Health Technology Assessment 2003; Vol. 7: No. 27

Evaluating non-randomised intervention studies

JJ Deeks
J Dinnes
R D'Amico
AJ Sowden
C Sakarovitch
F Song
M Petticrew
DG Altman

Health Technology Assessment
NHS R&D HTA Programme

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.

Evaluating non-randomised intervention studies

JJ Deeks[1*]
J Dinnes[2]
R D'Amico[1]
AJ Sowden[3]
C Sakarovitch[1]
F Song[4]
M Petticrew[5]
DG Altman[1]

In collaboration with the International Stroke Trial and the European Carotid Surgery Trial Collaborative Groups

[1] Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK
[2] Southampton Health Technology Assessments Centre, University of Southampton, UK
[3] NHS Centre for Reviews and Dissemination, University of York, UK
[4] Department of Public Health and Epidemiology, University of Birmingham, UK
[5] MRC Social and Public Health Sciences Unit, University of Glasgow, UK

* Corresponding author

Declared competing interests of authors: none

Published September 2003

This report should be referenced as follows:

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;**7**(27).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE.

NHS R&D HTA Programme

The NHS R&D Health Technology Assessment (HTA) Programme was set up in 1993 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 provide care in the NHS.

Initially, six HTA panels (pharmaceuticals, acute sector, primary and community care, diagnostics and imaging, population screening, methodology) helped to set the research priorities for the HTA Programme. However, during the past few years there have been a number of changes in and around NHS R&D, such as the establishment of the National Institute for Clinical Excellence (NICE) and the creation of three new research programmes: Service Delivery and Organisation (SDO); New and Emerging Applications of Technology (NEAT); and the Methodology Programme.

The research reported in this monograph was identified as a priority by the HTA Programme's Methodology Panel and was funded as project number 96/26/99.

The views expressed in this publication are those of the authors and not necessarily those of the Methodology Programme, HTA Programme 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 any recommendations made by the authors.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

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 and publisher have tried to ensure the accuracy of this report but do not accept liability for damages or losses arising from material published in this report. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.

M

Abstract

Evaluating non-randomised intervention studies

JJ Deeks,[1*] J Dinnes,[2] R D'Amico,[1] AJ Sowden,[3] C Sakarovitch,[1] F Song,[4] M Petticrew[5] and DG Altman[1]

In collaboration with the International Stroke Trial and the European Carotid Surgery Trial Collaborative Groups

[1] Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK

[2] Southampton Health Technology Assessments Centre, University of Southampton, UK

[3] NHS Centre for Reviews and Dissemination, University of York, UK

[4] Department of Public Health and Epidemiology, University of Birmingham, UK

[5] MRC Social and Public Health Sciences Unit, University of Glasgow, UK

[*] Corresponding author

Objectives: To consider methods and related evidence for evaluating bias in non-randomised intervention studies.

Data sources: Systematic reviews and methodological papers were identified from a search of electronic databases; handsearches of key medical journals and contact with experts working in the field. New empirical studies were conducted using data from two large randomised clinical trials.

Methods: Three systematic reviews and new empirical investigations were conducted. The reviews considered, in regard to non-randomised studies, (1) the existing evidence of bias, (2) the content of quality assessment tools, (3) the ways that study quality has been assessed and addressed. (4) The empirical investigations were conducted generating non-randomised studies from two large, multicentre randomised controlled trials (RCTs) and selectively resampling trial participants according to allocated treatment, centre and period.

Results: In the systematic reviews, eight studies compared results of randomised and non-randomised studies across multiple interventions using meta-epidemiological techniques. A total of 194 tools were identified that could be or had been used to assess non-randomised studies. Sixty tools covered at least five of six pre-specified internal validity domains. Fourteen tools covered three of four core items of particular importance for non-randomised studies. Six tools were thought suitable for use in systematic reviews. Of 511 systematic reviews that included non-randomised studies, only 169 (33%) assessed study quality. Sixty-nine reviews investigated the impact of quality on study results in a quantitative manner. The new empirical studies estimated the bias associated with non-random allocation and found that the bias could lead to consistent over- or underestimations of treatment effects, also the bias increased variation in results for both historical and concurrent controls, owing to haphazard differences in case-mix between groups. The biases were large enough to lead studies falsely to conclude significant findings of benefit or harm. Four strategies for case-mix adjustment were evaluated: none adequately adjusted for bias in historically and concurrently controlled studies. Logistic regression on average increased bias. Propensity score methods performed better, but were not satisfactory in most situations. Detailed investigation revealed that adequate adjustment can only be achieved in the unrealistic situation when selection depends on a single factor.

Conclusions: Results of non-randomised studies sometimes, but not always, differ from results of randomised studies of the same intervention. Non-randomised studies may still give seriously misleading results when treated and control groups appear similar in key prognostic factors. Standard methods of case-mix adjustment do not guarantee removal of bias. Residual confounding may be high even when good prognostic data are available, and in some situations adjusted results may appear more biased than unadjusted results. Although many quality assessment tools exist and have been used for appraising non-randomised studies, most omit key quality domains. Healthcare policies based upon non-randomised studies or systematic reviews of non-randomised studies may need re-evaluation if the uncertainty in the true evidence base was not fully appreciated when policies were made. The inability of

case-mix adjustment methods to compensate for selection bias and our inability to identify non-randomised studies that are free of selection bias indicate that non-randomised studies should only be undertaken when RCTs are infeasible or unethical. Recommendations for further research include: applying the resampling methodology in other clinical areas to ascertain whether the biases described are typical; developing or refining existing quality assessment tools for non-randomised studies; investigating how quality assessments of non-randomised studies can be incorporated into reviews and the implications of individual quality features for interpretation of a review's results; examination of the reasons for the apparent failure of case-mix adjustment methods; and further evaluation of the role of the propensity score.

Contents

List of abbreviations

BCG	bacille Calmette–Guérin vaccine
CASP	Critical Appraisal Skills Programme
CHD	coronary heart disease
CINAHL	Cumulative Index to Nursing and Allied Health Literature
CT	computed tomography
CV	covariate
DARE	Database of Abstracts of Reviews of Effectiveness
ECST	European Carotid Surgery Trial
ERIC	Educational Resources Information Centre database
HCT	historically controlled trial
IST	International Stroke Trial
log OR	logarithm of the odds ratio
LR	logistic regression
MA	meta-analysis
MI	myocardial infarction
MSK CRG	Cochrane Musculoskeletal Collaborative Review Group
NASCET	North American Symptomatic Carotid Endarterectomy Trial
NRS	non-randomised study
OR	odds ratio
PS	propensity score
QA	quality assessment
RCT	randomised controlled trial
RR	risk ratio
SD	standard deviation
SE	standard error
TB	tuberculosis
TENS	transcutaneous electrical nerve stimulation
TIA	transient ischaemic attack

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices in which case the abbreviation is defined in the figure legend or at the end of the table.

Executive summary

Background

In the absence of randomised controlled trials (RCTs), healthcare practitioners and policy-makers rely on non-randomised studies to provide evidence of the effectiveness of healthcare interventions. However, there is controversy over the validity of non-randomised evidence, related to the existence and magnitude of selection bias.

Objectives

To consider methods and related evidence for evaluating bias in non-randomised intervention studies.

Methods

1. Three reviews were conducted to consider:
   - empirical evidence of bias associated with non-randomised studies
   - the content of quality assessment tools for non-randomised studies
   - the use of quality assessment in systematic reviews of non-randomised studies.

   These reviews were conducted systematically, identifying relevant literature through comprehensive searches across electronic databases, handsearches and contact with experts.
2. New empirical investigations were conducted generating non-randomised studies from two large, multicentre RCTs by selectively resampling trial participants according to allocated treatment, centre and period. These were used to examine:
   - systematic bias introduced by the use of historical and non-randomised concurrent controls
   - whether results of non-randomised studies are more variable than results of RCTs
   - the ability of case-mix adjustment methods to correct for selection bias introduced by non-random allocation.

The resampling design overcame particular problems of meta-confounding and variability of direction and magnitude of bias that hinder the interpretation of previous reviews.

Results

Empirical comparisons of randomised and non-randomised evidence

Eight studies compared results of randomised and non-randomised studies across multiple interventions using meta-epidemiological techniques. The studies reached conflicting conclusions, explicable by differences in:

- whether data were sourced from primary studies or systematic reviews
- consideration of meta-confounding
- inclusion of studies of varying quality
- criterion for classifying discrepancies in results.

The only deducible conclusions were (a) results of randomised and non-randomised studies sometimes, but not always, differ and (b) both similarities and differences may often be explicable by other confounding factors.

Quality assessment tools for evaluating non-randomised studies

We identified 194 tools that could be or had been used to assess non-randomised studies. Around half were scales and half checklists, most were published within systematic reviews and most were poorly developed with scant attention paid to principles of scale development.

Sixty tools covered at least five of six pre-specified internal validity domains (creation of groups, blinding, soundness of information, follow-up, analysis of comparability, analysis of outcome), although the degree of coverage varied. Fourteen tools covered three of four core items of particular importance for non-randomised studies (How allocation occurred? Was the study designed to generate comparable groups? Were prognostic factors identified? Was case-mix adjustment used?). Six tools were thought suitable for use in systematic reviews.

Use of quality assessment in systematic reviews of non-randomised studies

Of 511 systematic reviews that included non-randomised studies, only 169 (33%) assessed study quality. Many used quality assessment tools designed for RCTs or developed by the authors themselves, and did not include key quality criteria relevant to non-randomised studies. Sixty-nine reviews investigated the impact of quality on study results in a quantitative manner.

Empirical estimates of bias associated with non-random allocation

The bias introduced by non-random allocation was noted to have two components. First, the bias could lead to consistent over- or underestimations of treatment effects. This occurred for historical controls, the direction of bias depending on time trends in the case-mix of participants recruited to the study. Second, the bias increased variation in results for both historical and concurrent controls, owing to haphazard differences in case-mix between groups. The biases were large enough to lead studies falsely to conclude significant findings of benefit or harm.

Empirical evaluation of case-mix adjustment methods

Four strategies for case-mix adjustment were evaluated: none adequately adjusted for bias in historically and concurrently controlled studies. Logistic regression on average increased bias. Propensity score methods performed better, but were not satisfactory in most situations. Detailed investigation revealed that adequate adjustment can only be achieved in the unrealistic situation when selection depends on a single factor. Omission of important confounding factors can explain underadjustment. Correlated misclassifications and measurement error in confounding variables may explain the observed increase in bias with logistic regression, as may differences between conditional and unconditional odds ratio estimates of treatment effects.

Conclusions

Results of non-randomised studies sometimes, but not always, differ from results of randomised studies of the same intervention. Non-randomised studies may still give seriously misleading results when treated and control groups appear similar in key prognostic factors. Standard methods of case-mix adjustment do not guarantee removal of bias. Residual confounding may be high even when good prognostic data are available, and in some situations adjusted results may appear more biased than unadjusted results.

Although many quality assessment tools exist and have been used for appraising non-randomised studies, most omit key quality domains. Six tools were considered potentially suitable for use in systematic reviews, but each requires revision to cover all relevant quality domains.

Healthcare policies based upon non-randomised studies or systematic reviews of non-randomised studies may need re-evaluation if the uncertainty in the true evidence base was not fully appreciated when policies were made.

The inability of case-mix adjustment methods to compensate for selection bias and our inability to identify non-randomised studies which are free of selection bias indicate that non-randomised studies should only be undertaken when RCTs are infeasible or unethical.

Recommendations for further research

- The resampling methodology utilised here should be applied in other clinical areas to ascertain whether the biases we describe are typical.
- Efforts should be focused on developing quality assessment tools for non-randomised studies, possibly by refining existing tools.
- Research should consider how quality assessment of non-randomised studies can be incorporated into reviews, and the implications of individual quality features for interpretation of a review's results.
- Reasons for the apparent failure of case-mix adjustment methods should be further investigated, including assessments of the generalisability of our results to risk assessments and epidemiological studies, and differences between conditional and unconditional estimates of treatment effects.
- The role of the propensity score should be further evaluated, and computer macros made available for its application.

Chapter 1
Introduction

The ultimate goal of the evaluation of healthcare interventions is to produce a valid estimate of effectiveness, in terms of both internal and external validity. Internal validity concerns the extent to which the results of a study can be reliably attributed to the intervention under evaluation, whereas external validity concerns the extent to which a study's results can be generalised beyond the given study context.

The randomised controlled trial (RCT) is widely regarded as the design of choice for the assessment of the effectiveness of healthcare interventions. The main benefit of the RCT is the use of a randomisation procedure that, when properly implemented, ensures that the allocation of any participant to one treatment or another cannot be predicted. The randomisation process makes the comparison groups equal with respect to both known and unknown prognostic factors at baseline, apart from chance bias.[1] RCTs also tend to benefit from so-called 'inherited properties', which generally mark them out as higher quality studies. These properties include the fact that they are prospective studies, with written protocols specifying, and thus standardising, important aspects of patient enrolment, treatment, observation and analysis.[2] RCTs are also more likely to employ specific measures to reduce or remove bias, such as blinded outcome assessment.

There are instances where non-randomised studies have either been sufficient to demonstrate effectiveness or they appear to have arrived at results similar to those of RCTs. However, where randomisation is possible, most agree that the RCT should be the preferred method of evaluating effectiveness.[2–4] The risks of relying solely on non-randomised evidence include failing to convince some people of the validity of the result, or successfully convincing others of an incorrect result.[3]

Nevertheless, several scenarios remain under which an RCT may be unnecessary, inappropriate, impossible or inadequate.[5] Examples include the assessment of rare side-effects of treatments, some preventive interventions and policy changes. Furthermore, there must be hundreds of examples of interventions for which RCTs would be possible but have not yet been carried out, leaving the medical and policy community to rely on non-randomised evidence. It is therefore essential to have an understanding of the biases that may influence non-randomised studies.

Types of non-randomised studies

A taxonomy of study designs that may be used to assess the effectiveness of an intervention is provided in *Box 1*. However, there is inconsistent use of nomenclature when describing non-randomised studies, and other taxonomies may apply different definitions to the same study designs. To attempt to avoid the problems of inconsistent terminology, six features can be identified that differentiate between these studies. First, some studies make comparisons between groups, whilst some simply describe outcomes in a single group (e.g. case series). Second, the comparative designs differ in the way that participants are allocated to groups, varying from the use of randomisation (RCTs), quasi-randomisation, geographical or temporal factors (cohort studies), the decisions of healthcare professionals (clinical database cohorts), to the identification of groups with specific outcomes (case–control studies). Third, studies differ in the degree to which they are prospective (and therefore planned) or retrospective, for matters such as the recruitment of participants, collection of baseline data, collection of outcome data and generation of hypotheses. Fourth, the method used to investigate comparability of the groups varies: in RCTs no investigation is necessary (although it is often carried out), in controlled before-and-after designs baseline outcome measurements are used, and in cohort and case–control studies investigation of confounders is required. Fifth, studies differ in the level at which the intervention is applied: sometimes it is allocated to individuals, other times to groups or clusters.

Finally, some studies are classified as experimental whereas others are observational. In experimental studies the study investigator has some degree of control over the allocation of interventions. Most importantly, he/she has control over the allocation

BOX 1 *Taxonomy of study designs to assess the effectiveness of an intervention*

Experimental designs
A study in which the investigator has control over at least some study conditions, particularly decisions concerning the allocation of participants to different intervention groups.

1. ***Randomised controlled trial***
   Participants are randomly allocated to intervention or control groups and followed up over time to assess any differences in outcome rates. Randomisation with allocation concealment ensures that on average known and unknown determinants of outcome are evenly distributed between groups.
2. ***Quasi-randomised trial***
   Participants are allocated to intervention or control groups by the investigator, but the method of allocation falls short of genuine randomisation and allocation concealment (e.g. allocated by date of birth, hospital record number, etc.)
3. ***Non-randomised trial/quasi-experimental study***
   The investigator has control over the allocation of participants to groups, but does not attempt randomisation (e.g. patient or physician preference). Differs from a 'cohort study' in that the intention is experimental rather than observational.

Observational designs
A study in which natural variation in interventions (or exposure) among study participants is investigated to explore the effect of the interventions (or exposure) on health outcomes.

4. ***Controlled before-and-after study***
   A follow-up study of participants who have received an intervention and those who have not, measuring the outcome variable both at baseline and after the intervention period, comparing either final values if the groups are comparable at baseline, or change scores. It can also be considered an experimental design if the investigator has control over, or can deliberately manipulate, the introduction of the intervention.
5. ***Concurrent cohort study***
   A follow-up study that compares outcomes between participants who have received an intervention and those who have not. Participants are studied during the same (concurrent) period either prospectively or, more commonly, retrospectively.
6. ***Historical cohort study***
   A variation on the traditional cohort study where the outcome from a new intervention is established for participants studied in one period and compared with those who did not receive the intervention in a previous period, i.e. participants are not studied concurrently.
7. ***Case–control study***
   Participants with and without a given outcome are identified (cases and controls respectively) and exposure to a given intervention(s) between the two groups compared.
8. ***Before-and-after study***
   Comparison of outcomes from study participants before and after an intervention is introduced. The before and after measurements may be made in the same participants, or in different samples. It can also be considered an experimental design if the investigator has control over, or can deliberately manipulate, the introduction of the intervention.
9. ***Cross-sectional study***
   Examination of the relationship between disease and other variables of interest as they exist in a defined population at one particular time point.
10. ***Case series***
    Description of a number of cases of an intervention and outcome (no comparison with a control group).

Adapted from CRD Report 4.[175]

of participants to intervention groups either using randomisation of participants, or haphazard allocation by alternation, dates of birth, day of the week or case record numbers. The key to a well-controlled experimental study is the concealment of the allocation schedule from the study investigators, such that the allocation of any given participant cannot be predicted.[6] When this is not ensured, major bias may be introduced. In observational studies, on the other hand, the groups that are compared are generated according to variation in the use of interventions that occurs regardless of the study. When allocation is determined largely by health professionals, the treatment decision is based not only on 'hard' data such as age, sex and diagnostic test results, but also on 'soft' data, including type and severity of symptoms, rate of development of the illness, and severity of any co-morbid conditions, which are rarely made explicit.[7] Allocation in observational

TABLE 1 *Sources of bias*

Source of bias	RCTs	Cohort studies
Selection bias	Randomisation	Control for confounders
Performance bias	Blinding (of participants and/or investigators)	Measurement of exposure
Attrition bias	Completeness of follow-up	Completeness of follow-up
Detection bias	Blinded outcome assessment	Blinded outcome assessment

studies may also be based on factors such as availability of care or geographical location. In observational studies, therefore, there are likely to be systematic differences in the case-mix of patients allocated to the intervention and comparison groups.

