsis of the practicalities in obtaining informed consent for clinical trials has been reported, stressing that this process 'should not be seen as an exercise in bureaucratic form filling, but as an essential part of the trial requiring time, insight and communication skills'.[48]

Box 39.2 Patient information necessary for informed consent

1. Diagnosis.
2. Available treatments and treatment on trial.
3. Potential risks and benefits of treatment.
4. Concept of a clinical trail (including randomisation, use of placebos, double-blind procedures).
5. Discomforts of inconveniences associated with assessments.
6. Number of follow-up visits or extra travel for trial.

Clinicians may see relaying the concept of a randomised controlled trial as admittance of ignorance about the best treatment for the patient, or may make ageist assumptions concerning the ability of older people to consent to a trial. An analysis of audio-taped consultations that involved discussion of participation in cancer trials suggest that the reality of securing informed consent falls short of the ideal.[49] Qualities generally considered important in obtaining informed consent such as shared decision making and clarity of information were adequately addressed in only a minority of consultations. There is no evidence that these qualities may be more important for older people but extra effort is likely to be required to ensure that those who are more vulnerable can make an informed choice.

However, the clinical trial design is complex and, even if explained carefully, patients may not understand fully enough to give true informed consent. A qualitative study, as part of a set of trials of the effectiveness of treatments for men with urinary retention and benign prostatic disease, found that, although information given was accurately recalled, subjects found the concept of randomisation difficult to accept and were confused by terms such as 'trial' and 'random' which have different meanings to lay and professional groups.[50] The ability to understand information about clinical trials, particularly the randomisation process, may well be correlated with level of education.[51]

A systematic review of literature on informed consent found evidence of impaired understanding of the informed consent information in older subjects and those with less formal education[52] and suggested that overly detailed consent forms may impede understanding. Ironically, external bodies such as ethics committees that have certain language requirements may, when applied universally, prevent basic comprehension among certain groups.[53] The Recruitment and Enrollment Assessment in Clinical Trials Study, part of the Cardiac Arrhythmia Suppression Trial (CAST), did not find education differences in enrollers and non-enrollers, although enrollers were more likely to have read the informed consent themselves and to have understood it.[23] An instrument to assess understanding of information given to ascertain informed consent for ambulatory trials has been developed, but its disadvantages are that it is study-specific and it was tested on relatively young and well-educated subjects.[54]

After the Second World War, the Nuremberg code required the 'voluntary consent of the human subject' in experimental research and that 'the person should have legal capacity to give consent'. The need to have sufficient understanding to give consent without coercion, when taken literally, e.g. by being required to pass a test of competency, would make research on the efficacy of treatments and management strategies for dementia patients, particularly those with advanced dementia, virtually impossible.[55] The increasing prevalence and incidence of dementia with advancing age may also pose problems for gaining informed consent more generally for trials, not just those specifically for dementia treatments.

Currently, informed consent is usually gained from proxies on behalf of dementia patients, although technically only the subject may provide consent to be entered into a trial. Within clinical care there has been encouragement for patients to prepare advanced directives or living wills to cover the eventuality that they may not have the capacity to agree to treatment being given or withdrawn. Although this might be seen as a solution for dementia research also, the strong motivational factors for individuals with clinical care are unlikely to be present for dementia research.[56] In addition, the number of people preparing living wills is still very small and often restricted to well-educated, higher social class groups. A more realistic future goal might be that people are encouraged to name proxies and state broad beliefs about research in advanced directives.

Rather than immediately approaching a proxy for consent with dementia patients, it may be best to promote the pragmatic view of decision-making capacity that if an individual appears competent then he or she is.[57] Dementia patients have been shown to be capable of understanding and differentiating the risk/benefit ratio between different treatments and of expressing their contentment with having a proxy make decisions on involvement in research, although the proxies themselves tended to be more protective with their relatives than with themselves.[57] A more pressing problem is the lack of suitable proxies to provide informed consent on behalf of patients, one trial of palliative care of patients with advanced dementia who had been hospitalised finding that almost half (72/146) of eligible patients could not be enrolled.[58] In only four cases was this due to the proxy refusing consent, the proxies themselves lacking the capacity to understand the protocol in 18% of cases and in almost one-third no functional proxy being found. None of the patients for whom a proxy could not be found had made a living will.

A further issue that needs to be considered is the level of confidence that can be placed on how the potential participant's interests are accurately represented by their caregiver in the role of proxy. A study of family members acting as proxies for nursing home residents unable to provide informed consent found that proxies' decisions were informed not just by the perceived disturbance to the resident, but also by their own beliefs about research in nursing homes.[59] Furthermore, nearly a third (17/55) of proxies provided consent even though they believed the participant would refuse if they could.

Individuals fulfilling the same trial entry criteria may have many different goals that will determine the likelihood of consenting to take part. During the informed consent process these need to be explored by research staff to maintain study interest and completion.[60] Ultimately, the most successful approach to acquiring informed consent is a flexible one that does not compromise ethical standards but is sensitive to the needs of older people being approached to take part in a clinical trial.

FOLLOW-UP

Although it may be more difficult to enrol older people into clinical trials they may be less likely to choose to withdraw than younger age groups.[61] However, their greater risk of developing comorbid conditions, cognitive impairment, and need for other drug treatments will mean that they may have to exit the study before the final outcome assessment. To some extent this can be planned for in advance by allowing for a realistic rate of loss to follow-up when calculating sample size.

We have argued that greater effort is needed to recruit older people but they are also likely to need more support during the recruitment and consent process. Ongoing support for older people once they have entered a trial is important in helping them comply with treatments and assessments according to the protocol. Provision of information about the trial should not be considered a 'once and for all' activity at the commencement of the trial, and opportunities to re-enforce the importance of the participants' role in the success of the trial (e.g. at interim assessments) should be exploited. More flexible timing of follow-up visits may prevent the unnecessary loss of data and by using correct procedures should not pose any problems for analysis.

Successful recruitment and retention has been achieved by the involvement of a gerontological nurse specialist to support participants with Alzheimer's disease throughout their trial involvement.[62] High levels of adherence are possible among older people in clinical trials and what evidence there is suggests that depressive symptoms and low self-rated health, rather than indicators of physical health, are risk factors for medication discontinuation.[63] An awareness of 'at-risk' groups for study medication non-adherence provides researchers with a target group that might be provided with more support during the trial. There is a suggestion that the use of community pharmacists in drug trials might assist treatment compliance.[64]

Missing data are still likely to be a problem. When appropriate, self-report measures could be substituted by information provided by a proxy. It is argued that this may be the only way of avoiding disenfranchising very frail older people from clinical research.[65] More complex data analysis techniques should be used to maximise the use of the data that are present. Some statistical packages for repeated measures data analysis – a common analysis for trials with regular follow-ups – ignore cases with data missing. Newer techniques such as multilevel modelling and random effects models can accommodate incomplete data. Finally, outcomes such as mortality that may be easy to measure and important for younger populations may, in older people, be valued less than quality of life and the ability to function independently.[66]

