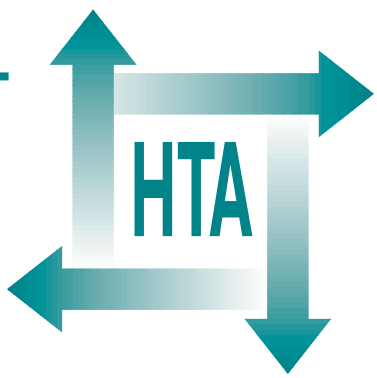
Health Technology Assessment 1998; Vol. 2: No. 13

Review

Choosing between randomised and non-randomised studies: a systematic review

A Britton
M McKee
N Black
K McPherson
C Sanderson
C Bain

Health Technology Assessment
NHS R&D HTA Programme



Standing Group on Health Technology

Chair: Professor Sir Miles Irving,
Professor of Surgery, University of Manchester, Hope Hospital, Salford †

Dr Sheila Adam,
Department of Health

Professor Martin Buxton,
Professor of Economics, Brunel University †

Professor Angela Coulter,
Director, King's Fund, London

Professor Anthony Culyer,
Deputy Vice-Chancellor, University of York

Dr Peter Doyle,
Executive Director, Zeneca Ltd,
ACOST Committee on Medical Research & Health

Professor John Farndon,
Professor of Surgery, University of Bristol †

Professor Charles Florey,
Department of Epidemiology & Public Health, Ninewells Hospital & Medical School, University of Dundee †

Professor John Gabbay,
Director, Wessex Institute for Health Research & Development †

Professor Sir John Grimley Evans,
Department of Geriatric Medicine, Radcliffe Infirmary, Oxford †

Dr Tony Hope,
The Medical School, University of Oxford †

Professor Howard Glennester,
Professor of Social Science & Administration, London School of Economics & Political Science

Mr John H James,
Chief Executive, Kensington, Chelsea & Westminster Health Authority

Professor Richard Lilford,
Regional Director, R&D, West Midlands †

Professor Michael Maisey,
Professor of Radiological Sciences, UMDS, London

Dr Jeremy Metters,
Deputy Chief Medical Officer, Department of Health †

Mrs Gloria Oates,
Chief Executive, Oldham NHS Trust

Dr George Poste,
Chief Science & Technology Officer, SmithKline Beecham †

Professor Michael Rawlins,
Wolfson Unit of Clinical Pharmacology, University of Newcastle-upon-Tyne

Professor Martin Roland,
Professor of General Practice, University of Manchester

Mr Hugh Ross,
Chief Executive, The United Bristol Healthcare NHS Trust †

Professor Ian Russell,
Department of Health, Sciences & Clinical Evaluation, University of York

Professor Trevor Sheldon,
Director, NHS Centre for Reviews & Dissemination, University of York †

Professor Mike Smith,
Director, The Research School of Medicine, University of Leeds †

Dr Charles Swan,
Consultant Gastroenterologist, North Staffordshire Royal Infirmary

Dr John Tripp,
Department of Child Health, Royal Devon & Exeter Healthcare NHS Trust †

Professor Tom Walley,
Department of Pharmacological Therapeutics, University of Liverpool †

Dr Julie Woodin,
Chief Executive, Nottingham Health Authority †

† Current members

HTA Commissioning Board

Chair: Professor Charles Florey, Department of Epidemiology & Public Health,
Ninewells Hospital & Medical School, University of Dundee †

Professor Ian Russell,
Department of Health, Sciences & Clinical Evaluation, University of York *

Dr Doug Altman,
Director of ICRF/NHS Centre for Statistics in Medicine, Oxford †

Mr Peter Bower,
Independent Health Advisor, Newcastle-upon-Tyne †

Ms Christine Clark,
Honorary Research Pharmacist, Hope Hospital, Salford †

Professor David Cohen,
Professor of Health Economics, University of Glamorgan

Mr Barrie Dowdeswell,
Chief Executive, Royal Victoria Infirmary, Newcastle-upon-Tyne

Professor Martin Eccles,
Professor of Clinical Effectiveness, University of Newcastle-upon-Tyne †

Dr Mike Gill,
Director of Public Health and Health Policy, Brent & Harrow Health Authority †

Dr Jenny Hewison,
Senior Lecturer, Department of Psychology, University of Leeds †

Dr Michael Horlington,
Head of Corporate Licensing, Smith & Nephew Group Research Centre

Professor Sir Miles Irving
(Programme Director), Professor of Surgery, University of Manchester, Hope Hospital, Salford †

Professor Alison Kitson,
Director, Royal College of Nursing Institute †

Professor Martin Knapp,
Director, Personal Social Services Research Unit, London School of Economics & Political Science

Dr Donna Lamping,
Senior Lecturer, Department of Public Health, London School of Hygiene & Tropical Medicine †

Professor Theresa Marteau,
Director, Psychology & Genetics Research Group, UMDS, London

Professor Alan Maynard,
Professor of Economics, University of York †

Professor Sally McIntyre,
MRC Medical Sociology Unit, Glasgow

Professor Jon Nicholl,
Director, Medical Care Research Unit, University of Sheffield †

Professor Gillian Parker,
Nuffield Professor of Community Care, University of Leicester †

Dr Tim Peters,
Reader in Medical Statistics, Department of Social Medicine, University of Bristol †

Professor David Sackett,
Centre for Evidence Based Medicine, Oxford

Professor Martin Severs,
Professor in Elderly Health Care, Portsmouth University †

Dr David Spiegelhalter,
MRC Biostatistics Unit, Institute of Public Health, Cambridge

Dr Ala Szczepura,
Director, Centre for Health Services Studies, University of Warwick †

Professor Graham Watt,
Department of General Practice, Woodside Health Centre, Glasgow †

Professor David Williams,
Department of Clinical Engineering, University of Liverpool

Dr Mark Williams,
Public Health Physician, Bristol

Dr Jeremy Wyatt,
Senior Fellow, Health and Public Policy, School of Public Policy, University College, London †

* Previous Chair
† Current members

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Choosing between randomised and non-randomised studies: a systematic review

A Britton[1]
M McKee[1]
N Black[1]
K McPherson[1]
C Sanderson[1]
C Bain[2]

[1] London School of Hygiene and Tropical Medicine, University of London, UK
[2] University of Queensland, Australia

Published October 1998

This report should be referenced as follows:

Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assessment* 1998; **2**(13).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE. Copies of the Executive Summaries are available from the NCCHTA web site (see overleaf).

NHS R&D HTA Programme

The overall aim of the NHS R&D Health Technology Assessment (HTA) programme is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and work in the NHS. Research is undertaken in those areas where the evidence will lead to the greatest benefits to patients, either through improved patient outcomes or the most efficient use of NHS resources.

The Standing Group on Health Technology advises on national priorities for health technology assessment. Six advisory panels assist the Standing Group in identifying and prioritising projects. These priorities are then considered by the HTA Commissioning Board supported by the National Coordinating Centre for HTA (NCCHTA).

This report is one of a series covering acute care, diagnostics and imaging, methodology, pharmaceuticals, population screening, and primary and community care. It was identified as a priority by the Methodology Panel and funded as project number 93/45/06.

The views expressed in this publication are those of the authors and not necessarily those of the Standing Group, the Commissioning Board, the Panel members or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for the recommendations for policy contained herein. In particular, policy options in the area of screening will, in England, be considered by the National Screening Committee. This Committee, chaired by the Chief Medical Officer, will take into account the views expressed here, further available evidence and other relevant considerations.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The editors have tried to ensure the accuracy of this report but cannot accept responsibility for any errors or omissions. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

Contents

List of abbreviations

ACE	angiotensin-converting enzyme
AIMS	APSAC (anisoylated plasminogen streptokinase activator complex) Intervention Mortality Study
(A)MI	(acute) myocardial infarction
ASAT	aspartate aminotransferase
ASSET	Anglo-Scandinavian Study of Early Thrombolysis
ATBC	Alpha-Tocopherol, Beta-Carotene and Cancer Prevention Study
BARI	Bypass Angioplasty Revascularisation Investigation
BHAT	Beta-blocker Heart Attack Trial
BIP	Bezafibrate Infarction Prevention
CABG	coronary artery bypass grafting
CABRI	Coronary Angioplasty versus Bypass Revascularisation Investigation
CASS	Coronary Artery Surgery Study
CHD	coronary heart disease
CI	confidence interval
CMF	chlorambucil, melphalan, flurouracil
CONSORT	Consolidation of Standards for Reporting Trials
CVD	cardiovascular disease
CVS	chorionic villus sampling
DCCT	Diabetes Control and Complications Trial
EA	early amniocentesis
EAST	Emory Angioplasty versus Surgery Trial
ECOG	Eastern Cooperative Oncology Group
EF	ejection fraction*
EORTC	European Organisation for Research and Treatment of Cancer
ERACI	Argentine Randomised Trial of Percutaneous Transluminal Coronary Angioplasty versus Coronary Artery Bypass Surgery in Multivessel Disease
GABI	German Angioplasty Bypass-surgery Intervention Trial
GBSG	German Breast cancer Study Group
GISSI	Italian group studies into treatment of myocardial infarction
hCG	human chorionic gonadotrophin
HINT	Holland Interuniversity Nifedipine/metoprolol Trial
ISIS	International Study of Infarct Survival
LHRH	luteinising hormone-releasing hormone
MASS	Medicine, Angioplasty or Surgery Study
MeSH	MEDLINE Search Heading
MRC	Medical Research Council
NIH	National Institutes of Health
NS	not significant*
NSAIDs	non-steroidal anti-inflammatory drugs
PTCA	percutaneous transluminal coronary angioplasty
RCT	randomised controlled trial
RITA	Randomised Intervention Treatment of Angina
RMITG	Recurrent Miscarriage Immunotherapy Trialists Group
RR	relative risk
SD	standard deviation
SHEP	Systolic Hypertension in the Elderly Program
SPRINT	Secondary Prevention Reinfarction Israeli Nifedipine Trial
TRACE	Trandopapril Cardiac Evaluation
TRENT	Trial of Early Nifedipine Treatment
WESDR	Wisconsin Epidemiologic Study of Diabetic Retinopathy

* *Used only in tables*

Executive summary

Background

Studies that compare healthcare interventions can be divided into those that involve randomisation of subjects between comparison groups, and those that do not. The former, in its commonest form the randomised controlled trial (RCT), is seen by many as the 'gold standard' as it should ensure that subjects being compared differ only in their exposure to the intervention being considered. The RCT has been criticised, however, with some arguing that design features tend to exclude many individuals to whom the results will subsequently be applied. Furthermore, practitioner and patient preferences may influence the outcome of treatment and cause the results to be misleading. These criticisms have led some to advocate the use of non-randomised designs.

Objectives

This review explored those issues related to the process of randomisation that may affect the validity of conclusions drawn from the results of RCTs and non-randomised studies.

Methods

The review was based on a series of systematic reviews involving structured searches of databases. Details of the methods used are described in the main report. Four research questions were addressed.

- Do non-randomised studies differ systematically from RCTs in terms of treatment effect?
- Are there systematic differences between included and excluded individuals and do these influence the measured treatment effect?
- To what extent is it possible to adjust for baseline differences between study groups?
- How important is patient preference in terms of outcome?

Results

Previous comparisons of RCTs and non-randomised studies

Eighteen papers that directly compared the results of RCTs and prospective non-randomised studies were found and analysed. No obvious patterns emerged; neither the RCTs nor the non-randomised studies consistently gave larger or smaller estimates of the treatment effect. The type of intervention did not appear to be influential, though more comparisons need to be conducted before definite conclusions can be drawn.

Several reasons emerged as to why RCTs might produce a greater or lesser estimate of treatment effect than non-randomised studies. A greater effect may occur in RCTs if patients receive higher quality care or are selected in a way that gives greater capacity to benefit. A lower estimate of treatment effect may occur if:

- patient selection produces a study population with less capacity to benefit than would be the case in non-randomised studies
- strong patient preference exists against a particular treatment in an unblind RCT, thus reducing the treatment effect
- non-randomised studies of preventive interventions include a disproportionate number of people with greater capacity to benefit
- publication bias exists; negative results are less likely to be published from non-randomised trials than from RCTs.

Exclusions

The number of eligible subjects included in the RCTs ranged from 1% to 100%. Reasons for exclusions may be medical (e.g. high risk of adverse events in certain groups) or scientific (selecting only small homogeneous groups in order to increase the precision of estimated treatment effects). Blanket exclusions (e.g. the elderly, women of childbearing potential) are also common in RCTs.

Large clinical databases containing detailed information on patient severity and prognosis have been used instead of RCTs, and where database subjects are selected according to the same **inclusion** criteria as RCTs, the treatment effects of the two methods are similar.

Participation

Most RCTs failed to document adequately the characteristics of eligible individuals who did not participate in trials. However, RCTs were more likely than non-randomised trials to include

university and teaching centres and this may have exaggerated the treatment effect measured in the RCTs.

Participation in RCTs differed between studies of treatment interventions (subjects tended to be less affluent, less educated and more severely ill and therefore had greater capacity to benefit from treatment) and those evaluating preventive interventions (more affluent, better educated and generally healthier and therefore had less potential to benefit than eligible subjects who declined to participate).

Adjusting for baseline differences

Adjustment for differences in baseline prognostic factors in non-randomised studies often changed the treatment effect size but not significantly; importantly, the direction of change was inconsistent. Most of the case studies were too small to draw conclusions but where this was possible, the superiority of one treatment over another was probably a function of the patients' clinical characteristics.

Patient preference

Only four papers directly addressed the role of patient preference on trial results. However, preference could account for some of the observed differences between RCTs and non-randomised studies.

Conclusions

Results of RCTs and non-randomised studies do not inevitably differ, and the available evidence suffers from many limitations. It does, however, suggest that it may be possible to minimise any differences by ensuring that subjects included in each type of study are comparable. The effect of adjustment for baseline differences between groups in non-randomised studies is inconsistent but, where it is done, it should involve rigorously developed formulae. Existing studies have generally been too small to assess the impact of such adjustment.

Implications for policy

While a high level of exclusion may have some advantages for those conducting an RCT, it also has important implications for policy. In particular, there is a risk of denial of effective treatment to those who might benefit but who have been excluded from the RCTs, and delay in obtaining definitive results because of low recruitment rate. In addition, there is a danger of unjustified extrapolation of results to other populations, and it is concluded that it should **not** be assumed that summary results apply equally to all potential patients.

Recommendations for research

Conducting research

- A well-designed non-randomised study is preferable to a small, poorly designed and exclusive RCT.
- RCTs should be pragmatic by including as wide a range of practice settings as possible. Study populations should be representative of all patients currently being treated for the condition.
- Exclusions for administrative convenience should be rejected.

Interpretation

- Heterogeneity of populations and interventions should be addressed explicitly. Practitioners should apply caution when extrapolating to populations that differ from those included in RCTs.
- For both study designs, authors should define their reference population, state the steps taken to ensure the study population is a representative sample or explain how it differs. They should also give details of patient and centre participation and the characteristics of eligible individuals who did not participate.
- Further research is required on patient characteristics, long-term follow-up, participation of centres and practitioners and patient preference.

Chapter 1
Introduction

Background

Practitioners, providers and purchasers increasingly are seeking to enhance the level of sophistication of commissioning health care. These activities require an understanding of the methods available for evaluating healthcare interventions and comparing institutions and policies. Of the principal methods available to measure effectiveness, one important distinction is between randomised controlled trials (RCTs) and non-randomised studies. The latter include quasi-experiments, natural experiments, and observational studies, which may be prospective or retrospective cohort studies or case–control studies. For the purposes of the present review, we have focused principally on prospective studies that were designed for the purposes of research.

Randomised and non-randomised designs each have their particular limitations. The objectives of this review are to examine systematically the methodological issues facing those who decide which approach to adopt in particular circumstances, or who seek to apply the results of such studies to formulating clinical practice and healthcare policy.

The potential contribution that RCTs and non-randomised studies can make to the evaluation of effectiveness has generated considerable controversy. At the outset, therefore, it is important to identify those areas about which there is consensus and those about which there is disagreement. It is widely agreed that a large, well-designed and conducted RCT, in which randomisation is undertaken in a way that ensures that allocation is actually random, will provide groups that can be expected to be comparable in every way except for the intervention. Consequently, it is reasonable to attribute any significant observed difference in outcome between the two groups to differences in the effect of the interventions, provided both the practitioner and patient are blinded and the outcome is assessed in a way that does not introduce bias. It is also widely accepted that in a non-randomised study, the comparison groups may differ, so that any observed difference in outcome between those receiving the intervention and those not may be due to differences in the characteristics of the two groups rather than the effectiveness of the intervention.

Beyond these two points, consensus breaks down. Some people believe that undetected and, potentially undetectable, differences between groups in non-randomised studies render such studies valueless. An eminent statistician has argued that scientific committees should "just say no" to non-randomised studies because of their "inherent bias".[1] Conversely, it has been argued that, in some circumstances, randomisation is unnecessary, inappropriate, misleading or impossible.[2] Thus, randomisation is unnecessary when the effect of an intervention is so dramatic that the contribution of unknown confounding factors can plausibly be ignored. Examples include penicillin for streptococcal infection and defibrillation for ventricular fibrillation. Randomisation may be inappropriate where a trial would have to be of disproportionate size and duration, and thus expense, if it were to detect very rare long-term adverse effects or the impact of policies designed to prevent rare events. An example is the case of benoxaprofen, an anti-inflammatory drug for which no adverse outcomes were detected in trials including over 3000 patients but which was subsequently found, from post-marketing surveillance, to have been associated with 61 deaths.[3]

Randomisation may be misleading where the process of random allocation may affect the effectiveness of the intervention. This can arise when the subjects cannot be blinded to the intervention because the intervention requires their active participation, which in turn will be affected by their underlying beliefs and preferences. An example would be a trial of the effectiveness of clinical audit in improving the quality of patient care, which would be complicated because effectiveness depends on the attitudes of the participating clinicians.[4] In some situations, experimentation may be impossible in practice as clinicians may not accept that there is uncertainty about the relative effectiveness of different interventions and thus deem such a trial to be unethical. Finally, in practice, trials may be less free from bias than previously supposed, with systematic differences arising from incomplete blinding during random-isation or assessment of outcome.[5,6]

These arguments have, in turn, been countered by the more enthusiastic advocates of RCTs who argue that problems of size and duration can be overcome given sufficient funding and that the argument that a particular RCT is unethical may, itself, be an unethical position if patients are otherwise subjected to unevaluated interventions.

A further argument addresses the issue of the generalisability of the results of RCTs.[7] Here, it is suggested that the process through which patients become involved in trials, including differential participation by centres, practitioners or patients, may limit the confidence with which the results can be applied in routine practice. It is argued that non-randomised studies, which often have more inclusive entry criteria and procedures, may include subjects that are more representative of the population to whom the results will be applied. Furthermore, the process of recruitment, in which only individuals willing to be randomised will be recruited, may introduce hidden biases that make subjects unrepresentative of the reference population.

These positions have been characterised by strongly held views based on limited empirical evidence and many of the arguments remain unresolved. Even where there is some measure of agreement that a particular issue, such as differing levels of participation or eligibility, had an effect on the results of an evaluation, there is often little common ground about how important such an effect is.

This review is a contribution to the debate. It seeks to define some of the key unresolved questions, to examine the evidence in support of the differing positions, and to suggest priorities for further research to help resolve continuing uncertainties.

Comparability of results of non-randomised studies and RCTs

It is known that when an intervention is assessed by both a non-randomised study and an RCT, the results obtained can differ (though it should also be noted, of course, that the results obtained from RCTs often differ from one another as do the results of non-randomised studies). It is claimed that results from non-randomised studies generally suggest larger treatment effect sizes than RCTs.[8,9] This argument appears to be based on two pieces of evidence. The first is a very small number of frequently quoted examples that directly compare the two methods. The second, and less direct line of evidence comes from studies of RCTs with varying quality of randomisation, which have shown that inadequate or unclear methods of treatment allocation exaggerate effects.[5]

Any differences are likely to reflect the interaction of several different factors. These have been set out in *Table 1* along with solutions that have been proposed to overcome them. It has been argued that non-randomised studies face greatest risk to their internal validity through allocation bias, though it may be possible to overcome this during analysis, by either risk adjustment or examination of comparable sub-groups. The internal validity of unblind RCTs may also be threatened by the consequences of patient preference, which may result in misleading estimations of treatment effect. Preference arms have been advocated as one possible solution to this problem.

Threats to external validity have been seen as of greater importance for RCTs, though the issue also arises in non-randomised studies. This can be due to restricted eligibility criteria, which can be addressed by expanding the criteria, or limited participation by centres or by patients and practitioners, which can be addressed by ensuring the design is pragmatic with less demanding consent requirements. Finally, potential subjects may not be invited to participate in the research because of practitioner preferences for one of the treatments or simply due to administrative oversight.

This review draws on these validity threats to examine the question of comparability from several directions.

TABLE 1 *Threats to validity of evaluative research and possible solutions*

	Threatening factors	**Proposed solution**
Internal validity	Allocation bias (risk of confounding)	Risk adjustment and sub-group analysis (analysis)
	Patient preference	Preference arms or adjustment for preference (design)
External validity	Exclusions (eligibility criteria)	Expand inclusion criteria
	Non-participation (centres/practitioners)	Multicentre, pragmatic design
	Not invited (practitioner preference or administrative oversight)	Encourage practitioners to invite all eligible patients
	Non-participation (patients)	Less rigorous consent procedures

Internal validity

Allocation bias

A major criticism of non-randomised studies is the possibility that groups being compared differ in prognostically important characteristics. The key question is whether the potential for allocation bias can be identified and, if present, whether it can be overcome by analysis rather than study design. Particular problems arise where large numbers of factors contribute to the risk of an adverse outcome.[10] Many of these factors may be unknown, sub-group analysis may fail because of small numbers in each sub-group, and adjustment models may be vulnerable to high levels of correlation between treatment and prognostic factors. In the present context, it is important to know whether attempts at adjustment for confounding can bring the results of non-randomised studies closer to those from RCTs, and if so, in which circumstances.

Preference

The possibility that a preference for a treatment will enhance its therapeutic effect (and conversely that preference for another treatment will dilute the effect of the treatment offered) has attracted interest.[11] The existence of patients' preferences for particular treatments can be studied without difficulty but their attributable therapeutic effects remain poorly quantified. RCTs, by necessity and desire, ignore preference effects and so may, as a result, underestimate the main treatment effects.

The difficulty is that if such preference effects exist, the ability to detect them reliably is always compromised by the serious possibility of confounding by selection.[12] People who tend to prefer something may well be different in some other consistent ways, plausibly related to prognosis, from those who do not.[13,14] Unfortunately, when people have strong preferences, randomisation becomes difficult;[15] one can never randomise between enthusiasts for a treatment and those who strongly reject it. An essential, but neglected, part of the research agenda is to disentangle the main physiological effect of a treatment from any possible benefit of preferences.

The idea that randomisation itself can give rise to biased results about outcome may seem surprising. However, when treatments are allocated randomly practitioners and patients are deprived of expressing a preference and, if choice and control are of therapeutic benefit, then a randomised comparison might provide a reduced estimate of the effectiveness of treatment. There are several plausible mechanisms for such an effect. These have been examined in research on the psychology of the placebo effect[16] as well as in research which demonstrates the role of a person's social or professional control over their lives in the aetiology of a condition such as coronary heart disease (CHD).[17] Consequently there is a need for systematic study of the role of choice and control and thus their impact on the results of RCTs.

External validity (exclusions and participation)

RCTs have been criticised because they often exclude many types of patients about whom clinicians seek advice on treatment. Most obviously, many exclude women,[18] the elderly,[19] those with strong preferences and those with multiple pathology or severe disease.[20] There may, however, be other important exclusions that are less obvious. Information on the characteristics of eligible patients and the proportion of all patients that they represent is usually not supplied. The impact that this can have was illustrated in a series of meta-analyses that examined the amount of heterogeneity among the results of randomised trials that could be ascribed to specific measurable characteristics of the study population.[21] Furthermore, the inclusion criteria also depend on the case-definition used, how the intervention is defined, and the features of the population in the area where the study is being undertaken.[22] A further concern about generalisability concerns the setting of studies, which may not be typical of those in which most people are treated. Some strategies have been developed to evaluate the effect of individual institutions in multicentre trials.[23] Strategies to determine the extent to which the results of studies may be generalised have been developed[24] but this issue requires further clarification. Consequently, it is necessary to examine the implications for generalisability of exclusion criteria and selective participation on the part of centres, patients and practitioners.

Developing a model

As a first step in bringing together these issues, we have developed a model that relates the people included in the different arms of a study to the population to whom the results will apply. The basic model is shown in *Figure 1*. It encompasses several, but not all of the issues being considered and it involves a number of simplifications. Nonetheless, it does help to illustrate the consequences of certain phenomena to those who are less familiar with the methodological issues concerning evaluative research. The reference population is defined by an envelope, with a vertical axis repre-

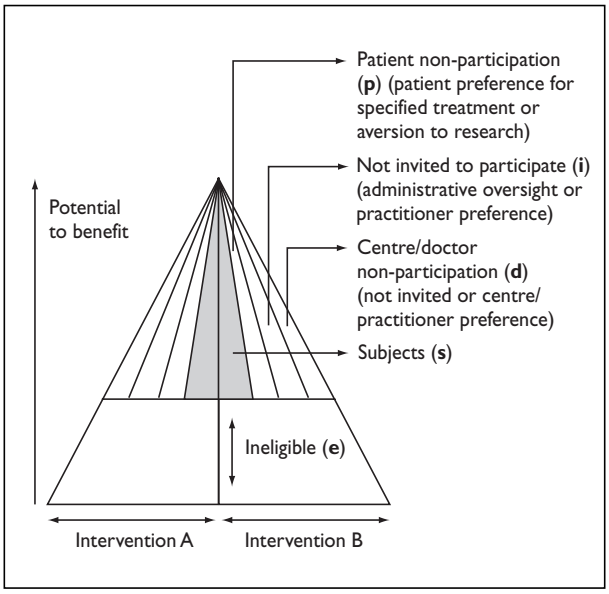


FIGURE 1 *Basic model*

senting capacity to benefit from the treatment in question. At some point, a lower threshold is reached below which the overall risks outweigh the benefits. As subjects are excluded or do not participate, the study population (designated **s** in *Figure 1*) becomes a progressively smaller subset of the reference population, raising the question of whether it is valid to apply results obtained from this sub-sample to the reference population. In the figure, the envelope is triangular, representing the common situation in which a small number of people have a large capacity to benefit, with larger numbers benefiting less. The envelope and the sub-groups within it can, of course, take many shapes depending on the condition and intervention.

One simplification is the assumption that the true treatment effect can be expressed in relation to **potential to benefit**, representing the balance of risks and benefits of the particular intervention. The nature of this 'potential to benefit' and how it might be measured, will vary widely for different interventions. Individual subjects at the top of the figure have most potential to benefit whereas those at the bottom have least and, most often, would actually have a net negative potential as risks would outweigh any potential benefit.

The outermost line delineates the population to which the results of an evaluation are intended to be applied in routine clinical practice. In an ideal situation, this population would be divided, randomly or otherwise, into two or more groups, each of which would be comparable with regard to every factor that can influence the outcome of the interventions. Each of the interventions being compared (one of which may be a placebo or watchful waiting) would then be administered to these groups.

In practice, not all members of the population will be included. Some will be excluded as they fall outside the eligibility criteria (**e**) defined by those who designed the evaluation. Some will be excluded because they are under the care of doctors or centres who have either not been invited or who have decided not to participate in the evaluation (**d**). Some potential subjects will be excluded as practitioners have a preference as to which intervention to use and therefore do not ask them to participate in the evaluation, or the eligible subjects are not invited simply due to an administrative oversight (**i**). Finally, some of those invited to participate will decline either because they do not wish to be in a research study or because they have a preference for one of the interventions (**p**). As noted above, this leaves the study subjects (**s**).

For an evaluation to have both internal and external validity, two conditions must be fulfilled. The first is that the groups remaining after these exclusions should be similar in terms of ability to benefit, that is, they have the same shape and position in the model, to ensure high internal validity. The second is that the gap between the subjects included and the overall population of interest should be as small as possible and at least represent the full range of potential to benefit, to ensure high external validity. One defence against the problem of external validity is for the investigators to define the population of interest in very restrictive terms. However, this is seldom done explicitly, and in practice trial results are applied to patients who would have been excluded. How these two conditions might be affected in different circumstances and how some hypotheses can be tested are discussed below.

A rigorous process of randomisation and high follow-up rates should ensure high internal validity, though such processes are susceptible to bias introduced by attempts to circumvent them.[5] In contrast, a non-randomised study can be more susceptible to differences between the groups. Consider a situation in which two interventions, A and B, were being compared by examining the results in two regions, X and Y. Region X treats patients almost exclusively with intervention A and region Y with intervention B. Only teaching hospitals are included in region X whereas all hospitals are included in region Y. The consequences are shown in *Figure 2*. The comparison

groups may have differing capacity to benefit and thus it will not be possible to attribute with certainty any difference in outcome to the treatments used.

A similar situation arises when the practitioners in one region have adopted different criteria for treating patients in the evaluation (*Figure 3*). Situations where clinicians are less willing to invite patients to participate in the evaluation or where the patients are less willing to participate once invited are variants on *Figure 2*. For example, *Figure 4* illustrates the consequence of greater levels of non-participation by centres, professionals or patients but where this does not affect the representativeness of the resulting sample. Conversely, *Figure 5* illustrates the risks of an unrepresentative sample, in which a disproportionate proportion of those with least ability to benefit do not participate.

Although not possible to show on a two-dimensional representation, there are further potential levels of complexity, such as whether the two interventions were administered at the same time, either in simple chronological terms, or in relation to the introduction of the intervention in the two regions.

In summary, therefore, it is hypothesised that an RCT may have advantages over non-randomised studies as regards internal validity, unless it can be shown that it is possible to adjust adequately for differences in case-mix or define sub-groups that are truly comparable. However, it is also hypothesised that inadequately blinded RCTs are vulnerable to preference effects and that generalisation of results to the entire population of interest, will be valid only when the subjects included in the study are representative of that population. This condition might not be achieved if, for example, only 'centres of excellence' agreed to participate in the evaluation, thus increasing **d**. Also it might not be achieved if eligibility criteria were defined in such a way as to exclude many of those for whom practitioners might consider the interventions in question appropriate, thus increasing **e**, or if particular groups of individuals were not invited to participate or chose not to accept, increasing **i** and **p**, respectively. It is hypothesised that these effects are likely to be greater with RCTs. The potential impact on effect size is illustrated in *Figure 6*.

Research questions

In this review we have attempted to investigate some of the threats to internal and external validity and to define the nature and magnitude of the

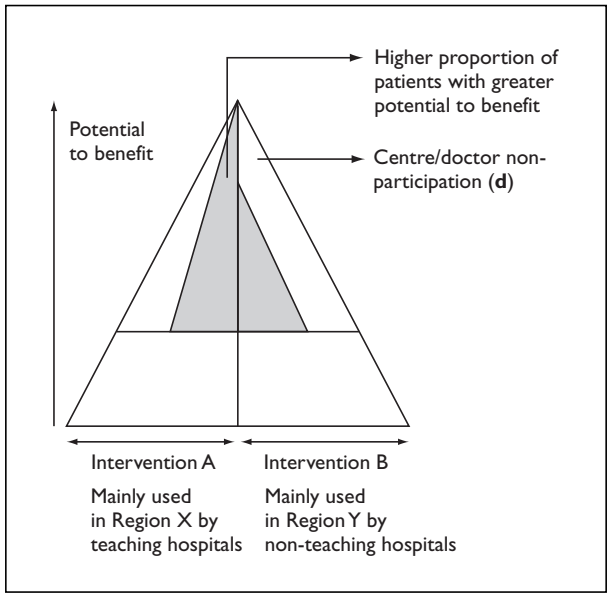


FIGURE 2 *Effects of differences in centre participation in a non-randomised study*

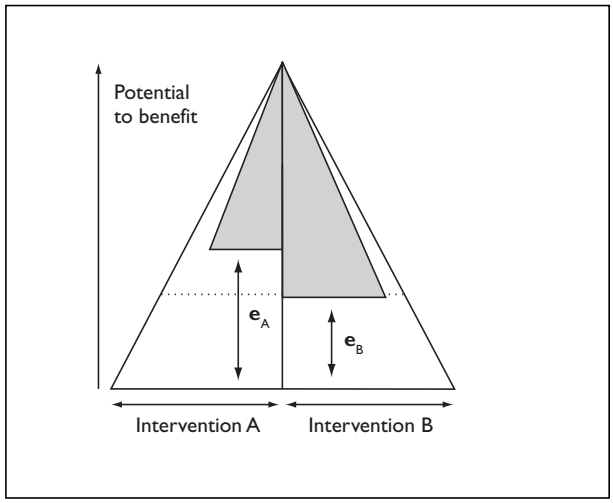


FIGURE 3 *Effects of differences in eligibility in a non-randomised study*

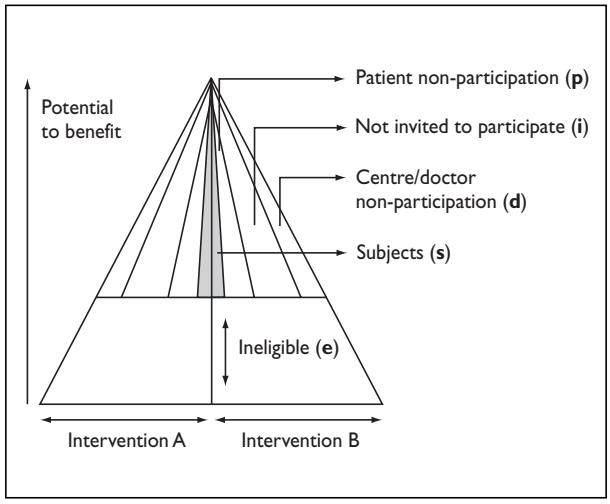


FIGURE 4 *Illustration of representative non-participation*

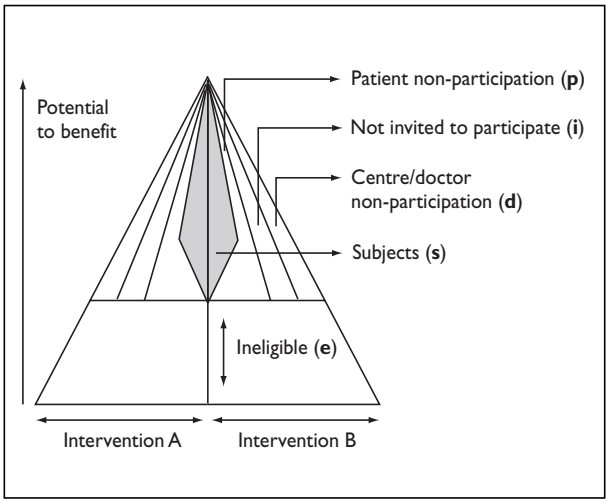


FIGURE 5 *Illustration of unrepresentative non-participation*

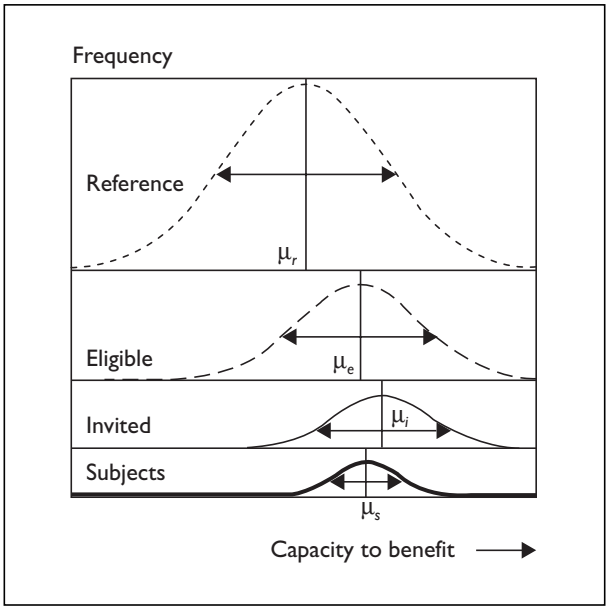


FIGURE 6 *Consequences for capacity to benefit of selective eligibility/invitation/participation*

parameters, **e**, **d**, **i** and **p**. We have confined our attention to issues that relate directly to the choice between RCTs and non-randomised studies and have not sought to examine other issues, such as timing, loss to follow-up, and methods of measuring outcome, as these apply equally to both designs regardless of how the interventions are allocated. Furthermore, this review should be read in conjunction with a related one that examines some of the issues of internal validity in more detail.[25]

Four research questions have been investigated in this review.

- Do non-randomised studies differ in the magnitude of the effect of a new intervention compared with an RCT, and in what circumstances?
- Do the parameters **e**, **d**, **i** and **p**, as threats to external validity, act consistently and, if so, what impact do these differences have on measures of effect? Although these threats can affect both RCTs and non-randomised studies, the present review is limited to their effect on RCTs. However, many of the issues relate more generally to participation in and exclusion from evaluative research. A related question is whether the consequences of restricted study populations have implications for generalisability and, specifically are the results of RCTs generalisable to routine clinical practice?
- In non-randomised studies, to what extent is it possible to compensate for potential allocation bias by adjusting for differences on the basis of measurable variables, so as to achieve results comparable to those in an equivalent RCT?
- How important are the preferences of subjects for a particular intervention and, if they are randomised to receive one that they would not have chosen, what is the potential effect on their outcome? Does this limit the practical value of treatment effect estimates from RCTs?

The review concludes by identifying ways in which any confirmed threats to validity might be overcome in order to define priorities for further methodological research.

Chapter 2
Methods

The four areas being addressed by this review are:

- the comparability of results from published non-randomised studies and RCTs
- the generalisability of results
- the ability to exclude confounding in non-randomised studies
- the role of patient preference.

Search strategy

For each of the four areas of study a review of relevant literature was undertaken, involving databses (MEDLINE, EMBASE, Science and Social Science Citation Index, and the Cochrane Library) using thesaurus terms and free text, as appropriate. For each topic, additional details of the search strategy and of the numbers of papers included at each step are given in appendix 1. Abstracts were initially screened by one of the team (AB) against a set of criteria (discussed in the corresponding chapters) and, where they met these criteria, full copies were obtained. In addition, cited references were obtained as were other papers identified through contact with other researchers. For each topic, all of the papers were then read independently by two members of the research team and relevant information extracted.

Review inclusion criteria

Randomised and non-randomised studies

The search terms used to identify studies that compared randomised and non-randomised studies are shown in appendix 1. The following inclusion criteria were applied to the search results.

- The results of the RCT must be compared with a non-randomised study, or the results of several RCTs combined compared with several non-randomised studies combined.
- The intervention must be the same and in similar settings.
- The control arms of the studies must receive similar therapy.
- There must be comparable outcome measures, preferably valid and reliable.

Generalisability of study results

Three aspects of generalisability are particularly relevant: eligibility criteria, participation of centres/practitioners and participation of subjects in trials. Systematic reviews of the literature were performed to assess the extent to which these have been shown to limit the generalisability of randomised trial results. The search terms used are shown in appendix 1. To explore the issue of centre participation, two recently completed systematic reviews that included both RCTs and non-randomised studies were examined to determine the extent to which participation differs by study design.

Patient preference in RCTs

This question was approached by means of a review of literature that had attempted explicitly to measure the effects of patient preference. An algebraic model was devised to quantify the possible bias that could be introduced by hypothetical preference effects.

Excluding confounding in non-randomised trials

In addition to examining existing research that had explicitly considered questions of exclusion, participation, and preference, a series of case studies were undertaken. In part these complemented the existing research in the other areas, but they were also designed to examine specifically the question of the effects of risk adjustment in non-randomised studies with respect to any differences in measured treatment effect compared with RCTs.

It was considered *a priori* that the nature of the intervention could raise specific issues so it would be important to try to encompass the spectrum of healthcare interventions. Discussion within the research team led to the identification of four broad categories of healthcare interventions that can be evaluated: surgical interventions, pharmaceutical interventions, organisational interventions, and preventive interventions. For each area a specific example was sought to illustrate the problems of confounding and

risk adjustment. Criteria for selection were the existence of at least one large, well-conducted non-randomised study and a randomised trial, or, preferably, a meta-analysis of RCTs, in which the treatment effect had been measured in a comparable manner. The following were selected:

- surgical intervention – coronary artery bypass grafting (CABG) versus percutaneous transluminal coronary angioplasty (PTCA)
- pharmaceutical intervention – calcium antagonists
- organisational intervention – stroke units
- preventive intervention – malaria vaccines.

Data extraction and synthesis

For each of the questions, the evidence was summarised and the implications for practice and research were discussed. Where evidence was lacking, as in the issue of preference, indirect evidence was used to develop a conceptual model that will enable future researchers to formulate appropriate questions.

Chapter 3

Comparing the outcome in RCTs and non-randomised studies

Introduction

When the same intervention is assessed in two separate studies, the results are rarely exactly the same. This may be because the settings, the subjects or the standards of care are slightly different, or simply due to chance. An added complication potentially arises when the allocation of subjects to treatments involves different processes. To what extent may the disparate results be due to these differences?

It has been noted in chapter 1 that studies in which subjects are randomly allocated to interventions often produce different estimates of treatment effect from those derived from non-randomised studies. In particular, the view is widely held that non-randomised studies tend to report larger estimates of treatment effects than those using random allocation.[26,27] The evidence most frequently cited in support of this comes from two papers published in the early 1980s. One compared patterns of risk factors in the comparison groups in studies of treatments for myocardial infarction (MI) and found that non-randomised and, to a lesser extent, unblinded randomised studies tended to have more subjects with a good prognosis in the group receiving the new treatment.[28] The second paper compared RCTs with studies using historical controls (non-parallel cohort studies) for six conditions, and concluded that non-randomised studies produced larger treatment effects, but found one for which a RCT did not.[29] However, as the authors noted, at least part of this difference may have been the result of some RCTs being too small to have the power to detect any effect, even if one existed, and there may have been some publication bias, with negative RCTs more likely to be published than negative non-parallel cohort studies. These papers leave several questions unanswered. Are the findings of the first paper a consistent phenomenon, and do the findings in the second paper, which relate to non-parallel cohort studies, and thus are susceptible to the effects of other contemporaneous changes, have any relevance to the now more common non-randomised design, the prospective cohort study?

The objectives of this chapter are:

- to describe the papers in which the results of RCTs and prospective cohort (non-randomised) studies have been compared
- to determine whether the effect sizes produced by RCTs are systematically greater or smaller than those from non-randomised studies
- to propose possible reasons for any differences in results obtained by the two methods.

Methods

The major electronic bibliographic databases were searched, using the strategy outlined in chapter 2. The number of papers identified is shown in *Table 2*.

TABLE 2 *Numbers of papers comparing RCTs and non-randomised studies*

Database	No. new papers identified in each search (cumulative %)	No. of papers read (cumulative %)
MEDLINE	106 (9)	61 (49)
EMBASE	584 (59)	29 (72)
Science Citation Index	442 (97)	29 (95)
Social Science Citation Index	26 (99)	2 (97)
Cochrane database	13 (100)	4 (100)
Total	**1171**	**125**

Inclusion criteria

The following inclusion criteria were adopted for this area of investigation.

- The results of a RCT must be compared with a non-randomised study, or the results of several RCTs combined compared with several non-randomised studies combined.
- The intervention must be the same and in similar settings.

- The control arms of the studies must receive similar therapy.
- There must be comparable outcome measures, preferably valid and reliable.

The 125 possible papers yielded 18 that met the criteria (14 single studies, four combined studies) and these are summarised in appendix 2. The treatment effect sizes, which had to be recalculated for the purposes of comparison are presented in appendix 3. Where possible a significance test was performed on the difference in the treatment effect sizes. (Two additional papers were identified after the initial draft of this report was completed.)

The type of intervention in each example was noted. The use of adjustment techniques to compensate for selection bias in non-randomised studies was also investigated.

Results

Single RCTs compared with single non-randomised studies

CASS Principal Investigators (1984)[30]

The Coronary Artery Surgery Study (CASS) included a randomised trial of CABG and medical therapy in the management of patients with mild or moderate stable angina pectoris or free of angina but with a documented history of MI. A total of 780 patients, from 11 institutions agreed to participate and were randomised to surgical or medical groups. A total of 1315 patients did not participate (69% refusal by physician, 28% by patient, 3% other) and were assigned to treatment in a non-random manner. The characteristics of the randomised and non-randomised patients were similar, with the notable exception of more extensive coronary artery disease in the randomised patients. All patients were followed for at least 46 months and their survival recorded.

Hlatky et al *(1988)*[31]

The findings of three major randomised trials of CABG were compared with predictions derived from the Duke Cardiovascular Disease Databank. Clinical characteristics of patients who met eligibility requirements for each of the three trials were used in multivariable statistical models to compare the observed 5-year survival rates in the randomised patients. A Cox's proportional hazards model was used to correct imbalance in known prognostic factors between groups in the non-randomised component.