Allocation to groups can also be based on patient choice, as in patient preference trials.[8] Individual preferences for one treatment above another may well imply differences in other consistent ways (potentially relating to prognosis), from those who do not hold such a preference.[9] In addition, preference for a particular treatment may enhance its therapeutic effect.

Sources of bias in non-randomised studies

The Cochrane Collaboration handbook has laid out the four main sources of systematic bias in trials of the effects of healthcare as being selection bias, performance bias, attrition bias and detection bias (*Table 1*). All of these biases can affect non-randomised studies and are discussed in this report. However, it is selection bias that we discuss in the greatest detail, and evaluate in new empirical studies, as it is the potential for selection bias that most clearly differentiates randomised from non-randomised studies.

Selection bias

The greatest distinction between the results of randomised and non-randomised studies is, as described above, the risk of **selection bias**, where systematic differences in comparison groups arise at baseline. The term selection bias can be misleading as it is used to describe both biased selection of participants for inclusion in a study (which applies to both experimental and observational studies) and biased allocation of patients to a given intervention (which occurs where randomisation is not used). The first type of selection bias is usually classified as an issue of external validity and is not discussed further in this report. Rather, we consider the second type, which is an issue of internal validity. It is sometimes referred to as case-mix bias, or confounding.

In non-randomised studies, selection bias will be introduced when participants chosen for one intervention have different characteristics from those allocated to the alternative intervention (or not treated). For observational studies the choice of a given intervention is largely at the discretion of the treating clinician (as would occur in normal clinical practice). The choice of an intervention under these circumstances will be influenced not only by a clinician's own personal preference for one intervention over another but also by patient preference, patient characteristics and clinical history. Sometimes the reasons for the choice of a treatment will be obvious, but at other times a clinician's treatment decision will be influenced by subtle clues that are not easily identifiable.[3] This may result in treatment groups that are incomparable, often with one intervention group 'heavily weighted by the more severely ill'.[10] According to Miettinen,[11] in clinical practice (and therefore in observational studies):

> "Interventions are commonly prompted by an indication, a state or event that signifies the prospect of an untoward outcome. Thus, by the very rationality of decisions to intervene, the treated tend to differ from the untreated with respect to their outlooks for the outcome criterion in efficacy assessment; there tends to be *confounding by the indication* – usually such that the treated tend to have less favourable outcome than the untreated."

In other words, when faced with a patient who may be eligible to receive a given intervention, the decision to treat will be influenced by some factor that in turn is related to the treatment outcome. This introduces systematic bias leading to either over- or underestimates of treatment effects, depending on the treatment decision mechanism.

Confounding by indication can take several guises and the term has been used to describe a number of situations. The original definition refers to the situation where an extraneous condition is both a

clinical indication for the application of the intervention under study and an indication for the outcome under study. For example, arrested labour is a factor that may directly contribute to the decision to introduce fetal monitoring, that is, arrested labour is an indication for the monitoring procedure. Any observational study would be expected to show a higher proportion of arrested labours among the monitored participants. Since arrested labour is also an indication for Caesarean section, we would expect a higher Caesarean frequency among monitored labours, even if monitoring had no effect.[12] Where the indication (arrested labour) influences the use of the intervention (fetal monitoring) and is a risk factor for the outcome (Caesarean section), there will be an imbalance in prognosis between the treatment groups being compared.

Confounding by severity (or by prognosis) may be considered as a special form of confounding by indication. It occurs where the severity of the extraneous condition influences both the introduction of the intervention and the outcome under study,[13] that is, any treatment reserved for the most ill will be associated with the worst outcomes (even where the treatment is effective). The underlying deterioration of the treated disease predicts the use of a given treatment, which in turn becomes associated with the progression of the disease.[14] For example, the frequent use of β-agonists predicts death from asthma because those with most severe asthma are more likely to be prescribed β-agonists.

Protopathic bias is a term coined by Horwitz and Feinstein[15] to describe situations where the first symptoms of a given outcome are the reason for treatment initiation: "Protopathic bias" occurs "when a pharmaceutical or other therapeutic agent is **inadvertently prescribed** for an early manifestation of a disease that has not yet been diagnostically detected" (our emphasis). For example, a drug given for abdominal pain may be wrongly associated with hepatic injury, as abdominal pain may be one of the prodromal symptoms.[13]

Unfortunately, in practice, the 'subtle clues' prompting the selection of a particular intervention for a given patient are often unrecognised and unrecorded. When discussing the use of observational databases, Byar[16] pointed out that "I have yet to see an analysis presented today (or in my whole life for that matter) that had really good information on why some patients got one treatment and others got another".

One approach to improving the design of non-randomised studies is the 'restricted cohort' design first developed by Horwitz and Feinstein[15] and later refined by the same group.[17] This approach involves restricting the eligibility criteria of cohort studies to those used in clinical trials, defining a 'zero-time' point from which patients are followed up, and using an approximation of intention-to-treat analysis. They have demonstrated, for example using a study of β-blocker therapy after acute myocardial infarction,[17] that similar results could be obtained to those reported by RCTs. This approach does not appear to have been widely taken up, and the original proposal has received some criticism.[3]

The degree to which non-random allocation methods are susceptible to selection bias and therefore may produce biased estimates of the effect of treatment is not clearly understood, although it seems likely that the potential for bias will vary between clinical areas. It is reasonable to expect that the comparability of the groups in terms of prognostic factors, and the extent to which prognosis influences both selection for treatment **and** treatment outcome will be of particular relevance. For example, in evaluations of childhood vaccines there are few indicators of prognosis that could be used to influence allocation, so randomisation may not be necessary (although there are dangers that allocation in clusters could be confounded by exposure to infectious disease, which means that in practice randomisation is recommended). By contrast, when patient factors could have a strong influence on allocation or where prognosis is strongly linked to outcomes (such as in cancer treatment), then randomisation is likely to be extremely important.

Other biases in non-randomised studies

Non-randomised studies are also susceptible to attrition, detection and performance bias. For example, **attrition bias** will occur if there are drop-outs, **detection bias** if the assessment of outcomes is not standardised and blinded, and **performance bias** if there are errors and inconsistencies in the allocation, application and recording of interventions (*Table 1*). All of these biases can also occur in RCTs, but there is perhaps potential for their impact to be greater in non-randomised studies which are usually undertaken without protocols specifying standardised interventions, outcome assessments and data recording procedures. A comprehensive assessment of the validity of both randomised and non-randomised studies requires the assessment of all dimensions of potential bias.

Case-mix adjustment methods

In the absence of information on factors influencing allocation, the traditional solution to removing selection bias in non-randomised studies has been to attempt to control for known prognostic factors, either by design and/or by analysis. Alternative statistical techniques for removing bias in non-randomised studies are briefly described by D'Agostino and Kwan[1] and include matching, stratification, covariance adjustment and the propensity score analysis (*Box 2*). However, all statistical techniques make technical assumptions (regression models typically assume that the relationship between the prognostic variable and the outcome is linear) and the degree to which they can adequately adjust for differences between groups is unclear.

Moses[18] lists three factors necessary for successful adjustment: (1) knowledge of which variables must be taken into account; (2) measuring those variables in each participant; and (3) using those measurements appropriately to adjust the treatment comparison. He further states that we are likely to fail on all three counts:[18]

> "We often don't understand the factors that cause people's disease to progress or not; even if we knew those factors, we might find they had not been measured. And if they were measured, the correct way to use them in adjustment calls for a theoretical understanding we seldom have."

BOX 2 *Commonly used methods of statistical adjustment*

Standardisation
Participants are analysed in groups (strata) which have similar characteristics, the overall effect being estimated by averaging the effects seen in each of the groups.

Regression
Relationships between prognostic factors and outcome are estimated from the data in hand, and adjustments calculated for the difference in average values of the prognostic factor between the two groups. Linear regression (or covariance analysis) is used for continuous outcomes, logistic regression for binary outcomes.

Propensity scores
Propensity probabilities are calculated for each participant from the data set, estimating their chance of receiving treatment according to their characteristics. Treatment effects are estimated either by comparing groups that have similar propensity scores (using matching or stratification methods), or by calculating a regression adjustment based on the difference in average propensity score between the groups.

Use of adjustment therefore assumes that researchers know which are the most important prognostic factors and that these have been appropriately measured. Further, it cannot address the problem of unknown or unmeasurable prognostic factors, which may play a particular role in confounding by indication. Full adjustment for confounding by indication would require information on prognostic factors influencing both disease progression and **treatment choice**.[14] Where there is an association between prognosis and treatment choice, it would seem that correlates of prognosis can only warn of a problem, not control for it. Furthermore, if the degree to which the markers are correlated with prognosis is different in diseased and non-diseased persons then the magnitude (and direction) of the resulting bias will be unpredictable in an overall analysis.[14]

Scope of this project

This report contains the results of three systematic reviews and two empirical studies all relating to the internal validity of non-randomised studies. The findings of these five studies will be of importance to researchers undertaking new studies of healthcare evaluations and systematic reviews, and also healthcare professionals, consumers and policy makers looking to use the results of randomised and non-randomised studies to inform their decision-making.

The first systematic review looks at existing evidence of bias in non-randomised studies, critically evaluating previous methodological studies that have attempted to estimate and characterise differences in results between RCTs and non-randomised studies.

Two further systematic reviews focus on the issue of quality assessment of non-randomised studies. The first identifies and evaluates tools that can be used to assess the quality of non-randomised studies. The second looks at ways that study quality has been assessed and addressed in systematic reviews of healthcare interventions that have included non-randomised studies.

The two empirical investigations focus on the issue of selection bias in non-randomised studies. The first investigates the size and behaviour of selection bias in evaluations of two specific clinical interventions and the second assesses the degree to which case-mix adjustment corrects for selection bias.

Chapter 2

Aims and objectives

This project has considered the methods and evidence-base for evaluating non-randomised intervention studies. It is restricted to designs 3–6 described in Chapter 1 (*Box 1*), and thus excludes case–control designs. The report first presents the findings from reviews of current evidence and existing tools for evaluating non-randomised studies, and the application of these tools in systematic reviews. It concludes with reports of two new empirical investigations of expected biases and methods for correcting for bias in non-randomised studies. The specific objectives addressed in each of the chapters are presented below.

Chapters 3–5 report the results of three separate reviews of the literature.

Chapter 3 reviews the current empirical evidence concerning bias associated with non-randomised studies of healthcare interventions. In particular, it considers the evidence:

- that there are inconsistencies between the results of randomised and non-randomised studies
- that non-randomised studies may be systematically biased
- that the results of non-randomised studies may be more variable than the results of RCTs
- that case-mix adjustment reduces systematic bias and variability in non-randomised studies.

Chapter 4 reviews existing tools for assessing the quality of non-randomised studies to:

- describe the content, development and usability of quality assessment tools for non-randomised studies
- appraise the degree to which the assessment tools evaluate specified quality domains
- identify the tools most suitable for use in systematic reviews of non-randomised studies, and to identify their limitations.

In Chapter 5, the use of quality assessment in systematic reviews is considered. The review aims to:

- evaluate the quality assessments made in systematic reviews that have included non-randomised studies
- investigate whether relationships between study quality and study results have been observed.

Chapters 6 and 7 report the results of new empirical investigations obtained through analysis of non-randomised studies generated by resampling participants from two large, multicentre randomised trials. The rationale and methodology are described in detail in Chapter 6.

The investigation reported in Chapter 6 evaluates:

- the degree of systematic bias introduced by the use of historical and concurrent controls in non-randomised studies of healthcare interventions and the impact that this has on study conclusions
- the degree to which results of non-randomised studies using historical and concurrent controls are more variable than those of randomised controlled trials and the impact that this has on study conclusions.

In Chapter 7, the investigation is extended to evaluate the ability of case-mix adjustment methods to correct selection bias introduced in non-randomised designs. Specifically it considers:

- whether case-mix adjustment methods can compensate for systematic bias introduced through the use of historical and concurrent controls
- whether case-mix adjustment methods can compensate for increased variability introduced through the use of historical and concurrent controls
- whether case-mix adjustment methods can correct for bias introduced through mechanisms akin to allocation by indication
- whether there are differences in the performance of stratification, logistic regression and propensity score methods for adjusting for selection bias.

Chapter 8 presents an overview of the findings from all sections of the project and makes recommendations concerning the evaluation of non-randomised studies and further research that should be undertaken.

Chapter 3

Review of empirical comparisons of the results of randomised and non-randomised studies

Introduction

Evidence about the importance of design features of RCTs has accumulated rapidly during recent years.[19–21] This evidence has mainly been obtained by a method of investigation that has been termed **meta-epidemiology**, a powerful but simple technique of investigating variations in the results of RCTs of the same intervention according to features of their study design.[22] The process involves first identifying substantial numbers of systematic reviews each containing RCTs both with and without the design feature of interest. Within each review, results are compared between the trials meeting and not meeting each design criterion. These comparisons are then aggregated across the reviews in a grand overall meta-analysis to obtain an estimate of the systematic bias removed by the design feature. For RCTs, the relative importance of proper randomisation, concealment of allocation and blinding have all been estimated using this technique.[20,21] The results have been shown to be consistent across clinical fields,[23] providing some evidence that meta-epidemiology may be a reliable investigative technique. The method has also been applied to investigate sources of bias in studies of diagnostic accuracy, where participant selection, independent testing and use of consistent reference standards have been identified as being the most important design features.[24]

The use of meta-epidemiology has also been extended from the comparison of design features within a particular study design to comparisons between study designs. In two separate HTA reports, the results of RCTs have been compared with those from non-randomised evaluations across multiple interventions to estimate the bias removed by randomisation.[25,26] However, the meta-epidemiology method may be inappropriate for between design comparisons due to:

- **meta-confounding**: the existence of other differences between randomised and non-randomised studies which could impact on their findings, and
- **unpredictability in the direction of effect**: there possibly being no overall systematic bias but biases acting unpredictably in different directions with varying magnitudes.

An alternative methodology for empirically investigating differences between study designs is introduced in Chapter 5. First, we review the evidence provided by meta-epidemiological comparisons of randomised and non-randomised studies of the importance of randomisation *per se*, and discuss weaknesses in the use of meta-epidemiology to make such a comparison. The comparisons are considered in regard to four particular issues, namely whether there is empirical evidence of:

- inconsistencies in findings between RCTs and non-randomised studies
- systematic differences in average estimates of treatment effects between RCTs and non-randomised studies
- differences in the variability of results between RCTs and non-randomised studies (between study heterogeneity), and
- whether case-mix adjustment in non-randomised studies reduces systematic bias and/or between study heterogeneity.

Methods

Reviews were eligible for inclusion if:

- they compared quantitative results between RCTs of an intervention and non-randomised studies of the same intervention
- they had accumulated, through some systematic search, results from several of these comparisons across healthcare interventions.

Reviews were identified from a search of electronic databases including MEDLINE, EMBASE and PsycLit, from the earliest possible date up to December 1999; from handsearches of *Statistics in Medicine, Statistical Methods in Medical Research, Psychological Bulletin* and *Controlled Clinical Trials*

and from contact with experts working in the field. The electronic search strategy used is provided in Appendix 1. Additionally, the searches carried out for other sections of the project (including searches from the Cochrane Library databases) were screened to identify suitable papers for inclusion. Difficulties with devising searches for methodological topics are discussed in Chapter 4.

The content, results and conclusions from each of the identified reviews were noted. In addition, the methodology of each review was critically assessed for potential weaknesses. Aspects considered were as follows:

1. Was the identification of included studies unlikely to be biased?
2. Did the RCTs and non-randomised studies recruit similar participants, use similar interventions and measure similar outcomes?
3. Were the RCTs and non-randomised studies shown to use similar study methodology in all respects other than the allocation mechanism?
4. Were sensible, objective criteria used to determine differences or equivalence of study findings?

Results

Eight reviews were identified which fulfilled the inclusion criteria; seven considered medical interventions and one psychological interventions. Brief descriptions of the methods and findings of each review are given below, with summary details given in *Table 2*. There is substantial overlap in the interventions (and hence studies) that were included in the reviews of medical interventions (*Table 3*).

Description of the identified reviews

Sacks, Chalmers and Smith[27]

Sacks and colleagues compared the results of RCTs with historically controlled trials (HCTs). The studies were identified in Chalmers' personal collection of RCTs, HCTs and uncontrolled studies maintained since 1955 by searches of *Index Medicus*, *Current Contents* and references of reviews and papers in areas of particular medical interest (full list not stated). Six interventions were included for which at least two RCTs and two HCTs were identified [cirrhosis with oesophageal varices, coronary artery surgery, anticoagulants for acute myocardial infarction, 5-fluorouracil adjuvant therapy for colon cancer, bacille Calmette–Guérin vaccine (BCG) adjuvant immunotherapy and diethylstilbestrol for habitual abortion (*Table 3*)]. Trial results were classified as positive if there was either a statistically significant benefit or if the authors concluded benefit in the absence of statistical analysis, otherwise as negative. For each of the six interventions, a higher percentage of HCTs compared with RCTs concluded benefit: across all six interventions 20% of RCTs showed benefit compared with 79% of the HCTs.

The authors also considered whether statistical adjustment for confounding factors in the HCTs reduced the discrepancy between the randomised and non-randomised results. They found that the adjusted studies showed nearly the same treatment effect as the unadjusted studies.

Kunz and Oxman[28] *and Kunz, Vist and Oxman*[29]

Kunz and Oxman searched the literature for reviews that made empirical comparisons between the results of randomised and non-randomised studies. They included the results of the six comparisons in Sacks and colleagues' study above, and results from a further five published comparisons [antiarrhthymic therapy for atrial fibrillation, allogenic leucocyte immunotherapy for recurrent miscarriage, contrast media for salpingography, hormonal therapy for cryptorchidism, and transcutaneous electrical nerve stimulation (TENS) for postoperative pain (*Table 3*)]. In some of the comparisons, RCTs were compared with truly observational studies and, in others they were compared with quasi-experimental trials. A separate publication of anticoagulants for acute myocardial infarction already included in Sacks and colleagues' review was also reviewed,[30] as was a comparison of differences in control group event rates between randomised and non-randomised studies for treatments for six cancers (which does not fit within our inclusion criteria).[31] The review was updated in 2002 including a further 11 comparisons, and published as a Cochrane methodology review.[29]

The results of each empirical evaluation were described, but no overall quantitative synthesis was carried out. The results showed differences between RCTs and non-randomised studies in 15 of the 23 comparisons, but with inconsistency in the direction and magnitude of the difference. It was noted that non-randomised studies overestimated more often than they underestimated treatment effects.