CONCLUSIONS

If clinicians and other professionals caring for older people are to provide optimal treatment and those receiving that care are to benefit from new advances in treatments, decisions need to be based on strong evidence of efficacy in older people. At present there is a lack of fit between the populations on which treatments are tested and those that present to services in need of treatment. We have discussed some of the reasons why older people have been and are still being excluded both explicitly and implicitly from trials. We cannot give any definitive solutions to ensure that older people are recruited and retained in sufficient numbers into trials, since the setting for the trial (community, nursing home, outpatient clinic) will influence the feasibility of different design options as well as the country in which the research is conducted. However, we list below important factors that are usually within the researcher's control that need to be considered when designing trials of future therapies that may ultimately be used by large numbers of older people:

- *Eligibility criteria should be wide*. Increasingly, trials are going to need to be more readily able to accommodate older people with co-morbid conditions and in receipt of a number of medications. This will ensure smaller random error, a wider applicability of results and a greater opportunity to test preplanned subgroup hypotheses.
- *Use multiple recruitment strategies*. In order to maximise the involvement for specific subgroups (such as the very elderly or ethnic minority elders) a range of recruitment strategies is likely to be needed. The relative success of these should be monitored so resources can be diverted as necessary.
- *Involve clinicians and service providers throughout*. The success of a clinical trial may depend on the willingness of service providers to be involved. Careful thought needs to be given to the attractiveness of a trial to this key group of gatekeepers and the extra workload that it will require of them.
- *Design the consent documentation and process carefully.* Ethical considerations are of paramount importance when potential participants are vulnerable. Consider whether and when consent by proxy is appropriate.
- *Minimise trial burden on participants.* Be discriminating in the amount of data required from participants for the trial to be a success. When possible, offer home assessments or provide transportation to clinics at times convenient to the participant and their caregivers. Consider other ways to offer ongoing support.
- *Be realistic*. Successful recruitment of older people to clinical trials takes time and a realistic level of attrition needs to be incorporated into the sample size calculation.

Finally, many clinical trials fail because of poor recruitment and lack of adherence to protocols. The problems outlined in this chapter may mean that these are particular issues for trials involving older people. There are useful lessons to be learnt from these experiences, yet by definition these are rarely shared in the published literature. Methodological issues that arise from others' successes and mistakes in carrying out clinical trials involving older people need to be aired and discussed in journals. If the quality of the evidence is improved then older people can expect to see an improvement in the quality of their health care.

REFERENCES

1. Statistical Bulletin (1995). *Statistics of prescriptions dispensed in the Family Health Service Authorities*: *England 1984–1994*.
2. Rochon PA, Berger PB, Gordon M. The evolution of clinical trials: inclusion and representation. *Can Med Associ J* (1998) **159**: 1373–4.
3. Avorn J, Gurwitz JH, Rochon P. Principles of pharmacology. In: Cassel CK, Leipzig R, Cohen

HJ, Larson EB, Meier DE, eds, *Geriatric Medicine: An Evidence-Based Approach*, 4th edn, New York: Springer (2003).

4. Taggart HMcA, Allardice JM. Fatal cholestatic jaundice in elderly patients taking benoxaprofen. *Br Med J* (1982) **284**: 1372.
5. ICH Harmonised Tripartite Guidelines. *Studies in support of special populations: Geriatrics (1993).*
6. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. *Br Med J* (1997) **315**: 1059.
7. Bayer A, Tadd W. Unjustified exclusion of elderly people from studies submitted to research ethics committee for approval: descriptive study. *Br Med J* (2000) **321**: 992–3.
8. Bartlett C, Davey P, Dieppe P, Doyal L, Ebrahim S, Egger M. Women, older persons, and ethnic minorities: factors associated with their inclusion in randomised trials of statins 1990 to 2001. *Heart* (2003) **89**: 327–8.
9. Mengis C, Aebi S, Tobler A, Dahler W, Fey MF. Assessment of differences in patient populations selected for or excluded from participation in clinical phase III acute myelogenous leukemia trials. *J Clin Oncol* (2003) **21**: 3933–9.
10. White HD, Barbash GI, Califf RM, Simes RJ, Granger CB, Weaver WD *et al*. Age and outcome with contemporary thrombolytic therapy. Results from the GUSTO-1 trial. *Circulation* (1996) **94**: 1826–33.
11. Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA* (1992) **268**: 1417–22.
12. Mitchell SL, Sullivan EA, Lipsitz LA. Exclusion of elderly subjects from clinical trials for Parkinson's Disease. *Arch Intern Med* (1997) **157**: 1393–8.
13. Gurwitz JH, Avorn J. The ambiguous relation between aging and adverse drug reactions. *Ann Intern Med* (1991) **114**: 956–66.
14. Wenger NK. Exclusion of the elderly and women from coronary trials. Is their quality of care compromised? *JAMA* (1992) **268**: 1460–1.
15. Yastrubetskaya O, Chiu E, O'Connell S. Is good clinical research practice for clinical trials good clinical practice? *Int J Geriatr Psychiat* (1997) **12**: 227–31.
16. Callahan CM, Hendrie HC, Tierney WM. The recognition and treatment of late-life depression: a view from primary care. *Int J Psychiat Med* (1996) **26**: 155–75.
17. Sokka T, Pincus T. Most patients receiving routine care for rheumatoid arthritis in 2001 did not meet inclusion criteria for most recent clinical trials or American College of Rheumatology criteria for remission. *J Rheumatol* (2003) **30**: 1138–46.
18. Yusuf S, Held P, Teo K. Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria. *Stat Med* (1990) **9**: 73–86.
19. Schneider LS, Olin JT, Lyness SA, Chui HC. Eligibility of Alzheimer's disease clinic patients for clinical trials. *J Am Geriatr Soc* (1997) **45**: 923–8.
20. Korczyn AD. Recruitment for drug studies in dementia. *Alzheimer Dis Assoc Disord* (2002) **16**: 283–4.
21. Finkel SI, Richter EM, Clary CM. Comparative efficacy and safety of sertraline versus nortriptyline in major depression in patients 70 and older. *Int Psychogeriatr* (1999) **11**: 85–99.
22. Krumholz HM, Gross CP, Peterson ED, Barron HV, Radford MJ, Parsons LS, Every NR. Is there evidence of implicit exclusion criteria for elderly subjects in randomized trials? Evidence from the GUSTO-1 study. *Am Heart J* (2003) **146**: 839–47.
23. Gorkin L, Schron EB, Handshaw K *et al*. Clinical trial enrollers vs. nonenrollers: the Cardiac Arrhythmia Suppression Trial (CAST) Recruitment and Enrollment Assessment in Clinical Trials (REACT) project. *Controlled Clin Trials* (1996) **17**: 46–59.
24. Whelton PK, Bahnson J, Appel LJ, Charleston J, Cosgrove N, Espland MA *et al*. Recruitment in the Trial of Nonpharmacologic Intervention in the Elderly (TONE). *J Am Geriatr Soc* (1997) **45**: 185–93.
25. Petrovich H., Byington R, Bailey G *et al*. (1991) Systolic Hypertension in the Elderly Program (SHEP) Part 2: Screening and recruitment. *Hypertension* (1991) **17**(Suppl II): II16–23.
26. Halbert JA, Silagy CA, Finucane P, Withers RT, Hamdorf PA. Recruitment of older adults for a randomized, controlled trial of exercise advice in a general practice setting. *J Am Geriatr Soc* (1999) **47**: 477–481.
27. Garrett SKM, Thomas AP, Cicuttini F, Silagy C, Taylor HR, McNeil JJ. Community-based recruitment strategies for a longitudinal interventional study: the VECAT experience. *J Clin Epidemiol* (2000) **53**: 541–8.
28. Silagy CA, Campion K, McNeil JJ, Worsam B, Donnan GA, Tonkin A. Comparison of recruitment strategies for a large-scale clinical trial in the elderly. *J Clin Epidemiol* (1991) **44**: 1105–14.
29. Stack JA, Paradis CF, Reynolds CFIII, Houck PR, Frank E, Anderson B, Mayo AL, Miller MD, Rifai AH, Perel JM, Kupfer DJ. Does recruitment method make a difference? Effects on protocol retention and treatment outcome in

elderly depressed patients. *Psychiat Res* (1995) **76**: 17–24.