Horwitz et al*(1990)*[32]

The results from a multicentre randomised trial (Beta-blocker Heart Attack Trial [BHAT]) were compared with a restricted non-randomised cohort (same eligibility requirements) and an expanded cohort (no eligibility restrictions), all from one US hospital. Adjusted mortality rates for the non-randomised cohorts were obtained using multiple logistic regression, including prognostic independent variables. The unadjusted and adjusted mortality rates in the restricted and expanded non-randomised patients were compared with those in the RCTs.

Paradise et al *(1984)*[33]

A comparison was made between patients who were randomised to surgical or medical treatment for severe throat infections at a US children's hospital and non-randomised patients at the same hospital whose parents declined to participate. The same eligibility criteria were used in both studies and the children's demographic and clinical characteristics were similar. No statistical adjustments were made. The average number of throat infections which occurred over the next 3 years were compared among the randomised treatment groups and the non-randomised treatments.

Paradise et al *(1990)*[34]

The efficacy of adenoidectomy in children with persistent otitis media was assessed by a randomised trial and also a non-randomised study of children assigned to treatment according to parental choice. The authors compared the outcomes in the two study designs in terms of recurrence of otitis media. The children's demographic and clinical characteristics were similar and no statistical adjustments were made for baseline differences.

Schmoor et al *(1996)*[35]

Patients in a randomised trial from 44 centres (German Breast cancer Study Group [GBSG]: trial 2) were compared with patients who were allocated treatment according to preference. Patients from another randomised trial (GBSG: trial 3) were also compared with non-participating non-randomised patients. All patients had to fulfil the same eligibility criteria to be included. A Cox's proportional hazards model was used to adjust for different baseline risks between the non-randomised treatment groups. The main outcome measure was recurrence-free survival time.

Yamamoto et al *(1992)*[36]

Parallel randomised and prospective non-randomised studies were conducted to assess

treatment of benign oesophageal stricture by two dilators. All patients attended one US clinic and fulfilled the same inclusion criteria. The non-randomised patients were allocated treatment at the discretion of the gastroenterologist. No significant differences were noted among patients assigned to receive Eder-Puestow dilation or balloon dilation, except that fewer patients with oesophagitis were allocated to balloon dilation. No statistical adjustments were made to compensate for this. Comparisons were made between the treatment gains in the randomised and non-randomised cohorts in terms of recurrence of dysphagia and proportion requiring redilation.

McKay et al *(1995)*[13]

Alcoholic patients from two US clinics were invited to participate in a randomised trial, comparing day-hospital rehabilitation with inpatient care. Those patients who took part in the trial were compared with those patients who self-selected their treatment (and who fulfilled the trial eligibility criteria). Comparisons of treatment groups showed that patients in day hospitals were older, had higher psychiatric severity, and had better overall employment status than those who were admitted. No statistical adjustments were made to address this imbalance when calculating the treatment effect in terms of alcohol and drug use at 3, 6 and 12 months follow-up.

Nicolaides et al *(1994)*[37]

A randomised trial was conducted at one UK hospital to compare the use of chorionic villus sampling (CVS) and early amniocentesis (EA) for foetal karyotyping. The authors compared the trial participants with patients who self-selected their procedure and fulfilled the trial eligibility. No significant differences were found between the women choosing CVS and those choosing EA, therefore no adjustments were made in the comparison. The main outcome measures were rates of foetal loss (total, induced and spontaneous).

Emanuel (1996)[38]

The author explored the notion that increased use of hospices and lower use of high-technology interventions for terminally ill patients produce significant cost savings. Included in this paper was a comparison of a randomised trial in Los Angeles (UCLA Veterans Administration Hospital) with the non-randomised National Hospice study. In both studies terminal cancer patients were given hospice care or conventional treatment and the per cent savings for hospice patients calculated. This estimate of saving was compared for the randomised and non-randomised patients, with some statistical adjustment made in the latter group for baseline differences.

Garenne et al *(1993)*[39]

The clinical efficacy of the standard Schwarz measles vaccine was investigated as part of an RCT in rural Senegal. The estimate of its efficacy was compared with the results of a national non-randomised campaign study. Vaccine efficacy was estimated by comparing vaccinated children with non-vaccinated children who had never had measles. Adjustments were made to control for intensity of exposure in both the randomised and non-randomised groups. In addition, the age at vaccination and time since vaccination were adjusted for in the non-randomised groups.

Antman et al *(1985)*[40]

A randomised trial was conducted to assess the efficacy of adjuvant doxorubicin for treatment of intermediate or high-grade sarcoma. The trial results were compared with the results of patients who were allocated treatment on the recommendation of a physician or by their own choice. These patients fulfilled the trial eligibility. Adjustments were made for known prognostic variables (location and stage) in the non-randomised patients, and the main outcome measure used in the comparison was time disease-free.

Shapiro and Recht (1994)[41]

In this review the evidence for late effects of adjuvant therapy for breast cancer was provided by a number of randomised and non-randomised trials. The effects of surgery only compared with additional radiation and chemotherapy were assessed in an RCT, and these can be compared with the results of two non-randomised studies. All three studies gave relative risks (RRs) of the occurrence of acute non-lymphocytic leukaemia after treatment.

Jha et al *(1995)*[42]

The cardiovascular protective properties of antioxidant vitamins were reviewed by calculating the RR reductions across non-randomised and randomised trials identified in a literature search. One RCT and one large cohort study with similar settings and outcome measures were compared with respect to treatment effect size. The Alpha-Tocopherol, Beta-Carotene and Cancer Prevention Study (ATBC) assessed the effect of taking vitamin E on mortality from cardiovascular disease. The US Nurses' Health Study recorded whether women were regularly taking vitamin E supplements and measured the risk reduction for death from cardiovascular disease and non-fatal MI. Adjustments

were made for a variety of dietary and clinical characteristics.

Several RCTs combined compared with several non-randomised studies

Pyorala et al *(1995)*[43]

This review included 11 RCTs and 22 non-randomised trials that investigated the use of luteinising hormone-releasing hormone (LHRH) and human chorionic gonadotrophin (hCG) as hormonal treatments of cryptorchidism. The studies were identified through a literature search of MEDLINE and the results of the combined randomised trials were compared with the results of the combined non-randomised studies.

The Recurrent Miscarriage Immunotherapy Trialists Group (1994)[44]

The efficacy of allogenic leukocyte immunotherapy for recurrent spontaneous abortion was reviewed in nine randomised and six non-randomised studies. No correction was made for the significant differences between treatment groups in the non-randomised studies. The main outcome measure was live birth rates.

Watson et al *(1994)*[45]

A meta-analysis of four randomised trials and six non-randomised studies evaluated pregnancy rates after the use of oil- or water-soluble contrast media during hysterosalpingography in infertile couples. Three of the non-randomised studies used historical controls as the water-soluble group. The odds ratios of pregnancy (oil- versus water-soluble contrast media) were compared for the randomised and non-randomised patients. No adjustments were described.

Reimold et al *(1992)*[46]

Six combined RCTs were compared with the results of six combined non-randomised studies. The studies assess the efficacy of antiarrhythmic therapy for chronic atrial fibrillation. In each study design the treatment benefit (quinidine compared with control arms) was calculated in terms of the percentage of patients in sinus rhythm and crude mortality rates. Insufficient information was available to provide an adjusted estimate of risk.

Do results obtained in RCTs and non-randomised studies differ in a consistent manner?

Seven of the 18 papers found no significant differences between treatment effects from the two types of study. Five of these seven had adjusted results in the non-randomised studies for baseline prognostic differences. The remaining 11 papers reported differences which are summarised in *Table 3*.

Yamamoto, comparing two interventions, obtained contradictory results from the two study types for one of two outcome measures.[36] The RCT indicated one type of dilation was associated with a lower need for a repeat procedure whereas the non-randomised study found the alternative procedure to be superior. The two procedures were, however, comparable in terms of recurrence of dysphagia.

Seven studies obtained differences in the same direction but of significantly different magnitude. In three, effect sizes were greater in the RCTs.

1. In the paper by Shapiro and Recht,[41] which focused on the frequency of an adverse effect, the RCT produced an RR of acute non-lymphocytic leukaemia after chemotherapy for breast cancer of 24.0, much greater than the figure of 3.7 in the non-randomised study.
2. Reimold and co-workers looking for a possible beneficial effect, also found a larger treatment effect in their combination of six RCTs than the six combined non-randomised trials.[46] For some outcome measures the difference in estimated benefit was substantial. The RCTs estimated a difference 25.5% (in favour of quinidine) for the percentage of patients remaining in sinus rhythm at 1 year. The non-randomised trials estimated the difference to be only 2.8%.
3. Nicolaides and co-workers found larger effect sizes for both beneficial and adverse effects of CVS compared with EA in the RCT than in the non-randomised group.[37]

In contrast, in four studies the estimated effect size was smaller in the RCT.

1. The CASS Principal Investigators, while finding better survival at 6 years following surgery rather than medical treatment for angina, reported a benefit of only 2% from the RCT but 4% in the non-randomised comparison.[30] Although the difference is only 2%, this is statistically significant due to the large numbers involved. In the paper, the authors emphasised the 5-year survival rates, where the two designs were in agreement, and concluded that there was no difference in the results obtained by the randomised and non-randomised studies.
2. In their review of vitamin supplements, Jha and co-workers also reported a greater benefit in the non-randomised study,[42] which reported a

TABLE 3 *Key points of studies finding differences between RCTs and non-randomised studies*

Study	Outcome	Summary of results of comparison	Adjustment for baseline differences in non-randomised element
Yamamoto *et al,* 1992 [36]	Proportion requiring repeat dilation	Two methods favoured different procedures	No
Shapiro and Recht, 1994 [41]	Occurrence of acute non-lymphocytic leukaemia after treatment	Effect found with both, greater with RCT	No details
Reimold *et al,* 1992 [46]	% remaining in sinus rhythm at 12 months	Effect found with both, greater with RCTs	No
Nicolaides *et al,* 1994 [37]	Survival	Effect found with both, greater (beneficial and adverse) with RCT	No
CASS Investigators, 1984 [30]	Survival	Effect found with both, smaller with RCT	No
Jha *et al,* 1995 [42]	Death from cardiovascular disease	Effect found with both, smaller with RCT	Yes
Pyorala *et al,* 1995 [43]	Descended testes	Effect found with both, smaller with RCT	No details
Horwitz *et al,* 1990 [32]	Mortality at 24 months (in expanded cohort)	See text	No
RMITG, 1994 [44]	Live births	RCT found effect, non-randomised did not	No
Emanuel, 1996 [38]	% savings	Non-randomised found effect, RCT did not	Yes
Antman *et al,*1985 [40]	Time disease free	RCT found no effect, non-randomised did not	No

RMITG = Recurrent Miscarriage Immunotherapy Trialists Group

31% RR reduction compared with 2% in the RCT, even though the non-randomised results were adjusted for baseline differences in the groups in terms of a wide range of prognostic variables. However, the confidence intervals (CIs) surrounding these estimates were large and the outcomes were not entirely comparable as the non-randomised study included risk of non-fatal MI which the RCT did not.

3. In their review of hormonal treatment for cryptorchidism, Pyorala and co-workers also concluded that the effect of treatment is overestimated in non-randomised studies.[43] This review involved pooling the results of RCTs and comparing them with pooled non-randomised results. Not only were the success rates of both hormonal treatments better in the non-randomised study than the comparative randomised studies, but also the benefit of LHRH over hCG was greater (14% difference in success rates in the non-randomised compared with 2% difference in the randomised study).
4. The fourth paper, by Horwitz reported a greater effect in non-randomised studies (see below).

In addition, in two reviews the non-randomised studies reported statistically significant effects whereas the RCTs found no such differences. In the RMITG paper no adjustments were made for the significant differences in composition of control and treatment groups in the non-randomised patients.[44] Significant differences were then found in the estimates of benefit by the two study designs, with the RCT showing significant benefits from immunisation while the non-RCT showed no benefit.

Emanuel's investigation into hospice savings[38] displayed some of the largest discrepancies between randomised and non-randomised studies, though the comparisons were fraught with complications which the author acknowledged. The randomised trial that was most comparable to the non-randomised study in terms of similar patients and outcome measures, showed no significant saving by type of terminal care. The National Hospice study, however, showed a significant saving of 34%. Other non-randomised studies mentioned in the paper showed a wide range of savings, from 68% to none.

There is some evidence of the impact of efforts to close any gap between the two methods. In the study by Antman and co-workers of chemotherapy for sarcoma, which found a treatment effect in the non-randomised study but not in the RCT, adjustment for baseline differences in the arms in the non-randomised study did not affect the results.[40] Horwitz and co-workers adopted a two-stage strategy.[32] They first limited the analysis to a subgroup of the non-randomised study that met the eligibility criteria for the RCT (the restricted cohort). This, however, had only a limited impact on the difference between the two study types. Subsequent adjustment of the restricted cohort to allow for baseline differences between the two arms did eliminate the difference, though similar adjustment using the entire sample in the non-randomised study (the expanded cohort) did not eliminate the difference.

The results of these papers display no consistent pattern and they certainly do not support the suggestion that non-randomised studies are intrinsically likely to produce larger estimates of treatment effects than RCTs. Frequently RCTs and non-randomised studies give comparable results. Either method can produce a greater effect and, when comparing two interventions, can produce contradictory results. However, the evidence reviewed here is extremely limited. It suggests that adjustment for baseline differences in arms of non-randomised studies will not necessarily result in similar effect sizes to those obtained from RCTs. In the terms of the model presented in chapter 1, this means ensuring that the consequences of differences in **e**, **d**, **i** and **p** are eliminated as far as possible.

There is no obvious pattern by intervention type, though slightly more surgical interventions found a bigger effect in the randomised trial. In addition, combination of results from more than one study did not appear to influence the direction or magnitude of the treatment effect.

As noted above, after the review was completed, two further relevant papers were published. A study by Stukenborg, examining the outcome of carotid endarterectomy, compared outcome in routine practice using Medicare data with what would be predicted from the results of trials.[47] He found that outcome was significantly worse among both patients with sufficient co-morbidity to have excluded them from the trials and among those treated in hospitals with perioperative mortality rates greater than what was found in the hospitals participating in trials. He concluded that estimates of efficacy of an intervention could only be applied to patients meeting the entry criteria of the trials and who are treated in hospitals that are representative of those participating in the trials.

The second study takes four meta-analyses, examining interventions in the health and education fields.[48] The authors sought to compare effect size in randomised and non-randomised studies. Pragmatically, they only included studies in which the intervention was compared with either a placebo or no intervention, where the data reported made it possible to calculate effect size, where the assignment method was clear, and where assignment was not pseudo-random. Individual studies were coded according to a wide range of parameters, including effect size, topic, design factors (assignment methods, difference in outcome measure between arms at entry, matching/stratification, total and differential attrition, activity in the control group – placebo or no treatment, control group drawn from same or different population, and self-selection or not of participants), and other factors that research suggested might influence effect size (publication status – published/unpublished, whether treatment standardisation occurred, mode of assessment – self or other, specificity of outcome assessment, exact or inexact estimate of effect size, and sample size). Using standardised effect size, in univariate comparisons, two topics (ability grouping of pupils in schools, prevention of drug use) showed greater effects in randomised studies while two (psychological interventions to improve postoperative outcomes, coaching for school tests) showed no difference. Using multiple regression, with exploration of the effects of outliers, interactions and transformations, the authors reported the emergence of certain relatively robust findings. The effect of assignment method hovered around the 5% significance level, suggesting a slightly larger effect size with randomised assessment, though the CIs would also be consistent with a very small increase in effect size with non-randomised studies. Variables significantly associated with

effect size were differences in the outcome measure between study arms pre-intervention, the use of passive controls (i.e. no treatment rather than a placebo), total and differential attrition rates, and self-selection of study arm.

The authors proposed implications for evaluative research. In non-randomised studies, as far as possible, subjects should not be permitted to self-select the arm into which they are entered. Large differences between arms in the outcome being measured at entry should be avoided using techniques such as matching on covariates or propensity scores and, if this is not possible, to adjust results for such differences.

Other comparisons

The papers discussed above compared evidence from different methods that examined specific interventions. In addition, some authors have looked more generally at the two approaches.

Colditz and co-workers attempted to relate study design to the magnitude of gains attributed to new therapies over old.[8] They analysed 113 reports of a heterogeneous group of interventions in a sample of medical journals and found that greater gains were found in non-randomised studies. This finding is supported by the companion paper on surgical interventions by Miller and co-workers.[9] They analysed 188 studies in leading surgical journals and found that the average gain (new techniques compared with conventional) was larger in non-randomised studies.

These results are consistent with earlier work published by Gilbert and co-workers who observed greater gains for the innovation among studies that used a non-random design compared with randomised trials.[49] They analysed 107 papers from a computerised bibliography that evaluated surgical and anaesthetic treatments.

Ottenbacher attempted a similar investigation by sampling 60 research articles (half randomised and half non-randomised) in two leading medical journals.[50] However, he found no significant difference in the treatment effects (measured by standardised mean differences).

Possible reasons for greater estimated treatment effects in RCTs

One possibility for a greater estimated effect size in an RCT is that care provided in the context of trials is better than that in routine clinical practice. Thus, any advantage conferred by the intervention will be magnified by the accompanying treatment package. This seems especially likely where the comparison is with placebo. It receives support from the finding by Stukenborg that outcomes are worse than those predicted by RCTs in hospitals with characteristics that have higher perioperative death rates than those participating in RCTs.[47]

Another possible explanation is that RCTs typically have specified eligibility requirements. This may produce a highly selected group of individuals for whom the new treatment is more likely to work. In non-randomised studies it is more common for all patients to be included, however poor their prognosis. In the context of the model in chapter 1, the sample in the RCT will be a smaller proportion of the reference population than in the non-randomised study and will contain a higher proportion of those with the best prognosis, as the area represented by **e** will be greater. In addition, it is possible that those with worse prognosis may either be excluded by those recruiting to RCTs, increasing **i**, or may even exclude themselves, by declining to participate in an RCT, increasing **p**. These issues will be examined in subsequent chapters.

On the evidence presented so far, the effect of differences in eligibility is supported by the paper on beta-blockers by Horwitz and co-workers.[32] The higher mortality rates in the expanded cohort were anticipated as this cohort included many patients with contraindications to beta-blockers, such as congestive heart failure, and more with conditions increasing their risk of death, such as angina pectoris. Given this cohort susceptibility bias, it is not surprising that such a large treatment effect was found. When the same eligibility criteria were applied to the non-randomised cohort, the treatment effect was found to be closer to that estimated in the RCT. This hypothesis also receives support from the study by Stukenborg.[47]

Possible reasons for greater estimated treatment effects in non-randomised studies

In non-randomised studies patients are typically allocated to treatments according to the professional opinion or preference of their practitioner and their own preferences. This assumes a belief that there is a larger therapeutic benefit from one of the treatments than from another (i.e. the state of individual equipoise which is deemed ethically necessary for an RCT does not exist). In these circumstances, where each patient is given the treatment that is considered most appropriate for their particular circumstances, one would expect a larger estimated treatment effect size to exist.

This may reflect explicit or implicit differences in eligibility (**e**) for the intervention(s) being studied, or differences in the likelihood of patients with worse prognosis either being invited or accepting one of the interventions being studied.

All other things being equal, patient preference might also have an effect. It may be that the practitioner is unsure of the relative merits of two treatments (individual equipoise) and would agree to enter the patient into a randomised trial. The patient, however, may have a strong preference for one treatment over another, and if given the desired treatment, has enhanced therapeutic benefit arising from this belief. Hence, a non-randomised study, where preferences are important and patients are free to choose their treatment, may lead to a larger estimate of treatment benefit.

The type of intervention studied may be important when interpreting the estimates of gain observed. If the intervention is preventive, for example, a vaccine or breast cancer screening, it can be misleading to compare the results of those who volunteer for treatment with those who do not respond to invitations. It has been shown that those volunteering to receive preventive treatments have different characteristics from other members of the population.[51] Typically they are more health conscious and adopt other preventive strategies. These could conceivably interact with the intervention being studied to magnify any effect. Consequently, in the context of the model, **p** will vary. Unless this is controlled for in the analysis, the reduced risk will be attributed solely to the therapeutic benefit of the treatment in question, leading to an over-estimate of the benefit of the intervention. However, this effect may, in some circumstances, have the opposite effect if those included have less scope to benefit; in such a situation, the 'less healthy' population in a non-randomised study will have greater opportunity for benefit. Again, all of these issues will be discussed later.

Finally, it is possible that there is a subtle publication bias in operation. It has been suggested that larger treatment effects are seen in non-randomised studies because these tend to be written-up, submitted and subsequently published only when large treatment effects are found.[49] Studies that find no significant differences between treatments are considered 'un-newsworthy'. RCTs, on the other hand, are more likely to be published, regardless of their findings. Thus, publication bias may operate differently for RCTs and non-randomised studies.

Possible reasons why estimated treatment effect may be similar in RCTs and non-randomised studies

Obviously this is most likely to occur when the two treatments being studied are actually of identical therapeutic benefit, and this appears to be the case in some of our seven papers that showed similar estimates. However, there are examples where both the RCT and the non-randomised study have shown similar estimates of benefit when the interventions being compared are not equally effective.

Risk adjustment to compensate for allocation bias (should it exist) in non-randomised studies may be sufficient to bring the results close to those of RCTs, so that the initially observed difference was due to confounding, which is subsequently adjusted for. However, attempts to use statistical techniques to overcome allocation bias may not be successful if relevant prognostic factors are not collected at baseline and adjusted for.[38]

In comparison papers, such as those cited above, the possibility of publication bias must be considered. For example, the authors of an RCT might be trying to demonstrate that their trial results are validated by the results of a non-randomised study. The authors may have therefore selected one particular non-randomised study which has similar estimates to their trial while ignoring other non-randomised studies that dispute their results and question its generalisability. However, in this review nine of the 18 papers were written by those who conducted the RCT, and none of these nine appear to have selected a non-randomised study that supported their findings rather than others that did not.

Summary

- Attempts were made to find all papers that have directly compared the results of RCTs and prospective non-randomised studies. Eighteen such papers were found and analysed.
- No obvious patterns emerged; neither the RCTs nor the non-randomised studies consistently gave larger estimates of the treatment effect size. The type of intervention did not appear to be influential, though more comparisons need to be conducted before definite conclusions can be drawn.
- Reasons have been identified that may explain why RCTs might produce a greater or a lesser estimated treatment effect. For example, a greater effect may be seen in an RCT where

patients receive higher quality care or have a greater ability to benefit, compared with those in non-randomised studies. Reasons for a lower RCT estimate include selection bias introduced by practitioners when deciding whom to treat, an enhanced response to treatment among those with strong preference for a particular treatment, the inclusion in non-randomised preventive studies of individuals who, by virtue of their health-related behaviour, have greater ability to benefit, and finally, possible publication bias.

- The outcome of non-randomised studies best approximated to the RCT results when both used the same inclusion and exclusion criteria, and potential prognostic factors were well understood, collected and differences between arms in the non-randomised study adjusted for.

Chapter 4
Exclusions

Introduction

The aim of the typical intervention study is to measure the average effect of a given treatment in a given group of patients. If the study is to provide a secure basis for future clinical decision-making, both treatment and patients must be unambiguously defined.

For patients, the starting point will be all those with the problem the treatment is designed to deal with. However, for an RCT to be ethical, there must be sufficient doubt about the relative value of the different treatment options. In general, RCTs do not take place in a vacuum of knowledge about the treatment of interest, and at the time of the trial there may be reasons for excluding some sub-groups of patient on **medical** or **ethical** grounds, such as prior evidence that the treatment is effective, or an unusually high risk of treatment complications. There can be no argument in principle about exclusions of this kind. The questions that arise here are about how widely the exclusion criteria should be drawn. How potentially adverse do the risks have to be, and how strong the evidence for this, before patients are excluded?

In many RCTs, patients are also excluded on **scientific** or **administrative** grounds. They may confuse the picture, dilute the power of the study, or prove administratively awkward or costly. This is the type of reasoning that underpins restricting a study to a particular demographic group (middle-aged white males are commonly cited), for example, for excluding patients who have co-morbidities that may lead to poor broad-spectrum outcomes such as mortality for reasons unconnected to the treatment in question (cancer patients in an RCT of treatment for heart disease, or *vice versa*), or excluding patients who are, or are expected to be, difficult to gain consent from, non-compliant, disruptive in the clinic or difficult to follow-up (examples from the literature include children, mentally ill individuals and drug users).

If the exclusion criteria are many and widely drawn, only a small proportion of the patients with the condition in question will be enrolled in the RCT. The question for study designers with a limited budget is whether it is possible to identify in advance subsets of patients who can be expected to have a relatively homogeneous response to treatment. If so, they have to make a choice somewhere between two extreme positions:

- studying one very homogeneous subset, ensuring the most precise result for the group in question, but leaving generalisation to other groups entirely to clinical judgement

or

- excluding as few patients as possible, giving estimates of effect that are more directly generalisable to the population of patients as a whole, and also some indication of heterogeneity of response, but which will have relatively wide CIs if response to treatment **does** turn out to be heterogeneous.

This chapter explores this issue and, specifically, seeks to determine:

- the extent to which subjects to whom interventions might be applied are excluded from RCTs
- the reasons cited for their exclusion
- the extent to which blanket exclusions, based on socio-demographic characteristics, are applied
- the implications of patterns of exclusions for attempts to generalise results of RCTs.

Although this chapter addresses the issue of exclusions from RCTs, it is also recognised that many of the same issues will apply to non-randomised studies.[52]

What proportion of patients with the relevant condition are excluded from RCTs?

A major problem with the issue of exclusion of patients who have a relevant condition is that few RCTs report on it, and most of the evidence is indirect. Papers identified that specifically address the issue of exclusion are listed in appendix 4. Some studies report on the numbers of patients 'screened'. One of the early studies of exclusion criteria[53] was based on RCTs in the 1979 inventory of the National Institutes of Health (NIH). Aggregating over 16 RCTs, 73% of those screened were deemed ineligible, a far larger source of 'loss' than the 15% eligible but not randomised (4% withdrawn by the

doctor, 4% patient refusals and 7% withdrawn by the investigators). The eligibility rate in the 16 RCTs ranged from 10% to 99%. In a more recent overview by Muller and Topol,[54] of eight RCTs of intravenous thrombolysis, the percentage of patients screened who were considered eligible ranged from 9% to 51%. Even some of the largest RCTs differed substantially (33% and 18%, respectively in the Italian group studies of treatment of MI (GISSI) and the Anglo-Scandinavian Study of Early Thrombolysis (ASSET)).

However, these kinds of result are difficult to interpret. The word 'screened' in this context seldom means **diagnostic** screening for the index condition; for example, it can mean surveillance of all hospital admissions to provide possible **candidates** for proper diagnostic screening, and variations in what is meant by screening will explain at least part of the very substantial variation in eligibility rates.

This is illustrated by *Table 4*, which draws largely on material from RCTs reported more extensively later in this report. For present purposes it is sufficient to note the generally high inclusion rates for vaccine RCTs in which there was no sickness- or risk-related selection prior to screening, and very low rates for primary prevention of hypertension in which those screened were again whole populations. Neither of these is surprising *per se*. More striking are the more moderate but very variable exclusion rates from a set of secondary prevention trials in which screening was typically based on chest pain on hospital admission, and for cardiac surgery, in which those screened were patients with multiple vessel disease.

Another set of studies has examined the proportions of patients under treatment that have been entered into RCTs. Lee and Breaux reviewed recruitment to RCTs among 1103 patients treated for cancer at one American centre in 1979.[55] Of the 400 for whom there was a relevant RCT in progress, 137 (34%) fell foul of at least one exclusion criterion, 118 were excluded because of physician or radiotherapist preference, and 21 patients refused.

Begg and co-workers examined 3534 patients in Eastern Cooperative Oncology Group (ECOG) hospitals in 1981, and found that 66% of them were ineligible for any protocol despite the fact that the available ECOG protocols cover the vast majority of the major tumour sites and stages of disease.[56] The remainder were accounted for as follows: 24% clinician refusal, 9% patient refusal, and 13% technically unsuitable, leaving 54% to be randomised.

Martin and co-workers found that in the Veterans Administration study of warfarin anticoagulants, 69% of the 2687 subjects screened were 'technically' ineligible, and 15% were excluded because of patient or physician preference, leaving 16% who were registered for the RCT.[57]

TABLE 4 *Percentage of eligible patients/subjects included in RCTs of selected interventions*

Malaria vaccines		**Calcium antagonists**		**CABG/PTCA**		**Hypertension**	
Alonso	98.2%	TRENT	48.20%	CABRI	4.60%	SHEP	1.06%
D'Alessandro	97.3%	SPRINT II	66.60%	EAST	16.40%	MRC 1992	3.49%
Sempertegui	91.8%	SPRINT I	49.70%	GABI	4.00%	MRC 1985	3.37%
Valero 1993	84.8%	Branagan	12.40%	Lausanne	7.90%		
Valero 1996	77.6%	Simes	approx. 30.0%	ERACI BARI	40.20% 16.30%		

Individual studies are referenced in subsequent chapters

BARI = Bypass Angioplasty Revascularisation Investigation
CABRI = Coronary Angioplasty versus Bypass Revascularisation Investigation
GABI = German Angioplasty Bypass-surgery Intervention Trial
EAST = Emory Angioplasty versus Surgery Trial
ERACI = Argentine Randomised Trial of Percutaneous Transluminal Coronary Angioplasty
MRC = Medical Research Council
SHEP = Systolic Hypertension in the Elderly Program
SPRINT = Secondary Prevention Reinfarction Israeli Nifedipine Trial
TRENT = Trial of Early Nifedipine Treatment

In the European Organisation for Research and Treatment of Cancer (EORTC) trial of observation versus radiotherapy for ductal carcinoma *in situ* of the breast, 60% of cases were ineligible for the study, 4% of the exclusions being a result of patient refusal.[58]

Taking the intervention rather than patients as the starting point, Barnett and co-workers examined more than 400,000 coronary by-pass operations that were carried out in the USA between 1971 and 1979.[59] Of these, 4% were entered in the CASS registry, and less than 0.2% were randomly assigned to surgery or control.

It will be clear from these results that a) a relatively small proportion of apparently relevant patients are included in RCTs, and b) that 'technical' eligibility criteria often provide a more effective barrier to recruitment than practitioner or patient preferences. For many common exclusion criteria, both medical/ethical and scientific/administrative factors are involved; for example, people with serious co-morbidities may be excluded both because the expected balance of benefit and risk may be adverse, and also because they add to the heterogeneity of likely treatment effect.

In the sections that follow, the main medical and scientific reasons for exclusions will be discussed with illustrative examples. This is followed by a review of some common 'blanket' categories of exclusion, based on age, sex and ethnicity.

Medical reasons for exclusion

High risk of adverse effects

There may be some groups of patients for whom the risk of treatment complications – not just the loss of potential benefit from the experimental treatment, but actual harm – are unusually high, and can be expected to outweigh the anticipated benefits. People at anaesthetic risk may be excluded from recruitment to RCTs involving surgery. Pregnant women are often excluded, for fear of harm to the foetus (e.g. BARI for CABG versus PTCA; TRENT for calcium antagonists). People who have had recent surgery are commonly excluded from RCTs of anticoagulation for fear of post-surgical bleeding.

One specific example of this is provided by chemotherapy RCTs, in which organ function may be at particular risk of toxicity from the drugs. Begg and Engstrom examined the eligibility criteria of all nine Phase III studies of adjuvant breast cancer in the USA in August 1985.[60] All required specific levels of serum creatinine, though the level varied from minima of 1.5 mg/dl and 2 mg/dl to 'normal'. Seven studies specified eligible levels of white blood count, six specified minimum bilirubin and two alkaline phosphatase; again, the levels required for eligibility varied between RCTs.

What is striking in retrospect is the extent of variability between RCTs in how exclusions of this kind are handled. Some involve specific criteria, which vary for no apparent reason between studies. Others rely on blanket exclusions often based on age (elderly, and so at risk of complications; female and of child-bearing age, and so potentially pregnant during the period of study, etc.). While these serve a variety of other purposes too, they may severely limit the generalisability of the study. Other studies again exclude fewer subjects at the recruitment stage, but 'lose' more of them later on in the process of the study, for reasons that are seldom adequately reported.

Benefit already established

If previous studies have provided good evidence that certain groups of patients can expect to benefit from the intervention in question, they should be excluded from a new RCT on ethical grounds. This is relatively unusual as a subject for review, but some relevant data are provided in the meta-analysis of antihypertensive drugs by Collins and co-workers.[61] Fourteen of the 16 trials had **maximum** levels of blood pressure as one of their entry criteria. These ranged from 104 mmHg to 130 mmHg, and not all of this variation was attributable to a downward secular trend as results of early RCTs of more severe patients became available.

Benefit known or believed to be very small or unlikely

The RCT should involve a definition of the conditions that the treatment is designed to remedy. These are often described as the inclusion criteria. One difficulty is that many conditions (e.g. hypertension, diabetes, impaired liver function) are indicated by an out-of-normal-range value for some physiological parameter. The issue of treatment threshold then becomes important. For example, how high does blood pressure have to be before there is a possibility that the treatment will provide worthwhile benefits?

Again, RCTs vary in their criteria. In the meta-analysis of antihypertensive drugs by Collins and co-workers[61] the minimum levels of blood pressure required for eligibility varied from 85 mmHg (published in 1978) to 115 mmHg (published in 1970). Again, this was not just the result of clinical

knowledge improving over time. RCTs published in 1985 and 1986 had minimum values of 90[62] and 105.[63] In different RCTs of calcium antagonists, the specified time since onset of chest pain was less than 6 hours, less than 24 hours, and unspecified.

Reduced levels of expected benefit

Some economists have argued that assessments of the benefits from health care should be measured in terms of effect on quality-adjusted life years. The argument is that someone with low life expectancy will derive little benefit from a given treatment, and particularly in cases where the treatment involves risk or adverse effect on quality of life in the short term, the balance of advantage will be altered if any benefits in the longer term are expected to be truncated. This could be one of the reasons why, in the review of breast cancer RCTs by Begg and Engstrom,[60] two of the nine RCTs excluded patients with a subsequent life expectancy of '< 10 years ignoring cancer'. Exclusions for concomitant heart disease varied from none to exclusion for severe angina or significant arrhythmias. Patients with cancer are commonly excluded from coronary prevention RCTs.

Some co-morbidities, while not increasing the risk of treatment complications or early mortality, can mask potential **functional** benefits. Again, such patients may be excluded from RCTs because at the margin the functional benefits of treatment are unlikely to outweigh the risks. For example, patients with extreme shortness of breath may derive benefits from hip replacement in terms of pain relief, but not of improved mobility. Impaired mental function may also affect the potential for functional benefit. This is a controversial area from the technical as well as the ethical point of view, and exclusion of people with limited life expectancy or co-morbidity tends to be justified on scientific rather than medical/ethical grounds.

Scientific reasons for exclusion

To increase the precision of the study

Apart from sample size, two factors can affect the precision or power of a study. Where the outcome for each patient can be measured on a scale such as change in blood pressure or symptom severity, heterogeneity of subjects is important. The less heterogeneous the patients in terms of treatment effect, the more precise the estimate of average effect can be for a given sample size, or perhaps more to the point, the smaller the sample necessary to detect an average effect of given size. Thus, one strategy is to draw the eligibility criteria very narrowly, in the expectation that this will result in a homogeneous set of subjects.

Where the outcome is measured in terms of proportions of subjects experiencing an adverse event, such as a MI or death, the baseline event rate is important. If baseline mortality is low, samples need to be large. Thus, their higher incidence of heart disease has provided the scientific justification for studying men rather than women.

The obvious disadvantage of the restrictive strategy is that it provides no direct data on the types of patient excluded, and risks either inappropriate inference by clinicians to wider groups in the absence of anything else to go on, or subsequent discrimination in treatment against the excluded groups. The other strategy is to draw the criteria wide. This will give an estimate of average effect that is more relevant to the whole reference population of patients, but in the presence of heterogeneity it will give less precise estimates of average effect, and may be materially misleading for some types of patient included in the study.

Researchers have tended towards the first of these strategies. One plausible reason is that it is less risky from a scientific point of view. At the study design stage, researchers generally know very little about how the treatment effect varies between patients and patient groups, that is, about heterogeneity. Restricting entry may ensure enough homogeneity to produce a 'significant' result, and this is better than running the risk of a non-significant result if the wider group turns out to be heterogeneous. Of course this requires good predictors of treatment effect; Yusuf and co-workers argue that these are seldom available and that, in practice, very little reduction in heterogeneity is achieved.[64]

However, others have argued that it is quite reasonable to generalise the results from such studies, particularly where the intervention is a purely biological process,[65] on the basis that there is some sort of class effect, and no reason to believe that the impact will vary.[66] There is an inconsistency here. If the treatment effect is the same for different patient groups, there is no point in limiting the study to, for example, white males. If the treatment effect is not the same, then one cannot generalise from the study patients to women and other ethnic groups. The underlying assumptions seem to be heterogeneity for the purposes of design, and homogeneity for the purposes of inference.

Begg and Engstrom speculate on the reasons for this phenomenon.[60] In the classical laboratory experiment, the effect of the variable of interest is isolated by controlling the values of other relevant variables. In the RCT this is unnecessary; confounding is controlled by randomising large enough numbers of patients. Nonetheless there may be a residual feeling that an RCT is a form of laboratory experiment, and so is best served by the use of homogeneous experimental units.

Yusuf and co-workers have argued strongly for the second strategy – for keeping entry criteria simple and wide.[64] This increases the speed of recruitment, leading to larger sample sizes for given cost, provides more rapid and more widely applicable results, and provides better opportunities for examining sub-groups.

To avoid bias

A great deal of attention has been given to the internal validity of RCTs. One of the common threats to validity is the introduction of bias through non-random losses of subjects at different stages of the study. A variety of strategies have been evolved to avoid such bias; analysis on the basis of intention-to-treat is one. However, many of the problems can be avoided by minimising losses at every stage after randomisation.

The key to this is that identification of eligible subjects is regarded as the starting point of this process. If subjects who are likely to drop out at different stages of the study can be excluded from randomisation, internal validity will be improved, but this has to be weighed against potential loss of generalisability.

Once a subject has been identified as eligible, the first hurdle is to obtain their **consent**. This raises the issue of the patient's competence to give it, and administrative costs in obtaining it in complicated cases. Potential problems in this area are commonly given as reasons for excluding children, the mentally ill or confused, and people with drug or alcohol problems. Lumley and Bastian point out that children in foster care cannot take part in RCTs, and yet one third of HIV-positive children in New York are in foster care.[66]

A second consideration is compliance with treatment. Some studies have excluded patients judged to have been insufficiently compliant during a run-in period (for example, the Physician's Health Study),[66] unable to follow instructions, and potentially disruptive in the clinic.[65] A trial of the effect of sodium and potassium on blood pressure in children also excluded candidates on the basis of inadequate compliance during a pre-randomisation run-in.[67] Of the 19,452 initially screened, 3223 were eligible according to the blood pressure criteria. Of these, 8.5% were ruled out on the basis of criteria including 'family with father as the primary single parent', and 'family with more than four children', but a further 85% were excluded because of failures of compliance during four pre-randomisation clinics, leaving only 243 to be randomised. Of course as Haynes and Dantes point out, exclusion of the non-compliant is appropriate in an efficacy RCT[68] (does the treatment do more harm than good in those who take it?), but not in an effectiveness trial (does the treatment do more harm than good to those to whom it is offered?).

The third main consideration is measurement of outcome, and avoidance of losses to follow-up. This has led to exclusion of:

- patients who live a long way from the treatment centre, are of no fixed abode, or are likely to be mobile[65]
- women of childbearing years, not on the grounds of therapeutic risk, but because they will 'not have time'[65,69]
- elderly patients, whose long-term outcomes may be 'censored' by competing events or endpoints not related to the study intervention. In the Begg and Engstrom[60] study of breast cancer RCTs, most had exclusions for concomitant heart disease, but these varied, encompassing, for example, severe angina and significant arrhythmias
- confused or mentally ill patients in RCTs involving subjective outcomes, and patients with impaired hearing in RCTs involving interviews
- patients who have an inadequate command of the language, in studies involving interviews or self-completed questionnaires.

In an editorial, Chalmers[70] mounted a strong attack on the practice of excluding patients in order to simplify interpretation of the results.

Common blanket exclusions

The elderly

It is common for the elderly to be specifically excluded from RCTs. As we have seen, this can be for a mixture of medical and scientific reasons. They may be more at risk of complications, or of an adverse balance between short-term risk and long-term benefit. They may be more at risk of

significant co-morbidity, which may mask or censor outcomes measurement. Thus, rather than evaluate each patient screened on their merits, which may be costly and unreliable, a blanket exclusion criterion is used, even though the use of age alone as a predictor of risk or benefit in this way is an extremely blunt instrument.

This seems to have been particularly widespread in cardiovascular RCTs. In the overview by Muller and Topol[54] of eight RCTs of intravenous thrombolysis, 31% overall were eliminated because they presented too late (typically more than 6 hours) after onset of symptoms, 13% because of specific contraindications such as a history of stroke or transient ischaemic attacks, and 14% because they were too old, and the author commented that most of the larger thromboembolytic trials have excluded patients aged over 75 years. Gurwitz and co-workers[71] reviewed 214 RCTs of specific pharmacotherapies for treatment of acute myocardial infarction (AMI) published between 1960 and 1991. Over 61% of the RCTs formally excluded people aged more than 75 years. This was particularly so of RCTs involving thrombolytic therapy and of invasive procedures. The fact that the mean ages of subjects in studies with and without formal age-based exclusions were similar suggests that tacit exclusion of older patients was going on even where this was not formalised. In the first decade studied, less than 19% of the RCTs involved age-based exclusions, but during the 1980s this figure had risen to 73%. In a review by Kannry and co-workers[72] 48% of RCTs of prevention of AMI excluded patients aged over 70 years, and 49% of 67 hypertension RCTs had an upper age limit.

Gurwitz and co-workers[71] challenged the justification for exclusion on grounds of age, pointing out that the average life expectancy of a 75-year-old was 11 years, and for an 85-year-old 6 years. They also argued that the concern about dilution of treatment benefits due to co-morbidity was more relevant to lengthy prevention studies than to care for AMI. In contrast, recent RCTs of treatment of high blood pressure have focused specifically on the elderly.[73,74]

Explicit exclusion from trials on the basis of age does not seem to be as widespread in cancer trials. In the review by Begg and Engstrom,[60] one RCT excluded patients under 16 years, another those over 70 years, and the remaining studies had no age restrictions. One study required patients to be ambulatory. In the Lee and Breaux[55] study of patients at one oncology centre, of the 137 patients who were ineligible for RCTs, 21 were ruled out on grounds of age, compared with 30 who were ruled out on the grounds of co-morbid or pre-existing conditions. In an earlier paper, Lee and co-workers[75] had argued that exclusion on grounds of age was unnecessary in lung cancer RCTs.

Non-steroidal anti-inflammatory drugs (NSAIDs), used by older people for treatment of arthritic pain, represented about 5% of all dispensed prescriptions in the USA in 1982. In a review of recruitment of older people to arthritis RCTs, Rochon and co-workers found that in 83 RCTs of NSAIDs, 19 specified both upper and lower age limits, 18 specified lower limits only, and none specified an upper age limit only; 46 specified neither.[76] However, only about 10% of the 9664 people were reported as being aged 50 years or more, about 2% as aged 65 years or more, and about 0.1% as aged 75 years or more. The authors felt able to comment that the proportion of the population who are treated the most in practice are generally omitted from trials of the same drugs.

Women

Gurwitz and co-workers pointed out that the traditional explanations for excluding women from clinical studies have included risk of teratogenicity, hormonal fluctuations, the protective cardiovascular effects of oestrogens, and reduced statistical power due to the less frequent occurrence of measured outcomes.[71] They also showed that in RCTs of treatment for AMI, an indirect effect of excluding elderly patients was to exclude a high proportion of women. In trials without age exclusions, 23% of the subjects were women as compared with 18% of the subjects in trials which had such exclusions. About 5% of the studies completely excluded women, and an additional 8% excluded women of childbearing potential. And yet in the USA, women surpass men in the number of annual deaths due to cardiovascular disease.[77]

McDermott and co-workers[78] reviewed 444 articles from *The Lancet, The New England Journal of Medicine* and *JAMA* during 1971, 1981 and 1991. In 1971, women were specifically excluded in 11% of the studies, in 1981 from 5% and in 1991 from 2%, but the subjects were all male in 13%, 9% and 7%, respectively. The percentages with a study question specific to men's health were 2%, 2% and 0.7%, respectively. Similar findings were reported by Bennett.[79]

Caschetta and co-workers[80] claim that the exclusion of pregnant women from RCTs is often unsatisfactory on scientific grounds, and Lumley and Bastian[66] argue that this is more to do with

limiting medico-legal liability than science. This point was also taken up Patterson and Emanuel[69] in their analysis of the problems of extending exclusion of pregnant women to excluding all women of childbearing age, whether or not they are 'at risk' of becoming pregnant. Moreno, in the discussion that followed this paper, points to the shadow of thalidomide, but argues that the policy of excluding women of reproductive potential is clearly disappearing.