Britton, McKee, Black, McPherson, Sanderson and Bain[25]

Britton and colleagues searched for primary publications that made comparisons between

TABLE 2 *Description of studies comparing results of randomised and non-randomised studies*

	Sacks, 1982[27]	**Kunz, 1998,[28] 2002[29] (figures in parentheses refer to original publication)**	**Britton, 1998[25]**	**MacLehose, 2000[26]**	**Benson, 2000[32]**	**Concato, 2000[33]**	**Ioannidis, 2001[34]**	**Lipsey, 1996,[35] 2001[36]**
Number of comparisons	6	23 (11)	18	14 (38 outcomes)	19	5	45	76
Other study designs included	Historically controlled trials (HCTs)	Quasi-experiments, historically controlled trials (HCTs), patient preference trials	Quasi-experiments, natural experiments, and prospective observational studies	Quasi-experimental and observational studies	Observational studies	Case–control and cohort studies	Quasi-experiments, cohort and case–control studies	Non-randomised comparative studies
Numbers of RCTs/other studies	50/56	263 (122)/246 (152)	46/41	31/68	83/53	55/44	240/168	Not stated
Summary of overall quantitative results	79% of HCTs found the therapy to be better than the control regimen, compared to 20% of RCTs	In 15 (9) of 23 (11) comparisons effects were larger in non-randomised studies, 4 (1) studies had comparable results, whilst 4 (2) reported smaller effects	Significant differences were found in 11 of 18 comparisons, but there were inconsistencies in the direction of the differences	In 14 of 35 comparisons the discrepancy in RR was <10%, in 5 comparisons it was >50%. Discrepancies were smaller in "fairer" comparisons	Only in 2 of 19 comparisons did the point estimate of observational studies lie outside the confidence interval for the RCTs	For the five clinical topics considered the average results of the observational studies were remarkably similar to those of the RCTs	ORs of RCTs and non-RCTs are highly correlated ($r = 0.75$). Differences beyond chance were noted in 16% of comparisons: two-fold discrepancies in ORs occurred in 33%	Mean effect sizes (SD) of 0.46 (0.28) for RCTs and 0.41 (0.36) for non-randomised studies. Mean difference in effect sizes was 0.5, but varied between –0.6 and +0.8 SDs

continued

TABLE 2 *Description of studies comparing results of randomised and non-randomised studies (cont'd)*

	Sacks, 1982[27]	**Kunz, 1998,[28] 2002[29]**	**Britton, 1998[25]**	**MacLehose, 2000[26]**	**Benson, 2000[32]**	**Concato, 2000[33]**	**Ioannidis, 2001[34]**	**Lipsey, 1996,[35] 2001[36]**
Conclusion about bias (from abstract)	Biases in patient selection may irretrievably weight the outcome of HCTs in favour of new therapies	On average non-randomised studies tend to result in larger estimates of treatment. It is not possible to predict the magnitude or even direction of possible selection biases and consequent distortions of treatment effects	Results of RCTs and non-randomised studies do not inevitably differ; it may be possible to minimise any differences by ensuring that subjects included in each type of study are comparable	Quasi-experimental and observational study estimates of effectiveness may be valid if important confounding factors are controlled for	There is little evidence that estimates of treatment effects in observational studies are either consistently larger or qualitatively different from those obtained in randomised controlled trials	The results of well-designed observational studies do not systematically overestimate the magnitude of treatment as compared with those in randomised controlled trials	Despite good correlation between randomised trials and non-randomised studies – in particular, prospective studies – discrepancies beyond chance do occur and differences in estimated magnitude of treatment effect are very common	Non-randomised designs may yield different observed effects relative to randomised designs, but the difference is almost as likely to represent an upward as a downward bias (from text)
Conclusion about variability	No comment	No comment	No comment	No comment	No comment	Observational studies had less variability in point estimates than RCTs of the same topic	Between study variability is smaller among RCTs than prospective non-randomised studies ($p = 0.03$) and all non-RCTs	($p = 0.07$)
Conclusion about case-mix adjustment	Adjustment of outcomes in HCTs for prognostic factors did not appreciably change the results	No comment	The effect of adjustment for baseline differences between groups in non-randomised studies is inconsistent	No comment	No comment	No comment	No comment	No comment

RR, risk ratio.

TABLE 3 *Clinical topics investigated in each of the medical reviews*

	Clinical topic	Sacks, 1982[27]	Kunz, 1998,[28] 2002[29]	Britton, 1998[25]	MacLehose, 2000[26]	Concato, 2000[33]	Benson, 2000[32]	Ioannidis, 2001[34]
1	Treatment of cirrhotic patients with oesophageal varicies	✓	✓					✓
2	Surgical versus medical treatment of coronary artery disease	✓	✓	✓	✓		✓	✓
3	Anticoagulants for acute myocardial infarction	✓	✓		✓			✓
4	5-Fluorouracil adjuvant therapy for colon cancer	✓	✓					✓
5	BCG adjuvant immunotherapy for malignant melanoma	✓	✓					✓
6	Diethylstilbestrol for habitual abortion	✓	✓					✓
7	Antiarrhythmic therapy for atrial fibrillation		✓	✓	✓			✓
8	Allogenic leucocyte immunotherapy for recurrent miscarriage		✓	✓				✓
9	Oil soluble contrast media during hysterosalpingography for infertile couples		✓	✓			✓	✓
10	Hormonal therapy in cryptorchidism		✓	✓				✓
11	TENS for postoperative pain		✓					✓
12	Matching of cancer control groups for six different cancers		✓					
13	β-Blockers for acute myocardial infarction			✓	✓		✓	
14	Surgical versus medical management of severe throat infections			✓	✓			
15	Adenoidectomy for persistent otitis media			✓			✓	
16	Chemotherapy, radiotherapy and tamoxifen in women with breast cancer			✓	✓			
17	Treatment of benign oesophageal stricture			✓				
18	Day hospital versus inpatient care for alcoholism			✓				
19	Chorionic villus sampling versus early amniocentesis for karyotyping			✓			✓	
20	Hospice versus conventional care for the terminally ill			✓				
21	Measles vaccine			✓				
22	Adjuvant doxorubin for treatment of sarcoma			✓				
23	Acute non-lymphatic leukaemia following adjuvant treatment for breast cancer			✓			✓	
24	Antioxidant vitamins for cardiovascular protection			✓				
25	CABG vs PTCA for coronary artery disease			✓			✓	
26	Calcium antagonists for acute coronary disease			✓			✓	
27	Stroke units			✓				
28	Malaria vaccines			✓				
29	Breast conservation versus mastectomy for breast cancer				✓			
30	Tamoxifen versus placebo for breast cancer				✓			
31	Formula supplementation for infant feeding				✓			
32	Folic acid supplementation for prevention of neural tube defects				✓			
33	5-Fluorouracil adjuvant therapy for stomach cancer				✓			
34	Mammographic screening				✓	✓		✓
35	BCG vaccine for prevention of tuberculosis					✓		✓
36	Treatment of cholesterol and death due to trauma					✓		
37	Treatment of hypertension and stroke					✓		
38	Treatment of hypertension and coronary heart disease					✓		✓
39	Intensive versus conventional insulin						✓	
40	Pneumatic retinopexy versus scleral buckling						✓	
41	Geriatric unit versus medical ward care (unclear for what condition)						✓	
42	Local versus general anaesthesia for endarcterectomy						✓	✓
43	Lithotripsy versus nephrolithotomy						✓	
44	Laser versus electrosurgical salpingostomy						✓	
45	Eder–Puestow versus balloon dilation						✓	
46	Hormone replacement therapy for osteoporosis		✓[a]				✓	
47	Calcium channel blockers for renal graft survival		✓[a]				✓	
48	Laparoscopic versus open appendectomy		✓[a]				✓	
49	Naltrexone for alcohol dependence							✓
50	Low-level laser therapy for osteoarthritis							✓
51	Vaccine for meningitis serotype A							✓

continued

TABLE 3 *Clinical topics investigated in each of the medical reviews (cont'd)*

	Clinical topic	Sacks, 1982[27]	Kunz, 1998,[28] 2002[29]	Britton, 1998[25]	MacLehose, 2000[26]	Concato, 2000[33]	Benson, 2000[32]	Ioannidis, 2001[34]
52	Microsurgical vs macrosurgical salpingostomy in subfertility							✓
53	Faecal occult blood screening for colorectal cancer							✓
54	Intrathecal therapy for tetanus							✓
55	Vitamin A supplementation							✓
56	Interferon for hepatitis C							✓
57	Trial of labour versus no labour for breech delivery							✓
58	Surgical treatments for incontinence							✓
59	Nonoxynol-9 spermicides for sexually transmitted diseases							✓
60	Safety of postpartum discharge							✓
61	High-dose diuretics as first-line treatment of hypertension							✓
62	Aspirin for prevention of hypertension in pregnancy							✓
63	Allogenic blood transfusion in cancer							✓
64	Subcutaneous versus intravenous heparin in deep vein thrombosis							✓
65	Low-protein diets in chronic renal insufficiency							✓
66	Corticosteroids for idiopathic facial nerve palsy							✓
67	Graduated compression stockings							✓
68	Aspirin for primary prevention of stroke							✓
69	Fixed nail plates versus sliding hip for femoral fractures							✓
70	Corticosteroids for stable COPD							✓
71	Treatment of hypertension in elderly people							✓
72	Selective decontamination of the digestive tract							✓
73	Cyclosporin withdrawal							✓
74	Prehospital thrombolysis							✓
75	T3-therapy in euthyroid depressed patients		✓[a]					
76	Chemotherapy for oesophageal cancer		✓[a]					
77	Autologous blood donation		✓[a]					
78	Prevention of adolescent pregnancy		✓[a]					
79	Educational interventions related to smoking and alcohol		✓[a]					
80	Educational interventions related to nutrition and weight		✓[a]					
81	Educational interventions related to other health behaviours		✓[a]					
82	Medical and surgical management of coronary artery disease		✓[a]					

[a] Indicates comparison added in review update.
CABG, coronary artery bypass graft; PTCA, percutaneous transluminal coronary angioplasty; COPD, chronic obstructive pulmonary disease.

single randomised and non-randomised studies (14 comparisons) and secondary publications (reviews) making similar comparisons (four comparisons). Both observational and quasi-experimental studies were included in the non-randomised category. They included all four of the secondary comparisons included in the review by Kunz and colleagues[28] (*Table 3*). The single study comparisons included studies where a comparison was made between participants who were allocated to experimental treatment as part of a trial and a group who declined to participate, and studies of centres where simultaneous randomised and patient-preference studies had been undertaken of the same intervention. The studies were assessed to ensure that the randomised and non-randomised studies were comparable on several dimensions (*Table 4*).

There were statistically significant differences between randomised and non-randomised studies for 11 of the 18 comparisons. The direction of these differences was inconsistent and the magnitude extremely variable. For some interventions the differences were very large. For example, in a review of treatments for acute

TABLE 4 *Details of methodology of empirical comparisons between randomised and non-randomised studies*

	Sacks, 1982[27]	**Kunz, 1998,[28] 2002[29]**	**Britton, 1998[25]**	**MacLehose, 2000[26]**	**Benson, 2000[32]**	**Concato, 2000[33]**	**Ioannidis, 2001[34]**	**Lipsey, 1996,[35] Wilson, 2001[36]**
Method of identification of comparisons	**Primary studies**: Personal systematic collection of RCTs and HCTs in specific fields of interest	**Secondary studies**: Electronic and manual search for studies comparing randomised and non-randomised studies	**Primary and secondary studies**: Electronic searches for studies comparing randomised and non-randomised groups	**Primary and secondary studies**: Electronic and manual searches for studies comparing randomised and non-randomised groups	**Primary studies**: Electronic search of Medline and CCTR for observational studies and matching RCTs	**Secondary studies**: Meta-analyses published in 5 leading journals that included non-randomised studies	**Secondary studies**: Meta-analyses located by electronic searches (MEDLINE and Cochrane library) and manual searches	**Secondary studies**: Meta-analyses located by electronic searches (Psychology and Sociology databases) and manual searches
Similarity of patients, interventions and outcomes	Same intervention for the same medical condition	2 studies controlled for clinical differences in participants and interventions, 6 did so partly and 7 did not at all	Same intervention, similar setting, similar control therapy, comparable outcomes measures	Same intervention. Similarity of eligibility, time period, co-morbidity, disease severity and other prognostic factors assessed	Restricted to interventions allocated by physicians to induce comparability	Not assessed	Not assessed	Impact of differences considered in multivariate analysis
Similarity of other study methods	Not assessed	5 papers judged partly comparable, 10 not comparable on double blinding, complete follow-up and other methodological issues	Not assessed	Blinding of outcome assessed	Not assessed	Analysis according to study quality mentioned in discussion, but no methods or results presented	Not assessed	Impact of assessments of study quality considered in multivariate analysis

continued

TABLE 4 *Details of methodology of empirical comparisons between randomised and non-randomised studies (cont'd)*

	Sacks, 1982[27]	**Kunz, 1998,[28] 2002[29]**	**Britton, 1998[25]**	**MacLehose, 2000[26]**	**Benson, 2000[32]**	**Concato, 2000[33]**	**Ioannidis, 2001[34]**	**Lipsey, 1996,[35] 2001[36]**
Method of summarising study findings	Vote counting of results classified as positive (either statistically significant or had positive conclusions if no statistical analysis)	No consistent summary: results of randomised and non-randomised groups described using a variety of measures	Results of randomised and non-randomised groups described using risk differences, RRs or ORs	RRs and risk differences calculated separately for MAs for randomised and non-randomised studies	Calculation of overall MA results and CIs (ORs for binary data, differences in means for continuous)	Calculation of overall MA results and CIs (RRs and ORs)	Calculation of fixed and random MA estimates expressed as ORs and log ORs	(1) Mean effect sizes of all randomised and all non-randomised studies. (2) distribution of differences in effect sizes between randomised and non-randomised in each MA
Method for comparing results between groups	Comparison of the percentage with positive results	No overall analysis presented	Statistical significance of the difference in effect sizes	Distribution of relative and absolute differences in results reported	Assessment of whether observational point estimate fell within 95% CI for RCTs	No overall analysis presented	Calculation of *Z*-scores for difference between treatment effects	Graph of distribution of differences of effect sizes
Criteria for comparing variability of results	Not assessed	Not assessed	Not assessed	Not assessed	Not assessed	Dispersion of points calculated without considering differences in sample size	Significance ($p < 0.1$) of tests of between study heterogeneity	Not assessed, although standard deviations of randomised and non-randomised studies presented (unadjusted for differences in sample size)

CCTR, Cochrane Controlled Trials Register; MA, meta-analysis.

non-lymphatic leukaemia, the risk ratio in RCTs was 24 compared with 3.7 in non-randomised studies (comparison 23 in *Table 3*).

The impact of statistical adjustment for baseline imbalances in prognostic factors was investigated in two primary studies, and in four additional comparisons (coronary angioplasty versus bypass grafting, calcium antagonists for cardiovascular disease, malaria vaccines and stroke unit care: comparisons 25–28 in *Table 3*). In two of the six comparisons there was evidence that adjustment for prognostic factors led to improved concordance of results between randomised and non-randomised studies.

MacLehose, Reeves, Harvey, Sheldon, Russell and Black[26]

MacLehose and colleagues restricted their review to studies where results of randomised and non-randomised comparisons were reported together in a single paper, arguing that such comparisons are more likely to be of 'like-with-like' than those made between studies reported in separate papers. They included primary studies and also reviews that pooled results from several individual studies. Of the 14 comparisons included in their report, three were based on reviews (comparisons 3, 7 and 25 in *Table 3*) and the rest were results from comparisons within single studies. The non-randomised designs included comprehensive cohort studies, other observational studies and quasi-experimental designs.

The 'fairness' or 'quality' of each of the comparisons made was assessed for comparability of patients, interventions and outcomes and additional study methodology (see *Table 4*). Although the authors did not categorise comparisons as showing equivalence or discrepancy, the differences in results were found to be significantly greater in comparisons ranked as being low quality.

Benson and Hartz[32]

Benson and Hartz evaluated 19 treatment comparisons (eight in common with Britton and colleagues[25]) for which they located at least one randomised and one observational study (defined as being a study where the treatment was not allocated for the purpose of research) in a search of MEDLINE and the databases in the Cochrane Library (*Table 4*). They only considered treatments administered by physicians. Across the 19 comparisons they found 53 observational and 83 randomised studies, the results of which were meta-analysed separately for each treatment comparison. Comparisons were made between the pooled estimates, noting whether the point estimate from the combined observational studies was within the confidence interval of the RCTs. They found only two instances where the observational and randomised studies did not meet this criterion.

Concato, Shah and Horwitz[33]

Concato and colleagues searched for meta-analyses of RCTs and of observational studies (restricted to case–control and concurrent cohort studies) published in five leading general medical journals. They found only five comparisons where both types of study had been meta-analysed [BCG vaccination for tuberculosis (TB), mammographic screening for breast cancer mortality, cholesterol levels and death from trauma, treatment of hypertension and stroke, treatment of hypertension and coronary heart disease (CHD) (*Table 3*)] combining a total of 55 randomised and 44 observational studies. They tabulated the results of meta-analyses of the randomised and the observational studies and considered the similarity of the point estimates and the range of findings from the individual studies. In all five instances they noted the pooled results of randomised and non-randomised studies to be similar. Where individual study results were available, the range of the RCT results was greater than the range of the observational results.

Ioannidis, Haidich, Pappa, Pantazis, Kokori, Tektonidou, Contopoulous-Ioannidis and Lau[34]

Ioannidis and colleagues searched for reviews that considered results of RCTs and non-randomised studies. In addition to searching MEDLINE they included systematic reviews published in the Cochrane Library, locating in total 45 comparisons. Comparisons of RCTs with both quasi-randomised and observational studies were included. All meta-analytical results were expressed as odds ratios, and differences between randomised and non-randomised results expressed as a ratio of odds ratios and their statistical significance calculated. Findings across the 45 topic areas were pooled incorporating results from 240 RCTs and 168 non-randomised studies. Larger treatment effects were noted more often in non-randomised studies. In 15 cases (33%) there was at least a twofold variation in odds ratios, whereas in 16% there were statistically significant differences between the results of randomised and non-randomised studies. The authors also tested the heterogeneity of the results of the randomised and non-randomised studies for each topic. Significant heterogeneity was noted for 23% of the reviews of RCTs and for 41% of the reviews of non-randomised studies.