30. Boult C, Boult L, Morishita L, Pirie P. Soliciting defined populations to recruit samples of high-risk older adults. *J Gerontol Med Sci* (1998) **53**(5), M379–84.
31. Allsop SJ, Gosney MA. Difficulties of recruitment for a randomized controlled trial involving influenza vaccination in healthy older people. *Gerontology* (2002) **48**: 170–3.
32. Leader MA, Neuwirth E. Clinical research and the noninstitutional elderly: a model for subject recruitment. *J Am Geriatr Soc* (1978) **26**: 27–31.
33. Ory MG, Lipman PD, Karlen PL, Gerety MB, Stevens VJ, Fiatarone Singh MA, Buchner DM, Schechtman KB, The FICSIT Group. Recruitment of older participants in frailty/injury prevention studies. *Prev Sci* (2002) **3**: 1–22.
34. Gill TM, McGloin JM, Gahbauer EA, Shepherd DM, Bianco LM. Two-recruitment strategies for a clinical trial of physically frail community-living older persons. *J Am Geriatri Soc* (2001) **49**: 1039–45.
35. Arean PA, Gallagher-Thompson D. Issues and recommendation for the recruitment and retention of older ethnic minority adults into clinical research. *J Consult Clin Psychol* (1996) **64**: 875–80.
36. Connell CM, Shaw BA, Holmes SB, Foster NL. Caregivers' attitudes toward their family members' participation in Alzheimer disease research: implications for recruitment and retention. *Alzheimer Dis Assoc Disord* (2001) **15**: 137–45.
37. Boles M, Getchell WS, Feldman G, McBride R, Hart RG. Primary prevention studies and the healthy elderly: evaluating barriers to recruitment. *J Community Health* (2000) **25**: 279–92.
38. Stevens T, Katona C, Manela M, Watkin V, Livingston G. Drug treatment of older people with affective disorders in the community: lessons from an attempted clinical trial. *Int J Geriatr Psychiat* (1999) **14**: 467–72.
39. Pope JE, Tingey DP, Arnold JM, Hong P, Ouimet JM, Krizova A. Are subjects satisfied with the informed consent process? A survey of research participants. *J Rheumatol* (2003) **30**: 815–24.
40. Chang B, Hendricks AM, Slawsky MT, Locastro JS. Patient recruitment to a randomized clinical trial of behavioural therapy for chronic heart failure. *BMC Med Res Methodol* (2004) **4**: 8.
41. Flicker F, Wark L. Recruitment strategies for randomised clinical trials in elderly Australians. *Med J Aust* (1997) **167**: 438–9.
42. Harris R, Dyson E. Recruitment of frail older people to research: lessons learnt through experience. *J Adv Nurs* (2001) **36**: 643–51.
43. Amthauer H, Gaglio B, Glasgow RE, Dortch W, King DK. Lessons learned: patient recruitment strategies for a type II diabetes intervention in a primary care setting. *Diabetes Educ* (2003) **29**: 673–81.
44. Cohen-Mansfield J. Recruitment rates in gerontological research: the situation for drug trials in dementia may be worse than previously reported. *Alzheimer Dis Assoc Disord* (2002) **16**: 279–282.
45. Gurwitz JH, Guadagnoli E, Landrum MB, Silliman RA, Wolf R, Weeks JC. The treating physicians as active gatekeeper in the recruitment of research subjects *Med Care* (2001) **39**: 1339–44.
46. Arean PA, Alvidrez J, Nery R, Estes C, Linkins K. Recruitment and retention of older minorities in mental health services research. *Gerontologist* (2003) **43**: 36–44.
47. Thompson MG, Heller K, Rody CA. Recruitment challenges in studying late-life depression: do community samples adequately represent depressed older adults? *Psychol Aging* (1994) **9**: 121–5.
48. Wager E, Tooley PJH, Emanuel MB, Wood SF. How to do it. Get patients' consent to enter clinical trials. *Br Med J* (1995) **311**: 734–7.
49. Brown RF, Butow PN, Ellis P, Boyle F, Tattersall MHN. Seeking informed consent to cancer clinical trials: describing current practice. *Soc Sci Med* (2004) **58**: 2445–57.
50. Featherstone K, Donovan JL. Random allocation or allocation at random? Patients' perceptions of participation in a randomised controlled trial. *Br Med J* (1998) **317**: 1177–80.
51. Pucci E, Belardinelli N, Signorino M, Angeleri F. Patients' understanding of randomised controlled trials depends on their education. *Br Med J* (1999) **318**: 875.
52. Sugarman J, McCrory DC, Hubal RC. Getting meaningful informed consent from older adults: a structured literature review of empirical research. *J Am Geriatr Soc* (1998) **46**: 517–24.
53. Barron JS, Duffey PL, Byrd LJ, Campbell R, Ferrucci L. Informed consent for research participation in frail older persons. *Aging: Exp Res* (2004) **16**: 79–85.
54. Miller CK, O'Donnell DC, Searight HR, Barbarash RA. The Deaconess Informed Consent Comprehension Test: an assessment tool for clinical research. *Pharmacotherapy* (1996) **16**(5): 872–8.
55. Agarwal MR, Ferran J, Ost K, Wilson KCM. Ethics of 'informed consent' in dementia research – the debate continues. *Int J Geriatr Psychiat* (1996) **11**: 801–806.
56. Sachs GA. Advance consent for dementia research. *Alzheimer Dis Associ Disord* (1994) **8**: 19–27.

57. Sachs GA, Stocking CB, Stern R, Cox DM, Hougham G, Sachs RS. Ethical aspects of dementia research: informed consent and proxy consent. *Clin Res* (1994) **42**: 403–12.
58. Baskin SA, Morriss J, Ahronheim JC, Meier D, Morrison RS. Barriers to obtaining consent in dementia research: implications for surrogate decision-making. *J Am Geriatr Soc* (1998) **46**: 287–90.
59. Warren JW, Sobal J, Tenney JH, Hoopes JM, Damron D, Levenson S, Deforge BR, Muncie HL. Informed consent by proxy. *New Eng J Med* (1986) **315**: 1124–8.
60. Schaeffer MH, Krantz DS, Wichman A, Masur H, Reed E, Vinicky JK. The impact of disease severity on the informed consent process in clinical research. *Am J Med* (1996) **100**: 261–8.
61. Carter WB, Elward K, Malmgren J, Martin ML, Larson E. Participation of older adults in health programs and research: a critical review of the literature. *Gerontologist* (1991) **33**: 561–5.
62. Williams DE, Vitiello RK, Bokan J, Prinz PN. Successful recruitment of elderly community-dwelling subjects for Alzheimer's disease research. *J Gerontol* (1988) **43**: M69–74.
63. Buist DSM, LaCroix AZ, Black DM, Harris F, Blank J, Ensrud K, Edgerton D, Rubin S, Fox KM. Inclusion of older women in randomized clinical trials: factors associated with taking study mediation in the fracture intervention trial. *J Am Geriatr Soc* (2000) **48**: 1126–31.
64. Bene J, Liston R. Clinical trials should be designed to include elderly people. *Br Med J* (1998) **316**: 1905.
65. Gregson BA, Smith M, Lecouturier J, Rousseau N, Rodgers H, Bond J. Issues of recruitment and maintaining high response rates in a longitudinal study of older hospital patients in England – pathways through care study. *J Epidemiol Community Health* (1997) **51**: 541–8.
66. Grimley Evans J. Evidence-based and evidence-biased medicine. *Age Ageing* (1995) **24**: 461–3.