Ethnic minorities

Svensson[81] found that only 20% of the 50 non-cancer studies in US populations published in *Clinical Pharmacology and Therapeutics* in 1984–86 published data on ethnicity or race. He was able to obtain the necessary information on another 25 studies from the investigators, and for the 35 studies found that 57% included black subjects, black people making up between 3% and 100% of the study groups. In several studies there were marked differences between the proportions of black people in the study population and that in the city where the study was conducted, with black people tending to be under-represented in the RCTs, but there were notable exceptions to this. On balance the evidence was against the hypothesis that American Blacks were over-represented in clinical RCTs due to the inner-city location of most university hospitals. Similar findings were reported by Moore and co-workers concerning interventions to treat HIV.[82]

Generalising from eligible to ineligible patients

How safe is it to generalise from men to women, from the middle-aged to the elderly, and more generally from those included in RCTs to those excluded? A number of researchers have discussed this from a theoretical point of view. Cowan and Wittes argue that the closer an intervention is to a purely biological process, the more confident we feel in extrapolating beyond the types of patients studied.[65] The implication seems to be that the main threat to extrapolation is between-group variations in compliance, and it seems to be a common starting point among researchers to take no variation in biological response between sub-groups as a kind of null hypothesis, to stand unless there is good evidence to the contrary.

Yusuf and co-workers[l] argue that the probability is low of reliably finding an unanticipated qualitative interaction (that is, differences in the direction of effect) in an RCT that has already excluded those in whom the treatment is clearly indicated or contraindicated.[64] However, there are plenty of examples of the scale of the effect varying between sub-groups.

- Several studies have suggested that older people experience more toxic effects from NSAIDs, and drug-induced gastrointestinal events are more common, and more likely to be fatal, in older people.
- The menstrual cycle can vary antidepressant effects; low-density lipoprotein has been suggested as a risk factor for heart disease in women, but much less so for men.
- Black hypertensive patients do not respond to beta-blockers and angiotensin-converting enzyme (ACE) inhibitors as well as white patients.

Outcome for trial and non-trial subjects

Horwitz and co-workers compared patients in the BHAT trial with those in two other cohorts.[32] They started by constructing a database of 2497 patients who had had an AMI and were admitted to the Yale-New Haven Hospital in 1979–82, with data from hospital records and postal questionnaires. The National Death Index was the primary source for data on mortality. From this database they constructed two cohorts. Both excluded patients aged less than 30 years or more than 74 years (444), and patients who did not meet the trial's criteria for MI (528) or had missing records (342). This is shown in *Table 5*. After eliminating 124 who had died before hospital discharge, the 'expanded' cohort consisted of 1059 patients. Then the trial's exclusion criteria were applied (contra-indications to beta-blockade, beta-blockade strongly indicated, and 'any condition likely to hinder or confuse follow-up or endpoint evaluation, such as malignant neoplasm or drug addiction') to produce a 'restricted' cohort of 622 patients. In terms of **baseline**, patients in the RCT were slightly younger than those who were ineligible, and the RCT included a higher proportion of men. In both the cohorts of ineligible patients, the majority were treated with beta-blockers, and those treated tended to be younger and to have had more severe infarcts than those untreated. In terms of **outcome**, patients in the restricted cohort had had more severe infarcts than patients in the expanded cohort. The 24-month mortality rates in the treated and untreated groups were similar for the restricted cohort and the trial patients, particularly after adjustment for age and severity of infarct. But in the expanded cohort, mortality rates in both 'arms' and per cent

TABLE 5 *Consequences of varying eligibility of subjects in beta-blocker trials*

	BHAT		Restricted database		Expanded database	
	Beta-blockers	No beta-blockers	Beta-blockers	No beta-blockers	Beta-blockers	No beta-blockers
Baseline						
No.	1916	1921	417	205	626	433
Mean age	55	55	57	60	58	60
Males (%)	84	85	75	73	74	68
Mild (%)	58	61	72	68	64	51
Outcome						
Crude						
2-year mortality	7.3	9.2	7.2	10.7	9.3	16.4
Reduction (%)	21		33		43	
Adjusted for age and severity						
2-year mortality (%)	7.3	9.2	7.6	9.7	10.2	14.4
Reduction (%)	21		22		29	

Source: Horwitz et al, *1990*[32]

reduction were greater than for the RCT. The authors concluded that observational studies of trial-eligible patients could produce results very similar to those from RCTs. Generalisation to excluded patients appeared to be more hazardous, at least in terms of effect size.

The study by Horwitz and co-workers was the only study found in which the relative effectiveness of two treatments was compared for the trial and non-trial cohorts. Other studies have compared outcomes for out-of-trial patients with in-trial patients in the conventional treatment arm, or with all trial patients for trials with no significant treatment effect, shifting the focus to whether baseline prognosis was similar for trial and non-trial subjects.

The Diabetes Control and Complications Trial (DCCT) Group[83] compared outcomes for the 200 patients in the conventional treatment group of their RCT of intensive treatment with 111 'trial-eligible' patients from a primary care database. (Only 12.5% of the database patients would have met the eligibility criteria; the main reasons for exclusion would have been duration of diabetes [38%], retinopathy status [55%], proteinuria [18%], age < 13 years or > 39 years, [25%] and hypertension [23%].) Trial patients tended to be older and have been older at onset of diabetes than database patients. Blood pressures and body mass indices were similar. At **baseline** they had lower HbA, and were more likely to be receiving hospital-based care, and two injections and blood glucose monitoring sessions a day rather than one. HbA values for the 'conventional treatment' group in the RCT and the ineligible group converged over the period of the RCT, possibly as a result of the spread of new methods of diabetic care from hospital to general practice. Trial patients had lower rates of progression of retinopathy than corresponding database patients (24% vs. 40%), and the development of gross proteinuria was less common.

Ward and co-workers describe a trial in which it was found that adjuvant chemotherapy for patients aged between 15 years and 74 years with operable stomach cancer had no significant effect.[84] They compared the 249 trial patients with 960 non-trial cases identified through a cancer registry, of whom 493 would have passed the trial eligibility criteria. Of the ineligible non-trial cases, 93 would have failed the stage criterion, and 212 the fitness criterion. As for Horwitz and co-workers, the RCT subjects were significantly younger than the eligible non-RCT group.[32] (The upper age limit had excluded 31% of all new cases from the RCT.) In terms of outcome, median survival in the non-RCT group was 9 months. With the ineligible patients excluded, this extended to 11 months compared with 13 months in the RCT group. This was not a significant difference, and the survival curves for the eligible non-trial and the trial patients were very similar.

The trial by Marubini and co-workers compared two approaches to breast cancer surgery and again found no significant difference in outcome.[85] They then compared the 352 trial patients with 1408 non-trial patients in a clinical database who had had conservative surgery after completion of the trial. Non-trial patients tended to have worse

outcomes, but they also had poorer diagnostic indicators such as larger tumours and more axillary node involvement. In this study there was no attempt to compare trial patients with 'eligible' database patients, but adjustment for covariate effects in a Cox model gave hazard ratios for the two original groups close to 1 for mortality, distant metastasis and contralateral breast cancer. Intra-breast tumour recurrence remained much more common in the non-trial group however.

Kober and co-workers described results relating to ACE-inhibitors after AMI.[86] The index condition for their trial was left-ventricular systolic dysfunction with wall-motion index ≤ 1.2. They compared the 1739 trial subjects with those screened for the trial (7001 consecutive enzyme-confirmed AMIs), and also the 859 who were screened, found to have the index condition, but not randomised. The main reasons for non-randomisation among those with the index condition were ACE inhibitor mandatory (18%), cardiogenic shock (12%), alcohol abuse, drug abuse or dementia (17%), and refusal of consent (25%). At baseline the trial subjects were younger than the non-randomised, more likely to be male, more likely to have had thrombolysis and less likely to have congestive heart failure. One-year mortality was 24% for the randomised subjects, 54% for the non-randomised with the index condition, and 23% for the larger group screened after AMI. For the 218 eligible patients who refused consent, 1-year mortality was 32%, so the difference was more to do with eligibility than consent. This study is difficult to interpret because again there was no separate analysis by treatment group, but the authors considered that the difference in mortality between the in-trial and out-of-trial mortality rates cannot be explained by a beneficial effect of the ACE inhibitor in half of the randomised patients, and that the results cannot readily be extrapolated beyond those randomised.

The Toronto Leukaemia Study Group[87] studied 272 consecutive patients with acute myeloblastic leukaemia admitted to 14 general hospitals in Toronto. Overall the remission rate was 44%, but excluding patients who were untreated (43), partially treated (31), aged over 70 years (58), pre-leukaemic (11) and who had had chemotherapy for a previous malignancy (7) brought the remission rate up to 85%. The first 130 patients received a different treatment regimen to the second 142 patients. Without exclusions, the remission rates for the two groups were 35% and 52%, respectively; with exclusions they were 78% and 91%, respectively. The age criterion seemed to have had the greatest impact on remission rate.

Excluded groups may be denied effective treatment

In 1990, age under 75 years and presentation within 6 hours of onset were common criteria for thrombolysis after AMI, reflecting the exclusion of older and later presentation patients from most of the relevant RCTs. Some indication of the resulting loss of potential benefit is given by *Table 6*, derived from Yusuf and co-workers,[64] but this does not tell the whole story. As Muller and Topol[54] emphasised, the number of lives saved per patient treated is far greater in the elderly (8/100 treated) than for 'protocol' patients (3.5/100) or late presenters

TABLE 6 *Entry characteristics and outcome: thrombolysis for MI*

Category of patient	Approximate % of MI patients	Approximate size of mortality reduction (%)	Recent large RCTs including such patients
ST elevation < 6 hours; age < 75 years	20–30	30–40	ISIS-2, GISSI, ASSET, AIMS
Other electrocardiography abnormalities at entry	10–15	10–20	ISIS-2, GISSI, ASSET
Over 75 years	10–5	20–5	ISIS-2, a few in GISSI
Onset between 6 and 12 hours	20–30	15–20	ISIS-2, a few in GISSI
Onset between 12 and 24 hours	15–20	15–20	ISIS-2

Source: derived from Yusuf et al, *1990*[64]

ISIS = International Study of Infarct Survival
GISSI = Italian study groups for treatment of myocardial infarction
ASSET = Anglo-Scandinavian Study of Early Thrombolysis
AIMS = APSAC Intervention Mortality Study

(1.5/100) because of differences in baseline risk. The ISIS-2 study, with no upper age limit, suggested a reduction of mortality from 37% to 20% for patients aged 80 years or more, compared with from 6% to 4% in patients aged less than 60 years. And yet of 3256 patients in hospital with MI in the Seattle area in 1988–89, 29% of those aged less than 75 years had intravenous thrombolysis compared with 5% of those aged 75 years or more.[88]

As Yusuf and co-workers[64] pointed out, in early beta-blocker RCTs, patients with heart failure were excluded because it was generally thought that treatment would be dangerous. In the few RCTs that included such patients the reduction of mortality was similar in those with and without heart failure.[89] The elderly also suffer from relative under-treatment with beta blockers.[90] Maynard and co-workers reported underutilisation of thrombolytic therapy in eligible women with AMI.[91]

The results of the quality-adjusted meta-analysis of treatment for hypertension by Holme and co-workers[92] suggested greater efficacy (as measured by the odds ratio for mortality) in the older patients, though this was not statistically significant. However, the SHEP trial,[71] which included only 1% of the patients screened, was very influential in this result.

RCTs based on highly selected groups can be ignored

One further problem with research results based on restricted groups of patients is they can be dismissed if they run counter to 'expert' views. An illustration of this is provided by a series of papers in *The New England Journal of Medicine* in 1987, following publication of the results of an RCT of extracranial–intracranial bypass surgery to reduce the risk of stroke, involving 71 centres.[59,93,94] The principal finding was that the procedure conferred no benefit overall, and may have been harmful for some sub-groups. The investigators went as far as suggesting that the procedure was never indicated and that third-party payers should not reimburse for it. However, the procedure's advocates protested that these conclusions were invalid because the numbers randomised (1377) were small compared with the numbers at the participating centres who had received surgery outside the trial. A committee was set up to investigate the matter, and after reporting data from the RCT organisers that 1439 of the patients screened had been ineligible, (with no informed consent for a further 475 and a doctor's decision against randomisation for 95) concluded that the investigators' conclusions were "too sweeping because the evidence is limited to the two groups of medically eligible surgical patients – those operated on within the randomisation and those operated on outside it – represented the same distribution of disease and of risk. This observation is especially compelling because the number of medically eligible patients operated on in the RCT was so large in relation to the randomised surgical group".

Discussion

It is obviously important that the exclusion criteria are explicit and satisfied, and an agreed part of the study design. Without this there can be no assurance that recruitment is subject to a consistent ethical code, and no firm basis for signalling exactly what kinds of patient the results can be directly applied to.

However, it seems that historically at least, the exclusion criteria have been drawn wider than they need have been, and that large proportions of patients with the 'index' condition have been excluded from RCTs. This has had three kinds of possible disadvantage:

- **false-positives** – inappropriate extrapolation of positive findings to patient groups excluded from RCTs
- **false-negatives** – denial of effective treatment to patient groups excluded from RCTs because of a reluctance to extrapolate from limited RCTs, or treatment protocols based strictly on the available evidence
- it has taken longer, and been more costly, to accumulate the sample sizes necessary to establish effectiveness.

It also seems that in many cases the justification for some of the exclusion criteria was weak, or at least over-cautious, and occasionally ethically and/or scientifically suspect.

The medical justifications relate to ensuring that there is genuine doubt about the effectiveness of treatment for the subjects of the RCT, and avoiding putting patients at undue risk. Judgement is involved here. How strong does the evidence for efficacy have to be before patients are excluded because the experimental treatment cannot be withheld? And how severe must the risk of adverse effects be? The large numbers of replicated RCTs that have made meta-analysis both possible and useful suggest that the answers are generally unclear, and the natural tendency, particularly in a litigious environment, is to err on the side

of caution, and use blanket or wide-ranging exclusion criteria, wider perhaps than in normal clinical practice. There is a risk that certain groups, such as the elderly and expectant mothers, may be treated on the basis of results from RCTs from which they were excluded.

The scientific reasons are to do with improving the precision of the study and avoiding bias. Yusuf and co-workers argue that in human subjects the heterogeneity in response to treatment is so great that any attempt to improve homogeneity by focusing on a particular stratum will have a very limited effect. To some extent the appropriateness of exclusions designed to exclude the non-compliant depends on whether the RCT is designed as an explanatory or 'efficacy' study, or a pragmatic one aimed at effectiveness. The evidence in this chapter suggests that, at present, many RCTs fall into the former category though those interpreting and applying their results treat them as if they are the latter. In the context of the model developed earlier, it is clear that many RCTs are quite unrepresentative of the population to which their results will be applied, with **e** accounting for up to 98% of the reference population in some cases. The central question is whether results from such highly selected groups can be generalised. This will be discussed in chapter 12.

Summary

- The percentage of potentially eligible subjects included in RCTs varies greatly, from 1% to almost 100%.
- The reasons cited for exclusion may be medical or scientific. Medical reasons include:
  - high risk of adverse effects
  - benefit already established
  - benefit known or believed to be very small or unlikely
  - benefits believed to be less than for others deemed eligible.

  Scientific exclusions include:
  - increased precision by inclusion of only those subjects with high probability of specified outcomes
  - reduction of subsequent drop-out, which may introduce bias if greater in one arm than the other.
- In addition, there are common blanket exclusions, including the elderly, women and ethnic minorities; these are often unjustified.
- Patients with a given condition in databases tend to have poorer prognoses that those included in trials. Prognosis for 'eligible' cases in databases will be more similar to trial subjects that for ineligibles.
- Adjustment for prognostic indicators can reduce differences between database and trial patients.
- High levels of exclusions can lead to:
  - unjustified extrapolation of results to other populations
  - denial of effective treatment to those who might benefit
  - delay in obtaining definitive results because of inadequate sample size.

Chapter 5
Participation

Introduction

Even if most potential subjects meet the eligibility criteria, the generalisability of research results still depends both on the extent to which the clinicians and treatment centres that participate are representative of those providing services, and the degree to which the patients who agree to participate are representative of all those in need of treatment. Participation bias may, therefore, arise in two ways.

First, the providers who agree to take part may be atypical (**d** in *Figure 1*): if they are interested in research they are more likely to work in teaching or specialist centres, which traditionally are better resourced than non-specialist facilities; they may have a special interest, experience and skill in treating such patients; and being based in a specialist centre, their case mix may differ from non-specialist centres. Second, patients who participate may differ from those who do not. Non-participation by eligible patients may arise from clinicians not inviting them to participate (practitioner preference or oversight, **i** in *Figure 1*) and from patients refusing to participate when invited (either because the patient prefers one particular intervention or because he or she does not wish to take part in any research study, **p** in *Figure 1*).

Although non-randomised studies may, and often do, fail both to involve a wide range of providers and to recruit all eligible patients, by virtue of the additional administrative and organisational workload involved, this problem is likely to be a greater threat to the conduct of randomised studies. We have therefore concentrated on reviewing the evidence derived from RCTs, though many of the issues will apply to any form of evaluative research.

The objectives of this chapter are:

- to assess the representativeness of providers of care (centres and clinicians) participating in randomised studies and to see how they differ from non-participating providers
- to determine the extent of any bias arising from participating providers being unrepresentative
- to assess the representativeness of patients participating in randomised studies and to see how they differ from non-participants, and
- to determine the extent of any bias arising from participating patients being unrepresentative.

The impact of selective participation by centres and clinicians

The initial search of the literature revealed no examples of studies that provide any direct evidence related to this issue. However, the study by Stukenborg cited in chapter 3 provided indirect evidence that centres participating in RCTs of carotid endarterectomy have lower perioperative mortality rates than those that do not.[47] In a similar study, Wennberg and co-workers examined Medicare data on perioperative mortality in the hospitals participating in two large studies of carotid endarterectomy, and found that treatment in a participating hospital was associated with a reduction of 15% in mortality compared with high volume non-participating hospitals, increasing to a reduction of 25% compared with average volume hospitals and 43% compared with low volume hospitals.[95]

The impact of selective participation by patients

The possible impact that selective patient participation might have on research findings has been recognised for several decades. Research in the 1970s examined the characteristics of volunteers for behavioural research and concluded that volunteers tended to be better educated, of higher socio-economic status, of higher intelligence, in greater need of social approval and be more sociable.[96]

The first review of clinical research considered 41 NIH RCTs.[53] Only 14 RCTs had documented the differences between patient participants and non-participants. The authors of those reports claimed their participants were similar socio-demographically to non-participants. However, in eight of the RCTs the participants were more severely ill than the non-participants. It also suggested that most non-participation was due to clinicians' refusal to invite certain patients rather than refusal by the patients themselves.

Reviewing a wide range of RCTs, Hunninghake and co-workers suggested that the nature of any

participation bias depended on the type of intervention being studied.[51] Similar to the findings from behavioural research, participants in RCTs of prevention strategies were more likely to be of higher socio-economic status, better educated, married and employed than non-participants. In contrast, this was not true for treatment RCTs (for which there was very limited information). They concluded that it is important to understand and address the differences in the rate at which socio-demographic sub-groups are willing to participate in clinical trials, and that more research is required to better assess which sub-groups are under-represented and why.

Evidence of under-representation of particular minority groups soon appeared. Reviewing 50 drug trials, Svensson obtained data on the racial mix of patients in 35 of them.[81] The proportion of black patients was less than their proportion in the population. He then focused on trials of antihypertensive medications for which a differential physiological response between black and white patients had been well documented. Of 15 trials, only seven reported the racial mix of participants and only one attempted to compare responses between the black patients and the white patients. Observations from neonatology provided additional evidence.[97] Walterspiel noted that in Texas, it seemed easier to obtain consent from black mothers than from Hispanic ones. In a comparison of neonates recruited to a trial he conducted, he found those recruited with consent were healthier than those recruited when consent was not required. He attributed this to his reluctance to invite the parents of very sick neonates because of the difficulty such encounters entailed.

Meanwhile, awareness of the rather low participation rates being achieved in many randomised trials was growing. In the 1980s, only 2% of eligible patients with breast cancer in the USA were entering trials.[98] Taylor and co-workers showed that this was largely due to clinicians' reluctance to raise issues of uncertainty with patients newly diagnosed with malignant disease.[99] The suggestion that clinicians rather than patients were the principal reason for non-participation in cancer trials was confirmed by Gotay.[100] Contrary to earlier reports, however, she found participants tended to be younger, in better health and of higher socio-economic status than non-participants.

Despite these reports, the call for improved documentation of recruitment to trials that Hunninghake had called for in 1987 had not resulted in widespread change. A *Lancet* editorial in 1992 commented that the reasons why some subjects refuse to participate in research has not been studied much by empirical researchers.[101] Despite this, some commentators were convinced of the existence of participation bias: "Whether it is demographic characteristics, behaviour, personality factors or health status, when an investigator looks for differences between research participants and non-participants, differences are generally, although not always, found".[96] And of its likely effect: "Differences between the individuals who participate in a trial and those who refuse can have ethical implications, as well as potential impact on the generalisability of the trial's conclusions".[66]

Methods

The strategy used to search electronic databases has been described in chapter 2. Papers were only considered if they gave details of baseline characteristics of participants in randomised trials, together with details of those who did not participate. This led to the identification of 20 relevant papers.

The details of participants' and non-participants' characteristics were extracted from the papers and tabulated. Where the authors had tested for statistical significance this was also included, with the test name if reported. If no such test had been performed and adequate information was available in the paper, an appropriate single statistical test was performed for this review. Multivariate analyses were not conducted.

Providers (centres and clinicians): representativeness of participants

The randomised and non-randomised trials included in two systematic reviews of elective surgical procedures were used to assess the representativeness of the centres and clinicians who take part in research.[102,103]

The review of laparoscopic cholecystectomy included 15 randomised and 21 non-randomised studies. Only two of the RCTs were multicentre (both included five centres), the other 13 being based in one centre. All 15 trials were carried out in university hospitals. In contrast, six of the non-randomised trials were multicentre: five were restricted to only two hospitals, and one included 19. Only eight of the 21 non-randomised studies were conducted in university hospitals.

The review of stress incontinence surgery included 11 randomised and 20 non-randomised studies. All of the RCTs were single-centre studies and all but

one were carried out in a university hospital. Only three of the non-randomised trials were definitely multicentre (it was unclear in another three studies) and they were restricted to only two hospitals. All but one were carried out in university hospitals.

Patients: participants versus non-participants

In view of the observations made by Hunninghake and co-workers,[51] studies were reviewed in two categories – treatment trials and prevention trials. Although many of the papers included the authors' assessment and interpretation of any participation bias, we based our conclusions on our own assessment of the data.

Treatment trials

The search revealed 16 randomised trials that reported entry characteristics of both participants and non-participants. They come from across the whole range of health care: cancer (five), coronary artery surgery (two), AMI (three), childhood asthma (two), tonsillectomy, low birthweight infants, mental illness, and nutrition (one each).

Details of each trial are shown in appendix 5.[30,33,35,104–119] Most studies report both socio-demographic and clinical characteristics. Not surprisingly, studies vary in the characteristics included. While all but two[35,110] consider patient age, some characteristics are considered in only one of the studies: private health insurance,[114] social support,[107] number of siblings.[33]

Participation was significantly more likely if the patient was male,[104–106,113,114] younger than average,[105–107,113,114] non-white,[33,109,112,114] less educated,[104,107,112] of lower socio-economic status,[33,107] a smoker,[30,107,113] had inadequate social support,[107] and had no private health insurance.[106] Other studies either did not examine the characteristic or found no statistically significant difference between participants and non-participants.

As regards clinical characteristics, participation was significantly more likely if the patient had more severe or advanced disease,[84,105,108–110,112,114] more comorbidity,[30,112] and poorer health status or quality of life,[104,107,112] While other studies generally found no statistically significant difference, in two studies participants had less rather than more severe disease.[111,113] None of these findings were related to the types of disease being studied.

Five studies have reported outcomes both for participants and non-participants (one found from this search and four from the search used in chapter 3) and an inconsistent picture emerges. In a randomised trial of total parenteral nutrition in malnourished surgical patients, the incidence of complications during the first 30 days was 25% among participants and 15% among non-participants.[111] The RR (non-participants/participants) adjusted for total parenteral nutrition use was 0.65 (95% CI: 0.44–0.95). The RR after 90 days was 0.60 (CI: 0.42–0.86). In other words, participants had a worse outcome consistent with the finding that those who participate in RCTs tend to be sicker than those who do not participate. Other studies reveal a more mixed picture. Patients who declined to participate in an RCT of coronary artery surgery either had a worse outcome (6-year mortality 12% vs. 10% among those treated medically) or there was no difference (8% for all surgical patients[30]). Two small RCTs, one of tonsillectomy[33] and one of adenoidectomy[34] revealed that for some measures the opposite was true. Finally, in an RCT of rehabilitation for alcoholic individuals, participants who were allocated to day care did consistently better than non-participants attending day care but the opposite was true for the in-patients.[13]

Prevention trials

The search revealed only four randomised trials of interventions aimed at promoting health or preventing disease. Two were aimed at reducing the prevalence of cardiovascular disease,[37,117] and one each aimed at breast cancer prevention,[118] and balance enhancement in the elderly.[119]

Details of each trial are shown in appendix 5. Most studies report both socio-demographic and clinical or physiological characteristics. Participants were more likely to be younger,[118] be of higher social status (in terms of income,[118] housing,[117,119] education,[118,119] or car ownership[117]), and believe in and adopt a 'healthy lifestyle' (e.g. non-smoker, regular exercise).[116,117,119] In only one study was there a difference in baseline health status between participants and non-participants – participants in the balance enhancement trial were in better health than non-participants.[119]

Not surprisingly, given the long-term nature of prevention, none of the studies reported any outcomes both for participants and non-participants.

Discussion

The evidence concerning centres that participate in evaluative research is extremely limited but it does show that outcomes can differ significantly

from those obtained in other centres. In the absence of further information, it is not possible to generalise this finding, though it does raise grounds for concern.

Practitioners who take part in evaluative research are predominantly based in university or teaching centres. While this is true both for randomised and non-randomised studies, the latter are more likely to be representative of typical clinical practice as they are more likely to include non-teaching centres (*Figure 7*). The impact that such a lack of representativeness has on the treatment effect size has not been investigated sufficiently to be able to draw any conclusion as to the importance of this observation.

The patients or people who agree to participate in randomised trials (few people refuse to participate in non-randomised) are also different from those who refuse. Participants in treatment trials tend to be less affluent, less educated and more severely ill than those who do not. In contrast, people participating in preventive trials are more affluent, better educated and are more likely to have adopted a 'healthy lifestyle' than those who decline (*Figure 8*). These findings are based on our own analysis of the data presented in published accounts of the study. Authors often ignored or discounted clear, statistical evidence that participation bias may have occurred – presumably because they felt it would undermine their findings.

Before considering the implication of these findings, the methodological limitations of our review need to be explored.

- We were limited by the paucity of studies available. Few randomised (or non-randomised) studies report on the numbers of the eligible subjects who were invited to participate and who subsequently agreed. This is probably because the final proportion is often rather small.

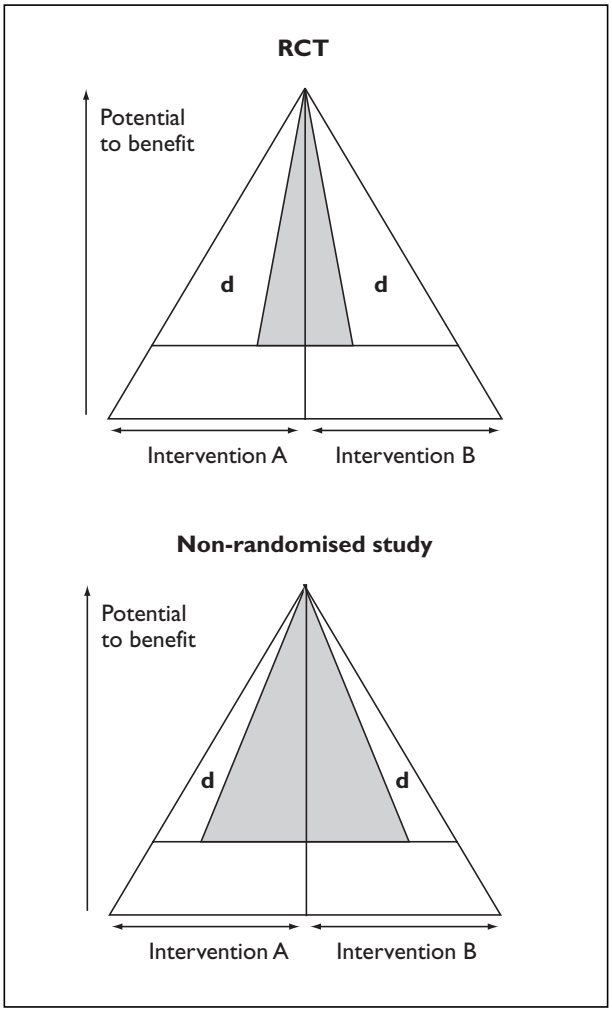


FIGURE 7 *Schematic effect of differences in centre participation*

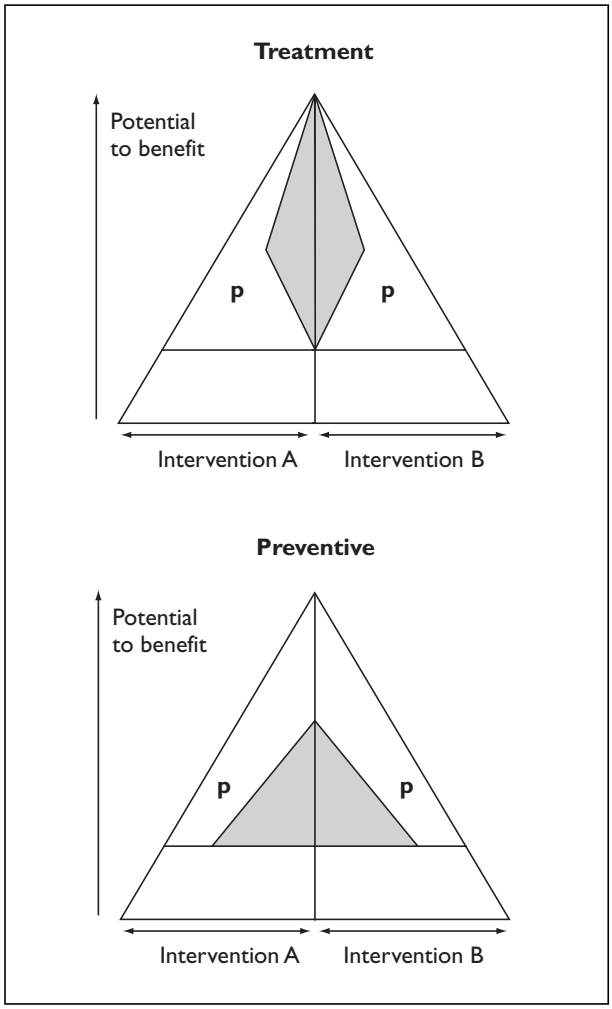


FIGURE 8 *Schematic representation of possible differences between participants in treatment and preventive trials*

- Even fewer studies present data to enable comparisons of participants and non-participants.
- The small sizes of many studies mean that some clinically significant differences may not reach statistical significance at the 5% level and are therefore ignored by investigators.
- Of those studies that have reported on participation bias, the non-participants included have sometimes been a self-selected sample of all non-participants. This may dilute any difference in characteristics from the participants.
- Only five studies present outcome measures and these suggest there is little consistency in the magnitude of any impact participation bias may have.
- Only two studies attempt multivariate analysis to explore for confounding.

So what are the implications of the finding that randomised trials are usually carried out in single centres which are unrepresentative of the centres in which most patients are treated? The answer is unclear as most trial reports have in the past failed to provide sufficient data on non-participants. We can, therefore, only postulate any likely impact. In treatment trials, unrepresentative participation may exaggerate the effect size of the intervention as the practitioners are likely to be 'better' than average and the patients are more severely affected and therefore have greater capacity to benefit from the treatment. In contrast, in prevention trials, the effect of the intervention may be underestimated as those who agree to participate are 'healthier' than average and therefore stand less chance of benefiting. Counteracting this, however, is the fact that those participating are also more likely to comply with the intervention and this will enhance any effect it has.

The existence of systematic differences in the nature of participation in preventive and treatment RCTs is strengthened by detailed consideration of one RCT. The trial of infant development[115] has, superficially, many of the characteristics of a study of prevention until it is realised that those at whom the intervention is targeted are differentiated from the general population by clearly defined medical characteristics, prematurity and low birthweight. Consequently, the study actually has many of the characteristics of a study of treatment. It is therefore interesting to note that participation was greatest among non-white groups, those attending non-University hospitals, and those with very low weight neonates.[115] In other words, participation bias in this study was similar to that seen in treatment trials.

The finding that participants in treatment trials are less educated and less affluent raises questions as to the extent to which fully informed consent is being obtained. If, as people become better educated and wield greater financial power they are less inclined to agree to be randomised, there is a suggestion that consent for randomisation may not always be clearly ethical. Another interpretation is that it will become increasingly difficult to mount randomised studies in affluent, well-educated societies where individual expectations include self empowerment and control of decisions such as how to be treated.

Finally, this brief review of participation bias has highlighted the need for full documentation of studies. The recent Consolidation of Standards for Reporting Trials (CONSORT) initiative taken by many leading biomedical journals will help enormously and provide future reviews of this issue with a wealth of evidence.[120] Until then, there is sufficient evidence to suggest participation both by providers and patients should be taken more seriously, and that it can introduce biases that may affect the reported effect sizes of interventions.

Summary

- Many RCTs fail to document adequately the characteristics of those who, while eligible, do not participate.
- Evaluative research is undertaken predominantly in university or teaching centres.
- Non-randomised studies are more likely than RCTs to include non-teaching centres but it is not possible, on the basis of available evidence, to know how important this is, though it is plausible that it will exaggerate any treatment effect.
- Participants in RCTs evaluating treatments tend to be less affluent, less educated, and more severely ill than those who do not participate.
- Those who participate in RCTs evaluating preventive interventions tend to be more affluent, better educated, and more likely to have adopted a healthy lifestyle than those who decline.
- The implications of these findings are speculative but it is plausible that RCTs of treatment may exaggerate treatment effects by including more skilled practitioners and participants with greater ability to benefit, while RCTs of prevention may underestimate effects as participants have less ability to benefit.
- These results raise questions about the nature of informed consent.

Chapter 6

Patient preference

Introduction

Randomising patients between a treatment and a control enables the reliable estimation of the average biological effect of an intervention, uncontaminated by confounding. This process **should** create a situation in which all other influences on prognosis are equal. In answering the question of whether the treatment works in a consistent way, the methodological complexities of detecting interaction are largely ignored.

An interaction, in contrast to a main effect of a treatment, describes the possibility of particular treatment effects that are different according to salient characteristics of patients. One possible kind of interaction is that between the physical and psychological influences on therapy. In this chapter we examine the possibility that individual preferences will themselves influence how well a treatment works. This can be characterised as the therapeutic effect of patient preference and is possibly strongly related to the well-recognised placebo effect.[121] For the purposes of this chapter, we have focused on what we describe as patient preference, though we recognise that practitioner preference is a strong determinant of the views of patients.

The ability to detect the effect of preference, if it exists, is compromised by the possibility of confounding;[12] people who tend to prefer something may be different in other ways that may plausibly be related to prognosis, from those who do not.[13,14] Obviously it is impossible to randomise between enthusiasm for a treatment and absolute rejection of it and, in circumstances where there are strong preferences, randomisation is beset with difficulties.[15] Furthermore, the reliable detection of interaction is particularly difficult because the number of patients required for a given statistical power is usually an order of magnitude greater than to detect a main treatment effect of similar magnitude. Thus, to detect a 10% improvement in survival usually requires over 1000 patients for a 90% power, but an interaction effect of around 10% between preferences and treatments will require several thousand for the same power. Moreover, it is also likely that many practitioners and patients with strong preferences will exclude themselves from RCTs[122] (unless an RCT is the only chance a patient has for receiving a new experimental treatment).

Non-randomised studies typically include patients who self-select, or their practitioner selects, their treatment. Conversely, RCTs, by definition, deny patients the ability to express their preferences. Depending on the strength of their preferences and whether the trial is blinded, it is conceivable that random allocation may have important repercussions on the results. RCTs might be expected to underestimate the benefit of treatments because the outcomes are less likely to be enhanced by any positive beliefs patients or carers may have.

Consequently, it is essential to be able to disentangle the physiological effect of a treatment from any possible effect of individual preference. Otherwise it may be necessary for evidence from RCTs to be treated with caution.

The objectives of this chapter are:

- to identify literature that investigates preference effects
- to develop a model that quantifies potential preference effects
- to explore the potential application of proposed means of addressing preference effects.

Evidence for preference effects

Irrefutable evidence of significant effects on outcome of patient preference for a particular treatment is sparse. Such effects are difficult to detect reliably so, not surprisingly, there have also been few attempts to identify them. A systematic attempt was made to find papers that have tried, by various methods, to study the possible impact of patient preferences. The search strategy outlined in chapter 2 was used and identified only four papers.

In a study by McKay and co-workers,[13] 48 alcoholic patients were randomly assigned to day hospital or inpatient rehabilitation. A total of 96 patients refused randomisation due to strong preferences and self-selected their treatment setting. Patients who were randomly assigned to treatment were

more likely to be African-American, to have received welfare in the previous 30 days, and to have reported more days of cocaine use in the previous 30 days. However, there were no significant differences between those patients who self-selected treatment setting and those who agreed to be randomised in terms of alcohol and drug use outcomes after treatment, or in terms of psychosocial outcomes.

Nicolaides and co-workers also looked at the characteristics of patients who were given the opportunity to select their treatment.[37] A total of 1870 women were offered participation in a randomised trial of CVS or EA. Of these, 488 agreed to be randomised, while 813 chose the procedure according to preference. There were no differences with respect to maternal age and weight, employment, cigarette smoking, previous obstetric history, vaginal bleeding, gestation at sampling, foetal crown-rump-length, and placental site. There were no significant differences between those who were randomised and those with preferences in the frequency of abnormal karyotypes or rates of foetal loss (total, induced or spontaneous).

A different approach was used by Torgerson and co-workers.[14] A total of 97 patients entered an RCT evaluating treatment for low back pain. Prior to randomisation the patients were asked their preferences. Fifty-eight patients preferred to be allocated to the exercise programme, while 38 were indifferent. One patient preferred conventional general practitioner management. Despite these stated preferences, no patient refused randomisation. Patients who preferred the treatment arm into which they were randomised had more confidence in its likely effectiveness. The indifferent patients had had back pain for longer than those who stated a preference, but the pain, on average, was not as severe as the preference group's pain. Unfortunately, no outcome measures after treatment have been reported yet.

In a study conducted by Fallowfield and co-workers,[123] an attempt was made to link preferences with psychological anguish. Of 269 women with early breast cancer, 31 were treated by surgeons who favoured mastectomy, 120 by surgeons who favoured breast conservation, and 118 by surgeons who offered a choice of treatments. Patients in this last group showed less anxiety and depression at 3 months and 12 months after surgery than those treated by other surgeons. Again, no details of outcome are given, though it has been reported that reduced anxiety and depression are associated with improved cancer outcome.

None of these papers managed to show conclusively whether preferences affect outcome. All that can be concluded from this rather limited empirical evidence is that preferences exist and that the characteristics of patients who chose their treatments may be different to those who agree to randomisation (as found in the chapter on participation). Whether these preferences work as an enhanced placebo effect and influence their biological outcome is not clear.

There is, however, indirect evidence for such effects. Medical treatment for the secondary prevention of CHD in the Coronary Drug Project appeared 10% more effective at delaying death if a placebo was 'properly' taken, than if not, in a double-blind RCT.[124] If drug compliance is considered a measure of some enthusiasm for the treatment, while accepting that other factors are also involved, then individual preferences (and/or expectations) may seem to have an important effect on outcome, which is not strictly pharmacological. The effect of compliance on mortality, even with a placebo, was very highly significant, even after adjustment for 40 potential confounding variables. In this case compliance is a proxy for preference in that the compliers believe the treatment to be effective and prefer to take it over nothing.

There are other studies that indirectly suggest that preference may have an important effect on outcome.[11] In a study of 28,000 adult Chinese-Americans,[125] Phillips and co-workers found earlier deaths among those with a combination of disease and birthyear considered by Chinese astrology and medicine to be ill-fated. They suggested that this phenomenon is partly a result of psychosomatic processes, where a strong belief has physiological consequences. In addition, there are comprehensive reviews of the placebo effect that add to this evidence.[126,127] An example cited in chapter 3 was the finding in the study by Heinsman and Shadish that, in a multiple regression model, the use of no treatment rather than a placebo in the control group was consistently associated with a greater relative benefit of the intervention.[48] Possible mechanisms by which the effects are exerted have been described, including the possible involvement of endocrine and immune systems.[128,129] The combined literature provides supportive evidence that psychological factors (whether they are preferences, beliefs or enthusiasm) can effect physiological outcome.

Consequently it is necessary to understand the situations in which such effects might be important. The most likely situation is where blinding is

difficult or impossible. In the absence of empirical evidence, this can be examined further by means of a theoretical model.[130]

A simple additive model

In its simplest form, the model is based on a condition for which two possible treatments, A and B, have been advocated. It is assumed that, with regard to the purely physiological effect, A benefits on average a proportion P of eligible people, and B a higher proportion P + *x*, in the absence of any effect of patient preference. Thus, to take an example where the measured outcome is 5-year survival, if P is 0.50 and *x* is 0.10 then on average 60% would be alive at 5 years on treatment B.

If a preference effect does exist, then having a preference for A would bestow an extra average advantage for treatment A of an amount *y*, giving P + *y* and a preference for B of a similar amount *y*, giving P + *x* + *y* for treatment B. Conversely, of those who prefer A, only P + *x* – *y* will benefit if given treatment B, and of those who prefer B, P – *y* will benefit if given A. These are postulated average interaction effects for patients among whom these treatments would be appropriate, and this simple model allows for a preference interaction even if the main effect of the new treatment is zero, that is if *x* = 0. These effects are summarised below:

Postulated treatment effects if:

	indifferent	prefer A	prefer B
on treatment A	P	P + *y*	P – *y*
on treatment B	P + *x*	P + *x* – y	P + *x* + *y*

If the proportion of the eligible population who prefer treatment A is α, while β prefer B and γ are indifferent, then it is required that $(\alpha + \beta + \gamma) = 1$. Clearly, the interaction between these effects might be more complicated, with multiplicative, graded or asymmetric interactions.

It can be shown (by subtracting the estimated mean effect in group A from that in group B) that the estimate of the attributable effect of treatment B over treatment A in a large well conducted randomised comparison will then be:

$$x + 2y\ (\beta - \alpha)$$

This is different from *x* (the true physiological effect) by an amount equal to $2y\ (\beta - \alpha)$ (the preference component) and hence such trials will only estimate *x* correctly either if *y* is zero (no effect of preference) or if $\beta = \alpha$ (an equal proportion prefer A as B). Obviously, the effect of random variation is ignored for simplicity. The next step is to examine how great such a 'bias' might be, under reasonable assumptions about *y* and $\beta - \alpha$.[131]

Considering the size of the difference in the proportions preferring the two treatments, if 35% prefer treatment B and 60% treatment A, the difference is 25%, (i.e. $(\beta - \alpha) = -0.25$). In the model, if the average 'physiological' effect of B over A (i.e. *x*) is 10% (let the effect of A alone be arbitrarily 50%) and if the preference advantage (i.e. *y*) is 5% then the treatment effects will be as follows:

Postulated treatment effects if:

	indifferent	prefer A	prefer B
on treatment A	50%	55%	45%
on treatment B	60%	55%	65%

In such a case, simple substitution indicates that a fair RCT will be wrong by 25% (i.e in this case: $2y\ [\beta - \alpha]$ is 25% of *x*). That is *x* (the 'physiological' effect) will be estimated as 0.75*x*, or if 60% prefer B and 35% A (i.e $[\beta - \alpha] = 0.25$), then the unbiased RCT estimate will be 1.25*x*. Either way these results would be wrong, as the estimated effect will be attributed to the treatment alone but will, in reality, reflect the distribution of preference effects.