Lipsey and Wilson[35] and Wilson and Lipsey[36]

Lipsey and Wilson searched for all meta-analyses of psychological interventions, broadly defined as treatments whose intention was to induce psychological change (whether emotional, attitudinal, cognitive or behavioural). Evaluations of individual components of interventions and broad interventional policies or organisational arrangements were excluded. Searches of psychology and sociology databases supported by manual searches identified a total of 302 meta-analyses, 76 of which contained both randomised and non-randomised comparative studies. Results were analysed in two ways. First, the average effect sizes of randomised and non-randomised studies were computed across the 74 reviews, and average effects were noted to be very slightly smaller for non-randomised than randomised studies. Second (and more usefully) the difference in effect sizes between randomised and non-randomised studies within each of the reviews was computed and plotted. This revealed both large over- and underestimates with non-randomised studies, differences in effect sizes ranging from –0.60 to +0.77 standard deviations.

Studies excluded from the review

Three commonly cited studies were excluded from our review.[37–39] Although these studies made comparisons between the results of randomised and non-randomised studies across many interventions, they did not match RCTs and non-randomised studies according to the intervention. Although they provide some information about the average findings of selected randomised and non-randomised studies, they did not consider whether there are differences in results of RCTs and non-randomised studies of the same intervention.

Findings of the eight reviews

The eight reviews have drawn conflicting conclusions. Five of the eight reviews concluded that there are differences between the results of randomised and non-randomised studies in many but not all clinical areas, but without there being a consistent pattern indicating systematic bias.[25,26,28,34,35] One of the eight reviews found an overestimation of effects in all areas studied.[27] The final two concluded that the results of randomised and non-randomised studies were 'remarkably similar'.[32,33]

Of the two reviews that considered the relative variability of randomised and non-randomised results, one concluded that RCTs were more consistent[34] and the other that they were less consistent.[33]

The two studies that investigated the impact of case-mix adjustment were in agreement, both noting that adjustment did not necessarily reduce discordance between randomised and non-randomised findings.[25,27]

Critical evaluation of reviews

The discrepancies in the conclusions of the eight reviews may in part be explained by variations in their methods and rigour, so that they had varying susceptibility to bias. We consider the weaknesses in these reviews under four headings.

Was the identification of included studies unlikely to be biased?

The studies used in all the reviews represent only a very small portion of all randomised and observational research. From *Table 3*, it is clear that the seven reviews in medical areas were each only based on a subset of known comparisons of randomised and non-randomised evidence. Even the largest review[34] did not include all comparisons identified in previous reviews. Correspondence has also cited several other examples of treatment comparisons where there are disagreements between observational and randomised studies.[40,41]

More important, could the comparisons selected for these meta-epidemiological reviews be a potentially biased sample? There are two levels at which publication bias can act in these evaluations: (a) selective publication of primary studies, which will affect all reviews, and (b) selective publication of meta-analyses of these studies, which will affect reviews restricted to secondary publications. Evaluations of publication bias have noted differences in the frequency of primary publication of randomised and observational studies, although the direction and magnitude of the differences vary between evaluations and the relationship to statistical significance is not known.[42] Similarly, the decisions made concerning the publication of meta-analyses that include non-randomised studies are likely to be influenced by the results of existing randomised controlled trials. The Concato review[33] may be the most susceptible to publication bias as it restricted study selection to meta-analyses combining randomised or observational results published in five general medical journals: *Annals of Internal Medicine, British Medical Journal* (*BMJ*), *Journal of the American Medical Association* (*JAMA*), *New England Journal of Medicine* (*NEJM*) and *The Lancet*. Therefore, the only studies eligible for inclusion were those where both authors and top journal editors had already decided that it was sensible to synthesise the

results of both observational and randomised studies. It seems highly likely that these decisions may relate to the similarity of the results of studies of different designs.

Did the RCTs and non-randomised studies recruit similar participants, use similar interventions and measure similar outcomes?

Discrepancies between the results of observational and randomised studies may be confounded by differences in the selection and evaluation of patient groups, in the formulation and delivery of treatments, in the use of placebos and other comparative treatments and in the methods used to maintain follow-up and record outcomes. For many interventions there may also be temporal confounding of study types, non-randomised studies typically being performed prior to the RCTs. Such **meta-confounding** will make it difficult to attribute a systematic difference directly to the use or non-use of a random allocation mechanism.

Six of the eight reviews noted this problem and incorporated features in their evaluations to reduce the potential for meta-confounding.[25–28,32,35] Two reviews made more stringent efforts to assess comparability than the others.[25,26] Britton and colleagues restricted the selection of studies to be similar in terms of intervention, setting, control therapy and outcome measure. MacLehose and colleagues assessed each comparison for the possibility of meta-confounding and found that the most susceptible (i.e. those with differences in eligibility criteria and time periods and no adjustment for severity of disease, comorbidity and other prognostic factors) had, on average, the largest discrepancies.

Were the RCTs and non-randomised studies shown to use similar study methodology in all respects other than the allocation mechanism?

Meta-confounding could also occur through differences in other aspects of study design, beyond the use of randomisation. For example, the results of RCTs are known to vary according to the quality of randomisation, especially the concealment of allocation at recruitment.[23] However, none of the reviews restricted the inclusion of RCTs to those with adequate concealment or on any other methodological basis. Only one review[26] assessed the comparability of randomised and non-randomised studies on any aspect of study quality (blinding). Discrepancies and similarities between study designs could be partly explained by differences in other unevaluated aspects of the methodological quality of the RCTs.

Similarly, there will be differences in the methodological rigour of the non-randomised studies. Importantly, the possible biases of non-randomised studies vary with study design. Only one review[27] restricted non-randomised studies to be of a single design (historically controlled studies). In all the others, RCTs were compared with non-randomised studies of a mixture of designs.

Were sensible, objective criteria used to determine differences or equivalence of study findings?

The manner in which results were judged to be 'equivalent' or 'discrepant' varied widely between the reviews and influenced the conclusions that were drawn. For example, Concato and colleagues[33] deemed that the randomised and non-randomised studies of mammographic screening (comparison 34 in *Table 3*) had remarkably similar results, whereas Ioannidis and colleagues[34] classified them as discrepant. Only in two reviews was the judgement made aggregated across the comparisons.[27,35] In all the other reviews each individual topic was classified as either equivalent or discrepant. Many of the comparisons made at the level of a clinical topic were based on very few data, for example, in the Ioannidis review[34] on average for each intervention five RCTs were compared with four non-randomised studies. Hence the absence of a statistically significant difference cannot be interpreted as evidence of 'equivalency' and clinically significant differences in treatment effects cannot be excluded. Conversely, the presence of a statistically significant difference does not indicate that a clinically important difference does exist. Four reviews[26,28,34,35] more usefully concentrated on describing the magnitude of the differences, all four noting substantial differences occurring in some, but not all, comparisons. The Concato review[33] subjectively classified all comparisons as being 'remarkably similar'.

The comparisons made in two reviews[33,34] of the relative variability of randomised and non-randomised results can be considered flawed owing to the criteria used to compare variation. Concato and colleagues considered the **range** of the point estimates from observational studies and RCTs.[33] This comparison was confounded by the different sample sizes used in observational and randomised studies, and the number of studies considered. On average, the RCTs in the Concato review were 25% smaller than the observational studies, hence greater variability in their results is

to be expected. It is also possible that the most extreme RCT results arose in particularly small randomised studies, and that there is an absence of equivalently small observational studies. Ioannidis and colleagues investigated the statistical significance of the heterogeneity.[34] This analysis is also dependent on the sample size and number of studies considered, hence the comparison between randomised and non-randomised studies may be confounded.

Discussion

The conclusions of the eight reviews are divergent, and as all the reviews have weaknesses, it is difficult to draw conclusions concerning the importance of randomisation from these investigations. The only robust conclusion that can be drawn is that in some circumstances the results of randomised and non-randomised studies differ, but it cannot be proved that differences are not due to other confounding factors. The frequency, size and direction of the biases cannot be judged reliably from the information presented.

One review reported the existence of a systematic bias, namely that historically controlled trials were more likely to conclude benefit of new therapies than RCTs.[27] However, as the conclusions were based on statistical significance and not size of effect, this may be confounded by sample size. Direct comparison of particular non-randomised designs with RCTs was not made in any of the additional seven reviews.

Two reviews raised concerns about the usefulness of case-mix adjustment methods in non-randomised studies.[25,27] Again, the validity of these conclusions is uncertain owing to the possibility of meta-confounding.

The efforts used to control confounding varied between the reviews. The potential for confounding in these comparisons requires detailed, clinically informed assessment of the similarity of populations, interventions and outcome assessments used in the RCTs and non-randomised studies. Six of the eight reviews are judged to be of poor quality in this regard, as they made no or little assessment of comparability. The reviews by Britton and colleagues[25] and MacLehose and colleagues[26] used the most comprehensive assessments to judge comparability, especially for the six comparisons they discussed in detail. For two of the four interventions considered in Britton and colleagues' report, the only conclusion that could be drawn was that due to differences in era, populations, dosage and length of follow-up, differences between randomised and non-randomised studies were to be expected. Britton and colleagues implied that for these clinical interventions a 'fair' comparison of randomised and non-randomised studies would not be possible. MacLehose and colleagues concluded their review by recommending that only comprehensive cohort studies could make fair comparisons between randomised and non-randomised evidence. Few such studies have ever been undertaken.

These investigations also raise concerns regarding the usefulness of meta-epidemiological investigations where there is (or could be) variability in the direction of bias. The meta-epidemiological method is designed to identify and estimate only systematic bias. Only if the selection of groups in non-randomised studies consistently led to over-representation of high- (or low-) risk participants in one group over another would the bias always be seen to act in the same direction. If selection bias in non-randomised studies arises as a result of haphazard variations in case-mix, there will be a mixture of under- and overestimates of the treatment effect. The results might all be biased, but not all in the same direction. In these circumstances, the difference between randomised and non-randomised studies would manifest as an increase in variance or heterogeneity of treatment effect (beyond that expected by chance) rather than (or in addition to) a systematic bias. This possibility is perhaps hinted at by the observation in five of the eight reviews that the differences observed between randomised and non-randomised studies did not act in a consistent manner. Ideally, a formal statistical comparison should aim to compare the heterogeneity in treatment effects, and not just the average treatment effects between randomised and non-randomised groups. None of the eight comparisons used a statistical model to compare results and, in any case, even the random effects method presently recommended for meta-epidemiological studies would be inadequate as it assumes that the variance of effects in all groups are equal.[22]

These major problems are in contrast to the successful application of the meta-epidemiological method to understand the importance of particular design features for RCTs.[20,21,23,43] The potential for meta-confounding in RCTs may be reduced if the types of participants, interventions and outcomes are more similar across trials than

across non-randomised studies, and there are fewer methodological characteristics of interest (randomisation, concealment of allocation, blinding and completeness of follow-up). Also, there are clear reasons to believe that the biases against which these design features protect are likely to act in favour of a particular treatment (the experimental treatment), leading to systematic bias, such that the problem of increased heterogeneity between study designs does not arise.

To take account of our concerns about the complexity of biases in non-randomised studies, we have first undertaken a review of quality domains and assessment methods used to consider the likelihood of bias in non-randomised studies (Chapters 4 and 5). Second, we have undertaken novel methodological assessments of the importance of two quality issues which are specific to non-randomised studies: method of allocation (Chapter 6) and use of case-mix adjustment (Chapter 7). Our methodological assessments have been designed to overcome the particular problems of meta-confounding and variability of direction and size of bias that plague the interpretation of these previous reviews.

Chapter 4

Evaluation of checklists and scales for assessing quality of non-randomised studies

Introduction

As discussed in Chapter 1, the RCT is widely regarded as the design of choice for the assessment of the effectiveness of healthcare interventions as it is theoretically most likely to produce an unbiased estimate of effect (has high internal validity). Nevertheless, there is an increasing recognition of the role of non-randomised studies for the assessment of effectiveness, either to support or generalise the results of RCTs or because RCTs simply do not exist and/or are not possible to conduct.

Regardless of the study designs available, the validity of any estimate of effectiveness is conditional on the quality of the study or studies upon which that estimate is based. The quality of a study has been defined as "the confidence that the trial design, conduct and analysis has minimised or avoided biases in its treatment comparisons".[44] This definition places the focus on questions of internal validity, and has been adopted for use in this report. Although many would argue that the main benefit from non-randomised studies lies in increasing the external validity of study findings, if we cannot establish the internal validity of non-randomised studies then the external validity of the results becomes largely irrelevant.

Study quality is a rather subjective concept, open to different interpretations depending on the reader. Regular readers of healthcare literature develop informal approaches to assessing quality, based on information picked up from a variety of sources, including previous experience and discussions with others. Where a study, or systematic review of studies, is to be used to evaluate the effectiveness of an intervention, more formal approaches to assessing quality are required. As Cooper[45] has pointed out, "almost every primary researcher and research reviewer begins an inquiry with some idea about the outcome of the enquiry", and these "predispositions toward a review's results can influence the reviewers' judgements about the methodological quality of a piece of research". Even short and apparently simple tools, such as the scale for assessing the quality of RCTs developed by Jadad and colleagues,[46] have been found to have low inter-rater reliability.[47] Further examples are cited by Cooper,[45] who concludes that there are two main sources of variance in evaluators' decisions: the relative importance that they assign to different research characteristics and their judgements about how well a particular study meets a design criterion.

A formal approach to quality assessment intuitively seems the best way of reducing the level of subjectivity that can be introduced. Approaches to quality assessment largely follow two main frameworks.[45,48] The first approach follows work in the social sciences field largely by Campbell and colleagues,[49,50] who outlined the various 'threats to validity' that can be encountered in experimental and quasi-experimental studies. This work originally focused on internal and external validity, but evolved over the years to cover 'statistical conclusion validity' and 'construct validity'. Within each category, a list of specific threats to validity was provided; the final version lists 33 threats to validity across the four categories.[50] Although perhaps not originally intended to be used as such, this work has been used as a basis for quality assessment. One or more of the threats to validity may be selected and reviewers assess whether or not they are present in any given study. The advantage of the approach is that relevant validity threats can be selected from an explicit list, which then contribute to an overall judgement of quality.

The second approach is the 'methods-description' approach, whereby the reviewer codes the objective characteristics of each study's methods as they are described by the primary researchers.[45] Among the first proponents of this approach were Chalmers and colleagues,[51] who outlined a detailed list of criteria with which to assess (and score) the quality of RCTs. The objective of this method is to provide an overall index of quality rather than to estimate the degree of bias present.[48] As with the threats to

validity approach, different reviewers may choose to list different methodological characteristics. However, one of the main advantages is that the coding of study characteristics does not require the same degree of judgement as when one is required to identify the presence of a threat to validity. For example, two reviewers might disagree on whether or not a study is low in power (threat to validity), and yet have perfect agreement when coding the separate components that make up the decision (e.g. sample size, study design, inherent power of the statistical test).[45]

Cooper[45] advocates that the optimal strategy for categorising studies is a mix of these two approaches, that is, coding all potentially relevant, objective aspects of research design as well as specific threats to validity which may not be fully captured by the first approach. Although this strategy does not (and could not) remove all of the subjectivity from the assessment process, it may be the best way of making assessments of quality as explicit and objective as possible. It should be noted that inter-rater reliability is likely to be further diminished where a judgement regarding the 'acceptability' of a given feature is required as opposed to identifying its presence or absence.[52]

Quality assessment tools developed for the healthcare literature have followed both approaches, but the majority are of the 'methods-description' variety, whether they took the form of a checklist or a scale. These tools provide a means of judging the overall quality of a study using itemised criteria, either qualitatively in the case of checklists or quantitatively for scales.[53] Alternatively, a component approach can be taken, whereby one or more individual quality components, such as allocation concealment or blinding, are investigated. However, as Moher and colleagues have pointed out, "assessing one component of a trial report may provide only minimal information about its overall quality".[53]

A common criticism of quality assessment tools is the lack of rationale provided for the particular study features that reviewers choose to code[52] and the inclusion of features unlikely to be related to study quality.[44,54] This may in part be due to the lack of empirical evidence for the biases associated with inadequately designed studies (although such evidence does exist to some extent for RCTs[21,55,56]). A further criticism of tools for assessing RCTs is lack of attention to standard scale development techniques,[44] to the extent that one scale[46] which was developed using psychometric principles has been singled out from other available tools. These principles involve the following steps as laid out by Streiner and Norman:[57] preliminary conceptual decisions; item generation and assessment of face validity; field trials to assess frequency of endorsement, consistency and construct validity; and generation of a refined instrument. However, as Jüni and colleagues have pointed out, following such principles does not necessarily make a tool superior to other available instruments.[23]

Quality assessment **scales** in particular have also been heavily criticised for the use of a single summary score to estimate study quality, by adding the scores for each individual item.[58] Greenland argues that the practice of quality scoring is the most insidious form of bias in meta-analysis as it "subjectively merges objective information with arbitrary judgements in a manner that can obscure important sources of heterogeneity among study results".[58] It has since been empirically demonstrated that the use of different quality scales for the assessment of the same trial(s) results in different estimates of quality.[21,59] Nevertheless, formal (or systematic) quality assessment, especially of RCTs, is increasingly common. A review by Moher and colleagues published in 1995 identified 25 scales for the quality assessment of RCTs.[44] Subsequent work by Jüni and colleagues has identified several more.[23,59]

In spite of these criticisms, it is largely agreed that the assessment of methodological quality should be routine practice in systematic reviews and meta-analyses. Although the majority of methodological work in this area has surrounded the assessment of RCTs, it is reasonable to suggest that if formal quality assessment of **randomised** controlled trials is important, then it is doubly so for **non-randomised** studies owing to the greater degree of judgement that is required. The largely observational nature of non-randomised studies leads to a much higher susceptibility to bias than is found for experimental designs, as discussed in Chapter 1.

A review of existing quality assessment tools for non-randomised intervention studies was conducted in order to provide a description of what is available, paying particular attention to whether and how well they cover generally accepted quality domains.

Methods

Inclusion criteria

To be considered as a quality assessment tool, a list

of criteria that could be (or had been) used to assess the methodological quality or validity of primary studies was required. A specific statement that the list of criteria was a scale or checklist intended to assess methodological quality was not required.