40

Clinical Trials in Rare Diseases

SIMON DAY
Licensing Division, Medicines and Healthcare products Regulatory Agency, London UK

INTRODUCTION

This chapter considers problems associated with clinical trials when there are very few patients available to study. Contrary to the hopes of many people, there are no special or specific methods (statistical or otherwise) that can be applied to the study of small populations. Indeed, if there were, there would be no reason why they could not be applied to more common diseases so that those diseases could also be studied using much smaller samples than are often used today. Not only do such methods elude us today but this situation is unlikely to change within the current paradigm of clinical trials and drug development. Should that paradigm change, then it may be possible to test new medicines for rare – and common – disorders by some other means. The ideas presented in this chapter might, therefore, be applicable to studies for many products in common diseases, although in cases where large numbers of patients *are* available for testing new products, some of the aspects presented may need less consideration. However, in situations where very few patients are available, clinical trials may benefit from close scrutiny of some of the issues presented. Arguments in favour of large clinical trials are easy to find (e.g. Peto *et al.*)[1] but others also advocate the benefits and incremental knowledge gained from small trials.[2] Little is specifically written to help researchers who genuinely have very few patients/volunteers to study. NASA has considered the problem for experimentation on astronauts and some guidance is available.[3]

The chapter considers different levels (or strengths) of evidence; the precise definition of the disease to be treated and the population of patients to be targeted; pharmacological considerations; duration of treatment and choice of endpoints; and comparator groups. Different aspects about the size of effects that may be looked for are discussed and the chapter ends with some comments on statistical (analytical) methods. Although many of these may be quite standard, some may be more applicable in difficult cases of small samples than they are in cases that are more common where large numbers of patients are available. Many issues cross-relate to others.

ADMINISTRATIVE ISSUES

As we are considering diseases in small populations, the number of patients available for

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green
© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2

clinical studies is often very limited. It is therefore important that as much information as possible about successful – as well as unsuccessful – clinical studies in such diseases becomes available and in the public domain.

TRIAL AND PATIENT REGISTERS

Usually positive studies are more likely to be published than negative or unequivocal studies. This is often referred to as 'publication bias'. However, even negative studies may contain important information: possibly to help avoid repeating mistakes made earlier or to caution against further testing/development of ineffective products. Unequivocal studies that do not conclusively show the presence of – or absence of – any effect are perhaps of least value but may still offer useful data to be included in subsequent meta-analyses. For these reasons, it is valuable to know the outcome of all studies and the publication of even negative study results should be encouraged. Initiating study registers for clinical studies in small populations might be a possible approach to help overcome the problems of publication bias. As soon as a study is registered (and thus the knowledge that such a study has been started is in the public domain), there might be interest in the outcome, interest in running complementary studies and interest in *not* needlessly replicating the same study. Such a register does not automatically lead to publication of negative results (some sponsors of negative studies might still not have an interest in such a publication and some journal editors may not wish to publish them) but it may encourage them to do so. However, the register may alert interested parties (other researchers, patient groups, individual patients, regulatory authorities and so on) to contact sponsors to find the results.

Setting up and maintaining such registers is usually costly – although, ironically, most often done by not-for-profit organisations. External access to such registers should be made possible as they may facilitate various aspects of planning future studies. If properly set up, patient registers might:

- give information helpful for planning future studies such as endpoints, clinically useful treatment effects, the variability of potential endpoints, etc.;
- help characterise and identify potentially important subgroups of patients;
- help with validating surrogate endpoints (in addition to epidemiological data).

Unfortunately, the conditions for establishing and maintaining patient registers vary between countries as they are governed by national laws.

COLLABORATIVE DEVELOPMENT

The limited size of the potential market might make it commercially unviable for pharmaceutical companies to develop treatments for rare diseases. Alternative approaches, e.g. joint sponsorships between pharmaceutical companies and not-for-profit organisations (governmental institutions, academia, charities, and so on), should be considered as early as possible. In oncology, for example, such approaches are widespread and very successful. Over 6000 patients are recruited annually into the many trials handled by the European Organization for Research and Treatment of Cancer (EORTC).[4] Other collaborations exist within, for example, the Sylvia Lawry Centre for Multiple Sclerosis Research (www.slcmsr.org).[5]

HOW CONVINCING IS THE EVIDENCE?

Hierarchies of evidence have been described[6] and although different authors may give slightly different perspectives, they generally place in order:

- meta-analyses of randomised controlled clinical trials
- individual randomised controlled trials
- meta-analyses of observational studies
- individual observational studies
- published case reports
- anecdotal case reports.

All such studies provide some information (even anecdotal case reports) and none should be ignored completely. However, the highest levels of evidence in drug development come from well-planned and executed comparative clinical trials, particularly trials that have minimised bias through appropriate allocation concealment,[7] blinding and randomisation. At their conclusion, the treatment effect would ideally be large and clinically significant confidence intervals for the effect narrow, and the effect size highly statistically significant. Well-planned and conducted meta-analyses of such trials provide even stronger evidence.

Generally, a larger sample size and a smaller variance will result in narrower confidence intervals and more extreme levels of statistical significance (i.e. smaller P-values). In addition, more extreme levels of statistical significance (although not affecting the width of the confidence interval) are obtained when larger effect sizes are observed. The chance of producing a 'statistically significant' result (whether the treatment is effective, *or not*) is increased by using a less extreme significance level (e.g. considering $P < 0.10$ rather than $P < 0.05$ as the threshold for 'statistical significance'). Of course, this does not change the information content of the data; simply changing the threshold for 'statistically significant' does not make the results more convincing. Furthermore, lessening the certainty of the conclusions does not imply lessening the quality of the trial. Note, also, that 0.05 is a common – but wholly arbitrary – threshold or 'cut-off' point. No such value is adequate to confirm that a treatment effect truly does exist. Finally, if the treatment truly is effective, then the chance of producing a 'statistically significant' result increases by increasing the power of the statistical significance test.

Since we are dealing with small or very small sample sizes, a traditional 'statistically significant' result is often not achievable simply by increasing the size of the study. Instead it may only be possible with the other options mentioned. A smaller variance usually requires a very homogeneous study population. This is also often difficult, especially in the case of rare diseases, where a limited – so more homogenous – population reduces the available sample size even further.

A large effect size is always desirable but the medical reality is that medium or small therapeutic effects are common. In serious diseases where no alternative treatment exists, such small benefits may still be valuable to patients.