If the difference in the proportions who prefer A or B is 50% then the size of the 'bias' from a randomised comparison rises to 50%, for the hypothetical values of *x* =10% and *y* = 5%. If, however, *y* is only 1% then the 'biases' in the results of RCTs will be reduced to 5% for a 25% difference in proportions with contrasting preferences, and 10% for a 50% difference. However if *y* = 10% (i.e. the role of preference is more profound than the physiological treatment effect) then the trials will be respectively 50% and 100% 'out', on average (i.e. the treatment effect will be estimated as 1.5*x* or 2.0*x*). This is potentially important if such large differences in the prevalence of preferences can be shown to be plausible.

The Coronary Drug Project observed a 26% 5-year mortality among non-adherers and 16% among adherers to placebo in the placebo arm of a double-blind RCT of drug prophylaxis. A value for *y* of 5% ($2y$ = 10% = 26–16) would thus be reasonable. Sixty-seven per cent of participants took more than 80% of the prescribed dose and 32% took less than this, a difference of 35%. If this is a reasonable representation of real preference effects, then trials of such interventions will overestimate the (absent) pharmacological role of such drugs by 3.5% in general.

If an unblinded or poorly blinded trial comparing placebo with a supposedly active new treatment is considered, where the benefits of the treatment are highly plausible but which in fact has no additional physiological benefit, the true difference in the prevalence of preferences might be large, with up to 90% preferring the new 'active' treatment, only 5% preferring the control and only 5% being indifferent. If such values are plausible, '$\beta - \alpha$' becomes 0.85 and hence (if $y = 0.05$) the bias is 8.5% in absolute terms. Consequently, if the natural history is such that 50% of subjects improve or survive regardless of treatment, such a trial would suggest that treatment improved this to 58.5%. This situation could arise where a new product received intensive press coverage due to high-profile marketing.

Discussion

The implications of these arguments are important. As clinical researchers are encouraged to randomise between successive new treatments, rather than comparing every new treatment to placebo,[132] the average net treatment effect in RCTs is thus minimised (as new treatments are typically less different from current treatment than from a placebo). Patients with a chronic disease, for which the current treatment is known not to be very effective, may be attracted to new treatments. Although Chalmers shows that new treatments are just as likely to be worse, as better, than their predecessors,[133] this message is often lost in the promotional activity of those with an interest in a new product, their enthusiastic clinical messengers, and patients anxious to try anything that may work.

Believing in a treatment does not, of course, necessarily enhance its effectiveness. Patients may prefer one treatment over another, not because they believe the outcome to be superior, but because they find the process of that preferred treatment more acceptable. However, if preferences do enhance outcome, the consequences for the uptake of new treatments may be important. If new treatments are favoured over established ones for conditions with poor prognosis (e.g. 50% survival), then the new treatments may gain in apparent effectiveness, even if they have no additional physiological benefit. There is then a tendency for this process to increment as evidence from each successive RCT may affect patient preferences, directly or indirectly,[134] such a process might accrue more and more expensive (and possibly unpleasant) treatments which are actually no better, in the sense of the postulated physiological mechanism, than the standard treatment.

Possible trial designs will be discussed in chapter 11, which may help resolve the situation.

Summary

- This review set out to identify evidence of the effect of preference on outcome. The argument that differences in patient preference can have a significant effect on the results of RCTs has been demonstrated theoretically but empirical evidence is awaited.
- Despite an exhaustive search, only four papers that addressed this directly were identified and they were either small or are yet to report full results. There is, however, considerable indirect evidence for such effects.
- At least in theory, this could have an important impact on results of RCTs, particularly where the difference between treatments being compared is small and could account for some observed differences between results of RCTs and non-randomised studies.

Chapter 7

Surgical interventions: coronary angioplasty and bypass grafting

Having reviewed previously published comparisons of randomised and non-randomised studies, we now present four cases of our own. The objective of these chapters is, primarily, to identify whether any difference in estimated effect size between RCTs and non-randomised studies persists after adjustment of the latter for baseline differences between groups.

Introduction

Our first example is from the surgical literature. Many surgical interventions have been the focus of both randomised and non-randomised studies, but few have outcomes that enable a comparison of methods. Revascularisation in patients with coronary artery disease is one such example, as death provides an unambiguous and sufficiently frequent outcome measure. There are two invasive procedures for this condition which have been widely discussed in the literature: CABG and PTCA. A recent meta-analysis[135] summarised the results of RCTs.[136–143] These RCTs will be compared with two of the most recent and largest non-randomised studies.[144,145] Our objectives are to determine how close the treatment effect estimates are in the two study designs and to examine the possible effects of selection bias and the ability to adjust for baseline differences between groups in non-randomised studies.

Methods

RCTs were identified initially from the meta-analysis and updated by communication with key researchers in this area and by a search of MEDLINE from 1990 to 1996 using the strategy given in chapter 2. Studies were limited to those in which patients who received CABG were compared with those who simultaneously underwent PTCA.

In addition to the references already identified in the meta-analysis, two further RCTs were found. One of these papers gave longer follow-up results for a RCT already included in the meta-analysis.[146] The other RCT is the largest to date[147] and was suggested by the author of the meta-analysis as being an essential addition to update his work. In total this gives nine RCTs.

Non-randomised studies were identified using the MEDLINE strategy described in chapter 2. As with randomised studies, this was supplemented by communication with key researchers.

Only one outcome measure, death at 1 year, was studied. The problem of baseline risk differences in the non-randomised studies was investigated by assessing the effects of statistical risk adjustment models and use of sub-group analysis.

Results

Details of the RCTs included in this review are shown in appendix 6 and the non-randomised studies are detailed in appendix 7. The study by Jones and co-workers was a large non-randomised prospective follow-up of patients receiving treatment at the Duke University Medical Centre between 1984 and 1990. The RCTs were published between 1992 and 1996. Four were multicentre and five single centre, and were conducted in Europe, North America and South America. The Medicare study by Hartz and co-workers included information on all Medicare patients receiving CABG or PTCA in the USA in 1985. These patients were all aged 65 years or older and thus differ from subjects included in the trials. A random sub-group of the patients (n = 2921) was identified and the MedisGroups method (a commercial severity adjustment system) of abstracting information was used to obtain key clinical findings (admission symptoms, history, physical examinations, etc.). Patients were classified as 'high-risk' or 'low-risk' according to the MedisGroups criteria. The results for all the Medicare patients were presented, as well as the random sub-group adjusted by MedisGroups.

The number of deaths and percentages that occurred during the first year of each of the studies (both randomised and non-randomised) are shown in *Table 7*, along with the RRs of dying for CABG

and PTCA. These RRs and the confidence limits are plotted in *Figure 9*.

The probability of surviving with CABG against the probability of surviving with PTCA is shown in *Table 8*, and the results are shown in a L'Abbé plot in *Figure 10*. The points to the far left of the graph represent the survival in the Medicare data, while the other non-randomised study is among the cluster of

TABLE 7 *Mortality at 1 year (unadjusted in non-randomised studies)*

Study name	Study type	No. deaths in first year		No. patients treated		RR
		CABG	PTCA	CABG	PTCA	CABG:PTCA
CABRI	RCT	14	21	513	541	0.70
RITA	RCT	6	9	501	510	0.68
EAST	RCT	4	7	194	198	0.58
GABI	RCT	9	4	177	182	2.31
Toulouse	RCT	2	3	76	76	0.67
MASS	RCT	0	1	70	72	0.00
Lausanne	RCT	0	1	66	68	0.00
ERACI	RCT	3	3	64	63	0.98
BARI	RCT	35	38	914	915	0.92
Duke*	Non-random	102.1	61.9	3890	2924	1.24
Medicare	Non-random	8407	2085	71,243	25,423	1.44
MedisGroups	Non-random	245	73	2063	858	1.40

* *Duke deaths are interpolated from whole study period (mean 5.3 years)*

RITA = Randomised Intervention Treatment of Angina
MASS = Medicine, Angioplasty or Surgery Study

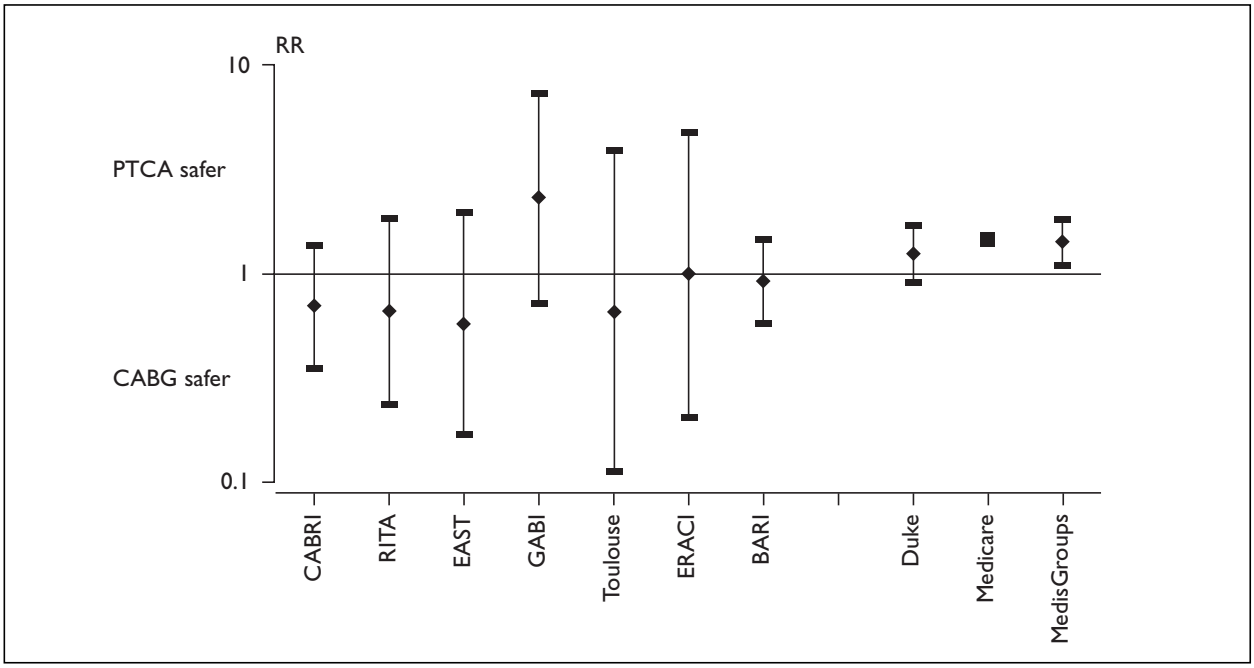


FIGURE 9 *RRs of mortality at 1 year: CABG versus PTCA*

RCTs to the right of the graph. The non-randomised data in these tables and figures are unadjusted for baseline differences between patients undergoing CABG and PTCA.

TABLE 8 *Survival following CABG versus survival following PTCA*

Study	Deaths in first year (%)		Survival (%)	
	CABG	PTCA	CABG	PTCA
CABRI	2.7	3.9	0.97	0.96
RITA	1.2	1.8	0.99	0.98
EAST	2.1	3.5	0.98	0.97
GABI	5.1	2.2	0.95	0.98
Toulouse	2.6	3.9	0.97	0.96
MASS	0	1.4	1.00	0.99
Lausanne	0	1.5	1.00	0.99
ERACI	4.7	4.8	0.95	0.95
BARI	3.8	4.2	0.96	0.96
Duke*	2.6	2.1	0.97	0.98
Medicare	11.8	8.2	0.88	0.92
MedisGroups	11.9	8.5	0.88	0.92

* *Duke deaths are interpolated from whole study period (mean 5.3 years)*

Baseline differences in non-randomised studies

Medicare patients (Hartz et al, *1992)*[145]

Comparisons of the CABG patients with the PTCA patients revealed substantial differences in baseline mortality risk. Factors significantly related to choice of procedure in a multivariate logistic regression analysis were: female gender, congestive heart failure, S_3 gallop, history of MI, history of CABG, history of PTCA, graft failure, diabetes mellitus, blood pH > 7.45 and pH < 7.35.

In general CABG patients were at a higher risk than PTCA patients. Overall the odds of a CABG patient having a high risk MedisGroups severity score compared with PTCA patients was 2.4.

The Duke database patients (Jones et al, *1996)*[144]

The PTCA patients were younger, on average, than the CABG patients and had a higher prevalence of AMI. Sixty-one per cent of the PTCA group had one-vessel disease and 10% had three-vessel disease, whereas for CABG the proportions were essentially reversed (10% and 56%, respectively).

Risk adjustment

Medicare patients (Hartz et al, *1992)*[145]

A Cox's proportional hazards model, adjusting for significant baseline differences, was used with the subset of Medicare patients (the MedisGroups patients) to compare the mortality risk for the two revascularisation procedures. The CABG patients had an even higher mortality risk than the PTCA patients after adjusting. The RR for mortality for CABG compared with PTCA patients was

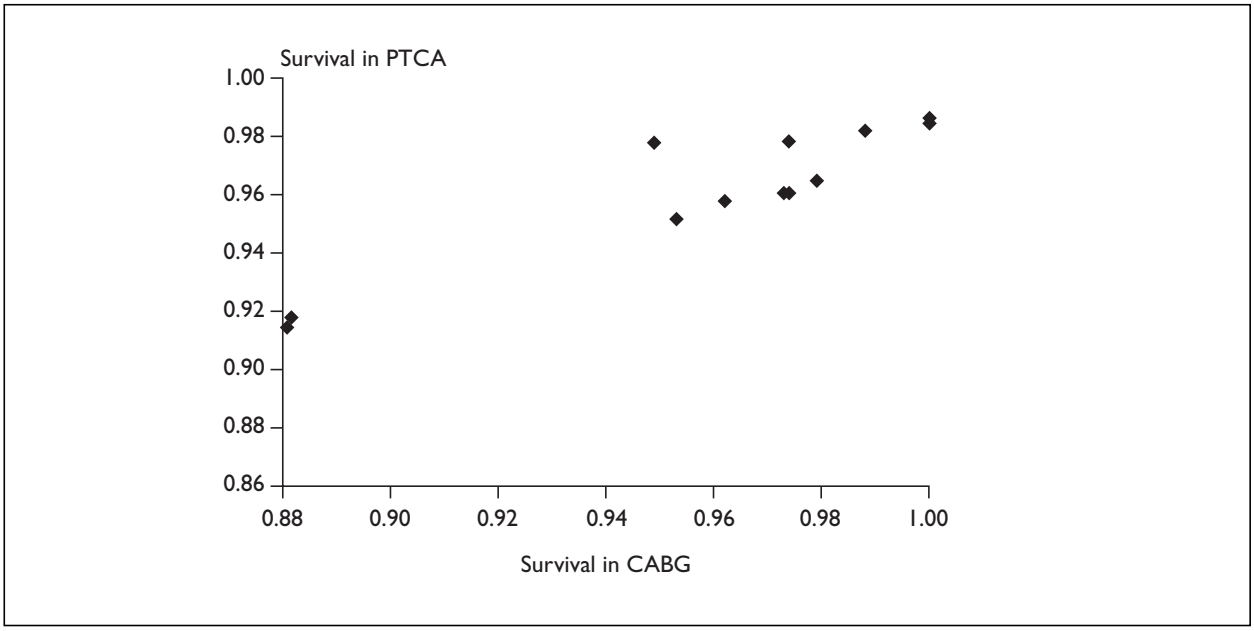


FIGURE 10 *Survival following CABG versus survival following PTCA*

1.72 ($p = 0.001$). This can be compared with the unadjusted rate of 1.44.

The CABG patients were at higher initial mortality risk and so adjustment measures should lower the RR compared with the crude rates of mortality after 1 year. This apparent anomaly may be partly explained by the fact that an important risk factor, the presence of left main coronary artery stenosis, was omitted. The authors concede that this is likely to be much more prevalent in the CABG patients than in PTCA patients.

The patients adjusted using MedisGroups were subdivided into a high-risk group (n = 506) and a low-risk group (n = 1856), according to initial severity risk. The adjusted RR of death following CABG compared with PTCA was much higher for the low-risk patients (RR = 2.15, $p = 0.0003$). The RR of death following CABG compared with PTCA did not differ significantly from 1 (0.90, $p = 0.69$).

Duke database patients (Jones et al, *1996)*[144]

The patients were classified according to a coronary artery disease score. This resulted in nine coronary anatomy groups representing a continuum of one-, two-, and three-vessel disease. A Cox's proportional hazards model, incorporating significant baseline variables (ejection fraction, age, coronary anatomy, co-morbidity, vascular disease, congestive heart failure, initial presentation) was used to compare the RRs of CABG and PTCA in the nine severity groups. The 5-year survival rates in each group, both observed and adjusted, are shown in *Table 9*, along with the RRs.

Hazard ratios were calculated from these adjusted survival rates to depict the relative benefits of CABG and PTCA. Unequivalent benefit of PTCA was seen for Groups 1 and 2. Suggestive benefit from PTCA was seen for Groups 3 and 4, but the difference does not reach statistical significance. Group 5 shows equivalent outcomes for both interventional procedures. Patients in Groups 6 through to 9 have greater survival when treated with CABG (*Figure 11*).

It is apparent that initial patient severity has a major bearing on the relative merits of CABG and PTCA. In order to enable a more valid comparison of the adjusted non-randomised results with the RCTs findings, an attempt was made to classify high-risk and low-risk patients. For the MedisGroups patients, the authors' definitions of 'high-risk' and 'low-risk' were used. For the Duke patients coronary anatomy groups, Group 1 was considered low risk and Group 9 was considered high risk. For the RCTs, the results were separated into patients with single-vessel disease (low risk) and patients with multi-vessel disease (high risk). This was taken from the meta-analysis by Pocock. These comparisons are shown in *Table 10*.

Discussion

In terms of crude numbers of deaths after 1 year, all but one of the RCTs (GABI) showed lower mortality following CABG. Conversely the non-randomised studies favour PTCA. However, given the large confidence limits around the estimates, the results are not strikingly different. The risk

TABLE 9 *The 5-year survival in each severity group, observed and adjusted (Duke data)*

Coronary anatomy group	Diseased vessels	CABG		PTCA		RR*	
		Observed	Adjusted	Observed	Adjusted	Unadjusted	Adjusted
1	1	0.97	0.92	0.95	0.96	0.60	2.00
2	1	0.89	0.91	0.93	0.94	1.57	1.50
3	2	0.92	0.90	0.91	0.92	0.89	1.25
4	2	0.81	0.90	0.87	0.90	1.46	1.00
5	1 and 2	0.86	0.89	0.85	0.88	0.93	0.92
6	2 and 3	0.87	0.89	0.80	0.85	0.65	0.73
7	3	0.85	0.88	0.74	0.80	0.58	0.60
8	3	0.81	0.86	0.83	0.75	1.12	0.56
9	3	0.83	0.85	0.61	0.68	0.44	0.47

* *RRs are CABG vs. PTCA*

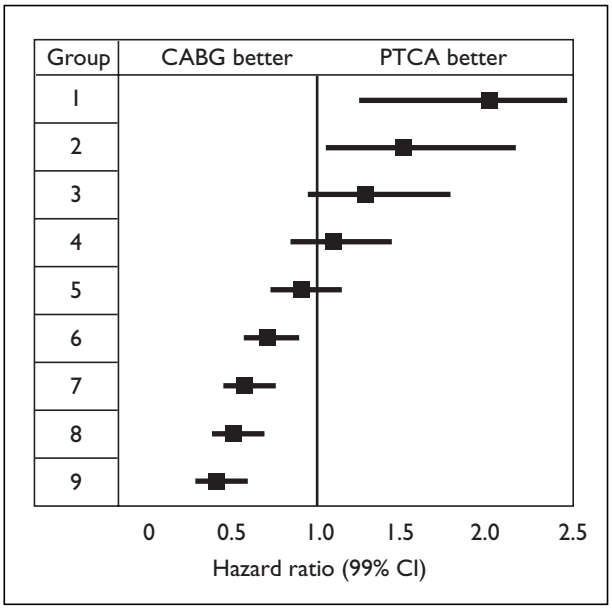


FIGURE 11 *Hazard ratios for the adjusted Duke data. Adapted with permission from Jones RH,* et al. *Long-term survival benefits of coronary artery bypass grafting and percutaneous transluminal angioplasty in patients with coronary artery disease.* J Thorac Cardiovasc Surg *1996;***111***:1013–25.*

TABLE 10 *RR (mortality in CABG versus PTCA) for high- and low-risk groups*

	RR of death	95% CI*	Significance
RCTs			
High (multi)	0.98	0.70–1.37	NS
Low (single)	0.82	0.38–1.78	NS
MedisGroups			
High	0.90		NS
Low	2.15		$p < 0.05$
Duke			
High (Group 9)	0.47		$p < 0.05$
Low (Group 1)	2.00		$p < 0.05$

* *95% CI cannot be calculated from the data available on the two non-randomised studies*

adjustment techniques do not seem to clarify matters even though there is strong evidence that differences in baseline characteristics have substantial prognostic importance. This may be a result of important risk factors being omitted from the model used by MedisGroups. Unfortunately this cannot be explored further as the MedisGroups algorithms are confidential for commercial reasons, though it is known that it involves abstracting very detailed information from case notes, including findings and results at several points during the patient's stay in hospital. It was not clear how the Duke database adjustments were made (though they must have been derived from a fitted model) and an overall adjusted RR was not given.

Whether one would expect the randomised results to agree with the non-randomised results can be questioned. The Medicare data were collected in 1985, while the RCTs were conducted largely in the 1990s. The techniques used changed considerably over this period; PTCA was only introduced in the UK in the late 1980s. Also, the non-randomised databases included many people for whom there must have been clear indications/contra-indications for one or other of the procedures and so they would not have been considered for entry into a randomised study. Therefore it is inappropriate to expect the RCT results to agree with the non-randomised findings.

In terms of the model described in chapter 1, the eligibility criteria (parameter **e**) varies considerably across the studies. Most trials randomised only patients with multi-vessel disease but one also included patients with single-vessel disease and two studies were confined to patients with single-vessel proximal left anterior descending artery disease. The studies gave limited data on patients who refused participation, but the percentage of eligible people who were actually randomised shows that there were considerable losses (range: 5–94%). We cannot say whether this would bias the estimates of treatment effect but in our model, participation (**p**), would vary widely from study to study.

The sub-group comparisons are, by necessity, crude. The definitions of 'high' and 'low' risk are different for each study. The Duke study focuses on angiographic findings to create the nine severity groups. The MedisGroups sub-groups were obtained from algorithms representing potential of organ failure, incorporating many clinical findings. Nevertheless, the two non-randomised studies give consistent results for the sub-groups, indicating that the least severe patients benefit more from PTCA.

In order to include the RCTs, a crude method of designating 'high' and 'low' risk was taken (the number of vessels involved, 'high' being multi-vessel and 'low' as single). The RRs of death (CABG vs. PTCA) for both single- and multi-vessel patients favoured CABG.

The BARI randomised trial also investigated the mortality in severity sub-groups. Patients were divided by stability of angina, left ventricular function, type of vessel disease, type C lesion and history of diabetes. For mortality at 5 years, the only significant difference occurred in the sub-group of patients with treated diabetes. Those assigned to PTCA had higher mortality than

those assigned to CABG (65.5% survival vs. 80.6% survival).

There are other limitations to this current comparison. In spite of the considerable size of some of the trials, mortality is rare and the number of events (deaths) reported is low. A more accurate and reliable comparison may have been achieved had there been sufficient data to compare mortality at 5 years follow-up.

Although the studies address the same clinical problem they differ in objectives, inclusion criteria, and follow-up, and these differences may contribute to the heterogeneity. It is not clear from the present review whether the difference between the estimates from randomised and non-randomised studies is attributable to:

- their differing methods of treatment allocation
- whether the patients they included are so vastly different in their initial profile risk that it would be unreasonable to expect the results to agree
- the use of retrospective data in one of the two non-random studies
- the crude methods for risk adjustment, or
- the different points on the PTCA learning curve at which the studies were conducted.

Summary

- Central measures of effect obtained from individual RCTs varied, both in magnitude and direction of relative benefit of each procedure, though the differences were not statistically significant.
- Notwithstanding problems of sample size, differences in results from individual RCTs and non-randomised studies would not be unexpected due to differences in the populations studied, and differences in the timings of studies.
- There is some evidence that the relative benefit of one procedure over another is related to the characteristics of the patient.

Chapter 8

Pharmaceutical interventions: calcium antagonists

Introduction

As in the previous example, it is likely that allocation bias will complicate the assessment of drug therapy when using a non-randomised design. There may be clear, prognostic reasons why one drug treatment is favoured over another, and thus it would be misleading to compare crude outcomes.

Calcium antagonists have been used in patients with cardiovascular diseases for about two decades. They are used to relieve angina pectoris or lower blood pressure. Nifedipine is a calcium antagonist that has been the subject of recent controversy as it has been suggested that it may be associated with an increased risk of mortality.[148–150] This drug lends itself as a suitable pharmaceutical example in which to study allocation bias because there is a recent large, high-quality non-randomised study that employs detailed risk-adjustment.[151] There is also a meta-analysis of 16 RCTs which combines results from over 8000 patients.[152] Mortality provides an unambiguous outcome measure in order to investigate risk adjustment in the non-randomised design, and there are well-described dosage regimens and degrees of baseline risk for a meaningful comparison.

Methods

As described in chapter 2, a MEDLINE search was conducted for 1990–97. In addition, an explosion technique was used whereby references from the best papers were also included. This led to the identification of a recent meta-analysis of 16 RCTs.[152] Interestingly, this paper was not found directly from the MEDLINE search strategy as it did not have 'randomised controlled trial' as a MEDLINE search heading (MeSH).

A cohort study of 11,575 patients was identified[151] in which the dosage of calcium antagonist was similar to that in nine of the RCTs in the meta-analysis.[153–161] The comparability of dosage is believed to be particularly important as the meta-analysis showed a significant association between high doses of the drug and increasing mortality. These nine low-dose RCTs will be compared with the results (adjusted and unadjusted) of the non-randomised study.

Results

Details of the RCTs included in this review are shown in appendix 8 and the non-randomised study in appendix 9. The nine randomised studies were published between 1984 and 1993 and were conducted in Europe, the Middle-East and South Africa. The length of follow-up ranged from 12 hours to 1 year. Six of the nine had follow-up shorter than 6 weeks, but the three largest trials, which included 85% of all the RCT patients, had follow-up of at least 6 months. The non-randomised study was conducted in 18 cardiology departments in Israel between 1990 and 1992. Patients were included in this study if they had undergone screening procedures for the Bezafibrate Infarction Prevention (BIP) Study and therefore had detailed medical records. Mortality data, after a mean follow-up period of 3.2 years (range: 2.0–4.6), for these patients were obtained by matching the patients' identification numbers with their life status in the Israeli Population Registry.

In the non-randomised study, the clinical characteristics of the treatment group and those not receiving calcium antagonists were similar, except that there were more patients with grades II to IV angina pectoris and hypertension in the calcium antagonist group, and more patients in the control group were receiving beta-blockers and digoxin (*Table 11*). A Cox's proportional hazards model was used to adjust for the baseline differences in three stages. The first adjustment considered age only; the second adjustment included age, gender and the prevalence of previous MI, angina pectoris, hypertension, New York Heart Association functional class, peripheral vascular disease, chronic obstructive pulmonary disease, diabetes and current smoking; and the third adjustment considered the additional effect of concomitant use of other medications.

The number and percentages of deaths that occurred in each set of RCTs (grouped according to dose of nifedipine) and all nine RCTs combined, and the non-randomised results (crude and risk-adjusted), are shown in *Table 12*, along with the risk ratios and confidence limits. *Figure 12* plots these risk ratios. *Table 13* shows the survival rates for controls and for patients receiving nifedipine, and these are then plotted in a L'Abbé plot (*Figure 13*).

Discussion

The unadjusted risk ratio from the non-randomised study showed a significantly increased mortality risk in the calcium antagonist group. The risk ratio is estimated to be 1.18. This is close to the estimate of 1.10 obtained from the combined randomised studies.

When adjustments are made for the differences in baseline risk in the non-randomised study, the

TABLE 11 *Baseline clinical characteristics in the non-randomised study of calcium antagonists*

	Calcium antagonist (n = 5843)	Control group (n = 5732)
Current angina		
None	30	49
I	32	30
II	34	19
III/IV	4	2
Drug therapy		
Beta-blockers	29	39
Digoxin	3	7
Diuretic drugs	16	16
Antiarrhythmic agents	5	7
Aspirin	56	58

TABLE 13 *Summary of survival in trials of calcium antagonists*

Study type	Calcium antagonist (mg/day)	Survival (%)	
		Calcium antagonist	Control
1 RCT	30	94.25	94.33
3 RCTs	40	93.21	93.74
1 RCT	50	91.08	91.3
4 RCTs	60	89.8	91.25
Total (9 RCTs)	**30–60**	**92.54**	**93.2**
Prospective cohort (unadjusted)	30–60	91.5	92.8

TABLE 12 *Summary of risk ratios for deaths in studies of calcium antagonists*

Study type	Study names	Calcium antagonist (mg/day)	No. deaths (%)		Risk ratio	95% CI	Adjustment
			Calcium antagonist	Control			
1 RCT	SPRINT 1	30	65 (5.75)	65 (5.67)	1.01	0.73–1.42	
3 RCTs	Branagan, Gordon, TRENT	40	157 (6.79)	146 (6.26)	1.09	0.87–1.35	
1 RCT	Sirnes	50	10 (8.92)	10 (8.70)	1.03	0.44–2.37	
4 RCTs	Walker, Erbel, SPRINT II HINT	60	123 (10.2)	105 (8.75)	1.18	0.93–1.50	
Total (9 RCTs)		**30–60**	**355 (7.46)**	**326 (6.80)**	**1.10**	**0.95–1.27**	
Prospective cohort	Braun	30–60	495 (8.5)	410 (7.2)	1.18 1.08 0.97 0.94	1.04–1.34 0.95–1.24 0.84–1.11 0.82–1.08	Unadjusted Age-adjusted Age, sex, history Age, sex, history, other medication

HINT = Holland Interuniversity Nifedipine/metoprolol Trial

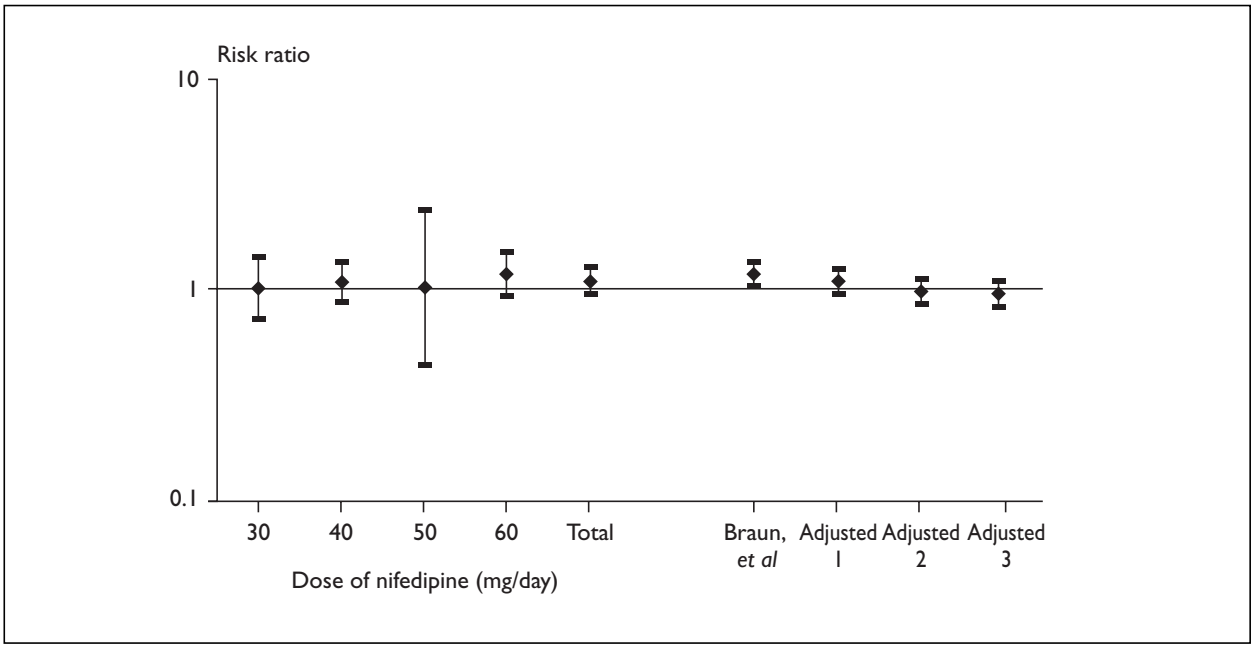


FIGURE 12 *Summary of risk ratios for death in studies of calcium antagonists*

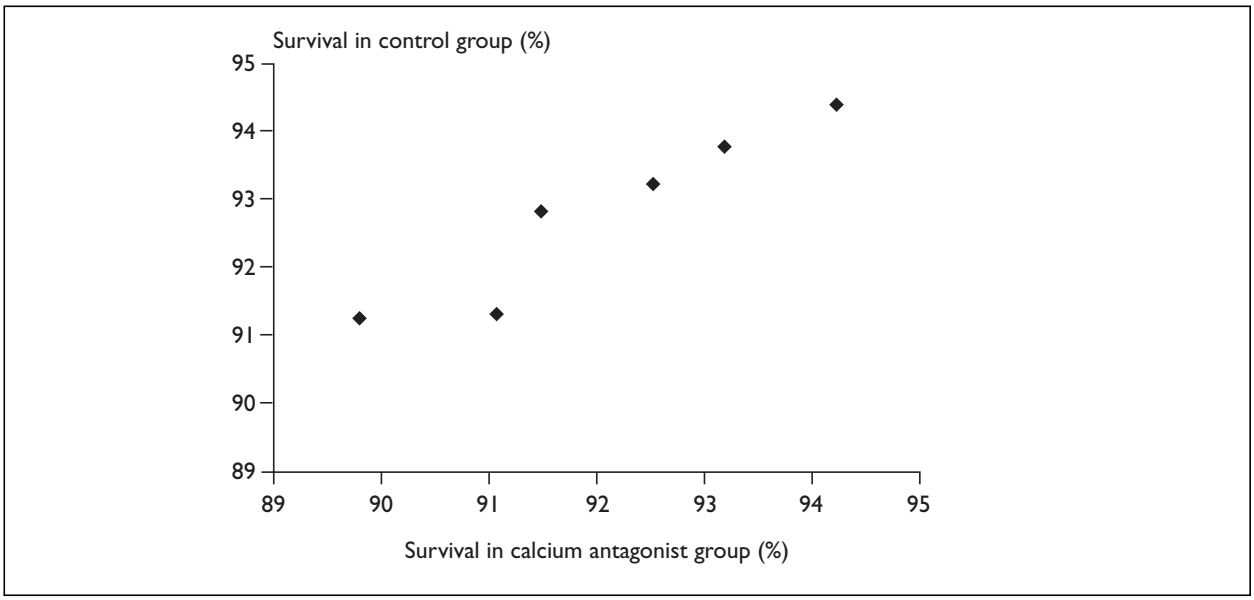


FIGURE 13 *Summary of survival in studies of calcium antagonists*

mortality risk associated with calcium antagonists is reduced and after the fullest adjustment the risk ratio fell to 0.94. However, this is still close to the estimate of 1.10 obtained in the RCTs, and both have confidence intervals including 1.

Whether we would expect the two estimates to be exactly the same can be questioned given that the follow-up lengths were so different and the randomised evidence has not been adjusted. Also, if the mix of dosages differed (between 30 and 60) then that could result in a difference. The longest follow-up in the randomised studies was 1 year, while the non-randomised study had a maximum follow-up of 4.5 years (mean 3.2 years). Mortality associated with calcium antagonists is likely to be associated with duration of follow-up. Given this, it is perhaps surprising how similar the survival rates are in *Table 13*.

In terms of the model described in chapter 1, eligibility (**e**), can be shown to vary considerably between RCTs and the non-randomised study, and even more markedly between the individual RCTs, as shown in the eligibility criteria listed in appendix 8. Six out of the nine RCTs gave some information on the participation rate (**p**) among those eligible. This

appeared to range from approximately 15% to 100%. What impact these potential biases have on the estimates of treatment effect, if any, is unclear because no information is given on the survival of those excluded or of the non-participants.

In this particular example, the non-randomised study was able to make use of data that were already collected and could follow a large number of people for up to 4.5 years to quantify side-effects. This has clear benefits over the small, short-term RCTs.

This example does not, unfortunately, provide adequate evidence about the ability to adjust for baseline differences in studies of pharmaceutical interventions. The unadjusted risk ratio, if anything, was slightly closer to the estimate from the RCT.

Summary

- The results obtained from RCTs and the unadjusted non-randomised study did not differ significantly.
- Although adjustment caused the results of the non-randomised study to diverge, again the change was not significant.

Chapter 9

Organisational interventions: stroke units

Introduction

Stroke units offer one of the few examples of organisational interventions for which there are both RCTs and non-randomised studies, with one of the latter incorporating differing levels of adjustment for case mix.

Methods

As noted in chapter 2, this section was originally undertaken by updating an earlier systematic review available in the Cochrane Library. After it was completed, a second systematic review was published by the same group.[162] This included new data from some of the studies included in the earlier review as well as data from as yet unpublished studies. The data in the second review were used as the basis of this analysis, supplemented with data on non-randomised studies identified in the searches specified earlier. The review contains data from RCTs comparing various combinations of dedicated stroke units, mixed assessment/rehabilitation wards, and general medical wards. To reduce heterogeneity, the present analysis is limited to those comparing dedicated stroke units and general medical wards.

Results

Eleven RCTs were included (appendix 10). For the trials for which information was available, the percentage of eligible patients included ranges from 34% to 99%, though the varying definitions used for both included and eligible make any meaningful interpretation of these figures impossible. A possible indirect measure of comparability of inclusions is the wide variation in the probability of survival in controls, from 54% to 94% (*Table 14*).

Figure 14 shows the odds ratios for survival at 1 year in stroke units compared with standard treatment for the 11 RCTs and two of the three non-randomised studies. The combined odds ratio for all the trials is 0.75 (0.61–0.75).

TABLE 14 *Evidence of inclusiveness of trials of stroke units*

Trial*	Probability of survival in controls	% of eligibles included	Odds ratio for death at final review: stroke unit vs. general wards (95% CI)
Goteborg-Ostra	0.94	Not given	1.27 (0.587–2.76)
Orpington 1993	0.94	67	0.90 (0.173–4.69)
Nottingham	0.87	Not given	1.10 (0.459–2.63)
Perth	0.80	Not given	0.64 (0.160–2.55)
Kuopia	0.78	Not given	0.67 (0.237–1.87)
Montreal	0.68	61	0.68 (0.318–1.47)
Trondheim	0.67	55	0.67 (0.371–1.21)
Edinburgh	0.65	99	0.82 (0.513–1.32)
Dover	0.61	34	0.82 (0.452–1.49)
Umea	0.59	Not given	0.92 (0.570–1.50)
Orpington 1995	0.54	Not given	0.28 (0.10–0.81)

* *Full references to studies are given in reference 162*

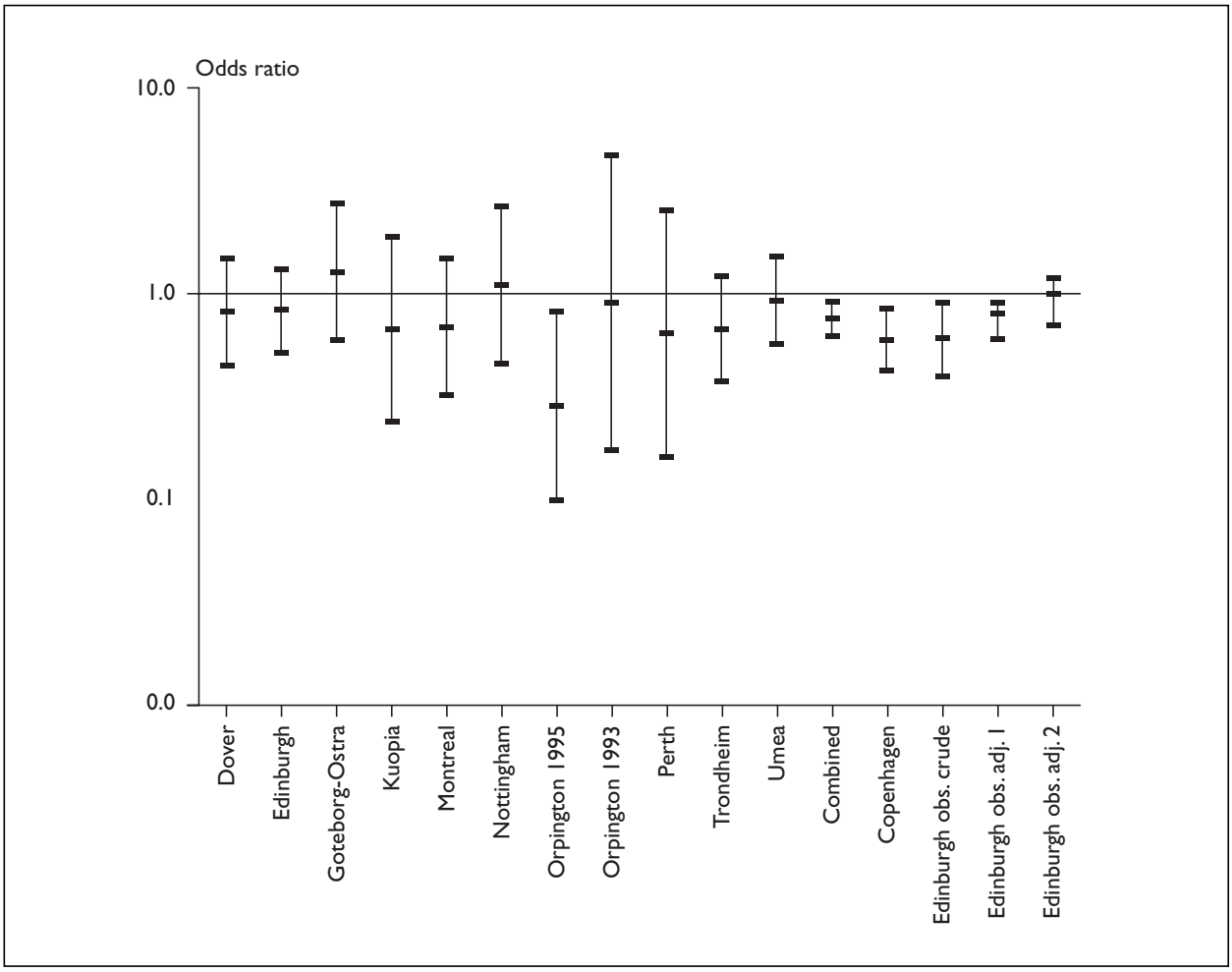


FIGURE 14 *Odds ratio for survival at 1 year: stroke units versus standard care*

Only two of the non-randomised studies were included as the third, undertaken in London and comparing two hospitals, only followed patients for 6 months and, thus, is not directly comparable with the others.[163] It reported no significant difference in survival between patients admitted to the hospital with the stroke unit and the one offering standard care. The authors argued that the failure to show an effect was most likely due to differences in case mix between the two groups, though despite having very detailed functional data, they were unable to identify any differences other than that those admitted to the stroke unit were more likely to be owner occupiers and were on average 3 years older than those admitted elsewhere.

Of the two included, one was undertaken in Copenhagen and compared patients admitted to two neighbouring hospitals, one with and one without a stroke unit.[164] The second, undertaken in Edinburgh, compared admissions before and after introduction of a stroke unit in one hospital.[165] It included detailed information on a wide range of prognostic factors, permitting adjustment for severity.

The Copenhagen study produced an odds ratio of 0.59. With the exception of the 1995 Orpington study, which included only 73 patients, this effect was greater than in any of the other RCTs. This value is also very close to that found before adjustment for differences in severity in the Edinburgh study.

There were no significant differences in baseline characteristics in the two groups in the Copenhagen study in terms of age, sex, marital status, nursing home residence and a range of cardiovascular and neurological parameters. The only difference was that those admitted to the stroke unit were more likely to have a history of hypertension.

The authors of the Edinburgh study adjusted first for age and sex (adj. 1 in *Figure 14*) and then for a wide range of very detailed prognostic variables including ability to lift arms, presence of diabetes, employment, eye opening, motor score, verbal

score, pre-stroke independence and blood pressure (adj. 2 in *Figure 14*). Progressive adjustment reduces the apparent benefit from the stroke unit so that, after the more detailed adjustment, it disappears. Adjustment was undertaken using a series of models, some of which had been derived from other data sets. Each gave similar results. These changes with risk adjustment suggest that the patients admitted before and after the introduction of the stroke unit differed considerably. The authors identify as a possible explanation the closure of the hospital's accident and emergency unit during the study, which led to patients with a different spectrum of severity being admitted.