These tools could exist either as individual publications in their own right or within the context of a systematic review or other type of review, such as methodological reviews that had used some form of tool.

The tool must have been (or must have the potential to be) applied to non-randomised studies of intended effect, that is, epidemiological studies or studies primarily aimed at the investigation of the side-effects of an intervention were excluded. Tools specifically designed to assess case–control and uncontrolled studies were excluded (case–control studies were excluded on the basis that the design is rarely used to examine **intended** effects). To provide as comprehensive a picture of current practice as possible, any tool that had been used to assess non-randomised studies was included in the review, even if it had been explicitly designed to assess only RCTs.

Both 'new' and 'modified' tools were included. 'Modified' tools were those based on a single existing tool. Tools that were stated to be based on more than one existing tool were considered to be 'new' tools. Tools which made no statement regarding the originality of the tool were **assumed** to be 'new' tools. This is likely to have led to an over-estimation of the number of unique tools in existence; however, it was not practical to check further on the origin of these tools.

Literature searches

In an attempt to identify the largest possible number of quality assessment tools, an extensive and comprehensive literature search from the earliest possible date up to December 1999 was carried out. This included searching a wide range of electronic databases (see *Table 5*). The search strategies used were developed via an iterative process, by which a series of strategies were suggested, amended and piloted, and the search results scanned to identify the proportion of relevant papers retrieved (see Appendix 1 for the sample search strategy). Owing to the nature of the searches, and the poor indexing of the studies, it was necessary to strike a balance between strategies that were less likely to miss any relevant papers, yet retrieved a 'manageable' number of citations. Similar problems with searching for methodological literature have been cited by previous HTA-funded projects.[60] The resulting lists of titles and abstracts were screened by two

TABLE 5 *Literature search results*

Source	Retrieved	Selected from screening[a]	Met inclusion criteria
MEDLINE (1966–99)	1897	149	26
EMBASE (1974–99)	639	11	0
PsycLit (1967–99)	1835	113	4
Science Citation Index (1981–99)	1078	45	5
Social Science Citation Index (1981–99)	502	11	0
Index to Scientific and Technical Proceedings (1990–9)	294	6	0
Applied Social Sciences Index and Abstracts (1987–99)	262	10	1
Educational Resource Information Centre database (ERIC) (1965–99)	699	29	3
British Education Index (1986–99)	11	0	0
Cochrane Review Groups			3
Citation searches	NA	NA	85
Database of Abstracts of Reviews of Effectiveness (DARE) (1994–9)	1109	131	75[b]
Other[c]	NA	NA	11
Total	8326		213

[a] Totals presented are those following de-duplication of search results, i.e. only the additional number of unique studies obtained from each source is presented.
[b] Number of included reviews that developed their own quality assessment tool or modified another.
[c] Includes the CRD and Cochrane Collaboration methodology databases, handsearching of a number of key journals (*Statistics in Medicine* (1984–98), *Controlled Clinical Trials* (1984–98), *Journal of Clinical Epidemiology* (1991–8), *Psychological Bulletin* (1994–9), *Psychological Methods* (1996–9) and the *International Journal of Technology Assessment in Health Care* (1985–99)) and contact with a number of methodological experts.

reviewers. The full papers for all records that potentially met our inclusion criteria were obtained.

Searches of primary electronic databases were supplemented with searches of registers of methodological research, citation searches for key papers, handsearching of key journals and contact with experts. The reference lists of all retrieved papers were also scanned to identify any additional tools.

Data extraction

A data extraction form for recording relevant information from each quality assessment tool was designed and piloted (see Appendix 2). Data were extracted by one reviewer and the completed data extraction form was double-checked against the original paper by a second reviewer. Any disagreements were resolved by consensus or by referral to a third reviewer where necessary.

For each tool, items relating to the following areas were recorded:

1. Descriptive information. The purpose for which the tool had been developed (*Box 3*); whether it was a new tool or a modification of an existing one; the type of study designs covered; the inclusion of design-specific questions; whether it was a scale or a checklist; for scales only, the weighting system adopted; the number of items (both generic and topic-specific); and whether the tool had been used in our identified systematic reviews.
2. Tool development. Information relating to how the items were generated; whether the selection of items was justified; and whether and how the validity and/or reliability of the tool was examined.
3. Tool content. To facilitate extraction of the content of the tools and allow comparison between tools, a taxonomy of 12 quality domains covering the major aspects of study quality was constructed *a priori* (*Table 6*). These domains covered internal validity, external validity and issues related to the quality of reporting. A thirteenth domain was used for ancillary issues not covered elsewhere. Within each domain, items that might be expected to appear in a quality assessment tool were specified. These domains and items were generated using a modified Delphi process amongst review team members. An initial list of domains and items generally known or believed to be important for the assessment of RCTs was drawn up. Additional items potentially relevant to the assessment of non-randomised studies were then added.

BOX 3 *Purposes for which a tool could be developed*

1. Planning a study
2. Assessing a statistical/methodological analysis
3. Evaluating a study when considering the practical application of the intervention
4. Evaluating/grading study recommendations when producing recommendations/guidelines
5. Peer reviewing
6. Reporting a study
7. Assessing studies included in a systematic review

Six of the 12 domains are mainly related to internal validity. Of these, we consider that the two most important for the evaluation of non-randomised intervention studies are the creation of the intervention groups (domain #5) and the comparability of the groups at the analysis stage (domain #9). For studies that are not randomised, it is important to know first how the allocations were performed (for example, according to clinician or patient preference) and second whether attempts were made to ensure that the groups were similar for key prognostic characteristics by design (such as through using matching). Furthermore, the attempts to identify all important prognostic factors and the use of case-mix adjustment to account for any differences between groups are thought to be particularly important in non-randomised studies. These two domains were considered core domains for the assessment of non-randomised studies, and within them four core criteria were specified (*Table 6*).

Other domains related to internal validity included blinding (of patients and investigators) (domain #6); the soundness of information about the intervention and the outcomes (domain #7); the adequacy of follow-up (domain #8); and the appropriateness of the analysis (domain #10). The soundness of information refers to the confidence one has that the patients actually received the intervention to which they were assigned and actually experienced the reported outcome(s) as a result of that intervention.

The majority of the items relating to the selection of the study sample (domain #2) were related to external validity. Although prospective or retrospective sample selection is more an issue of internal validity, the provision of explicit inclusion and/or exclusion criteria and the

TABLE 6 *List of quality domains and items*

#1	Background/context	1.1. Provision of background information 1.2. Problem/question clearly stated[b] 1.3. Study originality 1.4. Relevance to clinical practice 1.5. Rationale/theoretical framework
#2	Sample definition and selection	2.1. Retrospective/prospective[b] 2.2. Inclusion/exclusion criteria[b] 2.3. Sample size[b] 2.4. Selected to be representative[b] 2.5. Baseline characteristics described
#3	Interventions	3.1. Clear specification[b] 3.2. Concurrent/concomitant treatment 3.3. Feasibility of intervention
#4	Outcomes	4.1. Clear specification[b] 4.2. Objective and/or reliable 4.3. Selection of outcomes for relevance, importance, side-effects
#5	Creation of treatment groups[a]	5.1. Generation of random sequence[b] 5.2. Concealment of allocation[b] 5.3. *How allocation occurred*[b,c] 5.4. *Any attempt to balance groups by design*[b] 5.5. Description of study design 5.6. Suitability of design 5.7. Contamination
#6	Blinding	6.1. Blind (or double-blind) administration[b] 6.2. Blind outcome assessment[b] 6.3. Maximum potential blinding used 6.4. Testing of blinding
#7	Soundness of information	7.1. Source of information about the intervention[b] 7.2. Source of information about the outcomes[b]
#8	Follow-up	8.1. Equality of length of FU[d] for the two groups?[b] 8.2. Length of FU adequate? 8.3. Completeness of FU[b]
#9	Analysis: comparability	9.1. Assessment of baseline comparability[b] 9.2. *Identification of prognostic factors*[b] 9.3. *Case-mix adjustment*[b]
#10	Analysis: outcome	10.1. Intention-to-treat analysis[b] 10.2. Appropriate methods of analysis[b] 10.3. Pre-specified hypotheses
#11	Interpretation	11.1. Appropriately based on results[b] 11.2. Assessment of strength of evidence[b] 11.3. Application/implications[b] 11.4. Clinical importance and statistical significance 11.5. Interpretation in context
#12	Presentation and reporting	12.1. Completeness, clarity and structure[b] 12.2. Statistical presentation and reporting

[a] Underlined domains denote 'core' domains.
[b] Denotes items specified *a priori*
[c] Items in italics are 'core' items (see text).
[d] FU, follow-up.

representativeness of the sample are external validity issues. Issues relating to sample size may assess the possibility that the study was underpowered to detect a clinically important effect or that the sample size was not determined *a priori* but by the point at which the study results became significant (an issue of internal validity). As the sample size is accounted for statistically

when performing a meta-analysis (and the studies weighted accordingly), there is some debate around whether it should additionally be considered as a criterion of quality. However, where studies are not formally pooled, as in narrative reviews, there is a case for including an assessment of the adequacy of sample size as part of the quality assessment process.

Domains largely unrelated to study validity included reporting of the background and context of the study (domain #1), the specification of the interventions and outcomes assessed (domains #3 and #4) and issues related to interpretation and presentation (domains #11 and #12). These domains were judged to be more related to quality of reporting of a study.

Authors' quality items were allocated to our pre-specified items as far as possible, but in some cases additional items were added to accommodate all authors' items. At the analysis stage, the pre-specified items were considered to be 'key' criteria. Four of these criteria were designated 'core' criteria; these were the items that we considered the most relevant for the assessment of non-randomised studies. A total of 18 items were added *post hoc*. Additional items relating to validity as opposed to reporting were as follows: concurrent/concomitant treatment; objective and/or reliable outcome assessment; contamination; testing of blinding; and adequacy of length of follow-up. Other items relating to internal validity on a broader level included 'suitability of study design' and 'maximum blinding used'. The remainder of the additional items related more to presentation and reporting issues.

Tool selection process

Owing to the descriptive nature of the information extracted, the data from each study were tabulated and synthesised in a qualitative manner. It was anticipated that a large number of quality assessment tools and systematic reviews using quality assessment would be identified. In order to reduce the number of tools that were discussed in detail, a primary selection criterion was adopted, as follows.

'Top' tools

For analysis purposes, a 'good' quality assessment tool was deemed to be one that included **pre-specified items** from at least five of the six internal validity domains. The pre-specified items were used both because they had been specified *a priori* and because they were judged to be more directly related to internal validity than the majority of items added subsequently. The degree to which a tool met each of the six domains was displayed using the star plot facility in the statistical software package STATA, release 7.0 (STATA Corp., College Station, TX, USA). A star plot is constructed by assigning each of the six internal validity domains to a separate axis (*Figure 1*). Because the domains did not all have the same number of pre-specified items, each axis was scaled between 0 and 1. A tool that covered each domain, and each item within that domain, would achieve a symmetrical 'star' shape.

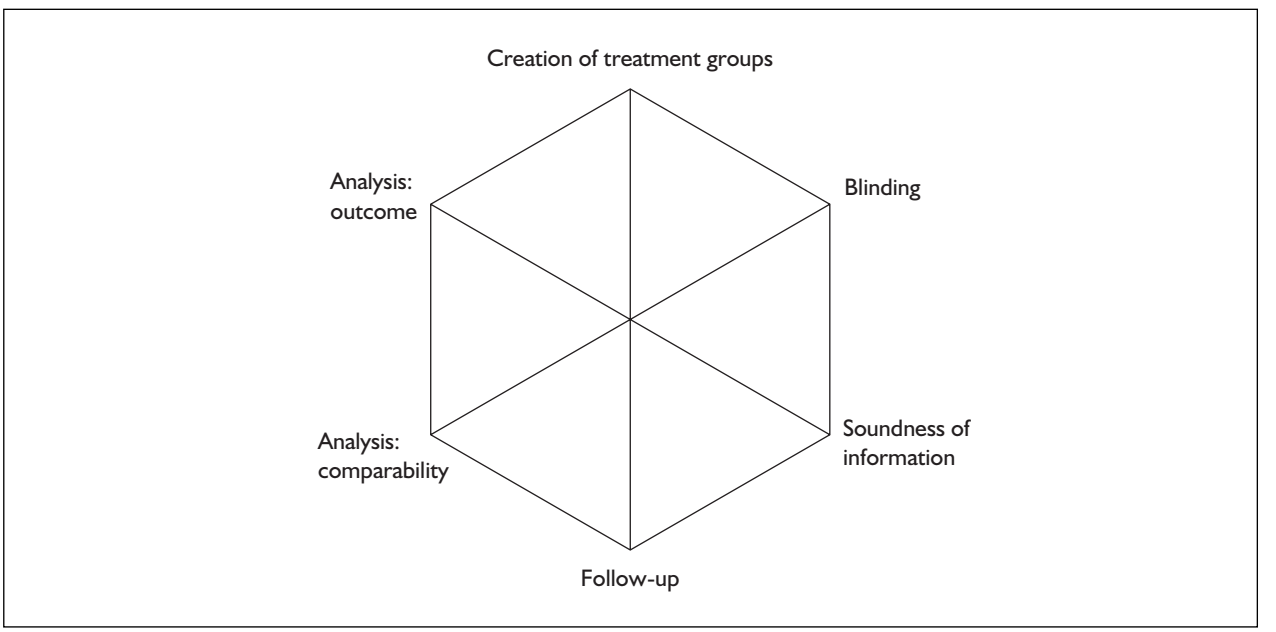

FIGURE 1 *Sample star plot*

The selected tools were described and broadly compared with those tools which did not meet the primary selection criterion, according to the number of times a tool had been used in our sample of reviews, the development of the tool and the purpose for which it had been designed. It was hypothesised that tools developed specifically for use in a systematic review, and particularly ones that were intended for the assessment of non-randomised studies, might be more practical, more focused on issues relevant to the assessment of non-randomised studies and more likely to provide a means of comparing quality across studies.

'Best' tools

Those tools which covered at least three of the four core items were considered to be the 'best' tools and were evaluated in more detail. Members of the project team used each tool at least twice, on one of three non-randomised studies. The first of these was a non-randomised trial of single versus multiple mammary artery bypass for patients undergoing coronary revascularisation,[61] where two surgeons each primarily conducted one of two types of surgery such that patient allocation occurred according to the surgeon from which the patient received their clinical consultation. The second study was a retrospective concurrent cohort study of pneumococcal vaccination, where allocation occurred according to clinician and/or patient preference.[62] The final study was an experimental before-and-after study on a single group of patients with diabetic nephropathy to examine the effect of restricting dietary protein on renal failure.[63]

For each tool, the reviewer completed an appraisal form to indicate how long it took to complete the tool, the ease of use of the tool, whether any items were ambiguous or difficult to interpret or were missing from the tool and any comments regarding the tool's suitability for use in a systematic review. The results of the appraisal are discussed narratively and examples of the types of questions included by the tool's authors are provided.

Results

General description

A total of 213 potential quality assessment tools were identified. On closer inspection, 13 of the cited papers did not present tools for quality assessment, and seven related to 'levels of evidence' to categorise studies by design as opposed to investigations of quality. Full details of the remaining 193 tools are provided in Appendix 3, and selected features are presented in *Table 7*. Throughout, tools are referred to either by the tool name or by the principal author. Eleven tools developed by the authors of systematic reviews did not provide a list of the quality items assessed and could not be included in the content analysis. Sixty of the 182 tools that did report the quality items were selected as 'top' tools, covering at least five of the six internal validity domains (*Table 8*). Of these, 14 met the criteria for 'best tools'. *Figure 2* provides a flowchart of the tool selection process.

Only three of the identified tools were unpublished,[64–66] with the remainder published in journals or book chapters. All three unpublished tools were identified through contact with experts in the field and were all selected as 'best tools'. Almost three-quarters of the tools (73%) were 'new' (or did not state the origin of the tool) and the remainder were modifications of a single existing tool. The most commonly modified tools were Chalmers and colleagues[51] and the Maastricht Criteria list,[67] the latter having been published several times by different authors,[68–73] but originating in 1989.[74,75]

Most (182) of the identified tools aimed to assess more than one study design. Only 23 of these included items specific to different designs; the rest included generic items to be applied to all designs. Eight of the tools were designed specifically to assess only RCTs but had been used to assess non-randomised studies, including the Chalmers scale[51] plus two modifications,[76,77] two versions of the Maastricht Criteria List,[70,71] plus three others.[46,78–80] The remaining four tools were designed specifically for cohort studies only [Anders,[81] Baker,[82] Critical Appraisal Skills Programme (CASP)[64] and Newcastle–Ottawa[66]].

Overall, around half of the tools were scales and half were checklists, although a higher proportion of the final sample of 'best tools' were checklists rather than scales (79%). Of those scales that reported the weighting scheme used, around one-third implemented an unequal weighting system. Most of those using unequal weighting did not develop an entirely new weighting scheme: of the 'top tools', nine were derivatives of the Chalmers scale[51] and three were based on the Maastricht Criteria list.[67] In spite of this, a variety of items received the highest weighting, including randomisation or the method of allocation used, allocation concealment, sample size and

TABLE 7 *Selected features of identified quality assessment tools*

Tool feature	Best tools (*n* = 14)	Top tools (*n* = 46)	Other tools (*n* = 133)	Total (*n* = 193)
Publication status				
Published	11 (79%)	46 (100%)	133 (100%)	190 (98%)
Unpublished	3 (21%)	0	0	3 (2%)
Type of tool				
Checklist	11 (79%)	18 (39%)	75 (56%)	104 (54%)
Scale	3 (21%)	28 (61%)	58 (44%)	89 (46%)
If scale				
Equal weighting	1 (33%)	14 (50%)	30 (51%)	45 (50%)
Unequal weighting	2 (67%)	14 (50%)	16 (29%)	32 (37%)
Not described	0	0	12 (20%)	12 (13%)
Tool origin				
New tool	12 (86%)	29 (63%)	101 (76%)	142 (74%)
Modified tool	2 (14%)	17 (37%)	32 (24%)	51 (26%)
Tool development				
Literature/consensus	6 (43%)	26 (57%)	41 (31%)	73 (37%)
Survey/expert panel	3 (21%)	3 (7%)	5 (4%)	11 (5%)
Not described	8 (36%)	17 (36%)	87 (65%)	112 (58%)
Validity/reliability				
Validity	1 (7%)	3 (7%)	5 (4%)	9 (5%)
Reliability	2 (14%)	22 (48%)	39 (29%)	54 (28%)
Tool purpose				
Study planning	0	2 (4%)	1 (1%)	3 (2%)
Statistical analysis	0	2 (4%)	13 (10%)	15 (8%)
Critical appraisal	5 (36%)	12 (26%)	22 (16%)	39 (20%)
Guidelines	0	1 (2%)	0	1 (1%)
Peer review	0	0	1 (1%)	1 (1%)
Reporting	1 (7%)	1 (2%)	3 (2%)	5 (3%)
Systematic review	8 (57%)	28 (61%)	93 (70%)	129 (65%)
Number of items				
Median (range)	22 (8–103)	17 (7–49)	10 (3–162)	12 (3–162)

description of the interventions. The basis for the weighting system adopted was rarely described.