Allowing a higher Type I error means a decrease in certainty but also a decrease in sample size. For example, a change from the traditional $\alpha = 0.05$ to $\alpha = 0.10$ can result in a 25% reduction in sample size. Similarly, reducing the power of a study from 90% to 80% can result in a similar reduction. Changing from a two-sided to a one-sided test or, possibly, to a non-inferiority (rather than superiority) test when an alternative active treatment exists, will also reduce sample size. In the case of non-inferiority testing, any doubts over the assay sensitivity of the study[8] will compromise the strength of evidence obtained. A big reduction in sample size (often more than half) is possible if a comparator group (active or placebo) is dropped, but a study without a control group essentially becomes an observational study and will, again, produce less convincing evidence.

In very rare diseases, the combination of single case studies may be the only way to accumulate evidence. In such situations, treatment regimens and data collection may still be carried out in a controlled manner and this will add weight to the evidence. This may be in the form of n-of-1 designs (see later), which have some benefits[9] but also drawbacks.[10,11] If careful consideration is given to the statistical analysis (including methods such as formal 'cumulative meta-analyses') then this will carry more strength than ad hoc pooling of several case reports. Overviews of individual case reports or of observational studies should still be considered with caution. A meta-analysis will not necessarily provide good evidence; if the individual studies have inherent biases within them, then a meta-analysis will merely provide a more precise – but equally biased – result.

DISEASE AND PATIENT DEFINITIONS

Homogeneity of the disease or study population will usually result in less variability of response to treatment. Hence, despite the rarity of the disease, it *might* be easier to demonstrate an effect of treatment in a restricted population of 'typical' cases. However, additional variation may be introduced by phenotypic and/or environmental heterogeneity, even in genetically identical cases of the same disease (e.g. in families with haemophilia). Even if demonstration of efficacy may be restricted to a specific subgroup of patients, that should not deter sponsors from collecting good-quality data from a broader population. Maitournam and Simon discuss the relative benefits (in terms of efficient experimental design) of studying 'all-comers' as opposed to targeting trials at those subgroups most likely to show an effect.[12]

HOMOGENEOUS CONDITIONS

Examples of homogenous conditions include clonal disorders such as chronic myelogenous leukaemia and acute promyelocytic leukaemia, which are both well defined in terms of their chromosomal aberrations and are relatively straightforward to diagnose.[13,14]

HETEROGENEOUS CONDITIONS

There are many conditions that fall under the term 'syndromes' that have a very mixed make-up of patients and presentation of patients.[15] Heterogeneity may be due to different pathophysiological mechanisms and, hence, adequate response to treatment may depend on the subcategory of disease. Such groups of patients may be naively regarded as coming from a single population but there may be hidden subgroups that are more, or less, likely to contain responders or non-responders.

The disease stage is another major and well-known contributor to variation in therapeutic response, particularly where early diagnosis and treatment may result in cure (e.g. in oncology). In very rare conditions, information may be lacking on many aspects of the disease including heterogeneity and natural course so that these types of prognostic factors may not be known.

POPULATIONS

It is often not possible to assess the influence of geographical location, medical practice, and so on, in studies of rare diseases because of the limited size of the population (which may sometimes be reduced to individual cases). It is often impossible to define homogeneous subgroups. In other situations, important subgroups may be well known: the paediatric population, for example, needs to be categorised because diseases and drug response may vary in different ages (pre-term newborns, term newborns, infants and toddlers, children, adolescents).

Even in paediatric studies, there can be considerable overlap in developmental (e.g. physical, cognitive and psychosocial) issues across the age categories. For efficacy, different endpoints may be established for paediatric patients of different ages, and the age groups might not correspond to the usual categories listed above. Lung function, for example, may have different pathophysiologies in different age groups[16] and may need to be measured in very different ways in different age/developmental groups.[17]

PHARMACOLOGICAL CONSIDERATIONS

Detailed knowledge of the pathophysiology of the disease and the pharmacology of the drug will facilitate the design of efficient clinical studies and will help dictate the amount of clinical data required.

Non-clinical pharmacology studies are of special importance for studying rare diseases and can frequently be used to inform the design of clinical studies. Such studies may also give important information regarding features such as dosing, dose frequency, route of administration, and so on, although investigation of these in people is still preferable. However, less human work may be needed if animal models are reliable.

For 'substitution studies' (e.g. hormone replacement), well-characterised short- and long-term consequences of the deficiency, and a clear understanding of the pharmacokinetics and pharmacodynamics of the compound, provide guidance for designing studies. It is possible that a within-patient comparison in a relentlessly progressive – and predictably progressive – disorder might provide sufficient data to support a benefit–risk assessment.

The credibility of study results may be enhanced if a clear chain of events can be identified (e.g. drug exposure to target occupancy, to dynamic measures, to clinical outcome). 'Black box designs', on the other hand, are much less convincing and will usually increase the data requirements to obtain robust and persuasive study results.

In very rare disorders, it is important that every patient contributes as much information as possible to help make a benefit–risk assessment possible. Therefore, the well-planned use of the best available techniques to obtain and analyse information is crucial. This applies throughout the study process from handling and analyses of biopsy material to pharmacokinetic and pharmacodynamic modelling – always considering (at the planning stage) the need for confirming preliminary results.

PURPOSE OF TREATMENT

THERAPEUTIC TREATMENTS

The majority of treatments for rare diseases (as with common diseases) fall into this category. The objective of a study should be to show superior efficacy (and/or safety, in cases of pharmacodynamic equivalence) and that the new treatment provides substantial benefit to patient care. Randomisation is a minimal requirement for a comparative trial, and should be introduced as early as possible in the development of new treatments.

PROPHYLACTIC TREATMENTS

It is important to distinguish between primary and secondary prophylaxis, although prophylactic approaches may often occur in combination with therapeutic treatment (e.g. replacement of coagulation factor in haemophilia). Vaccination is also a form of prophylaxis (and in some cases, the target disease will be very rare). Demonstrating efficacy of a prophylactic treatment can be more complex than for a therapeutic treatment, particularly in small populations. Major issues for studies evaluating prophylactic treatments are the sample size and the length of follow-up needed to demonstrate efficacy because one cannot usually identify a population in which the risk of the event(s) to be prevented within a reasonable study period is 100%. In some cases, using surrogate markers may help to enable such studies to be practically carried out.

DIAGNOSTIC TREATMENTS

General approaches, guidelines, and so on, used for therapeutic or prophylactic studies may not be applicable to the investigation of diagnostic tools and agents; different approaches may be needed. Diagnostic tools may be combined with therapeutic interventions (e.g. operative diagnosis of disease). In such cases, the relevant outcome should usually be a clinical endpoint rather than a successful, or accurate, diagnosis.

CHOICE OF ENDPOINTS

The objectives of a study should be reflected in the choice of endpoints used. The ultimate goal in therapeutics is to cure patients of a disease but this may only be possible in certain acute diseases (e.g. meningococcal meningitis). Less commonly (but with vaccines as an obvious example), the objective is to prevent disease occurring.

'CURE'

At one extreme, the endpoint may be complete 'cure' of disease; this may be possible in diseases such as simple infections or acute respiratory distress. In the latter case, the failure to cure a patient may result in that patient's death; in the former case, a less extreme outcome may result from treatment failure. In general, such endpoints can be relatively easy to measure but in others, ‘presumed’ cure may be followed (possibly years later) by recurrence. This is often the case in oncology. The endpoint and objectives need to be explicit in their definition of ‘cure’. Although ‘cure’ may often be easy to measure, sometimes the endpoint might be ‘time to cure’. In this case, the definition of ‘cure’ is even more important to standardise so that different investigators can all measure it in the same way. A study should either have fixed time points for assessment, or the ability to assess all patients in continuous time. Whichever is chosen should be described unambiguously. Bias may be introduced if one or other treatment group is monitored more closely or more frequently than another so that there is a greater chance of observing (or observing sooner) the outcome. Even if ‘cure’ or ‘all-cause mortality’ is the primary endpoint in a study, it may need to be supplemented with secondary, supporting endpoints including non-fatal adverse events and quality of life.