Discussion

Authors of several other studies on stroke patients, while not directly assessing outcome of stroke units, have also argued that differences in case mix create important problems for those undertaking non-randomised comparisons.[166] One example is the study of management by neurologists compared with care by general physicians, which reported a lower mortality rate among those treated by neurologists but also a lower rate of co-morbidities in this group.[167]

Although formal testing reveals insignificant evidence of heterogeneity among the RCTs, this test has only limited power. Examination of the percentage of possible patients excluded and the mortality in control groups suggests that the RCTs are comparing quite different populations. While noting the many limitations to such an approach, comparison of odds ratios with mortality in control groups does, at least, suggest the possibility of a relationship with those RCTs in which the population has a higher probability of mortality, and thus presumably more severe strokes, reporting a smaller advantage for stroke units (*Figure 15*). If true, this could explain the absence of an effect seen in the non-randomised London study and the adjusted results of the Edinburgh study, assuming that these studies were not subject to as many exclusions as in some of the RCTs. Two of the RCTs provide data that enable sub-group analysis that would, with caution, permit this hypothesis to be explored further. However, although both lack statistical power, neither offer much support for it. The Orpington trial showed a slightly greater, though non-significant effect in more severe and in older patients, and the Umea trial produced greater, and again non-significant, effects among older patients and those with a past history of cardiac disease or a previous stroke.

In summary, on the basis of the studies examined, RCTs are generally consistent with a moderate beneficial effect on mortality but unadjusted results from two non-randomised studies suggest

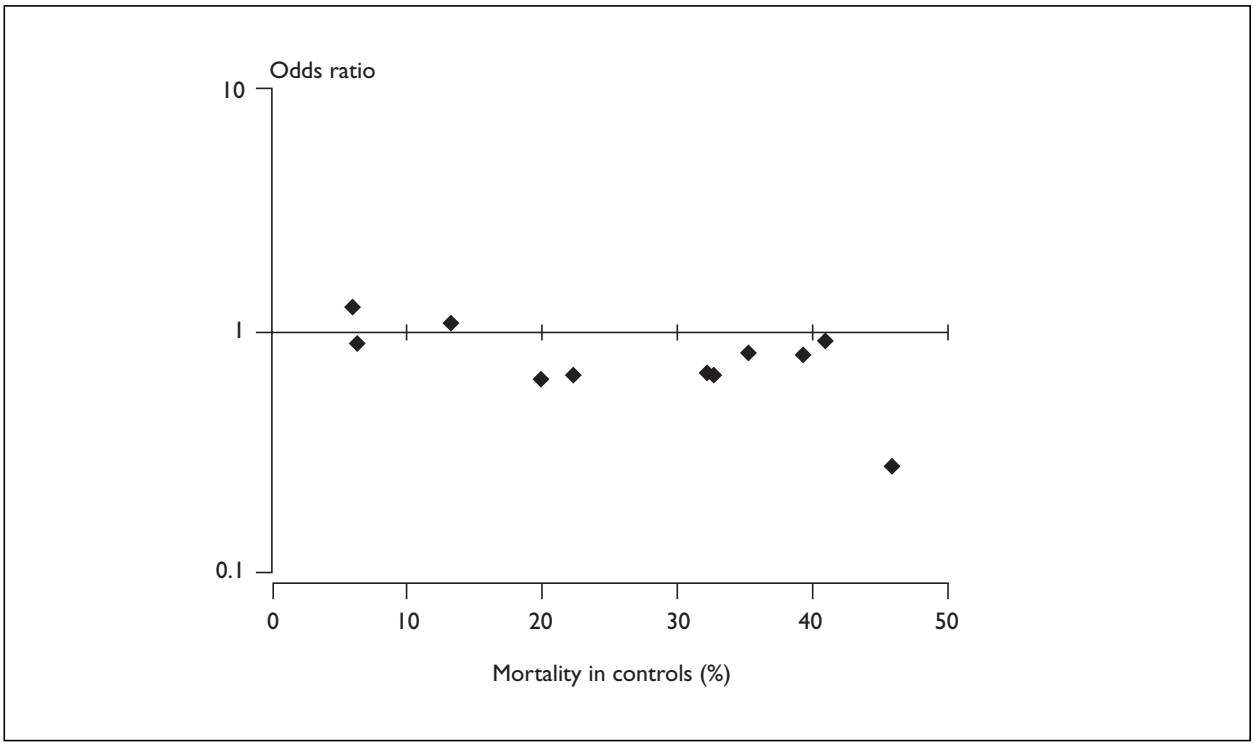


FIGURE 15 *Stroke units: association between mortality in controls and odds ratios for death*

a greater effect than that seen in trials, though when adjusted for case mix, the advantage associated with stroke units disappears. A third non-randomised study finds no difference between the two forms of care. There are several possible methodological explanations for the differences.

An organisational intervention, such as a stroke unit, may be especially susceptible to differences in the subjects included given the inevitable conflict between an unpredictable level of immediate demand and a service with a fixed capacity. This can be affected markedly by extraneous circumstances, such as the closure of the emergency department during the London study. Intuitively, such factors seem likely to give rise to short- and long-term fluctuations in the threshold for admission/inclusion, which in relation to the model will lead to differences over time in both **e** and **i**. There is some evidence from these studies not only that such effects exist but also, more tentatively, that they have an impact on the measured effect size. A further issue in these studies is that many patients are likely to have difficulty giving consent and it is not clear whether equally strenuous efforts were made to reduce the effect of differences in patient participation (**p**). However further exploration of this issue is constrained by the absence in most randomised studies of information on the percentage of eligible patients recruited. The non-randomised studies effectively adopt an intention-to-treat approach by comparing hospitals rather than the actual treatment received by patients. This will overcome some potential problems but, as a variable proportion of stroke patients are managed at home, it does not eliminate the effect of differences in eligibility (**e**) between the two hospitals, based on availability of resources.

Another potential reason for the observed effects after adjustment is a Hawthorne effect, with those involved in RCTs providing better care than would otherwise be the case, and thus multiplying the effect of particular components of care. In addition, the intervention has been treated as a 'black box' and it has not been possible to explore whether the term 'stroke unit' has changed its meaning in different places and over time.

In this case there is no evidence to support the contention that it is possible to close the gap between non-randomised studies and RCTs by means of adjustment using even quite detailed information on confounders. Furthermore, it is possible that results such as those in the Edinburgh study are due to over-adjustment for confounders. Perhaps all that can be said is that stroke units appear to be effective in reducing mortality for some patients but not all and further work is required, perhaps combining individual patient data from the various studies, to identify who will benefit most. Differences in results obtained in individual RCTs and in non-randomised studies are unsurprising in view of the apparent importance of patient characteristics.

Summary

- The results of RCTs of stroke units vary considerably, though, due to small sample sizes, these differences do not reach statistical significance.
- These differences may be due to chance, though there is also evidence of differences in populations included.
- Results of unadjusted non-randomised studies are consistent with those of RCTs.
- In one non-randomised study, adjustment caused divergence from the results of pooled RCTs but this was not statistically significant.

Chapter 10

Preventive interventions: malaria vaccines

Introduction

As shown in chapter 5, characteristics of participants in prevention RCTs are systematically different from those who decline the invitation and from the general population. Participants are, in general, healthier and wealthier. In RCTs patient characteristics should be distributed evenly across the trial arms so, while the participants in a preventive trial may be atypical of the general population, the effectiveness of the intervention should be measured in an unbiased way for that sub-group of the population. However, non-randomised studies of preventive interventions could face considerable selection bias especially when those who come forward for a preventive intervention are compared with the remainder of a population, as it is very likely that they will exhibit important baseline risk differences.

Methods

Immunisation was chosen as an example of a preventive intervention to examine some of the issues in interpreting results from the two study designs. The Cochrane database provided a meta-analysis of RCTs of human malaria vaccines.[168] As this is a recent comprehensive systematic review and included a MEDLINE search, no further RCTs were sought. Non-randomised studies were identified using the search strategy described in chapter 2.

By finding the references from these papers and those given in the Cochrane systematic review, a large non-randomised study was identified.[169] The study reported on the effectiveness of the SPf66 vaccination against malaria caused by *Plasmodium falciparum* and *Plasmodium vivax*. Only the *P. falciparum* results will be considered in this review.

The Cochrane review detailed five trials of SPf66 vaccine and its effectiveness against malaria due to *P. falciparum*. [170–174]

Many outcome measures were documented in the meta-analysis, but for the purposes of comparison with the non-randomised study, we focus here on the only outcome measure used in that paper, the incidence of malaria episodes during the follow-up period.

Results

Details of the five RCTs (published 1993–96) included in this review are shown in appendix 11 and the non-randomised study (1994) in appendix 12. Three of the five RCTs were conducted in South America and two in Africa, with follow-up ranging from 3.5 to 22 months. Two trials restricted their participants to children (one to infants aged 6–11 months and the other to 1–5-year olds), and the other three included everyone over a year old. All five trials were placebo controlled and involved three separate doses of vaccine (given in slightly different quantities and intervals).

The non-randomised study was conducted in 13 small villages in South Venezuela. All persons aged over 11 years were invited for screening. Those who attended and were eligible were given three separate doses of the vaccine and were followed-up 1 year later. The outcome in this treated group was compared with the remainder of the population.

The number of malaria episodes, the percentage experiencing an episode, and the odds ratio comparing vaccines to controls, are shown in *Table 15*. (For the non-randomised study, figures are unadjusted for baseline differences in the groups.) *Figure 16* plots these odds ratios. The percentage of people remaining malaria free in the vaccinated and control groups are plotted in *Figure 17*.

The SPf66 vaccine had the effect of reducing the number of malaria episodes in all the studies. The combined RCTs give an odds ratio of 0.62 (95% CI: 0.53–0.71). The unadjusted Noya study gives an odds ratio of 0.78 (95% CI: 0.52–1.17).

To assess 'allocation' bias, baseline characteristics of vaccinated and non-vaccinated subjects in the non-randomised study were compared. No material difference was observed in age, sex, or occupation. However, those receiving vaccination were over-represented in localities at higher risk of transmission. In an attempt to allow for this dissimilar malaria risk at baseline, rate ratios were calculated for the incidence in the 12 months subsequent to the third dose of vaccination, in relation to that observed during an equivalent calendar period

TABLE 15 *Summary of outcomes of evaluations of malaria vaccines*

Reference	Study type	Malaria episodes (%)		Odds ratio	95% CI
		Vaccine	Control		
Alonso, 1994[173]	RCT	73 (26.6)	102 (32.7)	0.75	0.53–1.07
D'Alessandro, 1995[174]	RCT	199 (63.0)	148 (64.1)	0.95	0.67–1.36
Sempertegui, 1994[172]	RCT	4 (1.7)	12 (5.2)	0.36	0.13–0.97
Valero, 1993[170]	RCT	168 (22.8)	297 (26.7)	0.52	0.42–0.64
Valero, 1996[171]	RCT	53 (8.4)	88 (14.1)	0.56	0.40–0.80
Cochrane meta-analysis	5 RCTs	497 (22.7)	647 (29.3)	0.62	0.53–0.71
Noya, 1994 (*P. falciparum*)[169]	Prospective cohort	46 (5.4)	56 (6.8)	0.78	0.52–1.17

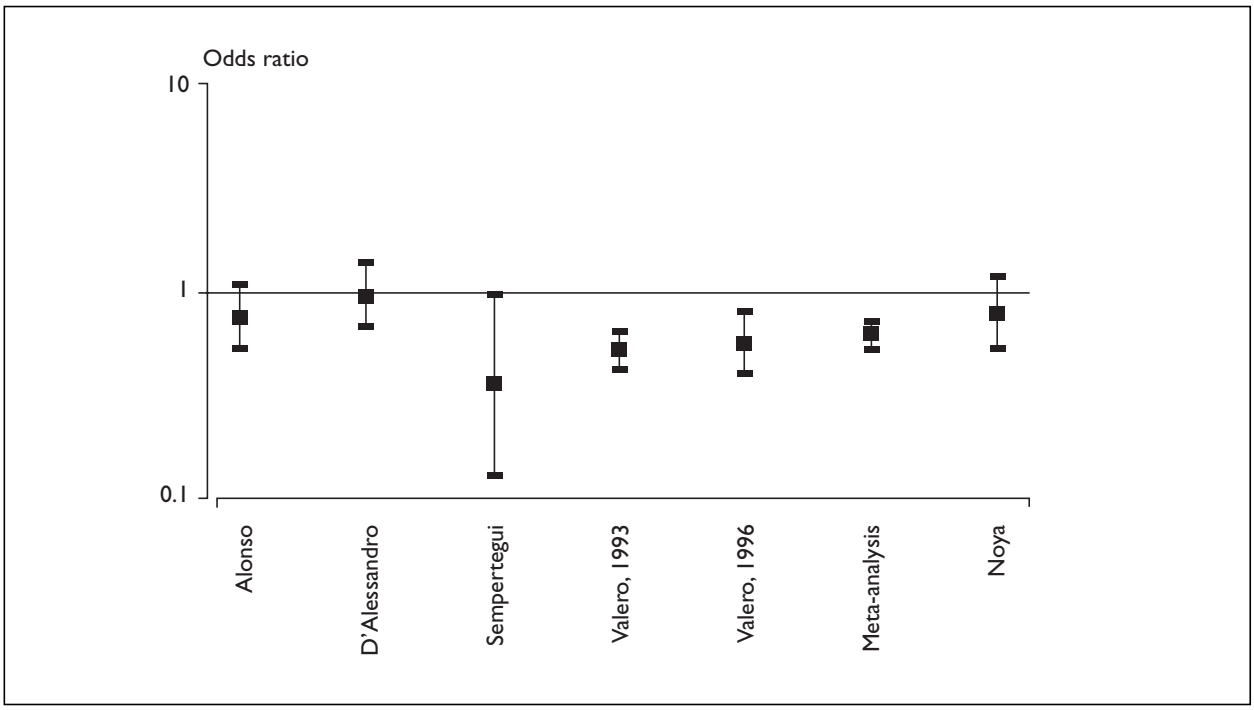


FIGURE 16 *Summary of outcomes of evaluations of malaria vaccines*

just before vaccination, and for comparable periods for the controls. The after- to before-vaccination incidence ratio of each group was used to derive an adjusted odds ratio, which was reduced to 0.45 (95% CI: 0.25–0.79).

Discussion

The study by Noya and co-workers illustrated how a non-randomised study design can have considerable selection problems when assessing preventive interventions. People who choose to accept vaccination are likely to present with different characteristics and risk of infection to those who decline vaccination. It could thus be misleading to compare the crude malaria incidence rates in the control group with those in the vaccinated group. Noya and co-workers addressed this selection problem by looking at the reduction in incidence of malaria in the two groups compared with their experience in the preceding 12 months: whereas the unadjusted odds ratio in the non-randomised study was slightly higher than that obtained in the Cochrane meta-analysis, after this attempt to control for baseline risk difference, the odds ratio fell to below the summary estimate from the RCTs.

None of these studies was flawless. The RCTs were not analysed on an intention-to-treat basis. Only

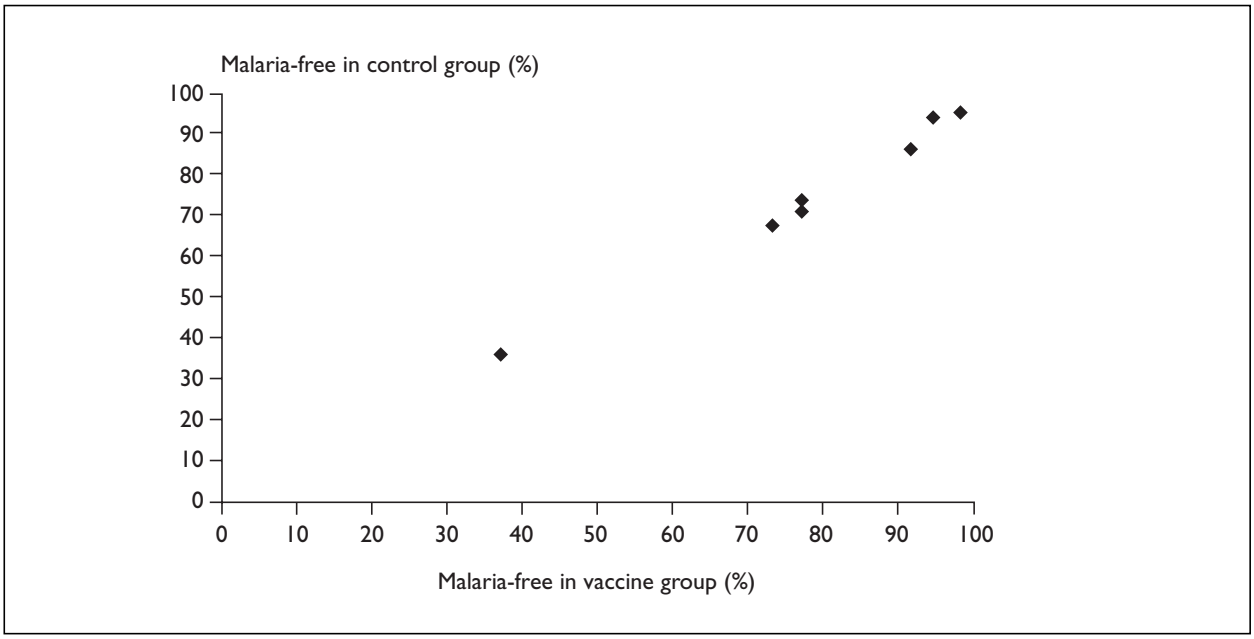


FIGURE 17 *Relationship between percentage of people malaria free in vaccine and control groups*

those who received all three doses of the vaccine (or placebo) were included in analyses. Compliance ranged from 60% to 94%, with some differences between treatment arms. It is likely that those who did not receive all three doses had different characteristics and risks of malaria. Furthermore, the differing endemicity of malaria in the areas where the studies were undertaken, the different lengths of follow-up and the different age groups targeted, bring into question whether a summary estimate of the RCT data has meaning.

While such variability in trial site and execution leaves uncertainties, the striking linear correlation between endemicity and relative vaccine effect seen in the RCTs suggests that here methodological foibles are overwhelmed by powerful biological forces at work. The vaccine appears to work best at lower disease burdens. Viewed in this light, the adjusted RR from the non-randomised data falls exactly as would be predicted on the plot of endemicity versus RR.

Thus, for this preventive intervention, after simple adjustment the non-randomised data are wholly concordant with the RCT findings; and most obviously so when heterogeneity is explored, rather than suppressed in a summary estimate. If the biological explanation proposed for the variation in effect is correct, it has clear implications for vaccine policy, and the non-randomised result has added to the coherence of the picture.

Unfortunately, it is improbable that this happy circumstance will hold for all, or even most, preventive interventions. In the vaccine evaluation, powerful and specific biological forces both limit the scope for risk factor imbalance, and make it plain when it does occur, and also obscure minor-to-moderate study flaws. This will not be the case for preventive interventions involving more personal, lifestyle-choice modifications (e.g. modifications of diet, activity, weight, and commitment to long-term hormone replacement therapy), where concomitant variations in attitudes and other behaviours will be many, often subtle, and of uncertain influence on quite modest effects on health risk. In such situations, absence of RCT data greatly limits the potential for making informed policy.

Summary

In this example, the treatment effect estimate from the non-randomised study is concordant with the estimates obtained from the randomised trials. However, this finding should not be generalised to all preventive interventions.

Chapter 11

Internal validity: lessons from comparisons of non-randomised studies and RCTs

The papers reviewed in chapter 3 comparing non-randomised studies with RCTs and the four case studies (chapters 7–10) provide insights into the issue of internal validity.

Do RCTs and non-randomised studies produce systematically different results?

The first hypothesis being tested is that, when evaluating the same intervention, non-randomised studies produce greater treatment effects for interventions versus placebos, or new versus old interventions. This is based on the contention that, in the absence of rigorous randomisation, a bias, conscious or otherwise, will creep into the process by which a particular patient receives one or other treatment and that this will have the effect of producing an imbalance so that those receiving the new treatment will have greater potential to benefit.

There are four main findings from the examination of papers in chapter 3, which compared the two methods.

- Within the limits of statistical significance, the results obtained by the two approaches are frequently similar.
- Any differences were most often, but not always, of similar magnitude of estimated treatment effect rather than in the same direction, and only rarely did the two approaches favour different interventions.
- Neither method consistently favoured intervention over placebo or new treatment over old.
- The differences in results between RCTs and non-randomised studies were frequently smaller than those between RCTs or between non-randomised studies.

These findings are supported by the four examples studied in chapters 7–10. In the case of CABG and PTCA, the RCTs tended to favour CABG and the unadjusted results of non-randomised studies favoured PTCA, though the wide confidence intervals make any firm conclusion impossible. The different methods used to examine calcium antagonists produced broadly similar results, as did studies of stroke units. The non-randomised study of malaria vaccines found a somewhat greater benefit than did the RCTs but the difference was not statistically significant.

There are several reasons for the failure to support the argument that a systematic bias arises from the use of one method rather than the other.

- As particular interventions tend to be studied using either RCTs or non-randomised studies, there are few examples of comparisons to study.
- Of those that do exist, confidence intervals are often wide, so it is not possible to say whether any differences are real or due to chance.
- Typically there are large differences in how the research is conducted using the two approaches, such as the population included, the setting, the practitioners, and how the outcomes are assessed. This review has produced considerable evidence of how much these factors may vary. One measure is the scale of the differences in the frequency of particular outcomes between control groups from different studies. Further evidence can be inferred from the results of the analysis of CABG and PTCA by Jones and co-workers.[141] This showed that for some patients the former is safer than the latter and for others, the converse, and that this is a function of initial severity. Consequently, a study that only included subjects from one part of this spectrum would inevitably produce different results, independently of the method used to allocate treatment (*Figure 18*).

A similar situation has been described by Rothwell, who re-analysed data from the European Carotid Surgery Trialists Collaborative Group[175] and showed how, when subjects are stratified according to risk at entry using an independently generated prognostic index, surgery confers a net survival benefit among high-risk patients but a net loss of life among low-risk patients.[176] He also showed how the benefit from aspirin used to prevent stroke is greater in those with the greatest initial risk. By inference, trials that had been less inclusive would have suggested greater RR reductions for

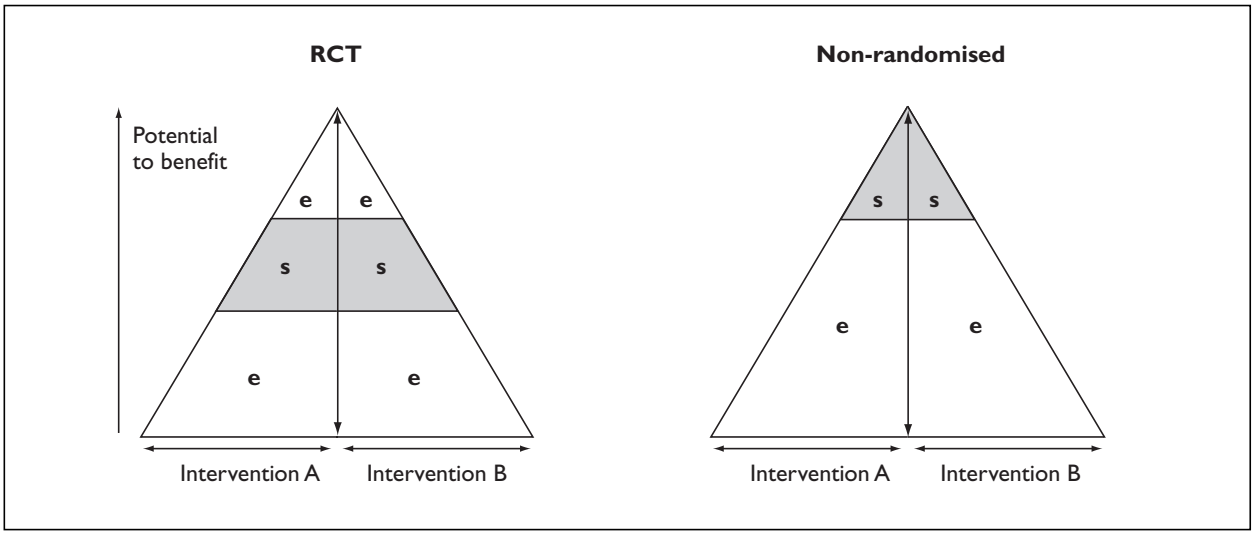


FIGURE 18 *Schematic representation of potential effect of studies on different populations (e = ineligible, s = subjects)*

these two interventions and even more inclusive studies might have found smaller overall benefits. (The contribution of regression to the mean to these effects is unclear.)

Further evidence of the importance of differences within samples comes from the paper by Horwitz and co-workers that examined beta-blockers.[32] Having found quite different results in RCTs and non-randomised studies they were able to produce reasonable agreement by excluding from the non-randomised group those not meeting the eligibility criteria for the RCT (*Figure 19*).

This problem is not confined to comparisons of the two approaches, as the finding that the survival rates in control groups of RCTs evaluating stroke units can vary almost two-fold suggests either that they are studying very different populations or that there are large differences in what is happening to those in the control groups.

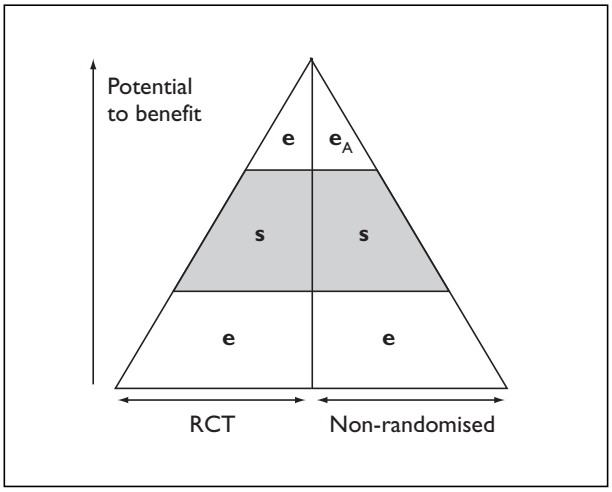


FIGURE 19 *Schematic representation of comparison by Horwitz* et al[32] *(e = ineligible, s = subjects, e_A = excluded from comparison)*

Ioannidis and Lau have provided further evidence from a range of studies of the problems arising when seeking to apply the results of trials undertaken on heterogeneous populations to individuals, but they also highlighted another problem:[177] if a population contains many individuals with little ability to benefit from the intervention in question and a few with a high ability to benefit, even though the overall sample size is large, the latter group may be very small and randomisation may not ensure that they are evenly distributed between the two arms. In this case, the disproportionate effect exerted may lead to serious bias.

It is recognised that, as so few comparisons have been published, there is considerable scope for publication bias in this field as in others, as authors seek to support a particular position. It is therefore inappropriate to place too much emphasis on the relative number of comparisons yielding particular findings. Instead, all that can be said is that the initial hypothesis is disproved and non-randomised studies do not always produce greater effect sizes than RCTs. However, the detailed study by Stukenborg[47] provided considerable evidence that if any effect of assignment *per se* exists, it is small.

Can adjustments be made for baseline differences in groups?

It is well established that non-random allocation of subjects can produce comparison groups that differ in terms of prognosis. This is seen in many of the studies cited in this review. This review has

examined whether adjustment for baseline differences between comparison groups in non-randomised studies will cause the results of such studies to converge with those of RCTs. It is recognised that baseline differences between arms of RCTs may also exist, either due to chance or, as is increasingly being realised, through imperfections in the randomisation process.[178] For the present purposes, however, we have accepted the commonly held view that randomisation results in completing comparable groups.

The papers cited in chapter 3 provided mixed answers. Antman was unable to reconcile the results of the different approaches to evaluating treatment for sarcoma.[40] Horwitz and co-workers were more successful, but only when they excluded those subjects that did not meet the eligibility criteria for the RCT.[32]

Our four case studies also produced a mixed picture. In the studies of CABG and PTCA, adjustment increased rather than decreased the differences between the results from the two methods of allocation. With calcium antagonists, stroke units and malaria vaccines the differences between the two methods were not statistically significant, though for the first two, there was a suggestion of a non-significant trend with progressive adjustment that tended first to close the gap but then to open it, but in the opposite direction.

The ability of the evidence reviewed to determine the potential for adjustment is extremely limited. Sample sizes are rarely adequate to detect differences with confidence even though trends are often apparent, and adjustment may be imperfect, such as the exclusion of left main stem disease in the model employed by Hartz and co-workers.[145]

Some of the papers examined in chapter 3 show how results can be affected by increasing the number of variables adjusted for and this was also seen in the example of stroke units, where adjustment eliminated a previously significant benefit. One can never be certain that adjustment has been sufficient or whether the inclusion of other, unmeasured variables might change the results further. This is analogous to comparisons of the performance of hospitals or practitioners in which rankings change as increasing numbers of prognostic variables are taken into account.[179]

As a minimum, adjustment methods should be developed in a rigorous fashion. This should be based primarily on an understanding of the biological and clinical issues involved but, in addition, a range of statistical issues must be considered. The potential statistical problems that may arise as well as the measures that should be taken to minimise them have been described by Concato and co-workers (*Table 16*).[180]

Another approach is to identify situations in which the threats to internal validity are minimised. Miettinen described 'confounding by indication' in which, those whom it is thought are most likely to benefit from a particular treatment are most

TABLE 16 *Problems with multivariate models of risk*

Problem	**Potential remedy**
Under- or over-fitting	Ensure > 10 outcome events per independent variable
Non-conformity with linear gradient	Check for linearity throughout range and stratify if necessary
Violation of proportional hazards	Check for proportionality throughout range and stratify or use time-dependent variables if necessary
Interactions	Include interaction terms, but be cautious about over-fitting
Variation in coding of variables	Specify how variables are coded and use them consistently
Selection of variables	Specify how variables are selected
Co-linearity	Select only one of several clinically similar variables or select using principal components analysis
Influential outliers	Be aware
Inadequate validation	Use independent sample/split sample/bootstrap sample

Source: Adapted from Concato et al[180]

likely to receive it.[181] He illustrated this with an example that appears to suggest that treatment with warfarin increases the risk of thrombosis 27-fold. Even after detailed adjustment, it continues to be associated with a nearly four-fold increased risk. Apparently, the ability of physicians to select patients at high risk of thrombosis surpasses the ability of definable risk factors to identify them. This, he contends, provides a strong argument against non-randomised studies of intended effects, as the combination of often subtle clues that practitioners use to identify those whom they expect to benefit most from an intervention often cannot be captured by the researcher. In contrast, however, seeking as yet unknown adverse effects of treatment is not a problem, as the physician deciding who will receive what treatment cannot possibly know the unknown (i.e. which patient factors make side-effects more likely).

Dealing with preference

It is necessary to understand the nature of the phenomenon of preference better and, in particular, where preferences are important and where they are not, and what the implications of this are on our understanding of the results of RCTs. For example, it seems plausible that interventions designed to affect behaviour might be much more susceptible to important patient preference effects than those with less avoidable outcomes such as death.

This could be examined by mounting RCTs from which the size of preference effects can be reliably measured. But this is difficult. Rucker postulated a two-stage design where randomisation between two groups (*Figure 20*) is described.[12] The two arms compare the outcome among no choices with patient preferences where they exist. However, even if it were possible for people with strong preferences to be recruited into such a trial (e.g. Torgerson *et al*,[14] demonstrated that it is possible, in one instance at least), the estimation of any preference effect would remain complex.

The problem is one of interpretation because, in this case, subtracting the means from the two randomised groups provides an estimate of a complex combined algebraic function of the main physiological effects and any preference effect (x and y, respectively). As has been emphasised, measuring the existence of main physiological effects is, in the known absence of preference effects, relatively straightforward, but estimating interactions such as y is difficult. This is true for the simple assumptions made here, but more complex assumptions quickly render the solution intractable. More complicated models could be imagined in which the effects of preference were multiplicative, graded, different for each treatment and/or asymmetric, but as these effects are poorly understood, the simplest possible theoretical effects are presented here.

Expected response rates:

Preference group	$P + y\,(\alpha + \beta) + x\,(\alpha + \gamma/2)$	R_P
Random: Group A	$P + y\,(\alpha - \beta)$	R_A
Random: Group B	$P + y\,(\alpha - \beta) + x$	R_B
Random: whole group	$P + x/2$	R_R

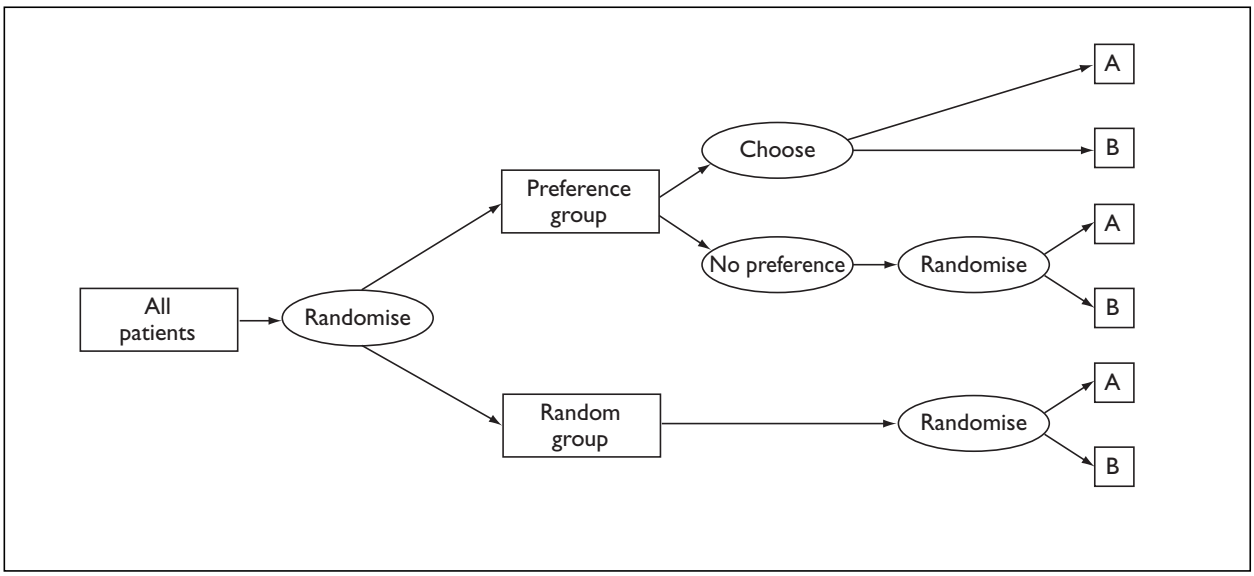


FIGURE 20 *Possible means of incorporating preference (Source: Rucker[12])*

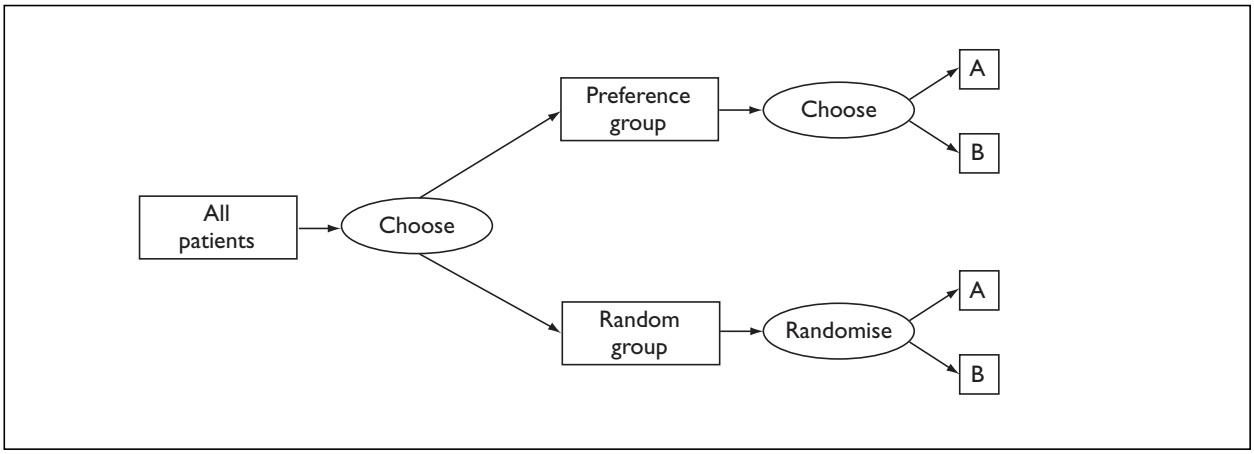


FIGURE 21 *Possible means of incorporating preference (Source: Brewin and Bradley[182])*

The physiological effect (x) can be estimated from the randomised arm ($R_B - R_A$), but with an unknown preference component based on imprecise estimates of the proportions α and β from the preference arm. A preference effect might thus be estimable from comparison of the results from the two arms, but the error structures are formidable. The algebra, which is highly laborious even on the simple model above confirms the difficulty of estimating these interactions reliably:

$$y = \frac{(R_P - R_R) + ((\beta + \gamma/2) - 1/2)(R_B - R_A)}{2\beta\,(1 - \beta + \alpha) + \gamma\,(\alpha - \beta)}$$

It may be easier to put into practice the trial design described by Brewin and Bradley[182] (*Figure 21*) but this will produce results for which a preference effect cannot be disentangled from the possible confounding arising from differences between patients with strong treatment preferences. Alternative methods[14] involve recording preferences before randomisation as a covariate and estimating a preference effect by including a y-type term in a regression equation estimating the main (x-type) effects. Unless the trial is enormous, such estimates will in general be very imprecise and probably too imprecise to distinguish them reliably from the main physiological effects.

To be able to interpret properly results from unblind RCTs it is essential to know that estimates of treatment effects are free from important preference components ($2y\,[\beta - \alpha]$), particularly as preferences can change while physiological effects may be less volatile. It is crucial, ultimately, to understand why a treatment works. Preference trials[183] can answer contemporary pragmatic questions about which treatment works best, incorporating both individual choice and their preference effects, but double-blind trials are more likely to control for any psychological effects and hence detect the physiological effects, only. Physiological effects cannot be reliably observed from unblind trials, since certainty that there are no preference effects is never justifiable experimentally.

Summary

- Within the limits of statistical significance, the results obtained by the two approaches are frequently similar and any differences are most often, but not always, of similar magnitude of the estimated treatment effect rather than in the same direction.
- The differences in results between RCTs and non-randomised studies are frequently smaller than those between RCTs or between non-randomised studies.
- Adjustment for baseline differences between arms in non-randomised studies should be explicit and rigorous.
- The risk of confounding may be less when unexpected rather that intended effects are being sought.
- Designs intended to detect preference effects have been proposed and, while they offer some advantages, formidable problems remain.

Chapter 12

External validity: a way forward?

Possible solutions

Eligibility, centre participation, invitation, and patient participation are considered together as many of the issues are similar. Having shown that each of these factors can give rise to samples that are unrepresentative of the populations from which they are drawn, the solutions are largely self-evident, if often difficult to implement. Inclusion criteria should be broad and, in particular, where exclusions are purely for administrative convenience or tradition, they should be abandoned. Efforts should be intensified to include settings to which the results are intended to be applied, though there may be formidable obstacles to doing this. A recent example highlights how apparently unrelated factors may influence this. Even among teaching centres in the USA, research is increasingly concentrated in those areas situated in less competitive healthcare markets, as pressures from healthcare funders squeeze research.[184] Changes to the funding of research in the NHS aim to avoid the same happening in the UK.[184]

The evidence concerning behaviour changes, either lay or professional, suggests that it may be difficult to overcome the failure (either deliberate or unintended) to invite certain patients to participate, or of patients to consent. Any strategy must address not only the lack of knowledge among both practitioners and patients, that patients often benefit simply from being in RCTs, and that a new treatment is as likely to be harmful as beneficial,[186] but also attitudes and practices. It should take into account the emerging body of research on patients' expectations of researchers and how recruitment might be improved.[187] In particular, where there is opposition to randomisation from practitioners, methods such as that advocated by Zelen, in which subjects are randomised before being asked to give consent should be considered (*Figure 22*).[188] This has been shown to increase participation by practitioners in certain circumstances, though it is difficult to predict how many subjects will be in each category and thus what sample size is required. Furthermore, this approach may raise ethical concerns for some people.

Notwithstanding the importance of these objectives, a question remains over whether it is possible, on the basis of what is known about the impact of these threats to external validity, to relate the results of restrictive RCTs to individuals with prognoses that differ from those included in an RCT. Two questions arise: the first is whether it is possible to identify trends in net benefit according to baseline prognostic features among those included in RCTs that can then be extrapolated to others not included. Although this review has focused on the differences between RCTs and non-randomised studies, evidence from those researchers that have sought to reconcile differences in effect size due to variation in baseline risk among participants in RCTs[189–191] also has implications for this review. Where sufficient information is available on participants at entry, it may be possible to identify trends in net benefit according to initial features.[21] This is analogous to stratified analyses. It is,

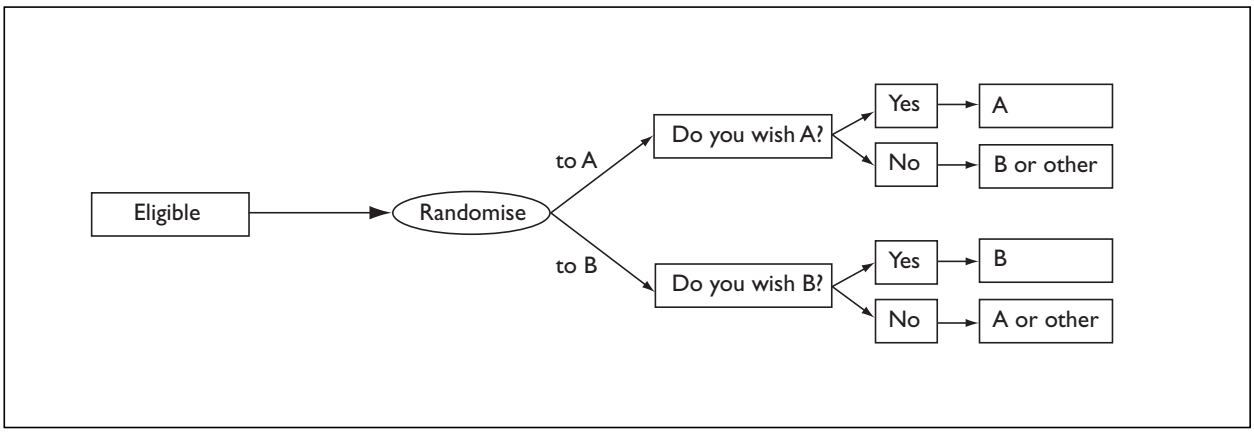


FIGURE 22 *Double consent randomised design (Source: Zelen[188])*

however, rarely possible to disentangle the effects of the possible factors, **e**, **i**, **d** and **p**, that may give rise to these differences. A superficially attractive alternative is to avoid cataloguing differences in particular prognostic factors and look at overall differences in risk, such as the relationship between treatment effect and the frequency of events in the control group or, as in the L'Abbé plot, to examine the relationship between the frequency of events n the treatment and control groups. Glasziou and Irwig have suggested such an approach in which the net benefit for a particular patient can be estimated from knowledge of the RR reduction achieved by an intervention, that patient's baseline risk, and a measure of harm from the intervention.[192] It assumes fixed adverse effects with increasing risk and a constant reduction in RR. Unfortunately, while these methods can provide some clues to identify studies requiring detailed examination, they do not support statistical analysis that would resolve this issue.[193]

The second and key question, however, is whether it is legitimate to extrapolate these trends to a wider population. This is reminiscent of debates about the limits of empiricism in the eighteenth century, when philosophers such as Hume concluded that no matter how many times an event was observed to have a particular effect, one could never be sure that it would always do so, in all circumstances and at all times. The present review has identified no evidence to challenge this view.

Summary

- Many of the solutions to problems of eligibility or participation are self-evident, such as removal of blanket exclusions.
- Implementing change may be difficult but some strategies, including methods of trial design, do exist.
- There is insufficient evidence to justify extrapolation to wider populations of results from limited samples on the basis of identifiable characteristics and there are well-established philosophical arguments why this might not be expected to be valid.

Chapter 13

Summary and conclusions

Overview

As noted in chapter 1, opinions about the relative merits of RCTs and non-randomised studies have tended to become polarised. To a considerable degree, the differing positions relate to the relative importance placed on the different sets of threats to either internal or external validity. While limited by the scarcity of relevant studies even after exhaustive searching, this review has managed to both confirm and refute some widely held beliefs that underpin these differing views. The four research areas are summarised below.

Do results from non-randomised studies and RCTs differ systematically?

The argument that non-randomised studies consistently favour an intervention above either placebo or no treatment, or a new treatment over an old one is not sustained. RCTs and non-randomised studies can produce different results but the direction of the difference is not consistent. In this review, in seven of 18 comparisons there was no significant difference in effect size according to the two methods. Furthermore, variation in results also occurs between RCTs and between non-randomised studies. This often reflects widely differing design features, which are sufficiently great as to preclude detection of any specific effect of the process of allocation.