The method used to develop the tool was rarely reported in detail. Most tools were based on some form of consensus, gleaned either from the methodological literature, existing quality assessment tools or expert opinion. A total of 11 tools used either surveys or panels of experts to select the quality items and piloted and/or circulated the list of items to experts for comment and revision.[46,66,83–91] Twenty tools gave no indication of how the items listed had been generated.

A very small minority of tools attempted to establish the validity of the tool (only four of the top 60 tools). Of these, two found face and content validity to be reasonable.[84,85] All four stated that they had examined criterion validity. Two of these compared their tools with that of Detsky and colleagues (a reduced version of the Chalmers scale for assessing the quality of RCTs):[92] one[84] found a significant difference in mean scores ($p < 0.001$), but a similar rank order of scores (Kendall's $W = 0.74$); the other[93] did not present the results of the comparison. Downs and Black[85] found a high correlation (Spearman correlation coefficient 0.90) between the total score from their tool (for both randomised and non-randomised studies) and that obtained for the Standards of Reporting Trials Group tool[94] (used on randomised studies only). Linde and colleagues[95] found similar results from their tool and that of Jadad and colleagues,[46] but did not carry out a statistical comparison.

Twenty-four of the top 60 tools assessed inter-rater reliability. Two of these stated only that agreement was mediocre or good and nine provided only percentage agreement between raters (range 70–94%). Where provided (13 studies), kappa or other correlation coefficients were generally >0.75 ($n = 12$), indicating good agreement. A few

TABLE 8 *Details of top 60 quality assessment tools (14 'best' tools denoted by shaded areas)*

Author[a]	Originality[b]	Type of tool[c]	No. of items[d]	NRS items?[e]	Used in reviews?[f]	Tool purpose[g]	Tool validity[h]	Tool reliability[i]	⩾5 IV domains[j]	Core items[k]	⩾5 domains and ⩾3 core items[l]
Antczak, 1986[76]	m; Chalmers, 1981[51]	s u	30	n	2	7	na	na	Y	0	
Audet, 1993[176]	m; Poynard, 1988[403]	s e	17	n	1	7	na	IRR	Y	1	
Bours, 1998[73]	m; van der Windt, 1995[71]	s u	18	y	n	7	na	IaRR	Y	1	
Bracken, 1989[104]	n	c	36	n	n	6	na	na	Y	4	Y
Cameron, 2000[83]	n	c	36	n	n	7	na	na	Y	0	
Campos-Outcalt, 1995[177]	n	s e	7	n	n	7	na	na	Y	1	
Carter, 1994[178]	m; Chalmers, 1981[51]	s u	12	n	n	7	na	na	Y	0	
CASP, 1999[64]	n	c	10	n	n	3	na	na	Y	3	Y
Chalmers, 1981[51]	n	s u	36	n	15	3	na	na	Y	0	
Cho, 1994[84]	m; Spitzer, 1990[105]	s e	31	n	n	3	FC, Cr	IRR	Y	2	
Cochrane MIG, 2002[179]	m; Chalmers, 1981[51]	s e	12	n	CC	7	na	na	Y	1	
Coleridge Smith, 1999[97]	n	c	49	y	n	7	na	IRR	Y	2	
Cowley, 1995[109]	n	c	17	n	n	7	na	na	Y	3	Y
Cuddy, 1983[180]	n	c	17	n	n	3	na	na	Y	2	
Dawson-Saunders, 1990[98]	n	c	49	y	n	3	na	na	Y	2	
de Oliveira, 1995[96]	m; Chalmers, 1981[51]	s u	30	n	1	7	na	IRR	Y	1	
de Vet, 1997[72]	m; ter Riet, 1990[68]	s e	15	n	n	7	na	IRR[67]	Y	2	
Downs, 1998[85]	n	s u	27	n	3	3	FC, Cr	IRR; T-R	Y	3	Y
DuRant, 1994[99]	n	c	103	y	n	3	na	na	Y	3	Y
Fowkes, 1991[107]	n	c	23	n	1	3	na	na	Y	3	Y
Friedenreich, 1993[100]	n	c	29	y	n	7	na	na	Y	2	
Gardner, 1986[181]	n	c	26	n	n	2	na	na	Y	2	
Glantz, 1997[182]	m; Chalmers, 1981[51]	s u	20	n	n	7	na	na	Y	0	
Gordis, 1990[101]	n	c	17	y	1	3	na	na	Y	2	
Greenhalgh, 1997[183]	n	c	11	n	n	3	na	na	Y	1	
Gurman, 1978[184]	n	s u	14	n	n	2	na	na	Y	1	
Guyatt, 1993[79]	n	c	12	n	1	3	na	IRR[185]	Y	0	
Hadorn, 1996[102]	m; US Task Force, 1989[378]	c	41	y	n	4	na	na	Y	3	Y
Kay, 1996[186]	n	s e	20	n	n	7	na	na	Y	0	
Koes, 1991[70]	m; ter Riet, 1990[68]	s u	17	n	3	7	na	IaRR	Y	0	
Kreulen, 1998[187]	m; Antzcak, 1986[76]	s u	17	n	n	7	na	IRR	Y	1	
Kwakkel, 1997[188]	n	s e	16	n	n	7	na	IRR	Y	1	
Lee, 1997[189]	m; Cook, 1979[50]	s e	7	n	n	7	na	IaRR	Y	2	
Levine, 1980[190]	n	s e	30	n	1	3	na	na	Y	1	
Linde, 1999[95]	n	s e	7	n	1	2/7	Cr	na	Y	0	
MacMillan, 1994[191]	m; Chalmers, 1981[51]	s u	9	n	n	7	na	IaRR	Y	1	
Maziak, 1998[192]	n	s e	10	n	n	7	na	IRR	Y	0	
Meijman, 1995[193]	m; Fowkes, 1991[107]	s e	11	n	n	2	na	no stats	Y	1	

continued

TABLE 8 *Details of top 60 quality assessment tools (14 'best' tools denoted by shaded areas) (cont'd)*

Author[a]	Originality[b]	Type of tool[c]	No. of items[d]	NRS items?[e]	Used in reviews?[f]	Tool purpose[g]	Tool validity[h]	Tool reliability[i]	⩾5 IV domains[j]	Core items[k]	⩾5 domains and ⩾3 core items[l]
Melchart, 1994[93]	n	s e	15	n	n	7	Cr	IaRR	Y	0	
Miller, 1995[87]	n	s u	12	y	1	7	na	IaRR	Y	2	
Moncrieff, 1998[195]	n	s e	30	n	n	3	na	IRR; IC	Y	1	
Morley 1996[196]	m; Chalmers, 1981[51]	s u	30	n	1	2/6	na	IRR	Y	2	
Mulrow, 1986[197]	n	c	13	n	n	7	na	IaRR	Y	2	
Newcastle–Ottawa[66]	n	s u	8	y	n	7	na	na	Y	3	Y
Nicolucci, 1989[77]	m; Chalmers, 1981[51]	s u	16	n	1	7	na	IaRR	Y	0	
Reisch, 1989[111]	n	s e	57	n	CC	1/3	na	IRR[114]	Y	3	Y
Salisbury, 1997[198]	n	c	45	n	n	7	na	na	Y	2	
Schechter, 1991[199]	n	c	30	n	n	3	na	na	Y	2	
Sheldon, 1993[200]	n	c	36	n	n	3	na	na	Y	2	
Spitzer, 1990[105]	n	c	20	n	1	7	na	na	Y	4	Y
Talley, 1993[106]	n	c	29	n	n	7	na	na	Y	0	
Thomas[65]	n	c	21	n	n	7	na	na	Y	3	Y
van der Windt, 1995[71]	m; Koes, 1991[70]	s u	17	n	1	7	na	IaRR	Y	1	
Vickers, 1996[201]	n	c	12	n	n	7	na	no stats	Y	1	
Vickers, 1995[110]	n	c	21	n	n	3	na	na	Y	3	Y
Weintraub, 1982[108]	m; Lionel, 1970[315]	c	47	n	n	1/2/3	na	na	Y	3	Y
Wilson, 1992[103]	n	c	10	y	n	7	na	na	Y	1	
Wingood, 1996[202]	n	c	16	n	n	7	na	na	Y	0	
Wright, 1995[203]	n	c	13?	n	n	7	na	na	Y	0	
Zaza, 2000[86]	n	c	22	n	n	7	na	na	Y	3	Y

[a] Name of tool or principal author.
[b] n, New tool; m, modification of existing tool.
[c] c, Checklist; s, scale, u, unequal weighting scheme, e, equal weighting scheme.
[d] Total number of items.
[e] Did the tool include items specific to non-randomised studies? n, no; y, yes.
[f] Has the tool been used in identified sample of systematic reviews?: n, no; CC, used in Cochrane reviews; or give number of reviews.
[g] What purpose was the tool designed for? 1, list for planning a study; 2, list for assessing a statistical/methodological analysis; 3, list for evaluating a study when considering the practical application of the intervention; 4, list for evaluating/grading study recommendations when producing recommendations/guidelines; 5, list for peer reviewing; 6, list for reporting a study; 7, list for assessing studies included in a systematic review.
[h] Was an attempt made to establish tool validity? na, not assessed; FC, face and content; Cr, criterion; Co, content.
[i] Was an attempt made to establish tool reliability? na, not assessed; IRR, inter-rater reliability; IaRR, intra-rater reliability; IC, internal consistency; T-R, test-retest; a, % agreement assessed.
[j] Were at least 5 internal validity domains covered? Y, yes.
[k] Number of core items covered.
[l] Were at least 5 internal validity domains and at least 3 of the 4 key items covered? Y, yes.

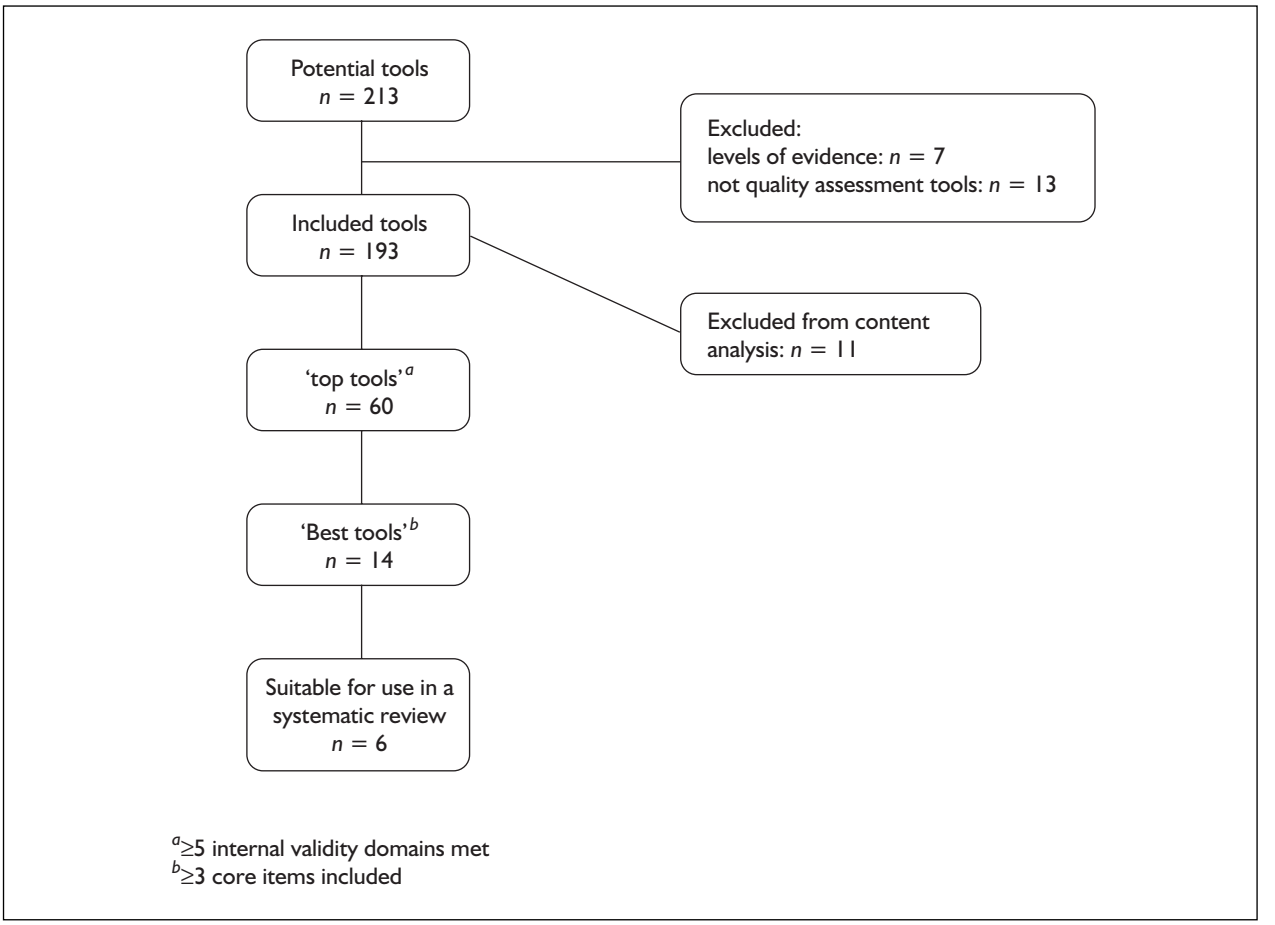

FIGURE 2 *Flowchart of tool selection process*

studies found the level of agreement to be dependent on the experience of the raters. de Oliveira and colleagues[96] achieved higher agreement for experienced raters, whereas Coleridge Smith[97] found higher agreement for two methodologists combined rather than clinicians alone or for a clinical/methodologist pairing. Other aspects of reliability were rarely examined.

The majority of the identified tools were designed for use in a systematic review (and most were in fact published within the context of a systematic review). Of the remainder, most were critical appraisal tools, designed to assist the reader when considering the usefulness of a particular intervention in their situation, rather than to be used to compare quality across studies.

The number of items in the tools ranged from 3 to 162, with a median of 12 items. Those tools selected as 'top tools' or 'best tools' had a higher median number of items compared with the unselected tools. Those with the largest number of items (>50) were more likely to have separate sections and questions according to study design (e.g. including items specific to RCTs, cohort studies, cross-sectional studies).

Possibly the most relevant aspect of these tools for application in practice is how well they cover questions of internal validity and, in particular, whether they covered the four core items. Eleven of the 193 tools did not specify the items included in the tools. Of the remainder, 60 tools covered at least five of the six internal validity domains. Individual star plots demonstrating how well, or to what extent, these tools covered each of the six internal validity domains are provided in *Figure 3*. These plots indicate a significant variation in how well these domains were covered. *Figures 4* and *5* provide information on how often each domain was missed and how well the tools covered each domain. The domain most commonly missed was soundness of information (38%), with only 37% of tools including an item relating to whether the interventions had been received by the study participants and only 25% including an item on whether or not the outcomes were attributable. All of the top 60 tools included one of the two

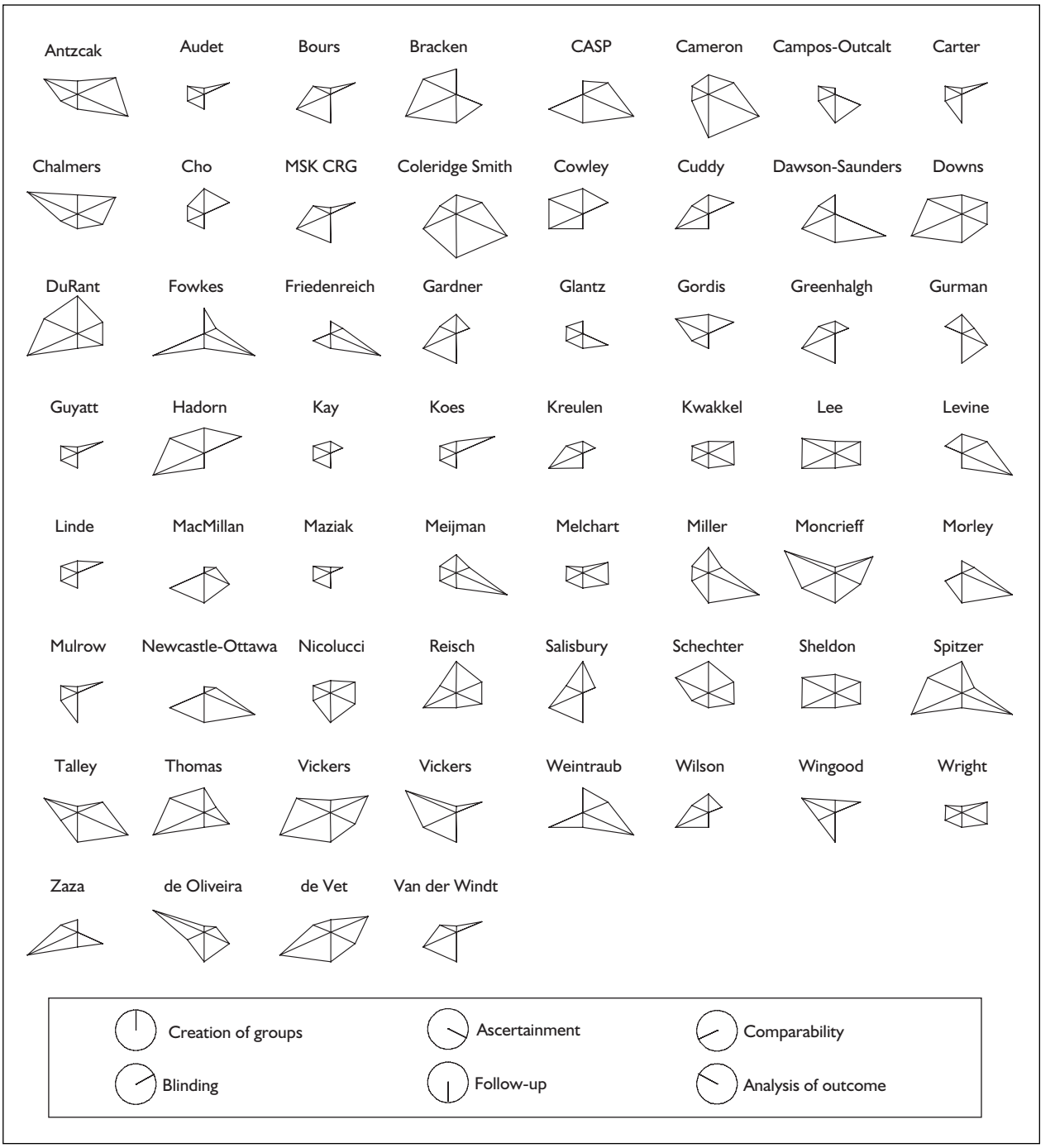

FIGURE 3 *Top 60 star plots*

pre-specified items in the follow-up domain (most commonly equality of length of follow-up between groups (88% of tools)).