SLOWING DISEASE PROGRESSION

Slowing disease progression is an intermediate level of endpoint and a measure of disease severity, or of disease progression, must be available. This should be validated as a tool for use in clinical trials (and not simply a diagnostic or epidemiological tool). In studies whose endpoint is time to progression or time to remission, adequate long-term follow-up of patients in a controlled (preferably blinded and randomised) way is important. It may be necessary to be able to identify whether a treatment does cause a particular (beneficial) outcome, or whether it just delays it.

Measurement scales should clearly distinguish between investigator-observed signs and patient-reported symptoms. Either, or both, may be acceptable – but they should be clearly distinguished. Validation of scales should be in patients with the same, or sufficiently similar, disease as those being treated in the trial. Validation should not be carried out as part of the clinical trial: a ‘valid’ scale should be both sensitive to change but also stable (or reproducible) in the presence of no change. A clinical trial cannot assess reproducibility. This, of course, presents a dilemma to those researching treatments for rare diseases – there may be insufficient patients for independent validation of rating scales. Validation of surrogate endpoints, however, may be possible from epidemiological data or from patient registers. These may provide rich sources of data from which to develop and validate potential surrogate markers of disease. Surrogate endpoints will always have the disadvantage of being difficult to relate to real clinical benefit and the size of benefit can be very difficult to estimate based on a surrogate endpoint.

Analysing ‘time to cure’ (or time to any event) can be statistically more efficient (and so need fewer patients) than simply analysing proportions at a specific time (e.g. five-year mortality). Conversely, it can be easier to understand the clinical significance of reported five-year mortality rates rather than the difference in median survival times. However, analysis using more sophisticated statistical models should not be seen as a bar to presenting simple summary statistics.

CLINICAL ENDPOINTS

Stroke is a good example of a clinically highly relevant endpoint (although it is not a rare disease). A stroke may severely impair a patient’s well-being due to subsequent aphasia and paralysis. Conversely, some clinical endpoints may need much more detailed validation and justification – for example, rhythm control in atrial fibrillation.

SYMPTOMATIC RELIEF

Relief of symptoms is a useful clinical endpoint – usually highly recognised by patients – but it may not reflect slowing true disease progression or delaying death. Pain management for arthritis is an example of benefit to patients, but one that does nothing to stop disease progression.

Similarly, pain management in end-stage cancer is beneficial to patients but does not impact on survival. As with surrogate endpoints, the measurement of symptoms should be based on well-validated scales and should clearly distinguish between investigator-observed and patient-reported signs and symptoms.

In contrast to endpoints such as 'cure', prolonged symptomatic relief and slowing of disease progression both usually imply long-term treatment. Hence, this should be studied in long-term trials. Short-term studies may be useful and efficient to determine if there is possible benefit for the treatment being studied (sometimes these are called 'proof of concept' studies). Long-term studies will be necessary to investigate cumulative dosing toxicity and possible tachyphylaxis. Tachyphylaxis may be managed by a change of dose or dose frequency but this also needs to be investigated in long-term studies – preferably by randomised comparisons. However, this is rarely done even in common diseases and it is very unlikely that it could be done for very rare diseases.

QUALITY OF LIFE

Clinical benefit may not necessarily be sufficient in the light of severe disability (such as neurological status following resuscitation or after an intracranial bleed). If quality of life is measured, it should always be assessed using scales validated for the particular indication being treated. Even with this restriction, it is unlikely that improvements in quality of life alone (i.e. in the absence of any other clinical benefit) would be sufficient to demonstrate the benefit of a new treatment. Quality of life data should be considered as supportive, and to help place the product in context with other available treatments. The complexity of measuring quality of life should not be underestimated; nor should the complexity of analysis and interpretation.[18]

BIOMARKERS AND SURROGATE ENDPOINTS

Assuming that a biomarker is a good surrogate endpoint requires it to be reasonably likely – based on epidemiologic, pathophysiologic, or other evidence – to predict benefit. For example, CD4 cell counts and HIV viral load are considered good surrogates for death or opportunistic infections in the evaluation of antiviral agents. Prediction in itself may not, however, be sufficient to attain the status of surrogate and a surrogate marker may not be sufficient to establish efficacy. Considerations should include:

- how closely changes in the surrogate endpoint are linked to causing changes in a clinical or symptomatic endpoint;
- how much risk is associated with the therapy;
- what other therapies are available for the same condition.

Biomarkers rarely offer sufficient final proof of clinical efficacy or long term benefit.

COMPARATOR GROUPS

Comparator groups (active and/or placebo) are used in clinical trials for numerous reasons. Amongst others, they:

- help to provide assay sensitivity of a trial (particularly placebo comparators);
- measure the effect of a treatment over and above 'regression to the mean';[19]
- measure the effect of a treatment over and above any 'placebo effect';[20]
- provide an estimate of the effect of a drug compared with existing therapies;
- help to assess the clinical relevance of the size of the observed effects of a drug.

Ideally, one wants to obtain an unbiased estimate of the effect of the treatment being investigated. This is true in studies in small populations as well as large trials for common diseases. Thus, in developing any treatment, a comparative trial will usually be preferable and may be necessary and all possibilities to run such trials should be evaluated. In serious and life-threatening diseases where no alternative treatments exist, there can be a tendency to grasp at any treatment seemingly offering some hope to patients. Anecdotal reports of patients responding may then make it ethically very difficult to justify subsequent controlled trials and may make it practically very difficult to persuade clinicians and patients to take part in trials. For these reasons, every attempt should be made to randomise patients from the very beginning of the therapeutic testing phase.

In general, there are two approaches to selecting control patients: internal controls or external controls, who may be historical or concurrent. The ideal is a comparative trial using an internal control group as there are several well-known problems inherent with historical (or other external) controls. Problems include (but are not limited to):

- the selection of appropriate controls
- comparability of medical conditions
- comparability of study design (treatment duration, concomitant treatment, and so on)
- comparability of endpoint assessment.

Thus, in comparisons using external controls it is often not possible to know whether or not they truly belong to the same patient population as that being studied. Historical controls often lead to biased estimates of treatment effects. The use of an internal control group as comparator certainly is the preferred option.

When using internal controls, patients should be randomly assigned to treatments and the randomisation codes should be concealed from study personnel to minimise any possible selection bias for any of the groups. If there are any strong prognostic factors for the outcome, then a stratified randomisation procedure might increase the efficiency of the trial and ensure greater credibility of the results by ensuring balance on these factors across the treatment groups. Although internal controls are the preferred option for comparative trials, under exceptional circumstances external controls may be acceptable. Such a situation might arise in indications where a treatment already exists but the use of a placebo control is still acceptable. Here, a three-armed trial would be the most informative but might be impossible to perform in a very small population. Thus, historical controls (using patients treated with the existing treatment) might, under exceptional circumstances, be necessary – and helpful – to demonstrate efficacy, safety, ease of administration, and so on, of the new treatment. In general, the absence of any control data is only likely to be adequate if the natural course of the disease is known beyond all reasonable doubt.