What happens to effect sizes when potential allocation bias is adjusted for?

The argument that, while groups in non-randomised studies may have different prognostic features at baseline, adjustment can cause the results to converge with those of RCTs is not sustained. The empirical evidence reviewed does not show that this invariably happens, though there are several reasons why this may be so. First, within the limits imposed by the power of many studies, there is often no significant difference between the results obtained by RCTs and non-randomised studies so there is no gap to close. Second, the quality of adjustment may be inadequate and thus says little about what could be achieved under optimal circumstances. Third, differences in other aspects (exclusions, participants, etc.) remain, so there is no reason to expect effect sizes to be the same.

Do threats to external validity affect generalisability to the reference population?

Chapters 4 and 5 have addressed the criticism that RCTs lack external validity in that those who meet eligibility criteria or who are either invited or agree to participate are significantly different from the population to whom the results of the study will be applied. It confirms that those who participate in RCTs, whether as practitioners or subjects, are frequently quite unrepresentative of the population to whom the results will be applied, though the consequences for policy are unclear.

Chapter 4 makes clear that, typically, a very small proportion of patients with a particular condition who could be included in RCTs are included, though precise figures are very difficult to ascertain because of a frequent lack of clarity about who might be included. The chapter also showed the extent to which important sub-groups of the population, and often those to whom the results of RCTs will be most commonly applied such as the elderly are systematically excluded from many trials.

Chapter 5 found considerable evidence that those individuals participating in studies are not truly representative even of those who are eligible to be included. Some potential subjects will be denied the opportunity to participate in trials simply by virtue of where they are treated. Evaluative research of all kinds is concentrated in teaching centres. The limited evidence available suggests that settings of non-randomised studies are slightly more representative. There is also evidence concerning those who pass the first hurdle and reach a centre in which a trial is being conducted. Here, it is necessary to differentiate trials of treatment and prevention. Those participating in RCTs of treatment tend to be less affluent, less educated, and more severely ill than those who do not. In contrast, in preventive trials, participants tend to be wealthier, better educated, and more likely to have adopted a healthy lifestyle. These findings can be interpreted as follows.

- Participation bias in preventive trials may increase the effect size on intermediate outcomes such as change in lifestyle, by exaggerating

adherence, but decrease the effect size on health outcomes such as morbidity and mortality as there is less potential to improve.
- Participation bias in treatment trials may increase the effect size because those included have greater potential to benefit.

The interplay of these factors is, however, complex and it is inappropriate to generalise. Each case should be examined individually.

What are the effects of preference?

The review showed how, in theory at least, elimination of patient preferences could play an important part in unblinded RCTs, particularly where the effect being sought is small. Indirect evidence for a preference effect was identified but the very limited body of empirical evidence precluded any firm conclusion on whether it actually does exist and whether it is important.

Implications for policy

While a high level of exclusion may have some advantages for those conducting an RCT, it also has important implications for policy. In particular, there is a risk of denial of effective treatment to those who might benefit but who have been excluded from the RCTs, and delay in obtaining definitive results because of low recruitment rate. These problems can also affect non-randomised studies but they are more likely to be a greater problem with RCTs. In addition, there is a danger of unjustified extrapolation of results to other populations, and therefore it is concluded that it should **not** be assumed that summary results apply equally to all potential patients.

Recommendations for research

The results of this review have several important implications for the ways in which evaluative research is conducted, interpreted, and reported.

Conducting research

This review has confirmed many of the observations made previously by others. RCTs are frequently too small to detect any effect that might exist and combination of results of small trials confronts the problem of heterogeneity, as the populations studied commonly vary quite considerably, judged by differences in event rates in control groups. It is easy to specify the ideal study design. This could be characterised as consisting of a very large RCT in which the settings and the subjects are representative of the population to which the results might be applied and those participating are equivocal as to which arm they are allocated. Furthermore, the study should be able to identify with confidence important factors, either related to patients or treatment settings, that will favour one treatment or another and thus facilitate the application of results to subgroups of patients with particular characteristics. In practice, this situation is rarely achieved, for a variety of procedural and practical reasons. Furthermore, we recognise that the perfect can become the enemy of the good and researchers should not abandon a question simply because they are unable to achieve perfection.

Considering the specific issues discussed in this report, we believe that there continue to be situations in which an RCT is not possible. We recognise that practical reasons for this, such as ethical concerns or cost, may not be sustainable in the eyes of some but they may be insurmountable. In these circumstances, we believe that if a well-designed non-randomised study is possible, it should be undertaken and will be preferable to a small, poorly designed and restrictive RCT. It will never be possible to know with certainty whether there has been adequate adjustment for baseline differences between groups but the approach used should adhere to the guidance set out in the previous chapter. A subsequent RCT may give a different answer but there may be several possible explanations for this, other than the method of allocating patients to interventions.

If an RCT is chosen, it should seek to include as wide a range of practice settings as possible and the study population should be representative of all patients currently being treated. Exclusions for administrative convenience should be rejected. It is recognised that, for many conditions, this may be complicated by the absence of information on the spectrum of severity of patients who are actually considered for treatment.

While arguing that RCTs should be as inclusive as possible, we recognise the need to avoid a situation in which subjects with quite different abilities to benefit from treatment are brought together to produce a single figure for effect size. In many cases it will be necessary to stratify the study population on the basis of either identified or suspected prognostic factors. This must obviously be done at the outset of the study and it is recognised that it will require larger numbers.

As far as possible, efforts should be made to maximise participation by practitioners and patients, and several suggestions as to how to do this have been made.

The issue of preference remains problematic. Designs incorporating preference arms exist but we have shown that it may still not be possible to differentiate treatment from preference effects. The only dependable solution is for trials to be absolutely blind to all concerned but, for many interventions, this will be impossible.

Interpreting research

This review has confirmed that those who participate in RCTs are often a highly selected group that have had to pass through a series of hurdles of eligibility, invitation and decision to participate. These factors act in different ways, in part depending on whether a study is of an intervention designed to treat those whose health is already impaired or to prevent illness or promote health. Certain key messages emerge:

- heterogeneity among studies, both in terms of the populations and the interventions studied should be addressed explicitly
- practitioners should use considerable caution when extrapolating results to populations that differ from those included in research studies
- when there is a difference in effect size between an RCT and a non-randomised study, differences in the study population or lack of power of one of the studies should be considered as well as differences in treatment allocation procedures.

Reporting research

An immediate priority is to improve the quality of reporting evaluative research. Throughout this review it has been noted that information on eligibility and participation is often lacking, though there is *prima facie* evidence that it varies widely. As a minimum, authors should define the population to whom they expect their results to be applied, what steps they have taken to ensure that the study population is representative of this wider population and any evidence of how it differs, the characteristics of those centres participating and any that declined, and the numbers and characteristics of those eligible to be included who either were not invited to do so or were invited and declined. As noted earlier, effective implementation of the CONSORT statement[120] will address this issue in part as it requires that authors report the number in the eligible population and the number not randomised for each reason. It does not, however, require information on the characteristics of those included and excluded.

Until these issues are better understood, such information should be collected and reported both for RCTs and non-randomised studies.

Further research

This review has identified many unanswered questions. In particular, it has highlighted the weakness of the evidence base on which many decisions must be made. Each of the examples studied suggest particular questions that could usefully be pursued but we believe that, as a priority, the following issues should be addressed.

Who is and is not included in RCTs and do any differences matter?

This review has shown that there are differences between those who do and do not participate in RCTs and that these differences could influence any effect detected. It also noted the very limited evidence available on which to make judgements on this issue. This is a relatively straightforward issue to address. Those commissioning research could require that organisers of RCTs collect information on relevant characteristics of the entire eligible population, including those eliminated at each stage. Consequently, each of the areas in *Figure 1* could be described and differences between **s**, **e**, **i**, **d** and **p** compared. Ideally, studies would follow-up all five categories of patient and compare their outcomes. Over several years this would provide a substantial volume of data that would enable many of the outstanding questions to be answered. Other research is needed to examine the effect of differences in participation by centres and practitioners.

What is the effect of patient preference?

We have concluded that the existence of important effects on therapeutic efficacy attributable to the preferences of patients is theoretically plausible. However, even if such effects exist, finding reliable empirical evidence is methodologically difficult and will require a major research effort.

Preference effects face a 'Catch 22' situation; large trials would be needed to reliably discern preference effects but it is not clear whether the current evidence of their importance is sufficient to convince funders of the need for such an effort

or whether such evidence can be obtained other than by mounting a large study.

There is, however, some scope for greater use of preference trials. Under this design, patients are carefully and fully informed about the relevant scientific uncertainties and encouraged to actively choose their treatment. Those with little or no preference for treatment are encouraged to accept randomisation. The systematic follow-up of such cohorts would offer the opportunity to establish through randomisation the physiological effects of treatment among those with no preference and to learn whether patients with apparently similar prog-nostic characteristics who actively choose their treatments have different outcomes than those predicted by randomisation.

Is it possible to design non-randomised studies that will produce valid and reliable results?

The literature reviewed showed that the results of RCTs and non-randomised studies do not inevitably differ but the available evidence suffers from many limitations. It does, however, suggest that it may be possible to minimise any differences by ensuring that subjects included in each type of study are comparable. The effect of adjustment for baseline differences between groups in non-randomised studies is inconsistent but, where it is done, it should involve rigorously developed formulae, as set out in the previous chapter. This hypothesis should be tested by evaluating several specific interventions, ensuring that both treatment and preventive interventions are included.

Acknowledgement

This review was supported by the NHS R&D Executive's Health Technology Assessment Programme. We are grateful to Professor Simon Thompson and Professor Tom Jefferson for many helpful comments.

References

1. Ellenberg JH. Selection bias in observational and experimental studies. *Stat Med* 1994;**13**:557–67.
2. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;**312**:1215–18.
3. Lancet. Opren scandal (editorial). *Lancet* 1983;**1**:219–20.
4. Black NA. The relationship between evaluative research and audit. *J Publ Health Med* 1992;**14**:361–6.
5. 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.
6. Majeed AW, Troy G, Nicholl JP, Smythe A, Reed MW, Stoddard CJ, *et al.* Randomised, prospective, single-blind comparison of laparoscopic versus small-incision cholecystectomy. *Lancet* 1996;**347**:989–94.
7. Office of Technology Assessment. Tools for effectiveness research. In: Identifying health technologies that work. Washington DC: Office of Technology Assessment, 1994.
8. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med* 1989;**8**:441–54.
9. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: Surgical. *Stat Med* 1989;**8**:455–66.
10. Iezzoni LI. Dimensions of risk. In: Iezzoni LI, editor. Risk adjustment for measuring health outcomes. AHSR, Ann Arbor, 1994.
11. McPherson K. The best and the enemy of the good: randomised controlled trials, uncertainty, and assessing the role of patient preference in medical decision making. The Cochrane Lecture. *J Epidemiol Community Health* 1994;**48**:6–15.
12. Rucker G. A two-stage trial design for testing treatment, self-selection and treatment preference effects. *Stat Med* 1989;**8**:477–85.
13. McKay JR, Alterman AI, McLellan T, Snider EC, O'Brien CP. Effect of random versus nonrandom assignment in a comparison of inpatient and day hospital rehabilitation for male alcoholics. *J Consult Clin Psychol* 1995;**63**:70–8.
14. Torgerson DJ, Klaber-Moffett J, Russell IT. Patient preferences in randomised trials: threat or opportunity. *J Health Serv Res Policy* 1996;**1**:194–7.
15. Lilford RJ, Jackson G. Equipoise and the ethics of randomisation. *J Roy Soc Med* 1995;**88**:552–9.
16. Redd WH, Anderson BL, Bovbjerg DH, Capenter PJ, Dolgin M, Mitnick L, *et al.* Physiologic and psychobehavioural research in oncology. *Cancer* 1991;**67**:813–22.
17. Marmot MG, Bosma H, Hemingway H, Brunner E, Stansfield S. Contribution of job control and other risk factors to social variations in coronary heart disease incidence. *Lancet* 1997;**350**:235–9.
18. Stephenson P, McKee M. Look twice. *Eur J Publ Health* 1993;**3**:151–2.
19. Heithoff KA, Lohr KN, editors. Effectiveness and outcomes in health care. Proceedings of an invitational conference by the Institute of Medicine Division of Health Care Services. Washington DC: National Academy Press, 1990.
20. Col NF, Gurwitz JH, Alpert JS, Goldberg RJ. Frequency of inclusion of patients with cardiogenic shock in trials of clinical therapy. *Am J Cardiol* 1994;**73**:149–57.
21. Thompson SG. What sources of heterogeneity in meta-analyses should be investigated? *BMJ* 1994;**309**:1351–5.
22. Bailey KR. Generalizing the results of randomised clinical trials. *Controlled Clin Trials* 1994;**15**:15–23.
23. Gray RJ. A Bayesian analysis of institutional effects in a multi-centre clinical trial. *Biometrics* 1994;**50**:244–53.
24. Davis CE. Generalizing from clinical trials. *Controlled Clin Trials* 1994;**15**:11–14.
25. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AMS. Comparisons of effect size estimates derived from randomised and non-randomised studies. In: Blavk N, Brazier J, Fitzpatrick R, Reeves B, editors. Methods for health service research: a state of the art guide. London: BMJ Publications, 1998.
26. Kleijnen J, Gotzsche P, Kunz RA, Oxman AD, Chalmers I. So what's so special about randomisation? In: Maynard A, Chalmers I, editors. Non-random reflections on health services research. London: BMJ Publications, 1997:93–106.
27. Duley L. Commentary: Sources of bias must be controlled. *BMJ* 1997;**315**:220.
28. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;**309**:1358–61.

29. Sacks H, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. *Am J Med* 1982;**72**:233–40.

30. CASS Principal Investigators and their Associates. Coronary artery surgery study (CASS): a randomised trial of coronary artery bypass surgery. Comparability of entry characteristics and survival in randomised patients and nonrandomised patients meeting randomization criteria. *J Am Coll Cardiol* 1984;**3**(1):114–28.

31. Hlatky MA, Califf RM, Harrell FE, Lee KL, Mark DB, Pryer DB. Comparison of predictions based on observational data with the results of randomised controlled clinical trials of coronary artery bypass surgery. *J Am Coll Cardiol* 1988;**1**:237–45.

32. Horwitz RI, Viscoli CM, Clemens JD, Sadock RT. Developing improved observational methods for evaluating therapeutic effectiveness. *Am J Med* 1990;**89**:630–38.

33. Paradise JL, Bluestone CD, Bachman RZ, Colborn DK, Bernard BS, Taylor FH, *et al.* Efficacy of tonsillectomy for recurrent throat infection in severely affected children. Results of parallel randomised and nonrandomised clinical trials. *N Engl J Med* 1984;**310**:674–83.

34. Paradise JL, Bluestone CD, Rogers KD, Taylor FH, Colborn DK, Bachman RZ, *et al.* Efficacy of adenoidectomy for recurrent otitis media in children previously treated with tympanostomy-tube replacement. Results of parallel randomised and nonrandomised trials. *JAMA* 1990;**263**:2066–73.

35. Schmoor C, Olschewski M, Schumacher M. Randomised and non-randomised patients in clinical trials: experiences with comprehensive cohort studies. *Stats Med* 1996;**15**:236–71.

36. Yamamoto H, Hughes RW, Schroeder KW, Viggiano TR, Dimagno EP. Treatment esophageal stricture by Eder-Puestow or balloon dilators. A comparison between randomized and prospective non-randomized trials. *Mayo Clin Proc* 1992;**67**:228–36.

37. Nicolaides K, de Lourdes Brizot M, Patel F, Snijders R. Comparison of chorionic villus sampling and amniocentesis for fetal karyotyping at 10–13 weeks' gestation. *Lancet* 1994;**344**:435–9.

38. Emanuel EJ. Cost savings at the end of life. What do the data show? *JAMA* 1996;**275**:1907–14.

39. Garenne M, Leroy O, Beau JP, Sene I. Efficacy of measles vaccines after controlling for exposure. *Am J Epidemiol* 1993;**138**:182–95.

40. Antman K, Amoto D, Wood W, Corson J, Suit H, Proppe K, *et al.* Selection bias in clinical trials. *J Clin Oncol* 1985;**3**:1142–7.

41. Shapiro CL, Recht A. Late effects of adjuvant therapy for breast cancer. *Monogr Natl Cancer Inst* 1994;**16**:101–12.

42. Jha P, Flather M, Lonn E, Farkouhn M, Yusuf S. The antioxidants vitamins and cardiovascular disease – a critical review of epidemiological and clinical trial data. *Ann Int Med* 1995;**123**:860–72.

43. Pyorala S, Huttunen NP, Uhari M. A review and meta-analysis of hormonal treatment of cryptorchidism. *J Clin Endocrin Metab* 1995;**80**:2795–9.

44. The Recurrent Miscarriage Immunotherapy Trialists Group. Worldwide collaborative observational study and meta-analysis on allogenic leukocyte immunotherapy for recurrent spontaneous abortion. *Am J Repro Immunol* 1994;**32**:55–72.

45. Watson A, Vail A, Vandekerckhove P, Brosens I, Lilford R, Hughes E. A meta-analysis of the therapeutic role of oil soluble contrast media at hysterosalpingography: a surprising result? *Fert Steril* 1994;**6**:470–7.

46. Reimold SC, Chalmers TC, Berlin JA, Antman EM. Assessment of the efficacy and safety of antiarrhythmic therapy for chronic atrial fibrillation: observations on the role of trial design and implications of drug-related mortality. *Am Heart J* 1992;**124**:924–32.

47. Stukenborg GJ. Comparison of carotid endarterectomy outcomes from randomized controlled trials and Medicare administrative databases. *Arch Neurol* 1997;**54**:826–32.

48. Heinsman DT, Shadish WR. Assignment methods in experimentation: when do nonrandomized experiments approximate answers from randomized experiments? *Psychol Methods* 1996;**1**:154–69.

49. Gilbert JP, McPeek B, Mosteller F. Progress in surgery and anesthesia: benefits and risks of innovative therapy. In: Bunker JP, Barnes BA, Mosteller F, editors. Cost, risks, and benefits of surgery. Oxford: Oxford University Press, 1977:124–69.

50. Ottenbacher K. Impact of random assignment on study outcome: an empirical examination. *Controlled Clin Trials* 1992;**13**:50–61.

51. Hunninghake DB, Darby CA, Probstfield JL. Recruitment experience in clinical trials: literature summary and annotated bibliography. *Controlled Clin Trials* 1987;**8**:6S–30S.

52. Horwitz RI, Feinstein AR. Improved observational method for studying therapeutic efficacy: suggestive evidence that lidocaine prophylaxis prevents death in acute myocardial infarction. *JAMA* 1981;**246**:2455–9.

53. Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. *BMJ* 1984;**289**:1281–4.

54. Muller DWM, Topol EJ. Selection of patients with acute myocardial infarction for thrombolytic therapy. *Ann Intern Med* 1990;**113**:949–60.

55. Lee JY Breaux SR. Accrual of radiotherapy patients to clinical trials. *Cancer* 1983;**52**:1014–16.

56. Begg CB, Zelen M, Carbonne PP, *et al.* Cooperative groups and community hospitals: measurement of impact on community hospitals. *Cancer* 1983;**52**:1760–67.

57. Martin JF, Henderson WG, Zacharaski LR, Rickles FR, Forman WB, Cornell CJ, *et al.* Accrual of patients into a multi hospital cancer clinical trial and its implications. *Am J Clin Oncol* 1984;**7**:173–82.

58. Fentiman IS, Julien JP, van-Dongen JA, van-Geel B, Chetty U , Coibion M. Reasons for non-entry of patients with DCIS of the breast into a randomised trial (EORTC 10853). *Eur J Cancer* 1991;**27**:450–2.

59. Barnett HJ, Sackett D, Taylor DW, Haynes B, Peerless SJ, Meissner I, *et al.* Are the results of the extracranial-intracranial bypass trial generalizable? *N Engl J Med* 1987;**316**:820–4.

60. Begg CB, Engstrom PF. Eligibility and extrapolation in cancer clinical trials. *J Clin Oncol* 1987;**5**:962–8.

61. Collins R, Peto R, MacMahon S, Herbert P, Fiebach NH, Eberlein KA, *et al.* Blood pressure, stroke, and coronary heart disease. Part 2, Short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. *Lancet* 1990;**335**:827–38.

62. Medical Research Council Working Party. MRC trial of treatment of mild hypertension: principal results. *BMJ* 1985;**291**:97–104.

63. Coope J, Warender TS. Randomised trial of treatment of hypertension in the elderly in primary care. *BMJ* 1986;**293**:1145–51.

64. Yusuf S, Held P, Teo KK. Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria. *Stat Med* 1990;**9**:73–86.

65. Cowan CD, Wittes J. Intercept studies, clinical trials, and cluster experiments: to whom can we extrapolate? *Controlled Clin Trials* 1994;**15**:24–9.

66. Lumley J, Bastian H. Competing or complementary? Ethical considerations and the quality of randomized trials. *Int J Tech Assess Health Care* 1996;**12**:247–63.

67. Gomez-Marin O, Prineas RJ, Sinaiko AR. The sodium-potassium blood pressure trial in children: design, recruitment and randomisation. *Controlled Clin Trials* 1991;**12**:408–23.

68. Haynes RB, Dantes R. Patient compliance and the conduct and interpretation of therapeutic trials. *Controlled Clin Trials* 1987;**8**:12–19.

69. Patterson WB, Emanuel EJ. The eligibility of women for clinical research trials. *J Clin Oncol* 1995;**13**:293–9.

70. Chalmers TC. Ethical implications of rejecting patients for clinical trials (editorial). *J Am Med Assoc* 1990;**263**:865.

71. Gurwitz JH, Nananda F, Avorn J. The exclusion of the elderly and women from clinical trials on acute myocardial infarction. *JAMA* 1992;**268**:1417–22.

72. Kannry JL, Chalmers TC, Orza TC, Reitman M, Brown D. Neglect of aged in clinical trials (abstract). *Controlled Clin Trials* 1989;**10**:348.

73. SHEP Cooperative Research Group. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. *JAMA* 1991;**265**:3255–64.

74. Dahlof B, Lindblom LH, Hansson I, Schersten B, Ekborn T, Wester PO. Morbidity and mortality in the Swedish trial in old patients with hypertension (STOP–Hypertension). *Lancet* 1991;**263**:3255–64.

75. Lee JY, Marks JE, Simpson JR. Age as a criterion for eligibility in a lung cancer clinical trial. *Cancer Clin Trials* 1982;**5**:449–52.

76. Rochon PA, Fortin PR, Dear KBG, Minaker KL, Chalmers TC. Reporting age data in clinical trials of arthritis. *Arch Intern Med* 1993;**153**:243–8.

77. Wenger NK. Exclusion of the elderly and women from coronary trials: is their quality of care compromised? *JAMA* 1992;**268**:1460–1.

78. McDermott MM, Lefevre F, Feinglass J, Reifler D, Dolan N, Potts S, *et al.* Changes in study design, gender issues, and other characteristics of clinical research published in three major medical journals from 1971 to 1991. *J Gen Intern Med* 1995;**10**:13–18.

79. Bennett JC. Inclusion of women in clinical trials – policies for population subgroups. *N Engl J Med* 1993;**329**:288–92.

80. Caschetta M, Chavkis W, McGovern T. FDA policy on women in drug trials (letter). *N Engl J Med* 1993;**329**:1815.

81. Svensson CK. Representation of American Blacks in clinical trials of new drugs. *JAMA* 1989;**261**:263–5.

82. Moore RD, Stanton D, Gopalan R, Chaisson R. Racial differences in the use of drug therapy for HIV disease in an urban community. *N Engl J Med* 1994;**330**:763–8.

83. Klein R, Moss SA and The Diabetes Control and Complications Trial Research Group. Comparison of the study populations in the diabetes control and complications trial and the Wisconsin epidemiologic study of diabetic retinopathy. *Arch Intern Med* 1995;**155**:745–54.

84. Ward LC, Fielding JWL, Dunn JA, Kelly KA for the British Stomach Cancer Group. The selection of cases for randomised trials: a registry survey of concurrent trail and non-trial patients. *Br J Cancer* 1992;**66**:943–50.

85. Marubini E, Mariani L, Salvadori B, Veronesi U, Saccozzi R, Merson M, *et al.* Results of a breast-cancer-surgery trial compared with observational data from routine practice. *Lancet* 1996;**347**:1000–3.

86. Kober L, Torp-Pedersen C and the TRACE study group. Clinical characteristics and mortality of patients screened for entry into the Trandolapril Cardiac Evaluation Study (TRACE). *Am J Cardiol* 1995;**76c**:1–5.

87. The Toronto Leukemia Study Group. Results of chemotherapy for unselected patients with acute myeloblastic leukemia: effect of exclusions on interpretation of results. *Lancet* 1986;**1**:786–8.

88. Weaver WD, Litwin PE, Martin JS, Kudenchuk PJ, Maynard C, Eisenberg MS, *et al.* Effect of age on the use of thrombolytic therapy and mortality in acute myocardial infarction. The MITI Project Group. *J Am Coll Cardiol* 1991;**18**:657–62.

89. Chadd K, Goldstein S, Byington R, Curb JD. Effect of propranalol after acute myocardial infarction in patients with congestive heart failure. *Circulation* 1986;**73**:503–10.

90. Montague TJ, Ikuta RM, Wong RY, Bay KS, Teo KK, Davies NJ, *et al.* Comparison of risk and patterns of practice in patients older and younger than 70 years with acute myocardial infarction in a two-year period (1987–1989). *Am J Cardiol* 1991;**68**:843–7.

91. Maynard C, Althouse R, Cerqueira M, Olsufka M, Kennedy JW. Underutilisation of thrombo-embolytic therapy in eligible women with acute myocardial infarction. *Am J Cardiol* 1991;**68**:529–30.

92. Holme I, Ekelund LG, Hjermann I, Leren P. Quality-adjusted meta-analysis of the hypertension/coronary dilemma. *Am J Hypertens* 1994;**7**:703–12.

93. Sundt TM. Was the international randomised trial of extracranial-intracranial arterial bypass representative of the population at risk? *N Engl J Med* 1987;**316**:814–16.

94. Goldring S, Zervas N. The extracranial-intracranial bypass study. *N Engl J Med* 1987;**316**:817–20.

95. Wenneberg DE, Lucas FL, Birkmeyer JD, Bredenberg CE, Fisher ES. Variation in carotid endarterectomy mortality in the Medicare population. *JAMA* 1998;**279**:1278–81.

96. Holden G, Rosenberg G, Barker K, Tuhrim S, Brenner B. The recruitment of research participants: a review. *Social Work Health Care* 1993;**19**:1–44.

97. Walterspiel JN. Informed consent: influence on patient selection among critically ill premature infants. *Pediatrics* 1990;**85**:119–21.

98. De Vita VT. Breast cancer therapy: exercising all our options. *N Engl J Med* 1989;**320**:527–9.

99. Taylor KM, Margolese RG, Soskolne CL. Physicians' reasons for not entering eligible patients in a randomized clinical trial of surgery for breast cancer. *N Engl J Med* 1984;**310**:1363–7.

100. Gotay CC. Accrual to cancer clinical trials: directions from the research literature. *Soc Sci Med* 1991;**33**:569–77.

101. Volunteering for research (editorial). *Lancet* 1992;**340**:823–4.

102. Downs SH, Black NA, Devlin HB, Royston CMS, Russell RCG. Systematic review of the effectiveness and safety of laparoscopic cholecystectomy. *Ann Roy Coll Surgeons (Engl)* 1996;**78**:241–323.

103. Black NA, Downs SH. The effectiveness of surgery for stress incontinence in women: a systematic review. *Br J Urol* 1996;**78**:497–510.

104. Barofsky I, Sugarbaker PH. Determinants of patients nonparticipation in randomised clinical trials for the treatment of sarcomas. *Cancer Clin Trials* 1979;**2**:237–49.

105. Hunter CP, Frelick RW, Feldman AR, Bavier AR, Dunlap WH, Ford L, *et al.* Selection factors in clinical trials: results from the community clinical oncology program physician's patient log. *Cancer Treat Rep* 1987;**71**:559–65.

106. Smith P, Arnesen H. Non-respondents to a post-myocardial infarction trial: characteristics and reasons for refusal. *Acta Med Scand* 1988;**223**:537–42.

107. Harth SC, Thong YH. Sociodemographic and motivational characteristics of parents who volunteer their children for clinical research: a controlled study. *BMJ* 1990;**300**:1372–5.

108. Vollmer WM, Hertert S, Allison MJ. Recruiting children and their families for clinical trials: a case study. *Controlled Clin Trials* 1992;**13**:315–20.

109. Williford WO, Krol WF, Buzby GP. Comparison of eligible randomized patients with two groups of ineligible patients: can the results of the VA total parenteral nutrition clinical trial be generalised? *J Clin Epidemol* 1993;**46**:1025–34.

110. Kaufmann CL, Schulberg HC, Schooler NR. Self-help group participation among people with severe mental illness. *Prevention in Human Services* 1994;**11**:315–31.

111. Stone JM, Laidlaw CR, Page FJ, Cooper I. Selection of patients for randomised trials: a study based on the MACOP-B vs CHOP in NHL study. *Aust N Z J Med* 1994:**24**:536–40.

112. Rogers WJ, Alderman EL, Chaitman BR, DiSciascio G, Horan M, Lytle B, *et al.* Bypass angioplasty revascularization investigation (BARI): baseline clinical and angiographic data. *Am J Cardiol* 1995;**75**:9C–17C.

113. van Bergen PFMM, Jonker JJC, Molhoek GP, van der Burgh PH, van Domburg RT, Deckers JW, *et al.* Characteristics and prognosis of non-participation of a multi-centre trial of long-term anticoagulant treatment after myocardial infarction. *Int J Cardiol* 1995;**49**:135–41.

114. Gorkin L, Schron EB, Handshaw K, Shea S, Kinney MR, Branyon M, *et al.* Clinical 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.

115. Constantine WL, Haynes CW, Spiker D, Kendall-Tackett K, Constantine NA. Recruitment in a clinical trial for low birth weight, premature infants. *J Dev Behav Pediatr* 1993;**14**:1–7.

116. Naslund GK, Fredrikson M, Hellenius ML, de Faire U. Characteristics of participating and non-participating men in a randomised, controlled diet and exercise intervention trial. *Scand J Prim Health Care* 1994;**12**:249–54.

117. Davies G, Pyke S, Kinmouth AL. Effect of non-attenders on the potential of a primary care programme to reduce cardiovascular risk in the population. *BMJ* 1994;**309**:1553–6.

118. Yeomans-Kinney A, Vernon SW, Frankowski RF, Weber DM, Bitsura JM, Vogel VG. Factors related to enrollment in the breast cancer prevention trial at a comprehensive cancer center during the first year of recruitment. *Cancer* 1995;**76**:46–56.

119. Pacala JT, Judge JO, Boult C. Factors affecting sample selection in a randomized trial of balance enhancement: the FICSIT study. *J Am Geriatr Soc* 1996;**44**:377–82.

120. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, *et al.* Improving the quality of reporting of randomized controlled trials: the CONSORT Statement. *JAMA* 1996;**276**:637–9.

121. O'Boyle CA. Diseases with passion. *Lancet* 1993;**342**:1126–7.

122. MacIntyre IMC. Tribulations for clinical trials. Poor recruitment is hampering research. *BMJ* 1991;**302**:1099–100.

123. Fallowfield LJ, Hall A, Maguire GP, Baum M. Psychological outcomes of different treatment policies in women with early breast cancer outside a clinical trial. *BMJ* 1990;**301**:575–80.

124. The Coronary Drug Project Team. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N Engl J Med* 1980;**303**:1038–41.

125. Phillips DP, Ruth TE, Wagner LM. Psychology and survival. *Lancet* 1993:**342**:1142–5.

126. Chaput de Saintonge DM, Herxheimer A. Harnessing placebo effects in health care. *Lancet* 1994;**344**:995–8.

127. Kleijnen J, de Craen AJM, Everdingen JV, Krol L. Placebo effect in double-blind clinical trials: a review of interactions with medications. *Lancet* 1994;**344**:1347–9.

128. Levy SM, Herberman RB, Lee J, Whiteside T, Beadle M, Heiden L, *et al.* Persistently low natural killer cell activity, age, and environmental stress as predictors of infectious morbidity. *Nat Immun Cell Growth Regul* 1991;**10**:289–307.

129. Redd WH, Silberfarb PM, Andersen BL, Andrykowski MA, Bovbjerg DH, Burish TG, *et al.* Physiologic and psychobehavioral research in oncology. *Cancer* 1991;**67**:813–22.

130. McPherson K, Britton AR, Wennberg JE. Are randomized controlled trials controlled? Patient preferences and unblind trials. *J Roy Soc Med* 1997;**90**:652–6.

131. McPherson K. Patients' preferences and randomised trials. *Lancet* 1996;**347**:1119.

132. Rothman KJ. Placebo mania. *BMJ* 1996;**313**:3–4.

133. Chalmers I. What is the prior probability of a proposed new treatment being superior to established treatments? *BMJ* 1997;**311**:74–5.

134. Strong PM. The ceremonial order of the clinic. London: Routledge & Kegan Paul, 1979.

135. Pocock SJ, Henderson RA, Rickards AF, Hampton JR, King SB, Hamm CW, *et al.* Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet* 1995;**346**:1184–9.

136. CABRI Trial Participants. First year results of CABRI (Coronary Angioplasty vs Bypass Revascularisation Investigation). *Lancet* 1995;**346**:1179–84.

137. RITA Trial Participants. Coronary angioplasty versus coronary artery bypass surgery: the Randomised Intervention Treatment of Angina (RITA) trial. *Lancet* 1993;**343**:573–80.

138. King SB, Lembo NJ, Weintraub WS, Kosinski AS, Barnhart HX, Kutner MH, *et al.* A randomised study of coronary angioplasty compared with bypass surgery in patients with symptomatic multi-vessel coronary disease. *N Engl J Med* 1994;**331**:1044–50.

139. Hamm CW, Reimers J, Ischinger T, Rupprecht HJ, Berger J, Bleifeld W, *et al.* A randomised study of coronary angioplasty compared with bypass surgery in patients with symptomatic multi-vessel disease. *N Engl J Med* 1994;**331**:1037–43.

140. Puel J, Karouny E, Marco F, Asson B, Galinier M, Elbaz M, *et al.* Angioplasty versus surgery in multi-vessel disease: immediate results and in-hospital outcome in a randomised prospective study. *Circulation* 1992;**86** (suppl 1):372.

141. Hueb WA, Bellotti G, Almeida S, Aries S, de Albuquerque C, Jatene AD, *et al.* The Medicine, Angioplasty or Surgery Study (MASS): a prospective randomised trial for single proximal left anterior descending artery stenosis. *J Am Coll Cardiol* 1995;**7**:1600–5.

142. Goy JJ, Eeckhout E, Burnand B, Vogt P, Stauffer J-C, Hurni M, *et al.* Coronary angioplasty versus left main internal mammary grafting for isolated proximal left anterior descending artery stenosis. *Lancet* 1994;**343**:1449–53.

143. Rodriguez A, Boullon F, Perez-Balino N, Paviotti C, Liprandi M, Palacios IF, *et al.* Argentine randomised trial of percutaneous transluminal coronary angioplasty versus coronary artery bypass surgery in multivessel disease (ERACHI): in-hospital results and 1-year follow up. *J Am Coll Cardiol* 1993;**22**:1060–7.

144. Jones RH, Kesler K, Phillips HR, Mark DB, Smith PK, Nelson CL, *et al.* Long-term survival benefits of coronary artery bypass grafting and percutaneous transluminal angioplasty in patients with coronary artery disease. *J Thorac Cardiovasc Surg* 1996;**111**:1013–25.

145. Hartz AJ, Kuhn EM, Pryor DB, Krakauer H, Young M, Heudebert G, *et al.* Mortality after coronary angioplasty and coronary artery bypass surgery (the National Medicare Experience). *Am J Cardiol* 1992;**70**:179–85.

146. Kosinski AS, Barnhart HX, Weintraub WS, Guyton RA, King SB. Five year outcome after coronary surgery or coronary angioplasty vs surgery trial (EAST) (abstract). *Circulation* 1995;**91**(suppl 1):1–542.

147. BARI Investigators. Comparison of coronary bypass surgery with angioplasty in patients with multi-vessel disease. *N Engl J Med* 1996;**335**:217–25.

148. Yusef S. Calcium antagonists in coronary artery disease and hypertention. *Circulation* 1995;**92**:1079–82.

149. Kloner RA. Nifedipine in ischemic heart disease. *Circulation* 1995;**92**:1074–8.

150. Opie LH, Messerli FH. Nifedipine and mortality. Grave defects in the dossier. *Circulation* 1995;**92**:1068–73.

151. Braun S, Boyko V, Behar S, Reicher-Reiss H, Shotan A, Schlesinger Z, *et al.* Calcium antagonists and mortality in patients with coronary artery disease: a cohort study of 11,575 patients. *J Am Coll Cardiol* 1996;**28**:7–11.

152. Furberg CD, Psaty BM. Calcium antagonists: antagonists or protagonists of mortality in elderly hypertensives? *J Am Geriatr Soc* 1995;**43**:1309–10.

153. Gordon GD, Mabin TA, Isaacs S, Lloyd EA, Eichler HG, Opie LH. Hemodynamic effects of sublingual nifedipine in acute myocardial infarction. *Am J Cardiol* 1984;**53**:1228–32.

154. Wilcox RG, Hampton JR, Banks DC, Birkhead JS, Brooksby IA, Burns-Cox CJ, *et al.* Trial of early nifedipine in acute myocardial infarction. The TRENT study. *BMJ* 1986;**293**:1204–8.

155. Walker LJE, MacKenzie G, Adgey AAJ. Effect of nifedipine on enzymatically estimated infarct size in the early phase on acute myocardial infarction. *Br Heart J* 1988;**39**:403–10.

156. Sirnes PA, Overskeid K, Pedersen TR, Bathen J, Drivenes A, Froland GS, *et al.* Evolution of infarct size during the early use of nifedipine in patients with acute myocardial infarction: The Norwegian Nifedipine Multicentre Trial. *Circulation* 1984;**70**:638–44.

157. Erbel R, Pop T, Meinertz T, Olshausen KV, Treese N, Heurichs KJ, *et al.* Combination of calcium channel blocker and thrombolytic therapy in acute myocardial infarction. *Am Heart J* 1988;**115**:529–38.

158. Goldbourt U, Behar S, Reicher-Reiss H, Zion M, Mandelzweig L, Kaplinsky E. Early administration of nifedipine in suspected acute myocardial infarction: the SPRINT 2 Study. *Arch Intern Med* 1993;**153**:345–53.

159. Holland Interuniversity Nifedipine/Metopropol Trial (HINT) Research Group. Early treatment of unstable angina in the coronary care unit: a randomised, double-blind, placebo controlled comparison of recurrent ischaemia in patients treated with nifedipine or metopropal or both. *Br Heart J* 1986;**56**:400–13.

160. The Israeli SPRINT Study Group. Secondary Prevention Reinfarction Israeli Nifedipine Trial (SPRINT). A randomised intervention trial of nifedipine in patients with acute myocardial infarction. *Eur Heart J* 1988;**9**:354–64.

161. Branagan JP, Walsh K, Kelly P, Collins WC, McCafferty D, Walsh MJ. Effect of early treatment with nifedipine in suspected acute myocardial infarction. *Eur Heart J* 1986;**7**:859–65.

162. Stroke Unit Trialists' Collaboration. Collaborative systematic review of the randomised trials of organised inpatient (stroke unit) care after stroke. *BMJ* 1997;**314**:1151–9.

163. Gompertz P, Pound P, Briffa J, Ebrahim S. How useful are non-random comparisons of outcomes and quality of care in purchasing hospital stroke services? *Age Ageing* 1995;**24**:137–41.

164. Jørgensen HS, Nakayama H, Raaschou HO, Larsen K, Hübbe P, Olsen TS. The effect of a stroke unit: reductions in mortality, discharge rate to nursing home, length of hospital stay, and cost. *Stroke* 1995;**26**:1178–82.

165. Davenport RJ, Dennis MS, Warlow CP. Effect of correcting outcome data for case mix: an example from stroke medicine. *BMJ* 1996;**312**:1503–5.

166. Hankey GJ, Dennis MS, Slattery JM, Warlow CP. Why is the outcome of transient ischaemic attacks different in different groups of patients? *BMJ* 1993;**306**:1107–11.

167. Horner RD, Matchar DB, Divine GW, Feussner JR. Relationship between physician specialty and the selection and outcome of ischaemic stroke patients. *Health Serv Res* 1995;**30**:275–87.

168. Graves P. Human malaria vaccines. In: Feng C, Garner P, Gelband H, Salinas R, editors. Tropical Disease Module of The Cochrane Database of Systematic Reviews. The Cochrane Library. The Cochrane Collaboration; Issue 3. Oxford: Update Software, 1996.

169. Noya O, Gabaldon Berti Y, Alarcon de Noya B, Borges R, Zerpa N, Urbaez JD, *et al.* Population-based clinical trial with the Spf66 synthetic Plasmodium malaria vaccine in Venezuela. *J Inf Dis* 1994;**170**:396–402.

170. Valero MV, Amador LR, Galindo C, Figueroa J, Bello MS, Murillo LA, *et al.* Vaccination with Spf66, a chemically synthesised vaccine, against *Plasmodium falciparum* malaria in Columbia. *Lancet* 1993;**341**:705–10.

171. Valero MV, Amador R, Aponte JJ, Narvaez A, Galindo C, Silva Y, *et al.* Evaluation of Spf66 malaria vaccine during a 22-month follow-up field trial in the Pacific coast of Columbia. *Vaccine* 1996;**14**(15):1466–70.

172. Sempertegui F, Estrella B, Moscoso J, Piedrahita L, Hernandez D, Gaybor J, *et al.* Safety, immunogenicity and protective effect of the SPf66 malaria synthetic vaccine against *Plasmodium falciparum* infection in a randomized double-blind placebo-controlled field trial in an endemic area of Ecuador. *Vaccine* 1994;**12**:337–42.

173. Alonso PL, Smith T, Schellenberg JR, Masanja H, Mwankusye S, Urassa H, *et al.* Randomised trial of efficacy of Spf66 vaccine against *Plasmodium falciparum* malaria in children in southern Tanzania. *Lancet* 1994;**344**:1175–81.

174. D'Alessandro U, Leach A, Drakeley CJ, Bennett S, Olaleye BO, Fegan GW, *et al.* Efficacy trial of malaria vaccine Spf66 in Gambian infants. *Lancet* 1995;**346**:462–7.

175. European Carotid Surgery Triallists Collaborative Group. MRC European Carotid Surgery Trial: interim results for symptomatic patients with severe (70–99%) or with mild (0–29%) carotid stenosis. *Lancet* 1991;**337**:1235–43.

176. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet* 1995;**345**:1616–19.

177. Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. *J Clin Epidemiol* 1997;**50**:1089–98.

178. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994;**272**:125–8.

179. Rockall TA, Logan RF, Devlin HB, Northfield TC. Variation in outcome after acute upper gastrointestinal haemorrhage. The national audit of acute upper gastrointestinal haemorrhage. *Lancet* 1995;**346**:246–50.

180. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariate models. *Ann Intern Med* 1993;**118**:201–10.

181. Miettinen OS. The need for randomization in the study of intended effects. *Stat Med* 1983;**2**:267–71.

182. Brewin CR, Bradley C. Patient preferences and randomised clinical trials. *BMJ* 1989;**299**:313–5.

183. Wennberg JE. What is outcomes research? In: Gelijns AC, editor. Medical innovations at the crossroads. Vol 1 Modern methods of clinical investigation. Washington DC: National Academy Press, 1990:33–46.

184. Cambell EG, Weissman JS, Blumenthal D. Relationship between market competition and the activities and attitudes of medical school faculty. *JAMA* 1997;**278**:222–6.

185. Culyer-AJ. Research Development Taskforce. Supporting research and development in the NHS. (Culyer Report). London: HM Stationery Office, 1994.

186. Chalmers I. Assembling comparison groups to assess the effects of health care. *J Roy Soc Med* 1997;**90**:379–86.

187. Corbett F, Oldham J, Lilford R. Offering patients entry in clinical trials: preliminary study of the views of prospective participants. *J Med Ethics* 1996;**22**:227–31.

188. Zelen M. Randomized consent designs for clinical trials: an update. *Stat Med* 1990;**9**:645–56.

189. Davey Smith G, Song F, Sheldon T. Cholesterol lowering and mortality: the importance of considering initial risk. *BMJ* 1993;**206**:1367–73.

190. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med* 1992; **1**:2077–82.

191. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized controlled trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA* 1992;**268**:240–8.

192. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *BMJ* 1995;**311**:1356–9.

193. Sharp SJ, Thompson SG, Altman DG. The relationship between treatment benefit and underlying risk in meta-analysis. *BMJ* 1996;**313**:735–8.