On average, about half of the items that we pre-specified were included in the tools (*Figure 5*). The domain with the best coverage was blinding. The two core domains, creation of groups and comparability of groups, had on average 42 and 55% of possible pre-specified items, respectively.

Six of the eight tools designed only for RCTs were included in the top 60. These were the Chalmers scale[51] and its derivatives,[76,77] the two versions of the Maastricht Criteria List[70,71] and the *JAMA* Users' Guide checklist.[78,79] The Jadad[46] and McMaster[80] tools did not include items in five of the six domains. In contrast, nine of the 23 tools aimed at more than one study design[73,87,97–103] and with specific items according to design, plus two of the four tools designed for cohort studies,[64,66] were included in the top 60 tools.

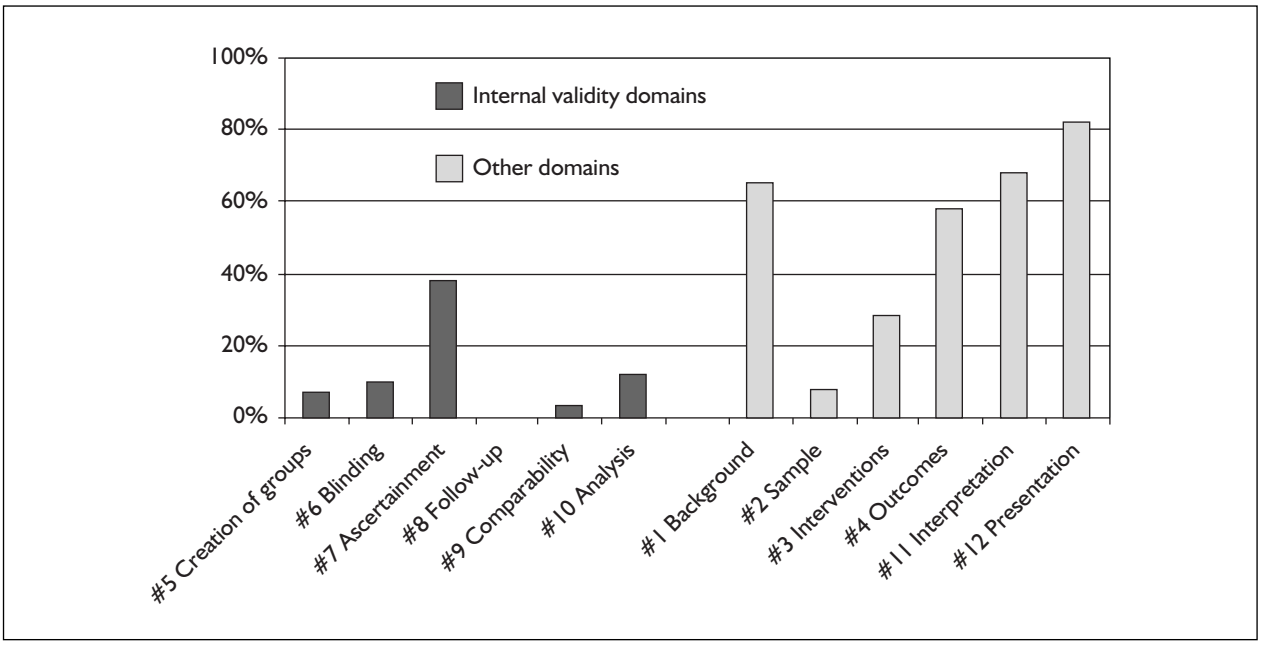

FIGURE 4 *Proportion of top 60 tools that missed each domain*

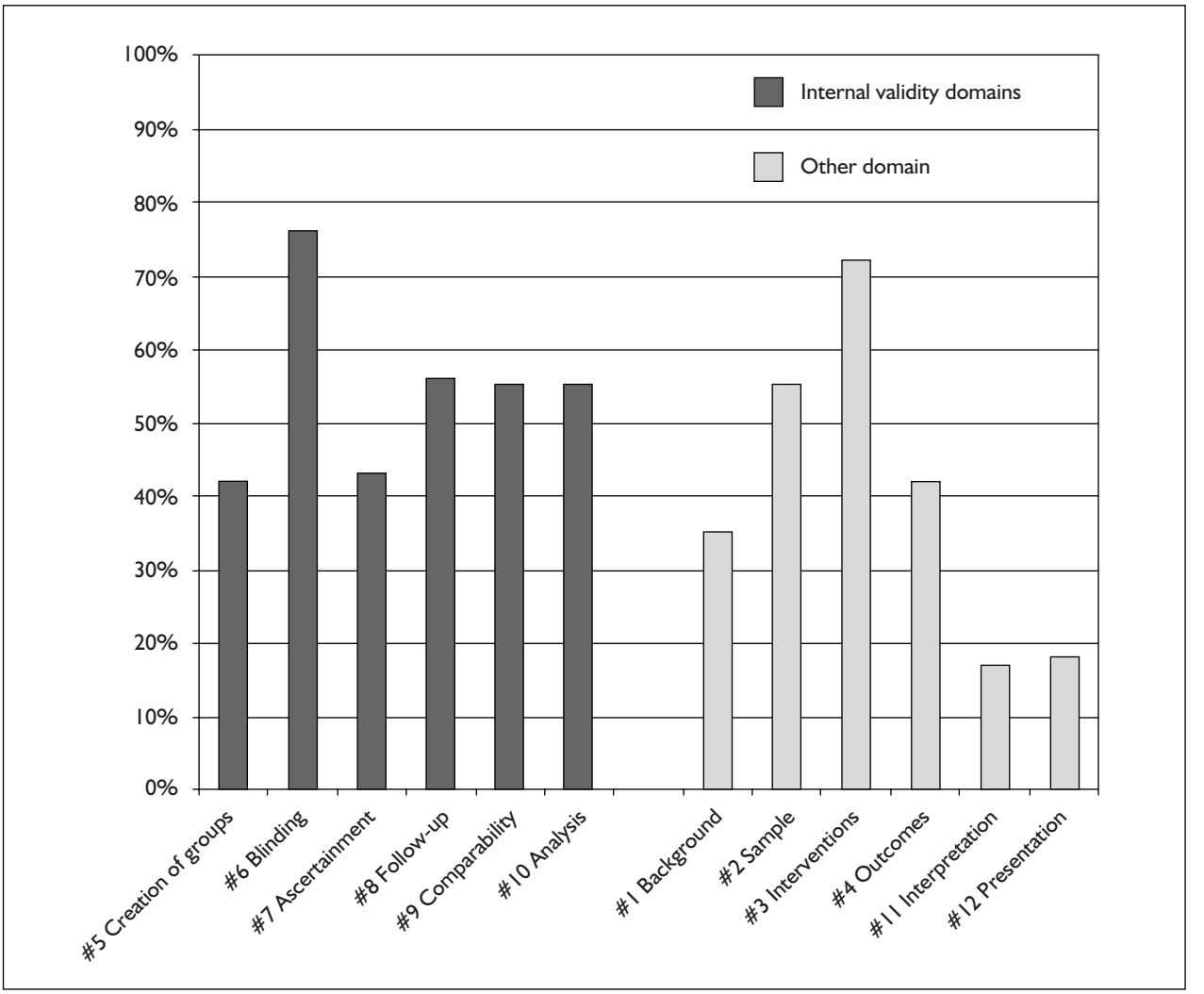

FIGURE 5 *Average proportion of pre-specified items met per domain (mean/maximum number of items)*

In terms of the four pre-specified core items, 15 of the 60 tools included none of the core items despite covering at least five domains, 16 covered one core item and 15 covered two items. The remaining 14 tools covered at least three core items and were considered to be the 'best' tools in our sample, two of which covered all four items.[104,105] It is interesting that of the six tools that included items in all six internal validity domains,[72,77,83,85,97,106] only one[85] included three of our four core items. None of the tools designed only for RCTs included three of the four core items.

'Best' tools

Fourteen tools were identified which covered at least five of the six internal validity domains and three of the four core items. *Tables 9* and *10* itemise the pre-specified items covered by each tool.

Amongst the top 14 tools, the internal validity domain with the poorest coverage was analysis (four tools with zero items – CASP,[64] Fowkes,[107] Newcastle–Ottawa[66] and Weintraub[108]), followed by blinding (missed by Bracken[104] and Zaza[86]) and ascertainment (missed by Cowley[109] and Hadorn[102]). The item most commonly missed was equal follow-up between groups (included by only two tools – Bracken[104] and Downs[85]). Only three tools asked about use of intention-to-treat analysis (Cowley,[109] Thomas[65] and Vickers[110]).

The two core domains were reasonably well covered. For the creation of groups domain, all of the tools except those specifically designed only for observational studies (Bracken,[104] CASP[64] and Newcastle–Ottawa[66]) included an item on randomisation, but only two tools specifically considered the use of allocation concealment (Downs[85] and DuRant[99]). Of the four core items, the most commonly missed item was that relating to how allocation occurred. Only eight tools included this item. Ideally, we were looking for an item that asked about how participants got into their respective groups, for example, was it by clinician or patient preference or was it spatial or temporal assignment. All of the tools except for Downs,[85] DuRant[99] and Hadorn[102] included the second pre-specified item – balancing of groups by design.

For the comparability of groups domain the two pre-specified items – identification of prognostic factors and use of case-mix adjustment – were missed by only two tools[109,111] and one tool,[108] respectively. All of the tools except the CASP tool[64] and the Newcastle–Ottawa tool[66] asked if baseline comparability had been assessed.

Our pre-specified items in the remaining six domains (*Table 10*) were, on the whole, not well covered, except perhaps for that relating to the selection of the study sample. Every tool included an item about the representativeness of the sample, and only four did not ask about the study inclusion/exclusion criteria (CASP,[64] Cowley,[109] Newcastle–Ottawa[66] and Thomas[65]). One of the items in this domain that is related to internal validity – retrospective or prospective selection of the sample – was included by only two tools, Cowley[109] and Reisch.[111]

The remaining five domains concerning the quality of study reporting were not well covered by the tools. The most commonly included item was one that considered clear specification of the interventions (nine tools). On the other hand, clear specification of the outcomes was included in only five tools.

Qualitative assessment of the 'best' tools

Of the best 14 tools, eight were judged to be unsuitable for use in a systematic review.[64,99,102,104,105,107,108,110] A description of which of the core criteria they covered and our assessment of them is provided in Appendix 4.

In summary, their unsuitability was largely related to the fact that they were not designed for use in a systematic review of effectiveness: one was published to guide the reporting of observational studies;[104] five were intended to help in the critical appraisal of research articles;[64,99,107,108,110] and one was developed for an epidemiological review.[105] Overall, these tools generally prompted some thinking regarding quality issues, but were not formatted in such a way as to allow an overall assessment of study quality or the comparison of quality across studies. Some[64,108] did conclude with a more general item requiring a judgement on the overall quality of the study, but little guidance was provided as to how this judgement should be made. The Hadorn tool[102] was intended for use in systematic reviews, but the assessors queried the inclusion of, or phrasing of, several of the items. For example, the emphasis on drug trials and use of placebos was felt to be overly specific.

Six quality assessment tools were judged to be potentially useful for systematic reviews,[65,66,85,86,109,111] although in several cases some modifications would be useful. All but one of

TABLE 9 *Selected features of identified quality assessment tools*

		Bracken[104]	CASP[64]	Cowley[109]	Downs[85]	DuRant[99]	Fowkes[107]	Hadorn[102]	Newcastle-Ottawa[66]	Reisch[111]	Spitzer[105]	Thomas[65]	Vickers[110]	Weintraub[108]	Zaza[86]
#5	**Creation of groups**														
	5.1. Generation of random sequence			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
	5.2. Concealment of allocation				✓	✓									
	5.3. How allocation occurred[a]	✓		✓	✓	✓		✓		✓	✓			✓	
	5.4. Balance groups by design[a]	✓	✓	✓			✓		✓	✓	✓	✓	✓	✓	✓
#6	**Blinding**														
	6.1. Blind (or double-blind) administration		✓	✓	✓	✓		✓		✓			✓	✓	
	6.2. Blind outcome assessment		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
#7	**Ascertainment**														
	7.1. Receipt of the intervention		✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
	7.2. Attributable outcomes	✓	✓				✓		✓		✓			✓	
#8	**Follow-up**														
	8.1. Equal follow-up between groups	✓			✓										
	8.3. Completeness of follow-up	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
#9	**Comparability**														
	9.1. Baseline comparability assessed	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
	9.2. Prognostic factors identified[a]	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
	9.3. Case-mix adjustment[a]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
#10	**Analysis**														
	10.1. Intention-to-treat analysis			✓								✓	✓		
	10.2. Appropriate analysis methods	✓		✓	✓	✓		✓		✓	✓	✓	✓		✓

[a] Items specfic to assessment of non-randomised studies.

TABLE 10 *Other domains: reporting and external validity*

		Bracken[104]	CASP[64]	Cowley[109]	Downs[85]	DuRant[99]	Fowkes[107]	Hadorn[102]	Newcastle-Ottawa[66]	Reisch[111]	Spitzer[105]	Thomas[65]	Vickers[110]	Weintraub[108]	Zaza[86]
#1	**Background**														
	1.1. Background information provided	✓				✓									
#2	**Sample**														
	2.1. Retrospective/prospective			✓						✓					
	2.2. Inclusion/exclusion criteria	✓			✓	✓	✓	✓		✓	✓		✓	✓	✓
	2.3. Sample size	✓			✓	✓	✓			✓	✓		✓	✓	
	2.4. Representative	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
#3	**Interventions**														
	3.1. Clear specification of interventions			✓	✓	✓		✓		✓	✓		✓	✓	✓
#4	**Outcomes**														
	4.1. Clear specification of outcomes			✓	✓			✓		✓			✓		
#11	**Interpretation**														
	11.1. Appropriately based on results	✓				✓					✓			✓	
	11.2. Assessed strength of evidence	✓	✓											✓	
	11.3. Application/implications	✓	✓			✓				✓					
#12	**Presentation**														
	12.1. Completeness, clarity, structure		✓		✓	✓				✓					

the tools in this group were specifically developed for use in a systematic review,[65,66,85,86,109] their main advantage being that the questions and responses were phrased in such a way as to force the reviewers to be systematic in their assessment of individual studies, and ensure that between-study quality can be compared.

The length and complexity of the tools varied, but most provided guides for completion of the tools and three of the tools were considered 'easy to use' in our assessment.[65,85,109] Although a certain degree of judgement is required for completion of any tool (including phrases such as 'appropriateness of allocation', 'truly representative' and 'adequately described'), the guides help the reader to interpret the questions and response options. In several of these tools,[66,85,109,111] the responses to individual questions could be used to contribute to an overall judgement of study quality, using either number of items fulfilled[66,109,111] or summary scores.[85] Two of the selected tools[65,86] used a 'mixed-criteria' approach, including specific factual questions about the study design and also requiring more general judgements as to the degree of bias present in a study.

A summary of each of the six selected tools follows, with details of how they covered the four core items provided in Appendix 5.

Cowley

Cowley[109] developed a quality assessment tool for use in a systematic review of total hip replacement. Separate items were provided for RCTs, comparative studies and uncontrolled case series. The tool provides a list of 13 questions for the quality assessment of comparative studies, split into two sections: 'key criteria' (seven items) and 'other criteria' (six items). No details of the method of item generation were provided. Studies were rated A, B or C, according to the number of 'key' and 'other' criteria that were met. Key criteria included the method of assignment of patients to groups, matching of groups for key prognostic variables, appropriate statistical analysis and specification of interventions and outcomes. Several items were topic-specific. No items for sample size and method of sample selection were provided. A mixture of items relating to both reporting and methodological quality were included. The authors did not attempt to establish the validity or reliability of the tool, and did not provide a guide to completion. However, we found the criteria quick and easy to apply, taking 5–15 minutes to complete the checklist. With some modifications, the Cowley tool was judged to be suitable for use in a systematic review.

Downs

Downs and Black[85] developed a scale to assess the methodological quality of randomised and non-randomised studies. The tool was used in two systematic reviews in our sample,[26,112] and appears to have been developed for such use. It provides a list of 27 questions to measure study quality, split into four sections: 'reporting' (10 items); 'external validity' (three items); 'internal validity – bias' (seven items); and 'internal validity – confounding (selection bias)' (seven items). Epidemiological principles and methodological literature were used to generate an initial version of the tool, which was then piloted by experienced epidemiologists and statisticians and a revised version produced.

The tool was found to be fairly comprehensive; however, questions regarding the allocation mechanism used relate only to randomised studies, and no items relating to baseline comparability were included. Nevertheless, it was found to be easy to use, with yes/no/can't tell answers and clear descriptions of how to score items. The tool authors found the validity and reliability of the tool to be reasonably high, except for the three items on external validity. Although relatively quick to complete (10–20 minutes), the tool is fairly long and a large number of questions relate to reporting as opposed to validity. The Downs tool was judged to be suitable for use in a systematic review, although some modification, such as the addition of the missing items (e.g. baseline comparability), may be warranted.