Different kinds of comparators may be considered. Where there is no recognised alternative treatment available, 'investigator's choice' or placebo may serve as possible comparators. The advantages and disadvantages of both these potential comparators have to be weighed carefully. Clinical trials are usually carried out in more than one centre and this can result in problems when using 'investigator's choice' as comparator, as different investigators may have different personal choices for best treatment. One solution (if there are only a few centres) is to stratify the randomisation by centre, or to group the different concepts of 'investigator's choice' and then stratify the randomisation accordingly. Another problem with 'investigator's choice' as a comparator is that such studies cannot usually be blinded so that the well-known problems with respect to a possible bias inherent in open-label studies have to be considered. If there is no recognised treatment alternative available, the use of placebo as a comparator instead of 'investigator's choice' might be acceptable. This might help reach a reliable conclusion as quickly as possible and thus might be beneficial and ethically quite acceptable. A further alternative is the comparison of a new treatment *in addition* to 'investigator's choice' (a so-called 'add-on' design). This approach allows a patient to be treated in the best way known and to gain a possible *additional* benefit from the new treatment. Such a treatment strategy is appealing and testing it in a randomised and double-blind trial can help to minimise possible sources of bias.

The situation is much more complex when a recognised treatment already exists so that using placebo as a comparator might be ethically unacceptable. If a new treatment offers only a small advantage over existing treatments, then it might be impossible to justify a placebo control, whereas an adequately powered active controlled trial might not be feasible because of the number of patients needed. However, in nearly all cases, an underpowered study – but one with concurrent, randomised controls – will be preferable to one with no controls at all. If the new treatment promises a substantial advantage over existing treatments (e.g. where the new therapy might cure patients but existing therapies may only improve symptoms), the obvious approach would be a comparative trial against the active control. However, in this situation a placebo-controlled trial, which would need far fewer patients, might also be ethical, partly because it would be relatively quick to complete.

If only active controlled studies are possible, then showing equivalence or non-inferiority may not be possible because assay sensitivity of the study cannot be assured and so obtaining convincing evidence of efficacy in these circumstances becomes extremely difficult. In such cases, only ‘superiority trials’ may be convincing.

THE SIZE OF TREATMENT EFFECT

Treatments that show large effects on endpoints of direct clinical benefit to patients are clearly preferred to those that only show small effects. However, in the absence of other treatments and in the absence of adverse effects, therapies offering small benefits may be useful to patients and to society as a whole. In the context of rare diseases, however, arguments used to justify public health benefits of small treatment benefits in very large populations (sometimes running into hundreds of millions) are unlikely to be convincing.

LARGE EFFECTS IN MANY PATIENTS

Treatments showing large treatment effects in a large number of patients are relatively easy to study and justify. Small studies may often be adequate.

LARGE EFFECTS IN FEW PATIENTS

Treatments may, however, be beneficial if they have large effects (possibly resulting in a ‘cure’ status) but only in a relatively small proportion of patients treated. Such treatments would be beneficial when the underlying prognosis is poor and spontaneous remission is very uncommon, or does not occur at all. However, in a disease that shows spontaneous remission, even a large effect (if in a very small proportion of treated patients) would be difficult to study and to justify as being clinically useful.

SMALL EFFECTS IN MANY PATIENTS

Some treatments may only have small or modest effects, but show those positive effects in a large proportion of treated patients. Such treatments may still be beneficial and worthwhile but it is particularly important to make clear the benefit to individual patients. In the presence of only small benefit (even if in a large proportion of patients), there can be a high potential for ‘minor’ adverse effects to outweigh any positive effects.

SMALL EFFECTS IN FEW PATIENTS

Treatments that show small benefit in only a small proportion of treated patients are the hardest to study. The potential benefit to an individual patient is well recognised but unless the types of patients that are likely to benefit can be prospectively identified, then – as a ‘policy’ – prescribing such a treatment to patients in a broad population will result in very few patients receiving any benefit. These types of treatment are only likely to be useful if the subset of patients who are more likely to benefit can be *prospectively* identified. As above, the benefits need clearly to outweigh the side effects, particularly since a large proportion of patients treated will not be expected to receive any benefit (but may be adversely affected).

STATISTICAL METHODS

As stated in the introduction, there are no special statistical methods applicable to studying treatments for rare diseases. It is certainly true that some statistical methods are only applicable to 'large' samples: for example, the chi-squared test for comparing proportions should be replaced by Fisher's exact test for small samples. However, this does not fundamentally address the problems of insufficient data for drawing reliable conclusions.

In terms of study design, 'n-of-1' trials are sometimes discussed as a solution to small patient numbers. Such n-of-1 trials have their place but only to establish the best treatment policy for a particular patient. The single inference from such a trial has little (or no) relevance to any other patient.[10] If several n-of-1 trials are carried out on different patients and they all begin to give the same answer, then this may begin to suggest a more widely applicable inference. However, a preplanned sequence of n-of-1 trials would be a more reliable (in terms of inter-'study' variability) and more efficient (in terms of fewer patients needed) strategy. A well-defined and preplanned sequence of n-of-1 trials is, in effect, the beginnings of the design of a crossover trial.

Sequential designs (see, for example, Whitehead)[21] require fewer patients (on average) than 'fixed-length' designs and so these may be an attractive option. Fully sequential designs are rarely used but group-sequential designs are quite common. One reason for not using fully sequential designs is the logistic (including data management) problem of updating the analysis after *every* patient. However, when there are only likely to be few patients, and their recruitment (and evaluation) will be greatly spread across time, these practical constraints may be easier to resolve.

Enrichment designs (see, for example, Liu)[22] that in some way exclude likely non-responders may be more acceptable for treating some very rare diseases than common ones. In trials for common diseases, it is important to know the 'pragmatic' answer to how a treatment will work when it is used widely in many different centres. When treating very rare conditions, since there may only be a few centres where patients will be treated, we may need to worry less about how inferences are affected by what happens if inclusion criteria or follow-up are not exactly as specified in the protocol. Finding a treatment for a highly specific disease that only works under very special conditions may be an acceptable goal.

More sophisticated statistical analyses can help to improve the efficiency of the data analysis but the acceptability and reliability of these methods has to be considered. As a theme throughout this chapter, if such methods are acceptable, then they should be acceptable in studies of common diseases as well as rare ones. Examples of more sophisticated analyses include use of covariates to model response to continuous variables such as 'age' rather than stratifying by 'age group'. This is only strictly valid if the modelling is correct and so treating age (for example) in categories might be considered a 'safe', but inefficient, analysis. In some situations it is well understood whether or not knowing this functional form is crucial; in others, it is less well understood. Analysing underlying continuous scales rather than responder rates will always be a more efficient statistical approach and it does not preclude presenting responder rates to help assess clinical significance.

Bayesian methods have also been suggested as helpful in the face of small data sets[23] and where useful prior information may exist.[24] Again, if such methods are acceptable, then they should be acceptable in studies of common diseases as well as rare ones – many people, however, still view them with caution. However, when trying to extrapolate results from adults to children (for example), much may be known about the treatment response in adults. These data might, therefore, form an obvious basis on which to base prior knowledge of the effect in children.