194. Stevens RS, Ambler NR. The Dover Stroke Rehabilitation Unit: a randomised controlled trial of stroke management. In: Rose FC, editor. Advances in stroke therapy. New York: Raven Press, 1982:257–61.

195. Stevens RS, Ambler NR, Warren MD. A randomised controlled trial of a stroke rehabilitation ward. *Age Ageing* 1984;**13**:65–75.

196. Garraway WM, Akhtar AJ, Hockey L, Prescott RJ. Management of acute stroke in the elderly; preliminary results of a controlled trial. *BMJ* 1980;**280**:1040–4.

197. Garraway WM, Akhtar AJ, Hockey L, Prescott RJ. Management of acute stroke in the elderly; follow up of a controlled trial. *BMJ* 1980;**281**:827–9.

198. Smith ME, Garraway WM, Smith DL, Akhtar AJ. Therapy impact on functional outcome in a controlled trial of stroke rehabilitation. *Arch Phys Med Rehab* 1982;**63**:21–4.

199. Wood Dauphinee S, Shapiro S, Bass E, Fletcher C, Georges P, Hensby V, *et al*. A randomised trial of team care following stroke. *Stroke* 1984;**5**:864–72.

200. Kalra L, Dale P, Crome P. Improving stroke rehabilitation: a controlled study. *Stroke* 1993;**24**:1462–7.

201. Kalra L, Dale P, Crome P. Do stroke units benefit elderly stroke patients? *Age Ageing* 1994;**23**(suppl 1):5.

202. Indredavik B, Bakke F, Solberg R, Rokseth R, Haaheim LL, Holme I. Benefit of stroke unit: a randomised controlled trial. *Stroke* 1991;**22**:1026–31.

203. Strand T, Asplund K, Eriksson S, Hagg E, Lithner F, Wester PO. A non-intensive stroke unit reduces functional disability and the need for long term hospitalisation. *Stroke* 1985;**16**:29–34.

204. Strand T, Asplund K, Eriksson S, Hagg E, Lithner F, Wester PO. Stroke unit care – who benefits? Comparisons with general medical care in relation to prognostic indicators on admission. *Stroke* 1986;**17**(3):377–81.

Appendix 1

Search strategy

Comparing the results of RCTs and non-randomised studies

Literature search

Literature searches were conducted using the following databases: MEDLINE, EMBASE, Science and Social Science Citation Index (BIDS) and the Cochrane Library.

MEDLINE 1966–96

Search strategy

No.	No. papers	Term
#1	7583	explode "RANDOMIZED-CONTROLLED-TRIALS"/ all subheadings
#2	89055	explode "RESEARCH-DESIGN"/ all subheadings
#3	2848	OBSERVATIONAL
#4	22599	COHORT
#5	1847	NON-RANDOM*
#6	3176	NONRANDOM*
#7	151	NATURAL EXPERIMENT*
#8	434	QUASI-EXPERIMENT*
#9	446	QUASI EXPERIMENT*
#10	121	NONEXPERIMENTAL
#11	57	NON-EXPERIMENTAL
#12	30781	#3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11
#13	106	#1 and #2 and #12

EMBASE 1980–96

Search strategy for TITLES

No.	No. papers	Term
#1	15499	(random*)@(TI)
#2	3187	(observational, non random*, nonrandom*, natural experiment*, quasi experiment*, nonexperimental, non experimental, cohort) @(TI)
#3	65	1 + 2

Search strategy for TITLES, ABSTRACT & KEYWORDS

No.	No. papers	Term
#1	21863	(randomized control*)@(TI, AB, KWDS)
#2	1181	(randomised control*)@(TI, AB, KWDS)
#3	22713	1, 2
#4	22353	(observational, non random*, nonrandom*, natural experiment*, quasi experiment*, nonexperimental, non experimental, cohort) @(TI, AB, KWDS)
#5	450	3, 4

Search strategy using THESAURUS TERMS

No.	No. papers	Term
#1	3500	(clinical trial)@ KMAJOR
#2	22353	(observational, non random*, nonrandom*, natural experiment*, quasi experiment*, nonexperimental, non experimental, cohort)@(TI, AB, KWDS)
#3	69	1 + 2

SCIENCE CITATION INDEX 1981–96

Search strategy for TITLES

No.	No. papers	Term
#1	39591	(random*)@(TI)
#2	5550	(observational, non random*, nonrandom*, natural experiment*, quasi experiment*, nonexperimental, non experimental, cohort)@(TI)
#3	442	1 + 2

SOCIAL SCIENCE CITATION INDEX 1981–96

Search strategy for TITLES

No.	No. papers	Term
#1	2537	(random*)@(TI)
#2	1739	(observational, non random*, nonrandom*, natural experiment*, quasi

		experiment*, nonexperimental, non experimental, cohort)@(TI)
#3	26	1 + 2

Cochrane Database of Abstracts of Reviews of Effectiveness

Search term 'observational' – 13 papers identified.

Additional search strategies

References from key articles were found

The Citation Index was used to identify work by key authors

Other research teams conducting systematic reviews for the NHS R&D Standing Group on Health Technology were contacted.

Interpretation of evidence

Tables were constructed to collate the findings from the papers that directly compared non-randomised and randomised studies. These were studied and hypotheses were generated to explain the findings. If queries arose, individual authors were contacted for elaboration.

Generalisability of study results

Three aspects of generalisability are particularly relevant: eligibility criteria, participation of centres/practitioners and participation of subjects in trials. Systematic reviews of the literature were performed to assess the extent to which these have been shown to limit the generalisability of randomised trial results.

Literature search

Eligibility

MEDLINE 1966–96

#1	1111	explode "eligibility-determination"/all subheadings
#2	7890	explode "randomized-controlled-trials"/all subheadings
#3	9	#1 and #2
#1	379	eligibility criteria
#2	859	inclusion criteria
#3	596	exclusion criteria
#4	2140	eligibility
#5	4987	eligible
#6	26497	inclusion
#7	13839	exclusion
#8	9458	recruitment
#9	24112	entry
#10	361	ineligible
#11	42	ineligibility
#12	79375	#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11
#13	7890	explode "Randomized-controlled-trials"/all subheadings
#14	459	#12 and #13
#1	379	eligibility criteria
#2	859	inclusion criteria
#3	596	exclusion criteria
#4	7890	explode "Randomized-controlled-trials"/all subheadings
#5	1795	#1 or #2 or #3
#6	79	#4 and #5
#1	7048	Exclusion*
#2	11371	Inclusion*
#3	4972	Eligib*
#4	7165	explode "randomized-controlled-trials"/all subheadings
#5	77005	Women
#6	13897	Gender
#7	77761	Female*
#8	22760	#1 or #2 or #3
#9	285	#8 and #4
#10	151963	#5 or #6 or #7
#11	44	#9 and #10
#1	7048	Exclusion*
#2	11371	Inclusion*
#3	4972	Eligib*
#4	7165	explode "randomized-controlled-trials"/all subheadings
#5	22760	#1 or #2 or #3
#6	285	#5 and #4
#7	24285	Elderly
#8	120954	Old*
#9	760158	Age*
#10	809047	#7 or #8 or #9
#11	161	#6 and #10

Centre Participation

To explore the issue of centre participation, two recently completed systematic reviews that included both RCTs and non-randomised studies were examined to determine the extent to which participation differs by study design.

Patient Participation

MEDLINE 1966–96

MeSH terms:

#1	1459	CLINICAL-TRIALS-METHODS in MJME
#2	1301	PATIENT-PARTICIPATION in MJME
#3	4	#1 and #2

#1	89055	explode "RESEARCH-DESIGN"/ all subheadings
#2	1301	PATIENT-PARTICIPATION in MJME
#3	34	#1 and #2
#1	3267	explode "PATIENT-PARTICIPATION"/ all subheadings
#2	7583	explode "RANDOMISED-CONTROLLED-TRIALS"/all subheadings
#3	31	#1 and #2

Textwords:

#1	12	Recruitment bias

Publication types:

#1	71766	PT = 'RANDOMIZED-CONTROLLED-TRIAL'
#2	11404	PT = 'CONTROLLED-CLINICAL-TRIAL'
#3	168593	PT = 'CLINICAL-TRIAL'
#4	83	NONPARTICIPATION
#5	76	NON-PARTICIPATION
#6	169	NONPARTICIPANTS
#7	112	NON-PARTICIPANTS
#8	168596	#1 or #2 or #3
#9	409	#4 or #5 or #6 or #7
#10	23	#8 and #9

EMBASE 1980–96

#1	8540	(clinical trials)@ Kmajor, Kminor
#2	302	(nonparticip*, non particip*)@TI, AB, KWDS
#3	17	#1 + #2

Science Citation Index 1981–96

#1	72	(nonparticip*, non particip*)@TI

Social Science Citation Index 1981–96

#1	81	(nonparticip*, non particip*)@TI

The role of patient preference in RCTs

This question was approached by means of a review of the literature that had attempted explicitly to measure the effects of patient preference.

Literature retrieval

MEDLINE 1966–96

MeSH terms:

#1	147	explode "research-design"/ all subheadings and patient* preference*
#2	7	Patient-participation in MJME and randomized-controlled-clinical-trials in MJME

Textwords:

#1	257	Patient* preference*
#2	4	Preference arm*
#3	2	Preference trial*
#4	15	Parental preference
#5	167	Patient* choice

Cochrane database

#1	56	"preference"

Interpretation of evidence

Tables were constructed to collate the findings from the papers that attempted to measure the effects of patient preference. An algebraic model was devised to quantify the possible bias that could be introduced by hypothetical preference effects.

Risk adjustment and the ability to exclude confounding in non-randomised studies

Interventions from across the spectrum of health technologies were selected as the methodological issues may differ. The following areas were chosen:

- surgical interventions
- pharmaceutical interventions
- organisational interventions
- preventive interventions.

For each area a specific example was sought to illustrate the problems of confounding and risk adjustment. Selection depended on the existence of a large, well-conducted non-randomised study that measured a treatment effect in a comparable way to a randomised trial, or preferably a meta-analysis of RCTs. The following were selected:

- CABG versus PTCA (surgical intervention)
- calcium antagonists (pharmaceutical intervention)
- stroke units (organisational intervention)
- malaria vaccines (preventive intervention).

Literature searches

CABG versus PTCA

RCTs were found using the following strategy.

MEDLINE 1990–96

Search	Papers identified	Term
#1	1394	CORONARY ARTERY BYPASS

		SURGERY
#2	6213	explode "CORONARY-ARTERY-BYPASS"/ all subheadings
#3	7697	explode "ANGIOPLASTY" / all subheadings
#4	6870	explode "RANDOMIZED-CONTROLLED-TRIALS"/ all subheadings
#5	19158	explode "CLINICAL-TRIALS"/ all subheadings
#6	6575	#1 or #2
#7	1237	#6 and #3
#8	19158	#4 or #5
#9	72	#7 and #8

Non-randomised studies were identified using the following strategy.

Search	Papers identified	Term
#1	1394	CORONARY ARTERY BYPASS SURGERY
#2	6213	explode "CORONARY-ARTERY-BYPASS"/ all subheadings
#3	7697	explode "ANGIOPLASTY" / all subheadings
#4	2167	OBSERVATIONAL
#5	17707	COHORT
#6	929	NON-RANDOM*
#7	1748	NONRANDOM*
#8	311	QUASI-EXPERIMENT*
#9	317	QUASI EXPERIMENT*
#10	76	NONEXPERIMENT*
#11	34	NON-EXPERIMENT*
#12	6575	#1 or #2
#13	1237	#12 and #3
#14	22560	#4 or #5 or #6 or #7 or #8 or #9 or #10 or #11
#15	76	#13 and #14

Calcium antagonists

RCTs were found using the following strategy.

MEDLINE 1990–97

Search	Papers identified	Term
#1	2400	Calcium antagonist
#2	5775	Nifedipine
#3	7165	explode "Randomized-controlled-trials"/ all subheadings
#4	20044	explode "clinical-trials"/ all subheadings
#5	7571	#1 or #2
#6	20044	#3 or #4
#7	181	#5 and #6

Non-randomised studies were identified using the following strategy.

Search	Papers identified	Term
#1	2400	Calcium antagonist
#2	5775	Nifedipine
#3	2293	observational
#4	18624	cohort
#5	976	non-random*
#6	1815	nonrandom*
#7	327	quasi-experiment*
#8	333	quasi experiment*
#9	78	nonexperiment*
#10	36	non-experiment*
#11	7571	#1 or #2
#12	23709	#3 or #4 or #5 or #6 or #7 or #8 or #9 or #10
#13	38	#11 and #12

Stroke units

Randomised trials were identified from a recent systematic review published in the Cochrane Library and supplemented with the search set out below. Subsequently, a more recent systematic review that updated that in the Cochrane Library was published and this was used to update the work.

RCTs were found using the following strategy.

MEDLINE 1990–96

Search	Papers identified	Term
#1	15686	Stroke
#2	53591	Unit
#3	87	Stroke Unit
#4	1877	explode "Cerobrovascular-disorders"/therapy
#5	4426	explode "Randomized-controlled-trials"/ all subheadings
#6	6526	"Random-Allocation"
#7	17035	"Double-blind-method"
#8	2208	"Single-blind-method"
#9	13945	explode "clinical-trials"/ all subheadings
#10	1948	#3 or #4
#11	38527	#5 or #6 or #7 or #8 or #9
#12	64	#10 and #11

Non-randomised studies were identified using the following strategy.

Search	Papers identified	Term
#1	15686	Stroke

#2	53591	Unit
#3	87	Stroke Unit
#4	1877	explode "Cerobrovascular-disorders"/therapy
#5	1942	observational
#6	15519	cohort
#7	807	non-random*
#8	1527	nonrandom*
#9	259	quasi-experiment*
#10	264	quasi experiment*
#11	64	nonexperiment*
#12	25	non-experiment*
#13	1948	#3 or #4
#14	19782	#5 or #6 or #7 or #8 or #9 or #10 or #11 or #12
#15	31	#13 and #15

Malaria vaccines

A recent systematic review of RCTs of malaria vaccine was found in the Cochrane Library. Non-randomised studies were found using the following strategy.

MEDLINE 1990–97

Search	**Papers identified**	**Term**
#1	200	"Malaria-vaccines"/all subheadings
#2	2293	Observational
#3	18624	Cohort
#4	976	Non-random*
#5	1815	Nonrandom*
#6	327	Quasi-experiment*
#7	333	Quasi experiment*
#8	78	Nonexperiment*
#9	36	Non-experiment*
#10	23709	#2 or #3 or #4 or #5 or #6 or #7 or #8 or #9
#11	6	#1 and #10

Electronic database literature retrieval

For each of the above literature searches, abstracts of potentially relevant literature were reviewed (by AB) to ascertain whether they met previously agreed criteria. Sub-sets were also reviewed by co-investigators to assess the reliability of this process. Full papers were retrieved when the abstract was judged to be pertinent.

Appendix 2

Results of RCTs and non-randomised studies

TABLE 17 *ONE RCT compared with ONE non-randomised study*

Study	Intervention type	Topic	Randomised (n)	Non-randomised (n)	Outcome measures	Adjustment	Randomised results	Non-randomised results
CASS Principal Investigators, 1984[30]	Surgical	Coronary artery surgery	390 medical 390 surgical	745 medical 570 surgical (declined participation)	Survival at 6 years	None	Medical: 90% Surgical: 92%	Medical: 88% Surgical: 92%
Hlatky *et al*, 1988[31]	Surgical	CABG	686 (Veterans Administration Cooperative Study)	719 (Duke database)	Survival at 5 years	Cox's proportional hazards model (see paper for factors)	Medical: 78% Surgical:83%	Medical: 80.9% Surgical: 85.5%
			767 (European Trial)	512 (Duke database)			Medical: 84% Surgical: 92%	Medical: 86.3% Surgical: 91.9%
			780 (CASS)	250 (Duke database)			Medical: 92% Surgical: 95%	Medical: 87.2% Surgical: 93.0%
Horwitz *et al*, 1990[32]	Pharmaceutical	Beta-blocker therapy	1916 beta-blockers 1912 no beta-blockers (from BHAT)	Expanded cohort: 626 beta-blockers 433 no beta-blockers	Mortality at 24 months	None	Beta-blockers: 7.3% No beta-blockers: 9.2%	Beta-blockers: 9.3% No beta-blockers: 16.4%
				Restricted cohort: 417 beta-blockers 205 no beta-blockers	Mortality at 24 months	None	Beta-blockers: 7.3% No beta-blockers: 9.2%	Beta-blockers: 7.2% No beta-blockers: 10.7%
				Expanded cohort	Mortality at 24 months	Age adjusted	Beta-blockers: 7.3% No beta-blockers: 9.2%	Beta-blockers: 9.8% No beta-blockers: 15.2%
				Restricted cohort	Mortality at 24 months	Age adjusted	Beta-blockers: 7.3% No beta-blockers: 9.2%	Beta-blockers: 7.6% No beta-blockers: 9.8%
				Expanded cohort	Mortality at 24 months	Age and severity adjusted	Beta-blockers: 7.3% No beta-blockers: 9.2%	Beta-blockers: 10.2% No beta-blockers: 14.4%
				Restricted cohort	Mortality at 24 months	Age and severity adjusted	Beta-blockers: 7.3% No beta-blockers: 9.2%	Beta-blockers: 7.6% No beta-blockers: 9.7%
Paradise *et al*, 1984[33]	Surgical	Tonsillect-omy	43 surgery 48 controls	52 surgery 44 controls (declined participation)	Average no. throat infections per person (year 1)	None	Surgical: 1.24 Control: 3.09	Surgical: 1.77 Control: 3.09
					Average no. throat infections per person (year 2)	None	Surgical: 1.61 Control: 2.66	Surgical: 1.18 Control: 2.50
					Average no. throat infections per person (year 3)	None	Surgical: 1.77 Control: 2.20	Surgical: 1.47 Control: 3.15
Paradise *et al*, 1990[34]	Surgical	Adenoidec-tomy	52 surgery 47 controls	47 surgery 67 controls (declined participation)	Otitis media days/total days in year 1	None	Surgical: 15.0% Control: 28.5%	Surgical: 17.8% Control: 23.3%
					Otitis media days/total days in year 2	None	Surgical: 17.8% Control: 28.4%	Surgical: 16.9% Control: 23.5%
					Mean no. episodes of supporative otits media (year 1)	None	Surgical: 1.06 Control: 1.45	Surgical: 0.90 Control: 1.39
					Mean no. episodes of supporative otits media (year 2)	None	Surgical: 1.09 Control: 1.67	Surgical: 0.59 Control: 1.35

continued

TABLE 17 contd *ONE RCT compared with ONE non-randomised study*

Study	Intervention type	Topic	Randomised (n)	Non-randomised (n)	Outcome measures	Adjustment	Randomised results	Non-randomised results
Schmoor *et al*, 1996[36]	Pharmaceutical	Breast cancer	GBSG trial 2 3 × CMF = 145 6 × CMF = 144	3 × CMF = 72 6 × CMF = 104	Treatment effect of chemo-therapy (RRs)	Cox's model factors: menopausal status; no. nodes; tumour size; tumour grade; oestrogen + progesterone receptor sites	3 × CMF: 1.00 6 × CMF: 0.90 (CI: 0.7–1.2)	3 × CMF: 1.00 6 × CMF: 0.90 (CI: 0.6–1.4)
			3 × CMF + tamoxifen = 93 6 × CMF + tamoxifen = 91	3 × CMF + tamoxifen = 42 6 × CMF + tamoxifen = 29 (declined participation)	Hormonal therapy (RRs)	Cox's model: menopausal status; no. nodes; tumour size; tumour grade; oestrogen + progesterone receptor sites	No tamoxifen: 1.00 With tamoxifen: 0.75. (CI: 0.5–1.04)	No tamoxifen: 1.00 With tamoxifen: 0.53 (CI: 0.3–0.8)
			GBSG trial 3 6 × CMF = 101 6 × CMF + radio-therapy = 98	6 × CMF = 88 6 × CMF + radio-therapy = 41 (declined participation)	Treatment effect of radio-therapy (RRs)	Cox's model: menopausal status; no. nodes; tumour size; tumour grade; oestrogen + progesterone receptor sites	6 × CMF: 1.00 6 × CMF + RT: 0.79 (CI: 0.5–1.3)	6 × CMF: 1.00 6 × CMF + RT: 0.76 (CI: 0.4–1.5)
Yamamoto *et al*, 1992[36]	Surgical	Peptic strictures	16 treatment 15 controls	58 treatment 34 controls (concurrent patients)	Recurrence of dysphagia	None	Treated: 69% Control: 80%	Treated: 88% Control: 94%
					Median time to recurrent dysphagia	None	Treated: 0.35 years Control: 0.26 years	Treated: 0.24 years Control: 0.24 years
					Proportion requiring redilation	None	Treated: 38% Control: 27%	Treated: 43% Control: 50%
					Median time to redilation	None	Treated: 1.2 years Control: 2.4 yearss	Treated: 1.4 years Control: 1.6 years
McKay *et al*, 1995[13]	Organisational	Male alcoholic rehabilita-tion	24 day patients 24 in-patients	65 day patients 31 in-patients (declined participation)	Mean no. of drinking days at 1 year	None	Day patients: 2.8 ± 4.6 In-patients: 6.7 ± 7.2	Day patients: 4.5 ± 7.6 In-patients: 7.0 ± 9.0
					% any days intoxicated (> 3 drinks) at 1 year	None	Day patients: 20.0 In-patients: 55.0	Day patients: 35.1 In-patients: 39.1
					% any days cocaine use at 1 year	None	Day patients: 10.0 In-patients: 20.0	Day patients: 12.3 In-patients: 30.4
					% treated in rehabilitation again at 1 year	None	Day patients: 15.0 In-patients: 10.0	Day patients: 12.3 In-patients: 30.4
					% entered detoxification at 1 year	None	Day patients: 0.0 In-patients: 5.0	Day patients: 1.8 In-patients: 13.0
Nicolaides *et al*, 1994[37]	Diagnostic	Amnio-centesis	238 EA 250 CVS	493 EA; 320 CVS (declined participation)	Survival	None	EA: 91.6% CVS: 95.2%	EA: 92.7% CVS: 93.1%
					Total foetal loss	None	EA: 8.4% CVS: 4.8%	EA: 7.1% CVS: 6.9%
					Spontaneous death	None	EA: 5.9% CVS: 1.2%	EA: 5.1% CVS: 3.1%
					Termination for chromosomal defect	None	EA: 2.1% CVS: 2.4%	EA: 1.8% CVS: 3.4%
					Termination with normal karyotype	None	EA: 0.4% CVS: 1.2%	EA: 0.2% CVS: 0.3%

continued

TABLE 17 contd *ONE RCT compared with ONE non-randomised study*

Study	Intervention type	Topic	Randomised (n)	Non-randomised (n)	Outcome measures	Adjustment	Randomised results	Non-randomised results
Emanuel, 1996[38]	Organisational	Hospice vs. conventional care	247 terminal cancer patients (UCLA Veterans Study)	5853 terminal cancer medicare patients (National Hospice Study)	Costs/ patient % Savings/ patient	Age, sex, cancer type, medical service	Hospice: $16,000 Conventional: $15,493 Approx. 3%	Hospice: $7719 Conventional: $11,729 34% ($p < 0.001$)
Garenne *et al*, 1993[39]	Prevention	Measles vaccine	740 standard Schwarz vaccine 348 controls	1224 standard Schwarz vaccine 4403 controls (National campaign study)	Case-contact efficacy Vaccine efficacy in terms of measles incidence	None Model to control for intensity of exposure (RCT and non-RCT), age at vaccination (non-RCT)	97.2 (CI: 91.3–98.1) 98.0%	92.5 (CI: 88.8–94.6) 97.9% (CI: 91.6–99.5)
Antman *et al*, 1985[40]	Pharmaceutical	Chemotherapy for sarcoma	20 doxorubicin 22 control	21 doxorubicin 27 control (not invited to participate)	Time disease-free Time disease-free	None Location and stage	No significant difference between treatment and control ($p = 0.81$) No significant difference between treatment and control ($p = 0.26$)	Trend toward advantage for treatment ($p = 0.12$) Treatment significantly better ($p = 0.03$)
Shapiro *et al*, 1994[41]	Surgical and pharmaceutical	Breast cancer	Fisher (1985) No details of numbers	Curtis (1984) No details of numbers Harvery (1985) No details of numbers	Occurrence of non-lymphocytic leukaemia after treatment Occurrence of non-lymphocytic leukaemia after treatment	No details No details	RRs: Surgery only: 2.6 Radiation: 10.3 Chemotherapy: 24.0 (significant)	RRs: Surgery only: 1.4 Radiation: 3.7 Chemotherapy: 8.1 (significant) RRs: Surgery only: 1.2 Radiation: 2.5 (significant)
Jha *et al*, 1995[42]	Prevention	Vitamins and cardiovascular disease (CVD)	14,564 vitamin E 14,569 none (ATBC trial)	11,342 vitamin E 75,903 none (Nurses' Health Study)	Death from CVD (+ non-fatal MI in non-RCT)	Age, smoking, alcohol, menopausal status, hormone use, exercise, aspirin, hypertenstion, cholesterol intake, diabetes, caloric intake, vitamin C and beta-carotene intake	RR reduction: 2% (–8–11%)	RR reduction: 31% (3–51%)

TABLE 18 *SEVERAL RCTs combined versus SEVERAL non-randomised studies combined*

Study	Intervention type	Topic	Randomised	Non-randomised	Outcome measures	Adjustment	Randomised results	Non-randomised results
Pyorala *et al*, 1995[43]	Pharmaceutical	Hormonal treatment of crypt-orchidism	11 RCTs combined: 872 boys with 1174 undescended testes	22 studies combined: 2410 boys with 4350 undescended testes	Success rate (descended testes)	No details	LHRH: 21% (CI: 18–24) hCG: 19% (CI: 13–25)	LHRH: 47% (CI: 43–50) hCG: 33% (CI: 31–35)
RMITG, 1994[44]	Prevention	Immuno-therapy for spon-taneous abortion	Nine RCTs combined: 240 treatment 209 controls	Six studies combined: 877 treatment 256 controls	Live birth rates	No details	Treatment: 61.7% Control: 51.7%	Treatment: 59.0% Control: 55.1%
Watson *et al*, 1994[45]	Diagnostic	Hystero-salpingo-graphy for infertile couples	Four RCTs combined: 287 treatment 513 controls	Six studies combined: 1072 treatment 734 controls	Pregnancy odds ratio (treatment vs. controls)	No details	1.89 (CI: 1.33–2.68)	1.92 (CI: 1.55–2.38)
Reimold *et al*, 1992[46]	Prevention	Chronic atrial fibrillation	Six RCTs combined: 373 quinidine 354 controls	Six studies combined: 471 quinidine 290 controls	% patients remaining in sinus rhythm at 3 months	No details	Quinidine: 69.4 ± 4.6% Control: 45.2 ± 5.9%	Quinidine: 44.3 ± 5.2% Control: 35.1 ± 6.4%
					% patients remaining in sinus rhythm at 6 months	No details	Quinidine: 57.7 ± 6.8% Control: 33.3 ± 6.9%	Quinidine: 27.2 ± 4.9% Control: 18.8 ± 6.2%
					% patients remaining in sinus rhythm at 12 months	No details	Quinidine: 50.2 ± 7.6% Control: 24.7 ± 6.8%	Quinidine: 13.7 ± 3.4% Control: 10.9 ± 5.1%
					Crude pooled mortality rate		Quinidine: 2.9% Control: 0.8%	Quinidine: 1.5% Control: 0.3%

Appendix 3

Comparison of effect sizes and directions

TABLE 19 *Papers comparing RCTs and non-observational studies of similar interventions*

Study	Outcome	Adjustment	Treatment effect size		Difference in effect sizes	Significance
			RCT	Non-randomised		
CASS Principal Investigators, 1984[30]	Survival	None	2%	4%	–2%	$p < 0.01$
Paradise *et al*, 1990[33]	Otitis media days/total days in year 1	None	13.5%	5.5%	8.0%	$p < 0.05$
	Otitis media days/total days in year 2	None	10.6%	6.6%	4.0%	NS
Yamamoto *et al*, 1992[36]	Recurrence of dysphagia	None	11%	6%	5%	NS
	Proportion requiring dilation	None	11%	–7%	18%	$p < 0.001$
Horwitz *et al*, 1990[32]	Mortality at 1 year (expanded cohort)	None	1.9%	7.1%	–5.2%	$p < 0.001$
	Mortality at 1 year (restricted cohort)	None	1.9%	3.5%	–1.6%	$p < 0.01$
	Mortality at 1 year (expanded cohort)	Age adjusted	1.9%	5.4%	–3.5%	$p < 0.001$
	Mortality at 1 year (restricted cohort)	Age adjusted	1.9%	2.2%	–0.3%	NS
	Mortality at 1 year (expanded cohort)	Age and severity adjusted	1.9%	4.2%	–2.3%	$p < 0.001$
	Mortality at 1 year (restricted cohort)	Age and severity adjusted	1.9%	2.1%	–0.2%	NS
Schmoor *et al*, 1996[35]	RR success with 6 × CMF (vs. 3 × CMF)	Yes	0.90	0.90	0.00	NS ($p = 0.99$)
	RR success with tamoxifen (vs. no tamoxifen)	Yes	0.75	0.53	0.22	NS ($p = 0.22$)
	RR success with radiotherapy (vs. no radiotherapy)	Yes	0.79	0.76	0.03	NS ($p = 0.94$)
Nicolaides *et al*, 1994[37]	Survival	None	3.6%	0.4%	3.2%	$p < 0.001$
	Total foetal loss	None	3.6%	0.2%	3.4%	$p < 0.001$
	Spontaneous death	None	4.7%	1.6%	3.1%	$p < 0.01$
	Termination for chromosomal defect	None	0.3%	1.6%	–1.3%	$p < 0.05$
	Termination with normal karyotype	None	0.8%	0.1%	0.7%	$p < 0.001$
Garenne *et al*, 1993[39]	Vaccine efficacy	Yes	98.0%	97.9%	0.1%	NS
RMITG, 1994[44]	Live birth rates	No details	10.0%	3.9%	6.1%	$p < 0.001$

For further details of papers and of effects measured see chapter 3

continued

TABLE 19 contd *Papers comparing RCTs and non-observational studies of similar interventions*

Study	Outcome	Adjustment	Treatment effect size		Difference in effect sizes	Significance
			RCT	Non-randomised		
Watson *et al*, 1994[45]	Pregnancy odds (treatment vs. control)	No details	1.89	1.92		NS
Reimold *et al*, 1992[46]	% patients remaining in sinus rhythm at 3 months	No details	24.2%	9.2%	15.0%	$p < 0.001$
	% patients remaining in sinus rhythm at 6 months	No details	24.4%	8.4%	16.0%	$p < 0.001$
	% patients remaining in sinus rhythm at 12 months	No details	25.5%	2.8%	22.7%	$p < 0.001$
	Crude pooled mortality rate	None	2.1%	1.2%	0.9%	$p < 0.001$
Hlatky *et al*, 1988[31]	Survival (VACS)	Yes	5%	4.6%	0.4%	
	Survival (European trial)	Yes	8%	5.6%	2.4%	
	Survival (CASS)	Yes	3%	5.8%	–2.8%	
Paradise *et al*, 1984[33]	Mean no. throat infections per person (year 1)	None	1.85	1.32	0.53	
	Mean no. throat infections per person (year 2)	None	1.05	1.32	–0.27	
	Mean no. throat infections per person (year 3)	None	0.43	1.68	–1.25	
Paradise *et al*, 1990[34]	Otitis media-present days/total in year 1	None	13.5%	5.5%	8.0%	
	Otitis media-present days/total in year 2	None	10.6%	6.6%	4.0%	
	Mean no. episodes of supporative otitis media (year 1)	None	0.39	0.49	–0.1	
	Mean no. episodes of supporative otitis media (year 2)	None	0.58	0.76	–0.18	
McKay *et al*, 1995[13]	Mean no. drinking days at 1 year	None	3.85	2.53	1.32	
	% any days intoxicated (> 3 drinks at 1 year)	None	35.0	4.0	31.0	
	% any days cocaine use at 1 year	None	10.0	18.1	–8.1	
	% treated in rehabilitation again at 1 year	None	–5.0	18.1	–23.1	
	% entered detoxification at 1 year	None	5.0	11.2	–6.2	
Jha *et al*, 1995[42]	Death from CVD (+ non-fatal MI in non-RCT)	Yes	RR reduction: 2%	RR reduction: 31%	–29%	
Pyorala *et al*, 1995[43]	Descended testes	No details	2%	14%	–12%	
Emanuel, 1996[38]	% savings	Yes	3%	34%	31%	

For further details of papers and of effects measured see chapter 3

Appendix 4
Exclusions

TABLE 20 *Studies identifying characteristics of exclusions*

Study	Topic	Method	Findings
Ward *et al,* 1992[84]	Adjuvant chemotherapy in operable stomach cancer	249 trial patients were compared with 960 non-trial patients identified through a cancer registry (and classed as potential trial candidates) 15–74 years with resected gastric carcinoma	Half of the non-trial group would have failed to pass one or more of the exclusion criteria. There was a moderate survival advantage for the trial patients over the non-trial group (p = 0.05), though the benefit was confined to the first 2 years. Removing the non-trial patients who were ineligible, the survival rate was very similar to the trial patients
Muller and Topol, 1990[54]	Thrombolytic therapy for acute myocardial therapy	All RCTs of intravenous thrombolysis in AMI and unstable angina, identified through MEDLINE 1980–90	The reported eligibility ranged from 9% to 51%. When all the trials were combined, the eligibilty was 33%. The actual proportion of patients in USA who received treatment was 18% – so another 15% should get treatment. Comparison between the studies with different eligibilities showed that some were perhaps too strict (e.g. patients over 75 years are often exluded, but in trials where they were included a 33% reduction in mortality was seen)
Williford, 1993[109]	Total perenteral nutrition for malnourished surgical patients	Randomised patients (n = 395) were compared with eligbile refusers (n = 233) and insufficientely malnourised patients who were excluded from the trial (index group, n = 1220)	Patients in the index goup were significantly more healthy than the eligible groups. Septic and non-septic complications were higher in the trial and eligible refuser groups than the excluded patients. The trial results should not be generalised to patients who do not meet the level of malnourishment in the trial patients
The Toronto Leukemia Study Group, 1986[87]	Chemotherapy for acute myeloblastic leukaemia patients	272 consecutive patients; the first 130 had one drug and the next 142 had another (i.e. not a randomised trial)	A difference in exclusion criteria would potentially have had a profound effect on the relative remission rates of the two groups. The remission rate for all group A patients (first drug) = 35%, and for drug B patients = 52%. With all the exclusions (see paper) these changed to 78% and 91%, respectively
Horwitz *et al,* 1990[32]	Beta-blocker therapy after MI	The results from BHAT were compared with a restricted cohort of patients (same eligibility) and also an expanded cohort (ignoring eligibility)	In the expanded cohort the mortality rates for both treatment groups (beta-blockers and none) exceeded the BHAT trial. This is not surprising as the expanded cohort included patients with what are often considered to be contraindications to beta-blockers. A larger treatment effect was also seen, probably due to the contraindications. When adjusted for age and clinical severity the treatment benefit decreased from 43% to 29% in the expanded cohort. This brings the teatment benefit closer to that found in BHAT (22%)
Kober and Torp-Perdersen, 1995[86]	Treatment after AMI	Patients in the TRACE study were compared with those who were screened for entry but were excluded from the study	Patients randomised and those excluded differed substantially. Patients excluded were older (73 years vs. 69 years) and more severely ill. The overall 1-year mortality was 23% in the randomised patients and 54% in those who were excluded
Garber *et al,* 1996	Cholesterol screening in adults	Published RCTs and meta-analysis of cholesterol reduction trials were compared with a model based on the Framingham Heart Study	The benefits of cholesterol reduction were greatest in the groups excluded from the trials, who also had the greatest underlying risk for CHD
DCCT, 1995[83]	Intensive diabetes treatment	Patients in the DCCT trial were compared with a cohort of patients in the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR)	Of the 891 patients aged 13–39 years (the trial target range in the WESDR data), only 39 (4.4%) would have met the eligibility criteria for inclusion in the DCCT primary prevention trial

Appendix 5

Participation

TABLE 21 *RCTs permitting assessment of patient participation: treatment trials*

				Results			
Study	**Topic**	**Sample size**	**Measures**	**Participants**	**Non-participants**	**Significance**	**Authors' conclusions**
Barofsky and Sugarbaker, 1979[104]	Five sarcoma trials	32 non-participants 71 participants	Sex	Female: 26.8%	53.1%	$p = 0.009$	No differences were found in the socioeconomic status of eligible patients who were participators or non-participators It is treatment and treatment-related factors that determine participation
			Race	White: 84.5%	84.4%	$p = 0.99$	
			Age (< 19, 20–49, 50+)	31%, 45%, 24%	16%, 59%, 25%	$p = 0.23$	
			Education	Graduate/prof. training: 8.5% Partial college: 25.4% High-school grad. 25.4% 10th and 11th grade: 21.1% < 9th grade: 19.7%	12.5% 12.5% 34.4% 31.2% 9.3%	$p = 0.28$	
			Surgery at NIH	68.8%	46.4%	$p = 0.055$	
			Feelings about treatment	Adriamycin/ cytoxin regret: 7.0%	30.8%	$p = 0.015$	
			Changes in life since illness	Reduced activities: 81.3% Reduced work status: 50% Fewer friends: 18.8% Reduced social activity: 54.2%	53.6% 53.6% 21.4% 39.3%	$p = 0.01$ $p = 0.76$ $p = 0.78$ $p = 0.21$	
CASS Principal Investigators, 1984[30]	CABG vs. medical therapy	1315 non-participants 780 participants	Mean age	51.2	50.9	NS	The randomised patients are not a special or atypical subset of those eligible for randomisation
			Male	90.3%	90.6%	$p = 0.813$	
			White	98.3%	98.7%	$p = 0.486$	
			Work full-time	67.6%	67.7%	$p = 0.94$	
			Non-exertional angina	4.7%	7.5%	$p = 0.003$	
			Present smoker	39.7%	32.9%	$p = 0.007$	
			Hypertension	31.1%	27.3%	$p = 0.07$	
			Diabetes mellitus	8.7%	6.4%	$p = 0.05$	
			Stroke history	2.1%	1.1%	$p = 0.06$	
			Beta-blockers	43.3%	54.6%	$p < 0.0001$	
			Q-wave MI	29.2%	23.6%	$p = 0.005$	
			Left main artery	1.8%	5.3%	$p = 0.000$	
			Proximal left anterior descending disease	31.5%	35.5%	$p = 0.06$	
			Surgical management	54 days mean wait	53 days mean wait	NS	
			ST depression	9.8%	7.3%	$p = 0.04$	

continued

TABLE 21 contd *RCTs permitting assessment of patient participation: treatment trials*

Study	Topic	Sample size	Measures	Results		Significance	Authors' conclusions
				Participants	Non-participants		
Paradise *et al*, 1984[33]	Tonsillectomy	96 non-participants 91 participants	Age	3–4 years: 13% 5–6 years: 23% 7–15 years: 64%	11% 33% 55%	$p = 0.30$	It seems reasonable to assume that in children meeting the same stringent criteria, tonsillectomy will produce effects comparable to those reported here
			Sex	Male: 44%	50%	$p = 0.41$	
			Race	Black: 11%	5%	$p = 0.15$	
			Throat history	7+ episodes per year: 34% 5 per year for 2 years: 11% 3 per year for 3 years: 55%	32% 18% 50%	$p = 0.42$	
			Infection-free tonsil size	1+: 9% 2+: 35% 3+: 46% 4+: 10%	10% 34% 47% 10%	$p = 0.97$	
			Siblings	None: 13% Younger only: 29% Older only: 44% Both: 14%	26% 38% 28% 8%	$p = 0.02$	
			Parents' profession	Executive/prof.: 18% Clerical/skilled: 33% Semi-skilled/unskilled: 13% Disabled/unemployed: 35% Other: 1%	25% 46% 13% 16% 1%	$p = 0.04$	
			Parents' tonsil history	Neither parent: 20% One parent: 36% Both: 41%	19% 51% 28%	$p = 0.19$	
			Referral source	Unknown: 3% Children's hospital: 47% Community doctor: 41% Parents: 12%	2% 48% 42% 10%	$p = 0.94$	
Hunter *et al*, 1987[105]	Cancer clinical trials	5949 non-participants 3229 participants	Gender	Female (eligible group): 57%	65.4%		Selection for participation affects the characteristics of patients under study and may impact the ability to generalise the results
			Age		Median 59 years	Median 61 years (eligibile)	
			Stage of disease		'Less early disease'	'More early disease'	
Smith and Arnesen[106]	Long-term oral anti-coagulation after MI	270 non-participants 1214 participants	Male	78%	69%	$p < 0.005$	The slightly diverting risk factor profiles of participants and non-consenting subjects in this trial are of yet unknown importance. A future follow-up study may settle this issue
			Ventircular fibrillation	4%	4%	NS	
			Atrial fibrillation	7%	5%	NS	
			Cardiac insufficiency	14%	17%	NS	
			Q-infarcts	67%	70%	NS	
			Age (year, +SD) – overall – males – females	 62 + 9 61 + 9 65 + 8	 64 + 8 63 + 9 67 + 6	 $p < 0.001$ $p < 0.001$ $p < 0.001$	
			ASAT max (IU/1; +SD)	184 + 122	186 + 126	NS	
			Cardiac size (ml/m; +SD)	504 + 93	506 + 105	NS	

continued

TABLE 21 contd *RCTs permitting assessment of patient participation: treatment trials*

Study	Topic	Sample size	Measures	Results: Participants	Results: Non-participants	Significance	Authors' conclusions
continued Smith and Arnesen, 1988[106]		178 non-participants	Previous MI	20%	28%	NS	
			Diabetes	7%	14%	NS	
			No-smokers	45%	51%	NS	
			Ex-smokers	32%	31%	NS	
			Current smokers	23%	17%	NS	
			Diuretics and/or digitalis	33%	38%	NS	
			Beta-blockers	48%	38%	NS	
Harth and Thong, 1990[107]	Childhood asthma (new drug vs. placebo)	42 non-participants 68 participants	Mothers' age	20–29 years: 85% > 30 years: 15%	90% 10%	p = 0.56 (Fisher's)	Parents who volunteer their children for medical research are significantly more socially disadvantaged and emotionally vulnerable
			Fathers' age	20–29 years: 69% > 30 years: 31%	76% 24%	p = 0.51 (Fisher's)	
			Marital status	Married/cohabiting: 92% Single/separated: 9%	95% 5%	p = 0.71 (Fisher's)	
			Mean no. children	2.0	2.0	p = 0.87 (t-test)	
			Birth order	Firstborn: 29% Later born: 71%	27% 72%	p = 0.83 (Fisher's)	
			Mothers' ethnicity	White British/ Australian: 97% Other: 3%	95% 5%	p = 0.64 (Fisher's)	
			Fathers' ethnicity	White British/ Australian: 87% Other: 13%	95% 5%	p = 0.20 (Fisher's)	
			Mothers' education	Primary: 63% Secondary: 22% Tertiary: 15%	55% 19% 26%	p = 0.33 (chi-test)	
			Fathers' education	Primary: 52% Secondary: 32% Tertiary: 16%	24% 31% 45%	p = 0.002 (chi-test)	
			Mothers' occupation	Prof./admin.: 6% Clerical/trade: 15% Labourer: 7% Home duties: 66% Other: 6%	14% 36% 2% 38% 10%	p = 0.02 (chi-test)	
			Fathers' occupation	Prof./admin.: 9% Paraprof./trade: 24% Clerical/sales: 24% Labourer: 33% Other: 6% Unemployed: 4%	31% 19% 17% 19% 12% 2%	p = 0.04 (chi-test)	
			Health problems	Present: 38% Absent: 62%	31% 69%	p = 0.42 (Fisher's)	
			Visits to doctor/clinic	At least weekly: 13% 2–3 times/month: 28% Once a month: 43% Few times/year: 13% Rarely: 3%	5% 7% 10% 48% 30%	p = 0.74 (Fisher's)	

continued

TABLE 21 contd *RCTs permitting assessment of patient participation: treatment trials*