In using a statistical model, we need to trade off assumptions for reliability. If more assumptions can be made (and provided that they are sufficiently correct) then fewer data may be needed to draw conclusions. The step from 'non-parametric' to 'parametric' methods is a simple example. Fewer assumptions are necessary for the validity of inferences from non-parametric methods than for parametric methods. Provided the assumptions necessary for the use of parametric methods can be sufficiently relied on then, in general, fewer patients will be needed to draw similar conclusions.[25] The humble *t*-test and slightly more sophisticated analysis of covariance are, in fact, surprisingly robust to assumed problems of small samples and non-continuous data.[26,27]

SUMMARY AND CONCLUSIONS

It cannot be stressed enough that there are no special methods for designing, carrying out or analysing clinical trials in rare diseases that are not applicable to trials in large populations. The opposite is not true and some analytical methods applicable to large samples may not be reliable or valid for small samples. The only place where some methods may be usable in small studies and not in large studies is where practical and operational aspects are important. Use of sequential methods and adaptive designs may not be practical in fast-recruiting studies, but may be when recruitment is very slow.

Registers of ongoing clinical trials may help overcome problems of publication bias. Patient registers may supply crucial data for characterising the natural course of disease but they can have many problems such as selection bias, (in)completeness and data quality.

All forms of evidence can be helpful but we need to be wary of bias in uncontrolled, observational settings. Randomised, blinded, controlled clinical trials and meta-analyses of them provide the highest levels of evidence; individual anecdotal case studies provide the lowest level.

Accepting less strong evidence allows trials to be carried out more easily (typically with fewer patients or with non-concurrent controls) but it is not clear if there is uniform agreement that lesser quality evidence is acceptable.

Homogeneous conditions will be easier to study than heterogeneous ones but even in genetically homogenous conditions there may be hidden phenotypes that may respond differently. Effects of medical practice, environmental and social conditions may add to heterogeneity.

Detailed knowledge of the pharmacology of a compound may help when designing studies. Pharmacology studies (pre-clinical and clinical) may help identify sources of heterogeneity in patients. Non-clinical pharmacology (which may not be constrained by patient numbers) may be particularly helpful in conditions with very few available patients.

Endpoints should be carefully chosen to be relevant and reliable. Surrogate endpoints may be acceptable but need to be fully validated and their relationship to the 'true' clinical endpoint should be clearly understood. Ironically, in rare diseases there may not be enough patients to validate properly and independently, a 'new' surrogate endpoint.

Realistic consideration should be given to the anticipated size of treatment effect. A distinction should be drawn between large vs. small effects and whether these effects are likely to be seen in a large or small proportion of patients. Trials should then be designed accordingly.

Efficient but reliable statistical (analysis) methods should be used. It seems even more important to use such methods to 'get the most out of the data' when there are very few data than when large, often simple trials can be performed.

ACKNOWLEDGEMENTS

I am grateful for discussion with, and ideas contributed by, the following colleagues: Rembert Elbers, Bernd Jilma, Bertil Jonsson, Wolfgang Köpcke and Peter Volkers.

REFERENCES

1. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* (1984) **3**: 409–20.
2. Powell-Tuck K, MacRae KD, Healy MJ, Lennard-Jones JE, Parkins RA. A defence of the small

clinical trial: evaluation of three gastroenterological studies. *Br Med J* (1986) **292**: 599–602.

3. Evans CH, Ildstad ST (eds). *Small Clinical Trials: Issues and Challenges*. Washington, DC: Institute of Medicine (2001).
4. Sylvester R. European Organization for Research and Treatment of Cancer (EORTC). In: Redmond C, Colton T, eds, *Biostatistics in Clinical Trials*. Chichester: John Wiley & Sons, Ltd (2001) 191.
5. Noseworthy J, Kapposs L, Daumer M. Competing interests in multiple sclerosis research. *Lancet* (2003) **361**: 350–1.
6. Jadad A. *Randomised Controlled Trials*. London: BMJ Books (1998).
7. Altman DG, Schulz KF. Concealing treatment allocation in randomised trials. *Br Med J* (2001) **323**: 446–7.
8. ICH Steering Committee. *ICH Harmonised tripartite guideline. Choice of Control group in Clinical Trials*. European Union, Japan and USA: International Conference on Harmonisation (1999).
9. Johannessen T. Controlled trials in single subjects: 1. Value in clinical medicine. *Br Med J*, (1991) **303**: 173–4.
10. Lewis JA. Controlled trials in single subjects: 2. Limitations of use. *Br Med J* (1991) **303**: 175–6.
11. Day S. Controlled trials in single subjects (letter to the Editor). *Br Med J* (1991) **303**: 522.
12. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Stat Med* (2005) **24**: 329–39.
13. Keating MJ. The chronic leukemias. In: Bennett JC, Plum F, *Cecil Textbook of Medicine*. Philadelphia: W B Saunders (1996) 925–35.
14. Appelbaum FR. The acute leukemias. In: Bennett JC, Plum F, *Cecil Textbook of Medicine*. Philadelphia: W B Saunders (1996) 936–40.
15. Magalini SI, Magalini SC. *Dictionary of Medical Syndromes*, 4th edn. Philadelphia: Lippincott-Raven (1997).
16. Pérez Fontán JJ, Haddad GG. Respiratory pathophysiology. In: Behrman RE, Kliegman RM, Jenson HB, eds, *Nelson Textbook of Paediatrics*, 17th edn. Philadelphia: Saunders (2004) 1362–73.
17. Haddad GG, Palazzo RM. Diagnostic approach to respiratory disease. In: Behrman RE, Kliegman RM, Jenson HB, eds, *Nelson Textbook of Paediatrics*, 17th edn. Philadelphia: Saunders (2004) 1375–9.
18. Fayers PM, Machin D. *Quality of Life: Assessment, Analysis and Interpretation*. Chichester: John Wiley & Sons, Ltd (2000).
19. Senn S. Regression to the mean (editorial). *Stat Methods Med Res* (1997) **6**: 99–102.
20. Stiger TR, Pun EFC. Placebo effect. In: Chow S-C, ed., *Encyclopedia of Biopharmaceutical Statistics*. New York: Marcel Dekker (2000) 402–7.
21. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Chichester: John Wiley & Sons, Ltd (1997).
22. Liu J. Enrichment design. In: Chow S-C, ed., *Encyclopedia of Biopharmaceutical Statistics*. New York: Marcel Dekker (2000) 185–8.
23. Tan S-B, Dear KBG, Bruzzi P, Machin D. Strategy for randomised clinical trials in rare cancers. *Br Med J* (2002) **327**: 47–9.
24. Weeden S, Parmar M, Freedman LS. Bayesian reporting of clinical trials. In: Geller NL, ed., *Advances in Clinical Trial Biostatistics*. New York: Marcel Dekker (2004) 161–88.
25. Siegel S, Castellan NJ. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn. New York: McGraw-Hill (1998).
26. Heeren T, D'Agostino R. Robustness of the two independent samples t-test when applied to ordinal scaled data. *Stat Med* (1987) **6**: 79–90.
27. Sullivan RB, D'Agostino RB Sr. Robustness and power of analysis of covariance applied to ordinal scaled data arising in randomized controlled trials. *Stat Med* (2003) **22**: 1317–34.

Index

Note: page numbers in *italics* refer to figures and tables.

Textbook of Clinical Trials, Second Edition. Edited by D. Machin, S. Day and S. Green

© 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01014-2