Study	Topic	Sample size	Measures	Results		Significance	Authors' conclusions
				Participants	Non-participants		
continued							
Harth and Thong, 1990[107]			Relation with doctor	Good: 66%	57%	$p = 0.06$	
			Poor: 34%	43%	(Fisher's)		
			Acupuncture/ chiropracty	Yes: 10%	7%	$p = 0.01$	
				No: 90%	93%	(chi-test)	
			Naturopath/ herbalist/iridologist	Yes: 16%	33%	$p < 0.001$	
				No: 84%	67%	(chi-test)	
			Church attendance	Never: 38%	69%	$p < 0.01$	
				Few times/year: 21%	17%	(chi-test)	
				Monthly: 7%	2%		
				Weekly: 34%	12%		
			No. close friends	None: 18%	0%	$p < 0.001$	
				One: 32%	7%	(chi-test)	
				Two: 14%	57%		
				Three: 12%	19%		
				Four: 12%	3%		
				Five or more: 12%	14%		
			No. people available to consult on important decisions	None: 38%	5%	$p < 0.01$	
				One: 25%	5%	(chi-test)	
				Two: 18%	7%		
				Three: 6%	43%		
				Four: 6%	26%		
				Five or more: 7%	14%		
			Desire for more people	Yes: 62%	14%	$p < 0.001$	
				No: 38%	86%	(Fisher's)	
			Cigarette smoking	Smokers: 66%	29%	$p < 0.001$	
				Non-smokers: 34%	71%	(Fisher's)	
			Alcohol consumption	Daily: 7%	2%	$p = 0.17$	
				Few times a week: 13%	5%	(chi-test)	
				Few times a month: 26%	19%		
				Rarely/never: 54%	74%		
			Use of analgesics	Most days: 7%	0%	$p = 0.001$	
				Once or twice a week: 7%	9%	(chi-test)	
				Once or twice a month: 16%	7%		
				Rarely: 64%	48%		
				Never: 6%	36%		
			Use of tranquillisers	Yes: 47%	21%	$p = 0.01$	
				No: 53%	79%	(Fisher's)	

continued

TABLE 21 contd *RCTs permitting assessment of patient participation: treatment trials*

				Results			
Study	**Topic**	**Sample size**	**Measures**	**Participants**	**Non-participants**	**Significance**	**Authors' conclusions**
Ward *et al*, 1992[84]	Stomach cancer	493 non-participants 217 participants	Age: 15–55 years 56–65 years 66–74 years	 30% 35% 35%	 19% 38% 43%	 $p = 0.001$	Comparison of the eligbile cases revealed no unequivocal evidence that the trial patients were a highly selected, good prognosis group
			Male	70%	69%	$p = 0.69$	
			Duration of symptoms: < 0.5 year > 0.5 year	 41% 59%	 54% 46%		
			Stage of disease: 2 3a 3bc	 15% 59% 26%	 18% 53% 29%	$p = 0.96$	
			Curative surgery Palliative surgery	74% 26%	72% 28%	$p = 0.61$	
			Complete excision Tumour left	79% 21%	80% 20%	$p = 0.80$	
			No liver metastasis Liver metastasis	95% 5%	94% 6%	$p = 0.35$	
			No peritoneal metastasis	95%	94%	$p = 0.55$	
			Peritoneal metastasis	5%	6%		
			Site of tumour: Upper Body Lower	 17% 29% 54%	 24% 31% 45%	$p = 0.02$	
			No. sites involved: 1 2 or more	 82% 18%	 71% 29%	$p = 0.002$	
			Gastrectomy: Total Proximal Distal	 27% 6% 67%	 21% 25% 54%	$p = 0.0001$	
			Serosal involvement: Negative Positive	 9% 91%	 6% 94%	$p = 0.17$	
			Lymph node involvement: Negative Positive	 21% 79%	 28% 72%	$p = 0.06$	
			Resection line involvement: Clear Involved	 74% 26%	 66% 34%	$p = 0.08$	
			Pathological stage: 2 3a	 16% 84%	 23% 77%	$p = 0.05$	
			Differentiation: Poor Well	 69% 31%	 61% 39%	$p = 0.07$	
			Size of tumour: 0–5 cm > 5 cm	 42% 58%	 59% 41% (NB: for some measures up to 28% of the non-participants had no information)	$p = 0.001$	

continued

TABLE 21 contd *RCTs permitting assessment of patient participation: treatment trials*

Study	Topic	Sample size	Measures	Results: Participants	Results: Non-participants	Significance	Authors' conclusions
Vollmer *et al*, 1992[108]	Management of paediatric asthma	810 non-participants 244 participants	Severity: Mild Moderate/severe	 28.3% 71.7%	 55.9% 44.1%		Study participation was higher in families of children with severe asthma
Williford *et al*, 1993[109]	Malnourished surgical patients	199 non-participants 395 participants	Male	99.0%	98.5%	NS	The study population was significantly diffferent from the elderly population at large, and this should be considered when attempting to generalise the results of this clinical trial to unselected elderly patients
			Race (white)	69.6%	78.9%	$p = 0.042$	
			Full activity	37.2%	46.2%	NS	
			Severely malnourished	12.8%	6.8%	$p = 0.002$	
			Mean age (SD)	62.5 (10.0)	62.3 (10.3)	NS	
			Mean no. days hospitalised (SD)	7.7 (12.2)	6.7 (9.9)	NS	
			Serum prealbumin (g/dl) (SD)	16.6 (7.8)	17.6 (7.7)	NS	
			Weight (kg) % ideal weight (SD)	66.2 (13.7) 96.2 (17.7)	70.6 (13.2) 101.6 (17.0)	$p < 0.001$ $p < 0.01$	
			Triceps skinfold (mm) (SD)	11.5 (6.2)	12.9 (7.4)	NS	
			Nutritional risk index (SD)	93.2 (6.2)	94.7 (6.0)	$p < 0.05$	
			Hand strength dynamometer (SD)	33.5 (10.0)	36.0 (9.6)	$p < 0.05$	
Kaufmann *et al*, 1994[110]	Self-help groups among people with mental illness	75 non-participants 15 participants	Gender (female)	67%	60%		The results from this study offer some understanding of the characteristics of self-help group participation among people with serious and persistent mental illness
			Schizophrenia	60%	53%		
			Schizoactive	7%	15%		
			Major affective	33%	32%		
			Education	12.4 ± 1.7	12.9 ± 1.8		
			Lifetime admissions	3.1 ± 1.3	3.0 ± 1.1		
			Lifetime months in hospital	16.4 ± 8.0	18.1 ± 8.0		
Stone *et al*, 1994[111]	Non-Hodgkin's lymphoma	32 non-participants 43 participants	Age (median)	59	59.5	NS	The comparison of on-study data of eligible non-trial patients showed very little difference between the two groups
			Histology: follicular, large cell diffuse, small cell diffuse, mixed cell diffuse, large cell large cell, immunoblastic	 0% 7% 16% 74% 2%	 6% 23% 23% 48% 0%	$p = 0.06$	
			Stage: I II III IV	 12% 33% 28% 28%	 0% 29% 16% 55%	$p = 0.02$	
			Internal prognostic index: low low-intermediate high-intermediate high	 41% 33% 10% 15%	 33% 25% 33% 8%	$p = 0.5$	
			Sex ECOG performance Diagnosis			NS NS NS	

continued

TABLE 21 contd *RCTs permitting assessment of patient participation: treatment trials*

Study	Topic	Sample size	Measures	Results: Participants	Results: Non-participants	Significance	Authors' conclusions
Rogers *et al*, 1995[112]	PTCA vs. CABG	2013 non-participants 1829 participants	Mean age (years)	61.5	61.6	NS	Patients eligible for randomisation but not randomised had baseline characteristics generally similar to the randomised cohort, thus ensuring that the randomised patients were representative of all patients meeting the study eligibility criteria
			Gender	Males: 73%	74%	NS	
			Race	White: 90%	94%	$p < 0.01$	
				Black: 6%	4%	NS	
				Other: 3%	2%	NS	
			Education	Grade school: 21%	12%	$p < 0.01$	
				High school: 50%	48%	NS	
				College/tech. school: 18%	20%	NS	
				College grad.: 10%	19%	NS	
				Other degree: 1%	1%	NS	
			Health history	MI: 55%	51%	$p < 0.05$	
				Congestive heart: 9%	5%	$p < 0.01$	
				Diabetes: 25%	22%	$p < 0.05$	
				Family coronary history: 50%	49%	NS	
				Hypertension: 49%	48%	NS	
				Smoking: 71%	67%	$p < 0.05$	
				Stable angina: 30%	32%	NS	
				Unstable angina: 64%	61%	NS	
			Quality of life	Excellent: 8%	12%	$p < 0.01$	
				Very good: 22%	25%	NS	
				Good: 40%	39%	NS	
				Fair: 23%	18%	NS	
				Poor: 7%	6%	NS	
			Activity prior to any event	Sedentary: 15%	12%	$p < 0.05$	
				Mild: 36%	34%	NS	
				Moderate: 42%	46%	NS	
				Strenuous: 7%	8%	NS	
			Body mass index (kg/m^2)	28.0	27.9	NS	
			No. without MI in 6 weeks	1280	1434	NS	
			Total cholesterol obtained	89%	85%	$p < 0.01$	
			Type C lesions	None: 66%	63%	$p < 0.05$	

continued

TABLE 21 contd *RCTs permitting assessment of patient participation: treatment trials*

Study	Topic	Sample size	Measures	Results: Participants	Results: Non-participants	Significance	Authors' conclusions
van Bergen *et al*, 1995[113]	Long-term anticoagulation after MI trial	587 non-participants 350 participants	Age (mean ± SD)	59 ± 11	64 ± 11	$p < 0.0001$	Participants of our multicentre trial differed significantly from eligible non-participants with respect to important prognostic factors, and subsequently, survival
			Male	84%	64%	$p < 0.0001$	
			Previous MI	8%	12%	$p = 0.048$	
			Killip > 1	34%	31%	$p = 0.28$	
			Total atrioventricular block	4%	6%	$p = 0.229$	
			Risk factors:				
			current history	59%	45%	$p < 0.0001$	
			family history	6%	5%	$p = 0.380$	
			diabetes	9%	10%	$p = 0.455$	
			hypertension	20%	21%	$p = 0.741$	
			thrombolysis	33%	29%	$p = 0.245$	
			Medication at discharge:				
			diuretics	17%	35%	$p < 0.0001$	
			ACE inhibitors	11%	17%	$p < 0.005$	
			beta-blockers	36%	39%	$p = 0.406$	
			anticoagulants	50%	77%	$p < 0.0001$	
			Echocardiology performed:	57%	63%	$p = 0.061$	
			end diastolic volume > 55 mm	18%	25%	$p = 0.055$	
			akinesia	80%	81%	$p = 0.854$	
			Left ventricular echocardiology performed:	11%	13%	$p = 0.314$	
			end diastolic volume > 55 mm	8%	5%	$p = 0.568$	
			akinesia	84%	83%	$p = 0.882$	
			Coronary angiography:	11%	14%	$p = 0.187$	
			1-vessel	54%	59%	$p = 0.589$	
			2-vessel	33%	25%	$p = 0.356$	
			3-vessel	10%	10%	$p = 0.915$	
Schmoor *et al*, 1996[35]	Breast cancer clinical trial	247 non-participants 473 participants (GBSG trial 2)	Menopausal status	Pre: 42% Post: 58%	43% 57%	$p = 0.83$ (chi-test)	With respect to the Comprehensive Cohort Study design, comparing prognostic factors of randomised and non-randomised patients made it possible to investigate whether the former were representative of all eligible patients
			No. involved lymph nodes	≤ 3: 57% 4–9: 30% > 9: 13%	53% 30% 17%	$p = 0.30$ (chi-test)	
			Tumour size	< 20 mm: 28% 21–30 mm: 41% > 30 mm: 31%	24% 43% 33%	$p = 0.51$ (chi-test)	
			Tumour grade	I: 12% II: 66% III: 22%	12% 63% 25%	$p = 0.64$ (chi-test)	
			Oestrogen receptor	> 20 fmol: 60% < 20 fmol: 40%	65% 35%	$p = 0.18$ (chi-test)	
			Progesterone receptor	> 20 fmol: 59% < 20 fmol: 41%	63% 37%	$p = 0.28$ (chi-test)	
			Tumour location	Lateral: 66% Medial: 34%	57% 43%	$p = 0.02$ (chi-test)	

continued

TABLE 21 contd *RCTs permitting assessment of patient participation: treatment trials*

Study	Topic	Sample size	Measures	Results: Participants	Results: Non-participants	Significance	Authors' conclusions
continued Schmoor *et al*, 1996[35]		129 non-participants 199 participants (GBSG trial 3)	Menopausal status	Pre: 38% Post: 62%	41% 59%	$p = 0.60$ (chi-test)	
			No. involved lymph nodes			$p = 0.56$ (chi-test)	
				≤ 3: 61%	59%		
				4–9: 27%	32%		
				>9: 12%	9%		
			Tumour size	< 20 mm: 30%	30%	$p = 0.85$ (chi-test)	
				21–30 mm: 41%	38%		
				> 30 mm: 29%	32%		
			Tumour grade	I: 12%	11%	$p = 0.77$ (chi-test)	
				II: 60%	57%		
				III: 28%	32%		
			Oestrogen receptor	> 20 fmol: 61%	53%	$p = 0.15$ (chi-test)	
				< 20 fmol: 39%	47%		
			Progesterone receptor			$p = 0.11$ (chi-test)	
				> 20 fmol: 55%	46%		
				< 20 fmol: 45%	54%		
			Tumour location	Lateral: 55%	64%	$p = 0.09$ (chi-test)	
				Medial: 45%	36%		
Gorkin *et al*, 1996[114]	Antiarrhythmia medications after MI	139 non-participants 260 participants	Sex	Male: 84%	64%	$p < 0.001$	Multivariate analyses of patient factors revealed that having a higher income and being disabled from work were the stongest predictors of participation
			Age	60.6 ± 10.0	63.9 ± 9.7	$p = 0.002$	
			Job status:			$p = 0.007$	
			employed	36%	30%		
			homemaker	4%	12%		
			disabled	17%	10%		
			retired	39%	47%		
			unemployed	4%	1%		
			Medical insurance status:			$p = 0.002$	
			none	14%	2%		
			private	49%	57%		
			Medicare	17%	24%		
			Medicare and private	19%	17%		
			Veterans association	1%	0%		
			Ventricular tachycardia on Holter:			$p = 0.025$	
			present	25%	14%		
			absent	75%	86%		
			Race	White: 80%	91%	$p = 0.054$	
			Education			NS	
			Marital status			NS	
			Smoking status			NS	
			Depression level			NS	
			Functional status			NS	
			Perceived support	9.7 ± 2.0	9.3 ± 1.9	$p = 0.013$	
			Life stress			NS	

TABLE 22 *RCTs permitting assessment of patient participation: prevention trials*

Study	Topic	Sample size	Measures	Results		Significance	Authors' conclusions
				Participants	Non-participants		
Constantine *et al*, 1993[115]	Infant Health and Development Program (for low birthweight, premature infants)	317 non-participants 985 participants	Research site	Harvard: 14.0%	33.0%	$p < 0.01$	These differences were not large enough to have any practical effect on the representiveness of the study sample to the population, and had no effect on the comparability of the treatment groups
			Birthweight	< 1501 g: 26.0% 1501–2000 g: 37.3% ≥ 2011 g: 36.7%	18.9% 33.8% 47.3%	$p = 0.002$	
			Maternal race	Black: 52.5% Hispanic: 10.7% Caucasian: 36.9%	31.6% 7.9% 60.4%	$p < 0.001$	
			Gender	No details	No details	NS	
			Single vs. multibirth	No details	No details	NS	
			Maternal age	No details	No details	NS	
			Maternal education	No details	No details	NS	
Naslund *et al*, 1994[116]	Diet and exercise trial	27 non-participants 158 participants	Age	No details	No details	NS	None of the personality dimensions studied was associated with willingness to join the intervention trial. Nor was there any relationship between willingness to participate in the trial and the demographic variables investigated
			Marital status	No details	No details	NS	
			Children	No details	No details	NS	
			Occupation	No details	No details	NS	
			% workers	27%	20%	NS	
			Smokers	No details	No details	NS	
			Exercise	No details	No details	NS	
			Body mass index	No details	No details	NS	
			Blood pressure	No details	No details	NS	
			Serum cholesterol	No details	No details	NS	
			Serum-triglycerides	No details	No details	NS	
			Family history of CHD	No details	No details	NS	

continued

TABLE 22 contd *RCTs permitting assessment of patient participation: prevention trials*

Study	Topic	Sample size	Measures	Results: Participants	Results: Non-participants	Significance	Authors' conclusions
Davies *et al*, 1994[117]	Family Heart Study	608 non-participating families (information was obtained for 106 families) 1448 participating families	Housing	Owner occupier: 90% Tenant: 10%	83% 17%	$p = 0.04$ $p = 0.2$	With the exception of smoking, our results generally do not support the concern that screening programmes are attended by those who need them least
			Employment status	Employed: 84% Unemployed: 5% Retired: 5% Sick or other: 5%	83% 9% 5% 3%		
			Access to car	None: 11% One: 53% More than one: 35%	19% 40% 35%	$p = 0.008$	
			Personal history of disease	Men: 5% Women: 2%	5% 0%	$p = 0.8$ $p = 0.4$	
			Family history of disease	Men: 27% Women: 30%	34% 33%	$p = 0.3$ $p = 0.9$	
			Life-long smoker	Men: 31% Women: 54%	21% 38%	$p > 0.2$ $p = 0.2$	
			Current smoker	Men: 24% Women: 21%	28% 36%	$p = 0.2$ $p = 0.02$	
			Former smoker	Men: 36% Women: 24%	42% 26%	$p > 0.2$ $p > 0.2$	
			Systolic bp (mmHg)	Men: 137.2 Women: 128.1	137.7 128.0	$p > 0.2$ $p > 0.2$	
			Diastolic bp (mmHg)	Men: 86.2 Women: 80.8	85.7 879.4	$p > 0.2$ $p > 0.2$	
			Cholesterol (mmol/l)	Men: 5.72 Women: 5.54	5.91 5.39	$p > 0.2$ $p > 0.2$	
			Glucose (mmol/l)	Men: 5.35 Women: 5.31	5.45 5.41	$p > 0.2$ $p > 0.2$	
			Body mass index	Men: 26.1 Women: 25.2	25.8 25.6	$p > 0.2$ $p > 0.2$	
Yeomans-Kinney *et al*, 1995[118]	Breast Cancer Prevention Trial	127 non-participants 105 participants	Current age	< 50 years: 51.4% > 50 years: 48.6%	33.9% 66.1%	$p = 0.07$	These findings support the view that recruitment efforts for chemo-prevention trials should address barriers specific to their circumstances
			Marital status	Currently married: 71.4% Formerly married: 19.0% Never married: 9.5%	70.9% 24.4% 4.7%	$p = 0.262$	
			Education	High school or less: 12.4% Post-high school: 32.4% College grad.: 55.2%	22.8% 29.1% 40.8%	$p = 0.113$	
			Household income	< $35,000: 20.6% $35,000–50,000: 21.6% > $50,000: 57.8%	32.2% 16.1% 51.7%	$p = 0.135$	
			Currently employed	Yes: 37.1% No: 62.9%	48.8% 51.2%	$p = 0.098$	
			Gail risk score (mean ± SD)	18.2 ± 9.87	15.7 ± 8.51	NS	
			% with prior breast biopsy	55%	48%	NS	
			Family history of breast cancer	Mother: 65% Sister: 34%	65% 34%		

continued

TABLE 22 contd *RCTs permitting assessment of patient participation: prevention trials*

Study	Topic	Sample size	Measures	Results: Participants	Results: Non-participants	Significance	Authors' conclusions
Pacala *et al*, 1996[119]	Fatalities and injuries in the over 75s	139 non-participants (responded to non-respondent survey) 110 participants	Gender	Female: 41.8%	65.5%	$p < 0.001$	Recruiting older subjects by mail to studies of rigorous interventions can produce significant selection biases that may limit the population to which the results may be generalised
			Age:			$p = 0.260$	
			75–79 years	60.9%	52.2%		
			80–84 years	27.3%	29.1%		
			85+ years	11.8%	18.7%		
			Married	60.9%	42.5%	$p = 0.002$	
			Living alone	32.1%	45.2%	$p = 0.030$	
			Living in a house	81.8%	63.2%	$p < 0.001$	
			Education:			$p < 0.001$	
			0–11 years	6.4%	15.7%		
			12 years	21.1%	39.6%		
			13–15 years	25.7%	23.1%		
			16+ years	46.8%	23.6%		
			Health and functional status:				
			good/excellent	92.2%	79.4%	$p = 0.006$	
			no biopsychosocial	76.7%	60.7%	$p = 0.009$	
			not susceptible	90.3%	78.1%	$p = 0.0012$	
			exercise	69.9%	50.3%	$p = 0.002$	
			Current smoker	7.80%	8.80%	$p = 0.774$	
			Balance, gait and falls: walks independently	96.4%	81.8%	$p < 0.001$	
			gait better/ no change	62.5%	69.3%	$p = 0.204$	
			balance better/ no change	72.1%	73.0%	$p = 0.918$	
			fell in past year	30.8%	23.7%	$p = 0.216$	
			not afraid of falling	51.0%	51.1%	$p = 0.984$	

Appendix 6

RCTs comparing CABG and PTCA

TABLE 23

Study	Methods	% of eligibles randomised	Single or multi-vessel	Exclusions	No. patients		Follow- up	Outcomes
					CABG	PTCA		
CABRI, 1995,[136] Europe	RCT multicentre	4.60	Multi	Left main coronary disease or severe triple vessel. EF < 0.35, AMI in previous 10 days, previous CABG/PTCA, 76 years+, other life-shortening conditions	513	541	1 year	Mortality, revascularisation, medication, angina
RITA, 1993[137] UK	RCT multicentre	4.80	Single: 456 Multi: 555	Left main stem disease, previous CABG/PTCA, other life-threatening comorbidity	501	510	4.7 years	Mortality, MI, revascularisation, angina, exercise, employment
EAST, 1994[138] USA	RCT single centre	46.65	Multi	2 months chronic occlusions of bypass-able vessels serving viable myocardium, left main disease, EF < 0.25%, MI in previous 5 days, other life-threatening illness	194	198	3 years	Death, MI, ischaemic defects, other clinical and angio-graphic status measures, need for additional revascularisation
GABI, 1994[139] Germany	RCT multicentre	Approx. 66	Multi	75 years+, left-main disease, 30% stenosis, 50% left ventricular circumference in jeopardy, previous PTCA/CABG, MI in previous 4 weeks	177	182	1 year+	Length of hospital stay, in-hopsital and 1-year mortality, MI, angina, exercise capacity
Toulouse, 1992[140] France	RCT single centre	Not given	Multi	Not given	76	76	2.8 years	Revascularisation and death
MASS, 1995[141] Brazil	RCT single centre	Not given	Single	Unstable angina, prior infarction, significant valve disease, cardiomyopathy or prior open heart surgery	70	72	3.2 years	MI, revascularisation, stroke, mortality, exercise, angina, employment
Lausanne, 1994,[142] Switzerland	RCT single centre	94%	Single	Unstable angina, previous MI, abnormal creatine kinase activity, left stenosis 50+, EF < 50%, 60 years +	66	68	3.2 years	Death, MI, revascularisation, angina, exercise
ERACI, 1993[143] Argentina	RCT single centre	42%	Multi	Dilated ischaemic cardiomyopathy, severe left main trunk stenosis, severe 3-vessel disease plus depressed EF, severe valvular heart disease or hypertrophic disease, evolving AMI, limited life expectancy	64	63	3.2 years	Death, MI, revascularisation costs
BARI, 1996[147] USA and Canada	RCT multicentre	48%	Multi	Published elsewhere	914	915	Not given	In-hospital and 5-year mortality, MI, stroke, revascularisation

Appendix 7

Non-randomised studies comparing CABG and PTCA

TABLE 24

Study	Methods	Single or multi-vessel	Exclusions	No. patients		Follow-up	Outcomes
				CABG	PTCA		
Duke database, 1996[144] USA	Prospective database from one medical centre	Single and multi	50% left main stenosis, previous CABG/PTCA, 3+ or 4+ mitral regurgitation	3890	2924	Max: 10 years	Mortality
Medicare, 1992[145] USA	National data on all Medicare patients and detailed clinical data on a random sample (MedisGroups)	Single and multi	None	National: 71,243 Sub-group: 2063	National: 25,423 Sub-group: 858	Mean 18 months	Mortality

Appendix 8

RCTs evaluating calcium antagonists

TABLE 25

Study	Methods	Participants	% of eligibles randomised	Intervention	Exclusions	No. patients		Dosage mg/day	Follow-up
						Calcium antagonist	Control		
Gorfon *et al*, 1984[153] South Africa	RCT single blind	Patients with history and electro-cardiographic changes character-istic of acute transmural MI, admitted within 12 hours of onset of chest pain	No details	10 mg of nifedipine sublingually Dose repeated every 6 hours for 24 hours Control patients: 80 mg of furosemide if pulmonary artery wedge pressure > 18 mmHg	Sustained ventricular arrhythmias; systolic arterial pressure < 90 mmHg; current antiarrhythmic, beta-blocker, digitalis or calcium antagonist therapy	13	13	40	12 hours
TRENT, 1986[154] UK	RCT (A)	Patients aged 18–70 years, admitted to hospital within 24 hours of onset of chest pain	78–100	10 mg of nifedipine four times a day	Pregnancy, or ability to get pregnant within next 4 weeks; arterial blood pressure < 100 mmHg systolic or 50 mmHg diastolic immediately before administration; heart rate > 120/min immediately before administration; severe heart failure; known serious renal or hepatic dysfuntion; current calcium channel blocking drugs	2240	2251	40	1 year
Walker *et al*, 1988[155] UK	RCT (A)	Patients admitted with suspected MI within 6 hours of onset of chest pain	Approx. 77%	10 mg nifedipine or placebo every 4 hours sublingually for 24 hours, then every 4 hours orally for another 24 hours	Age over 75 years; systolic blood pressure < 85 mmHg	106	120	60	48 hours
Sirmes *et al*, 1984[156] Norway	RCT (A)	Patients with severe chest pain for at least 30 min. continuing on admission and admitted within 12 hours of onset	Approx. 89	10 mg nifedipine five times a day for 2 days and then 10 mg four times a day for 6 weeks	Age < 35 or > 75 years; use of calcium antagonist within last 48 hours; other serious disease; inability to attend 6 week follow-up	112	115	50	6 weeks
Erbel *et al*, 1988[157] Germany	RCT (A)	Patients admitted to hospital within 6 hours of onset of pain	No details	2 × 10 mg capsules of nifedipine sublingually, or 2 × 10 mg capsules of placebo. Then 20 mg nifedipine or placebo orally three times a day	Long period of resuscitation; history of allergy to streptokinase; previous cerebrovascular accident; surgery in preceding 10 days; history of peptic ulcer; history of bleeding problems	74	75	60	20 days
Goldbourt, 1993[158] Israel	RCT (A)	Patients aged 50–79 years who presented with suspected AMI	Approx. 93 randomised to first stage	Nifedipine, 6 × 10 mg per day for 6 days. Then 15 mg four times a day for 6 months	Sbp < 90 mmHg; known intolerance of nifedipine; heart disease other than coronary; previous heart surgery or AMI; other major disease; anginal syndrome in month before current AMI; history of hypertension; functional capacity class II or higher in month before; anterior site of presenting AMI; maximal level of serum lactate dehydrogenase > 3 × normal	680	678	60	6 months

A – concealment of allocation secure

continued

TABLE 25 contd

Study	Methods	Participants	% of eligibles randomised	Intervention	Exclusions	No. patients		Dosage mg/day	Follow-up
						Calcium antagonist	Control		
HINT, 1986[159] The Netherlands	RCT (A)	Patients with suspected AMI; a history of angina at rest in previous 12 hours, lasting > 15 min; a history of MI or unstable angina; at least 50% narrowing of major coronary artery	No details	Nifedipine, 6 × 10 mg a day plus metoprolol placebo, or metoprolol, 2 × 100 mg a day plus nifedipine placebo, or both drugs. If beta-blockers taken > 3 days, placebo or nifedipine 6 × 10 mg a day	Age > 70 years; new Q-wave; AMI within 1 week; treatment with nifedipine; heart rate < 50 or > 120; systolic blood pressure > 170 mmHg, diastolic blood pressure > 100 mmHg; anaemia; heart failure; congenital or valvar heart failure; cardiomyopathy; pulmonary or other non-cardiac disease; previous trial participation	341	327	60	48 hours
SPRINT 1, 1988[160] Israel	RCT (A)	Patients aged 30–74 years with a recent MI	47.9	Nifedipine, 30 mg a day	Presence of Prinzmetal's variant angina, non-CHD; previous cardiac surgery or pace-maker implantation; severe pulmonary hypertension; uncontrollable congestive heart failure preceding in recent AMI; persistent hypotension; complete left-bundle-branch block; cerebrovascular accident; malignant disease; renal or hepatic failure; alcoholism or psychiatric disorder	1130	1146	30	1 year
Branagan *et al*, 1986[161] Ireland	RCT (A)	Patients under 70 years, admitted to hospital within 6 hours of onset of chest pain	15.3	Nifedipine, 10 mg orally every 6 hours	Women of childbearing age; heart rate > 110 or < 60; complete heart block heart failure requiring diuretic therapy or vasodilator; inotropic or mechanical support; systolic blood pressure < 85 mmHg, ventricular fibrillation as the initial rhythm; already on nifedipine	60	68	40	1 month

A – concealment of allocation secure

Appendix 9

Non-randomised study evaluating calcium antagonists

TABLE 26

Study	Methods	Participants	No. patients		Follow-up	Outcomes	Adjustments
			Calcium antagonist	Control			
Braun *et al*, 1996[151] Israel	Data were collected in patients screened for the BIP Study, in 18 cardiology departments 1990–92. Mortality data were obtained by matching patients' identification with life status in Israeli Population Registry	Patients with an established diagnosis of chronic artery disease. See paper for baseline characteristics	5843 Diltiazem: 57% Nifedipine: 34% Verapamil: 6% Combination: 3%	5732	Mean 3.2 years (range: 2.0–4.6)	Mortality rates	Age Gender Previous MI Angina pectoris Hypertension New York Heart Association Functional Class Peripheral vascular disease Chronic obstructive pulmonary disease Diabetes Current smoking Concomitant medications

Appendix 10

RCTs evaluating stroke units

TABLE 27

Study	Methods	Participants	Exclusions	Interventions	Outcomes
Dover, 1984[194,195]	RCT (B)	Stroke patients up to 9 weeks after stroke onset (majority within 3 weeks)	Patients only eligible if investigated by general physicians and deemed "fit for and needing rehabilitation". Numbers eligible not stated	Stroke rehabilitation ward vs. general medical and geriatric medical wards	Death, functional status, place of residence and length of stay in hospital up to 1 year after stroke
Edinburgh, 1980[196–198]	RCT (B)	Stroke patients (moderate severity) within 7 days of stroke onset	Not stated	Stroke rehabilitation ward (acute care and rehabilitation) vs. general medical wards	Death, functional status, place of residence and length of initial hospital admission up to 1 year after stroke
Montreal, 1984[199]	RCT (B)	Unselected stroke patients within 7 days of stroke onset	443 screened for entry Exclusions: incorrect diagnosis (40); onset > 7 days (31); transient ischaemic attack (89); previous cerebrovascular accident with residual disability (35); no motor or sensory involvement of limbs (39); "placement in hospital" (33); non-resident of Montreal (3); refused consent (1)	Mobile stroke team vs. conventional care on general medical wards	Death, functional status, place of residence and length of initial hospital stay up to 6 weeks after stroke
Orpington, 1993[200,201]	RCT (B)	Stroke patients who had survived 2 weeks	377 eligible Exclusions: sub-dural haematoma (2); brain tumour (7); died before trial entry (79); mild and discharged before trial entry (37). 67% included	Stroke rehabilitation ward vs. general medical and geriatric wards	Death, functional status, place of residence and length of initial hospital stay assessment/ rehabilitation at end of follow-up
Trondheim, 1991[202]	RCT (A)	Stroke patients within 7 days of stroke onset	373 eligible Exclusions: missed by researchers (16); unconscious (42); symptoms > 1 week (12); patients living in nursing homes (21); patients with sub-dural or sub-arachnoid haemorrhage or brain tumour (?); living in other district (15); stroke unit full (47). 220 (< 59%) included	Acute/rehabilitation stroke unit vs. general medical wards	Death, functional status, place of residence and length of stay in hospital/institution up to 1 year after stroke
Umea, 1985[203,204]	Quasi-RCT (C) Treatment allocation according to bed availability	Stroke patients within 7 days of stroke onset Non-intensive stroke unit vs. general medical wards	Included "all meeting admission criteria for stroke unit"	Death, functional status, place of residence and length of initial stay in hospital up to 1 year after stroke	Treatment allocation according to bed availability

A – concealment of allocation secure; B – concealment possibly insecure; C – not adequately concealed

Appendix 11

RCTs evaluating malarial vaccines

TABLE 28

Study	Methods	Participants	% of eligibles randomised	Intervention	Exclusions	No. patients		Follow -up
						Vaccine	Control	
Alonso, 1994[173] Tanzania	RCT (A)	586 children, aged 1–5 years living in southern Tanzania, an area of intense perennial transmission	89.7	Three doses of SPf66, 2 mg per dose, at weeks 0, 4 and 25. Placebo: aluminium hydroxide + tetanus toxoid on the same schedule	History of allergies leading to medical consultation and treatment; acute conditions that lead to hospital admission; 'unsuitable' chronic conditions; packed cell volume < 25%	274	312	1 year
D'Alessandro, 1995[174] The Gambia	RCT (A)	547 infants aged 6–11 months living in the Gambia, an area of seasonal malaria with moderate transmission	96.7	Three doses of Spf66 (1 mg per dose) given at weeks 0, 4 and 26. Placebo: imovax polio given on same schedule	Weight for age < 60%; chronic diseases	316	231	3.5 months
Sempertegui, 1994[172] Ecuador	RCT (B) age stratified (five levels)	537 adults and children over 1 year old, living in La T, Ecuador, an area highly endemic for malaria	No details	Three doses of SPf66 adsorbed to aluminium hydroxide (2 mg in 0.5 ml for > 5 years, 1 mg in 0.25 ml in < 5 years) on days 0, 30 and 180. Placebo: tetanus toxoid adsorbed onto aluminium hydroxide for the first dose, aluminium hydroxide alone for the second and third doses	Age under 1 year; pregnancy; history of allergy; acute infection; renal; cardio-vascular or endocrine chronic diseases; refusals	259	278	1 year
Valero, 1993[170] Columbia	RCT (A)	1548 inhabitants of La Tola, Columbia, aged over 1 year. This area has perennial transmission with fluctuating incidence	81.5% received first dose, 62.0% received all three doses	Three doses of SPf66 on days 0, 30 and 180. 0.5 ml if > 5 years, 0.25 ml if < 5 years (adsorbed onto aluminium hydroxide, 4 mg/ml). Placebo: tetanus toxoid for first dose, saline in aluminium hydroxide for second and third doses	Prgnancy; children under 1 year; history of allergy or another 'unsuitable' acute or chronic condition	738	810	1 year
Valero, 1996[171] Columbia	RCT (A) stratified for age (five levels) and sex	1257 residents aged 1–86 years, from 14 villages along Rio Rosardio, Columbia, an area endemic for malaria	85.7% received first dose, 68.9% received all three doses	Three doses of Spf66, adsorbed to aluminium hydroxide (2 mg per dose, 1 mg for < 5 years) on days 0, 30 and 180. Placebo: tetanus toxoid at first dose, aluminium hydroxide at second and third doses	Pregnancy; those with parasitaemia; history of allergy; acute infection; renal, cardiovascular or endocrinological chronic disease	634	623	22 months

A – concealment of allocation secure; B – concealment possibly insecure

Appendix 12

Non-randomised study evaluating malarial vaccines

TABLE 29

Study	Methods	Participants	Intervention	Exclusions	No. patients		Follow-up
					Vaccine	Control	
Noya, 1993[169] South Venezuela	The eligible population were invited to attend a vaccination clinic. Attendees were compared with the rest of the population	Persons aged over 11 years, living in 13 small villages in South Venezuela, an endemic area for malaria with seasonal transmission	Three doses on days 0, 20 and 112. Each dose consisted of 4 mg/ml SPf66 adsorbed to aluminium hydroxide	Pregnant women; severe health problems; mental disorders; alcoholism; history of allergies or debilitating diseases; immunosuppressive drugs; and those with clinical symptoms suggestive of malaria	852	825	1 year

Health Technology Assessment panel membership

This report was identified as a priority by the Methodology Panel.

Acute Sector Panel

Chair: Professor John Farndon, University of Bristol †

Professor Senga Bond, University of Newcastle-upon-Tyne †

Professor Ian Cameron, Southeast Thames Regional Health Authority

Ms Lynne Clemence, Mid-Kent Health Care Trust †

Professor Francis Creed, University of Manchester †

Professor Cam Donaldson, University of Aberdeen

Mr John Dunning, Papworth Hospital, Cambridge †

Professor Richard Ellis, St James's University Hospital, Leeds

Mr Leonard Fenwick, Freeman Group of Hospitals, Newcastle-upon-Tyne †

Professor David Field, Leicester Royal Infirmary †

Ms Grace Gibbs, West Middlesex University Hospital NHS Trust †

Dr Neville Goodman, Southmead Hospital Services Trust, Bristol †

Professor Mark P Haggard, MRC †

Mr Ian Hammond, Bedford & Shires Health & Care NHS Trust

Professor Adrian Harris, Churchill Hospital, Oxford

Professor Robert Hawkins, University of Bristol †

Dr Gwyneth Lewis, Department of Health †

Dr Chris McCall, General Practitioner, Dorset †

Professor Alan McGregor, St Thomas's Hospital, London

Mrs Wilma MacPherson, St Thomas's & Guy's Hospitals, London

Professor Jon Nicholl, University of Sheffield †

Professor John Norman, University of Southampton

Dr John Pounsford, Frenchay Hospital, Bristol †

Professor Gordon Stirrat, St Michael's Hospital, Bristol

Professor Michael Sheppard, Queen Elizabeth Hospital, Birmingham †

Dr William Tarnow-Mordi, University of Dundee

Professor Kenneth Taylor, Hammersmith Hospital, London

Diagnostics and Imaging Panel

Chair: Professor Mike Smith, University of Leeds †

Professor Michael Maisey, Guy's & St Thomas's Hospitals, London *

Professor Andrew Adam, UMDS, London †

Dr Pat Cooke, RDRD, Trent Regional Health Authority

Ms Julia Davison, St Bartholomew's Hospital, London †

Professor Adrian Dixon, University of Cambridge †

Mr Steve Ebdon-Jackson, Department of Health †

Professor MA Ferguson-Smith, University of Cambridge †

Dr Mansel Hacney, University of Manchester

Professor Sean Hilton, St George's Hospital Medical School, London

Mr John Hutton, MEDTAP International Inc., London

Professor Donald Jeffries, St Bartholomew's Hospital, London †

Dr Andrew Moore, Editor, *Bandolier* †

Professor Chris Price, London Hospital Medical School †

Dr Ian Reynolds, Nottingham Health Authority

Professor Colin Roberts, University of Wales College of Medicine

Miss Annette Sergeant, Chase Farm Hospital, Enfield

Professor John Stuart, University of Birmingham

Dr Ala Szczepura, University of Warwick †

Mr Stephen Thornton, Cambridge & Huntingdon Health Commission

Dr Gillian Vivian, Royal Cornwall Hospitals Trust †

Dr Jo Walsworth-Bell, South Staffordshire Health Authority †

Dr Greg Warner, General Practitioner, Hampshire †

Methodology Panel

Chair: Professor Martin Buxton, Brunel University †

Professor Anthony Culyer, University of York *

Dr Doug Altman, Institute of Health Sciences, Oxford †

Professor Michael Baum, Royal Marsden Hospital

Professor Nick Black, London School of Hygiene & Tropical Medicine †

Professor Ann Bowling, University College London Medical School †

Dr Rory Collins, University of Oxford

Professor George Davey-Smith, University of Bristol

Dr Vikki Entwistle, University of Aberdeen †

Professor Ray Fitzpatrick, University of Oxford †

Professor Stephen Frankel, University of Bristol

Dr Stephen Harrison, University of Leeds

Mr John Henderson, Department of Health †

Mr Philip Hewitson, Leeds FHSA

Professor Richard Lilford, Regional Director, R&D, West Midlands †

Mr Nick Mays, King's Fund, London †

Professor Ian Russell, University of York †

Professor David Sackett, Centre for Evidence Based Medicine, Oxford †

Dr Maurice Slevin, St Bartholomew's Hospital, London

Dr David Spiegelhalter, Institute of Public Health, Cambridge †

Professor Charles Warlow, Western General Hospital, Edinburgh †

* Previous Chair
† Current members

continued

continued

Pharmaceutical Panel

Chair: Professor Tom Walley, University of Liverpool †

Professor Michael Rawlins, University of Newcastle-upon-Tyne*

Dr Colin Bradley, University of Birmingham

Professor Alasdair Breckenridge, RDRD, Northwest Regional Health Authority

Ms Christine Clark, Hope Hospital, Salford †

Mrs Julie Dent, Ealing, Hammersmith & Hounslow Health Authority, London

Mr Barrie Dowdeswell, Royal Victoria Infirmary, Newcastle-upon-Tyne

Dr Tim Elliott, Department of Health †

Dr Desmond Fitzgerald, Mere, Bucklow Hill, Cheshire

Dr Felicity Gabbay, Transcrip Ltd †

Dr Alistair Gray, Health Economics Research Unit, University of Oxford †

Professor Keith Gull, University of Manchester

Dr Keith Jones, Medicines Control Agency

Professor Trevor Jones, ABPI, London †

Ms Sally Knight, Lister Hospital, Stevenage †

Dr Andrew Mortimore, Southampton & SW Hants Health Authority †

Mr Nigel Offen, Essex Rivers Healthcare, Colchester †

Dr John Posnett, University of York

Mrs Marianne Rigge, The College of Health, London †

Mr Simon Robbins, Camden & Islington Health Authority, London †

Dr Frances Rotblat, Medicines Control Agency †

Mrs Katrina Simister, Liverpool Health Authority †

Dr Ross Taylor, University of Aberdeen †

Dr Tim van Zwanenberg, Northern Regional Health Authority

Dr Kent Woods, RDRD, Trent RO, Sheffield †

Population Screening Panel

Chair: Professor Sir John Grimley Evans, Radcliffe Infirmary, Oxford †

Dr Sheila Adam, Department of Health*

Ms Stella Burnside, Altnagelvin Hospitals Trust, Londonderry †

Dr Carol Dezateux, Institute of Child Health, London †

Dr Anne Dixon Brown, NHS Executive, Anglia & Oxford †

Professor Dian Donnai, St Mary's Hospital, Manchester †

Dr Tom Fahey, University of Bristol †

Mrs Gillian Fletcher, National Childbirth Trust †

Professor George Freeman, Charing Cross & Westminster Medical School, London

Dr Mike Gill, Brent & Harrow Health Authority †

Dr JA Muir Gray, RDRD, Anglia & Oxford RO †

Dr Anne Ludbrook, University of Aberdeen †

Professor Alexander Markham, St James's University Hospital, Leeds †

Professor Theresa Marteau, UMDS, London

Dr Ann McPherson, General Practitioner, Oxford †

Professor Catherine Peckham, Institute of Child Health, London

Dr Connie Smith, Parkside NHS Trust, London

Dr Sarah Stewart-Brown, University of Oxford †

Ms Polly Toynbee, Journalist †

Professor Nick Wald, University of London †

Professor Ciaran Woodman, Centre for Cancer Epidemiology, Manchester

Primary and Community Care Panel

Chair: Dr John Tripp, Royal Devon & Exeter Healthcare NHS Trust †

Professor Angela Coulter, King's Fund, London *

Professor Martin Roland, University of Manchester *

Dr Simon Allison, University of Nottingham

Mr Kevin Barton, East London & City Health Authority †

Professor John Bond, University of Newcastle-upon-Tyne †

Ms Judith Brodie, Age Concern, London †

Dr Nicky Cullum, University of York †

Professor Shah Ebrahim, Royal Free Hospital, London

Mr Andrew Farmer, Institute of Health Sciences, Oxford †

Ms Cathy Gritzner, The King's Fund †

Professor Andrew Haines, RDRD, North Thames Regional Health Authority

Dr Nicholas Hicks, Oxfordshire Health Authority †

Professor Richard Hobbs, University of Birmingham †

Professor Allen Hutchinson, University of Sheffield †

Mr Edward Jones, Rochdale FHSA

Professor Roger Jones, UMDS, London

Mr Lionel Joyce, Chief Executive, Newcastle City Health NHS Trust

Professor Martin Knapp, London School of Economics & Political Science

Dr Phillip Leech, Department of Health †

Professor Karen Luker, University of Liverpool

Professor David Mant, NHS Executive South & West †

Dr Fiona Moss, Thames Postgraduate Medical and Dental Education †

Professor Dianne Newham, King's College London

Professor Gillian Parker, University of Leicester †

Dr Robert Peveler, University of Southampton †

Dr Mary Renfrew, University of Oxford

Ms Hilary Scott, Tower Hamlets Healthcare NHS Trust, London †

* Previous Chair
† Current members

Health Technology Assessment 1998; Vol. 2: No. 13

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

ISSN 1366-5278