Newcastle–Ottawa

The Newcastle–Ottawa tool[66] was designed for use in an epidemiological systematic review, and can be used as either a scale or a checklist. It was developed using a Delphi process to define variables for data extraction, was then tested on systematic reviews and further refined. Separate tools were developed for cohort and case–control studies, although only the tool for cohort studies is considered here. The tool contains eight items, categorised into three groups: selection, comparability and outcome. For each item a series of response options are provided. For example, for the consideration of the representativeness of the exposed cohort, response options include truly representative, somewhat representative, selected group or no description of exposed cohort. A star system is used to allow a visual semi-quantitative assessment of study quality, such that the highest quality studies are awarded a maximum of one

star for each item within the selection and outcome categories and a maximum of two stars for comparability.

Items relating to the appropriateness of the analysis would make the tool more comprehensive in coverage. Further modifications would be necessary to adapt the tool for use in reviews of effectiveness of interventions rather than reviews of causation. For example, an item on method of allocation of participants to groups would be useful. (Such modifications to the tool have in fact been made for use in a systematic review of arterial revascularisation.[113]) No information was provided on the reliability or validity of the tool. Our reviewers found it relatively easy to use, taking only 5–10 minutes to complete. With the aforementioned caveats, the Newcastle–Ottawa tool was judged to be suitable for use in a systematic review.

Reisch

Reisch and colleagues[111] developed their tool to facilitate the evaluation of the design and performance of therapeutic studies and to teach critical appraisal skills. It is currently used by the Cochrane Inflammatory Bowel Disease Group. The tool was based on 'accepted research standards' and lists a total of 57 items grouped into 12 categories including study design, sample size, randomisation and comparison groups. The tool aims to evaluate any study design and is very comprehensive in its coverage of internal validity. It is organised such that a systematic approach to answering each question is ensured. Response options include yes, no, unclear or unknown, and not applicable.

The criteria considered most important were designated as primary criteria (a total of 34 items). The selection of some of the items, however, may be questionable regarding their importance for internal validity. For example, statement of the purpose of the study is more a reporting issue than one of validity. On the other hand, other items selected are very relevant, including both 'randomisation claimed and documented' and 'use of either prognostic stratification prior to study entry or retrospective stratification during data analyses'. The tool is also useful for reviews including non-randomised studies in that it asks whether a randomised design would have been possible. However, some of the criteria are rather too specific to pharmaceutical studies and would require modification for more general use. The inter-rater reliability of the tool was found to be high when assessed in a later study,[114] especially when only the primary criteria were considered (Pearson $r = 0.99$). However, we found the tool to be rather long, taking approximately 25 minutes to complete. Overall the Reisch tool was judged suitable for use in a systematic review, although a shorter version would be more useable.

Thomas

The tool produced by Thomas[65] was developed to assess the methodological quality of studies. It aims to cover any study design and includes 21 items separated into eight components: selection bias, study design, confounders, blinding, data collection methods, withdrawals and dropouts, intervention integrity and analysis. The method of generation of the items included was not reported. A list of response options for each item is provided. Following completion of the tool, reviewers are asked to give an overall rating of the study (strong/moderate/weak) for each of the first six components in order to provide a global quality rating.

This tool is able to deal with both randomised and non-randomised studies, including items on both randomisation and control for confounders, although methods of allocation other than randomisation were not considered. The validity and reliability of the tool were not reported. We found it easy to use, taking 10–15 minutes to complete, with a comprehensive guide for completion provided. The Thomas tool was judged to be suitable for use in a systematic review.

Zaza

The tool by Zaza and colleagues[86] was designed for use as part of a data collection instrument for systematic reviews to inform the US 'Guide to Community Preventive Services: Systematic Reviews and Evidence-based Methods'. Items were generated from a review of methodological literature, and expert opinion was solicited on pilot versions. The tool aims to cover any study design, with 22 quality items, grouped into six categories: descriptions, sampling, measurement, data analysis, interpretation of results and other. It was developed as part of a much larger standardised data abstraction form and detailed instructions for completion and cross-referencing to other parts of the data extraction form are provided.

The items included are not specific to any particular study design but are phrased in such a way that they can be applied to any study design. The authors provide explicit decision rules and

explanation of responses to aid completion. The concept behind the tool seems a good one, but it is difficult to complete in isolation from the rest of the data extraction form. Despite the instructions for completion, some of the items may be too generic and difficult to interpret. For example, allocation of participants to groups is considered in the 'interpretation of results' section under the item 'considering the study design, were appropriate methods for controlling confounding variables and limiting potential biases used'. This item aimed to cover randomisation, restriction, matching, etc. The complexity of the tool meant that it took up to 30 minutes to complete. It may be that the ease of use would increase with practice. The Zaza tool may be suitable for use in systematic reviews but does require a good understanding of validity issues.

Discussion

A total of 213 potential quality assessment tools were identified for inclusion in this review. Overall the tools were poorly developed, with almost no attention to standard principles of scale development.[57] Almost without exception, the items included in the tools were based on so-called 'standard criteria' gleaned from methodological literature, clinical trial or epidemiology textbooks or a review of other quality assessment tools. Most tools did not provide a means of assessing the internal validity of non-randomised studies, and several were aimed specifically at only randomised trials. Only 60 (30%) included items related to five out of six internal validity domains. Of these, 14 were sufficiently comprehensive in their content coverage to be considered in detail. In order to be selected, these tools had to include at least three of four pre-specified core items:

- how allocation occurred
- any attempt to balance groups by design
- identification of prognostic factors
- case-mix adjustment.

These were selected as core items because for non-randomised studies it is important to know, first, how study participants got into the intervention groups, that is, was it by clinician or patient preference, or was it according to spatial or temporal factors and, second, what specific factors influenced the selection of patients into each group. Given that study investigators rarely report the latter information, the identification of prognostic factors that influence outcome (as opposed to allocation) was included as a core item. However, covering these items does not necessarily make the tools useful. For example, asking for a description of the method of allocation[109,111] does not force a judgement about whether that method was appropriate or unlikely to introduce bias.

Other than their content coverage, the top 14 tools did not stand out from the remaining tools in terms of their development, which was often vaguely reported, or the investigation of validity or reliability.

A relatively informal assessment of the usefulness of the top 14 tools identified six that were potentially useful for systematic reviews.[65,66,85,86,109,111] The main advantage of these tools over the remaining eight was the phrasing of the items. On the whole, these force the reviewer to be systematic in their study assessments and attempt to ensure that quality judgements are made in the most objective manner possible. Four of the tools attempted to cover most study designs using the same questions, which could be the most useful approach for a systematic review which incorporates several different study designs. This approach did appear reasonably successful, although it may often be more appropriate to think of quality issues on top of a study design 'hierarchy'. For example, there seems to be little point in being overly concerned with blinding levels when comparing an RCT with a before-and-after study. Furthermore, in many cases a full assessment of study quality may require additional context-specific questions to cover aspects of external validity that would not be included in a generic quality assessment tool, for example, items relating to the quality of delivery of an intervention or quality of outcome measurements.

Many of the tools, including some in the top 14,[85,111] contained several items unrelated to methodological quality. Although it is important to distinguish the quality of a study from the quality of its report, one of the identified problems with scales for assessing RCTs is the inclusion of items relating to issues such as reporting quality.[59]

Those tools which followed the lines of Cooper's 'mixed-criteria' approach,[45] requiring objective facts regarding a study's design followed by a quality judgement (e.g. Thomas[65]), may prove to be the most useful for systematic reviews. Such tools make as explicit as possible the facts regarding a study's methods that should underlie a reviewer's judgement. Some tools were found that ignore judgements entirely and others

provoke thinking around quality issues but do not provide a means of being systematic (e.g. tools to help critical appraisal of studies). Both of these approaches seem equally unacceptable.

One of the main advantages of our review is that we identified a large number of quality assessment tools, developed for use in a wide range of fields and published in a variety of formats. Although it would not be possible to identify all available tools, we did attempt to be as systematic as possible. Our search covered a wide range of databases and included both published and unpublished material. As a result, we are likely to have identified a large proportion of available quality assessment tools. In addition, all tools were assessed in a systematic manner, by two independent reviewers.

Nevertheless, several criticisms could be put forward. First, we could be criticised for being over-inclusive as we did not require the tools to state that their purpose was the methodological quality assessment of non-randomised studies. We required only that the tool had been or could potentially be used to assess such studies. This resulted in the inclusion of a wide range of tools, including some originally developed for RCTs. We feel that this approach can be justified as our aim was to provide a comprehensive picture of how non-randomised studies are being assessed and what other authors consider to be important quality features. Furthermore, we had no *a priori* reason to assume that tools developed for RCTs would not be suitable for use with non-randomised studies, with or without adaptation.

Second, the list of quality items against which we assessed the included tools was of our own generation, and as such, could be questioned. Our strategy was to take everything that is known about quality for randomised studies and add what we know to be different about non-randomised studies. It should be noted that the aim of the taxonomy was to provide an aid to data extraction of the identified tools as opposed to providing an exhaustive or comprehensive list of quality items. The alternative would have been to extract details of all the items from each quality assessment tool and then try to group them, but the sheer volume of tools, and the use of varying terminology and question phrasing, would have made this impossible.

Furthermore, the selection process that we used to identify the 'best' tools (i.e. items in five out of six internal validity domains plus at least three of four pre-specified core items) meant that those in the final sample of 14 would not necessarily have the best coverage of the six domains. However, these selection criteria were chosen first to reflect our *a priori* position that degree of selection bias is a fundamental issue that needs to be addressed by quality assessment tools (hence the selection of the four core items) and second because a more restrictive policy, for example requiring items in all six domains, would have led to the final assessment of only six tools for their 'usability' in systematic reviews.

Another criticism that could be made relates to our assessment of the 'usability' of the best tools. This was based on their application to only three non-randomised studies and could benefit from repetition on a larger sample of studies.

Future research should include:

- further appraisal of the quality criteria that we selected
- consideration of the creation of a new tool or the revision of an existing one
- more detailed examination of tool 'usability', for example, tools may be more or less useful according to field or type of intervention.

Chapter 5

Use of quality assessment in systematic reviews of non-randomised studies

Introduction

As we have discussed in some detail in previous chapters, the principal difference between randomised and non-randomised studies lies in the latter's considerable susceptibility to selection bias. Concealed randomisation specifically removes the possibility of selection bias or confounding in RCTs, i.e. any differences between the groups are attributable to chance or to the intervention, all else being equal.[20] For non-randomised studies, confounding between groups is likely. Whatever the method of allocation, the reasons for the choice of an intervention can be influenced by subtle clues that are not easily identifiable but which may relate, for example, to the patient's prognosis.[3] Treatment selection may be confounded by differences in case-mix, but also by use of concomitant healthcare interventions or variation in the outcome assessment process, for example where allocation is influenced by geographical location or temporal differences. Any assessment of the degree of selection bias introduced by non-random methods of allocation is of primary importance for systematic reviews of non-randomised studies.

Empirical evidence for the impact of assessing study quality on the results of systematic reviews is limited. However, work on RCTs generally indicates that low-quality trials can lead to a distortion of true treatment effects.[20,21,59,115,116] Furthermore, it has been shown that the choice of quality scale can dramatically influence the interpretation of meta-analyses, and can even reverse conclusions regarding the effectiveness of an intervention.[59] This lends support to the need for careful choice of a quality assessment tool and for consideration of the impact of study quality on a review's conclusions.

Recent guidelines for the reporting of meta-analyses of observational studies[117] recommend the "assessment of confounding, study quality and heterogeneity" to be clearly reported in the methods section of these reviews but moreover that "thorough specification of quality assessment can contribute to understanding some of the variations in the observational studies themselves". The results of quality assessment can be used in a systematic review in several ways,[92] including forming inclusion criteria for the review or primary analysis; informing a sensitivity analysis or meta-regression; weighting studies; or highlighting areas of methodological quality poorly addressed by the included studies and the impact of this on the review's conclusions.

The objectives of the study described in this chapter are to estimate the extent to which quality assessment has been used in systematic reviews of non-randomised studies and can adequately identify the sources of potential bias, and to describe the methods used to incorporate the results of the quality assessment into the conclusions of a review.

Methods

Study selection

We sought systematic reviews evaluating the intended effect of an intervention and that included at least one non-randomised study.

Data sources

The NHS Centre for Reviews and Dissemination Database of Abstracts of Reviews of Effectiveness (DARE) was used to identify reviews. This is a database of quality-assessed systematic reviews identified by handsearching key major medical journals, regular searching of electronic bibliographic databases and scanning 'grey' literature since 1994 (further details about DARE can be found at http://agatha.york.ac.uk/darehp.htm). All reviews entered in DARE up to December 1999 were screened for inclusion. Further searches of primary databases with the aim of identifying additional systematic reviews not indexed in DARE were not carried out; however, any such reviews identified from the other searches for the project (see Chapters 3 and 4) were assessed for inclusion.

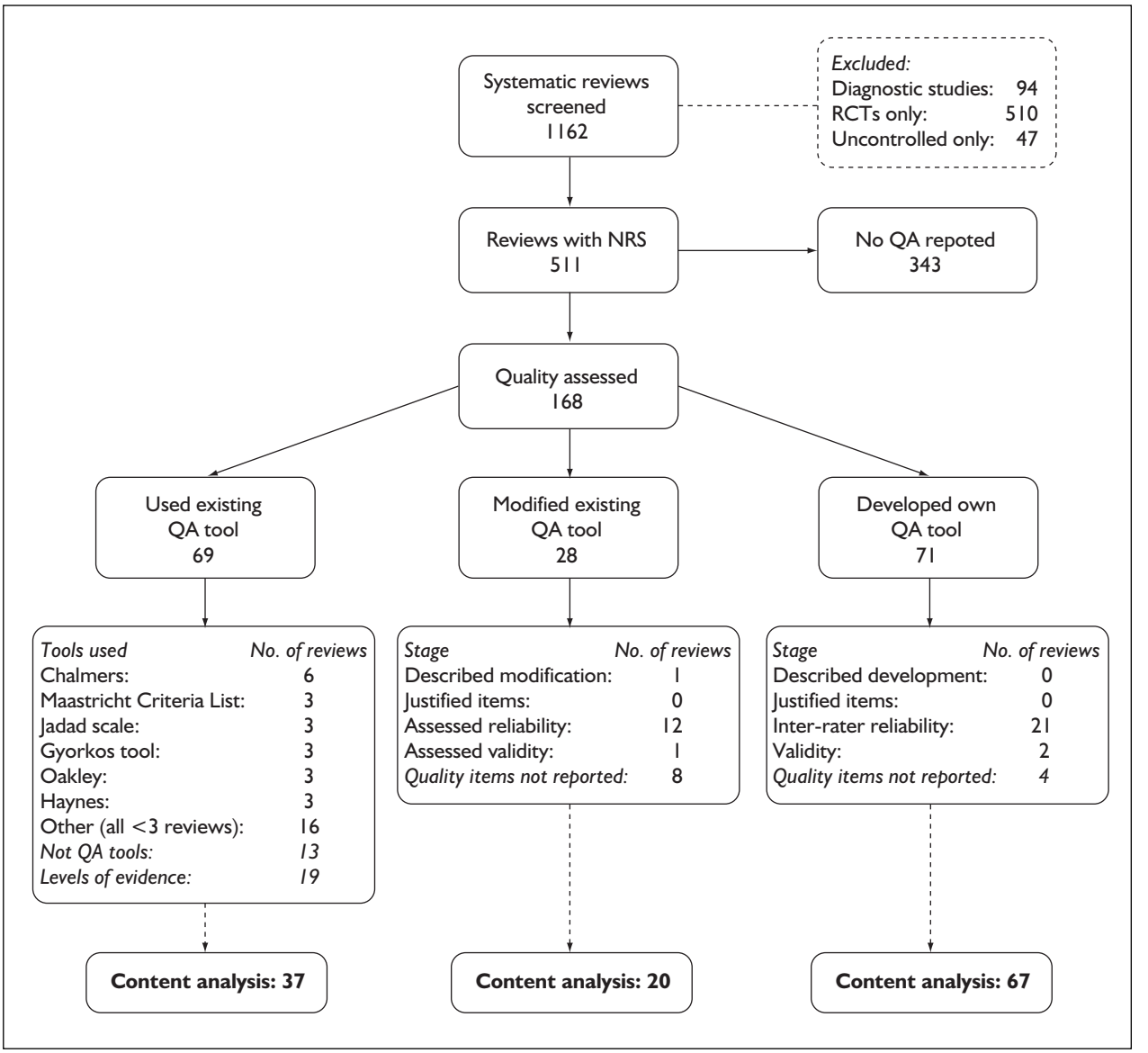

FIGURE 6 *Flowchart of systematic reviews. QA, quality assessment; NRS, non-randomised study.*

Analysis

A data extraction form for recording relevant information from each systematic review was designed and piloted. Information relating to the quality assessment tool used and the content of the tool were extracted in the same way as described in Chapter 4. Again, of the six internal validity domains we considered the creation of groups and comparability of groups at baseline to be central to the evaluation of non-randomised studies and within these we regarded four criteria as core to assessing selection bias.

We extracted information on how the results of the quality assessment were reported and subsequently used in the study syntheses.

The full systematic reviews were pre-screened independently by two reviewers. Those meeting the inclusion criterion were data extracted by one reviewer and the completed data extraction forms checked against the full paper by a second reviewer. Any disagreements were resolved by consensus or by referral to a third reviewer if necessary.

Results

Of the 1169 systematic reviews identified, 511 appeared to include at least one non-randomised study, of which 168 (33%) claimed to have carried out quality assessment of the included studies (see *Figure 6*). A total of 332 reviews (65%) included at

TABLE 11 *Breakdown of reviews including non-randomised studies*

Study designs included	No. of reviews	No. using quality assessment	Proportion using quality assessment (%)
NRS only	32	5	16
NRS + RCTs	216	99	46
NRS + RCTs + uncontrolled	116	34	29
NRS + uncontrolled	41	9	22
Not clearly reported	106	21	20
Total	511	168	33

NRS, non-randomised study.

least one RCT and in 87 reviews a term such as 'controlled trial' was used to describe the studies so that RCTs were not clearly distinguished from non-randomised studies (*Table 11*). Those reviews which included both RCTs and non-randomised studies were more likely to have conducted quality assessment (conducted in 46% of cases) than those with non-randomised studies only (16%) or those also including uncontrolled studies. Full details of the review methods are provided in Appendix 6 and review results in Appendix 7.

Source of assessment tools

Of the 168 reviews claiming to have carried out quality assessment, 69 (41%) stated they had used existing quality assessment tools, 28 (17%) made modifications to existing tools and 71 (42%) developed their own tool (*Figure 6*).

Closer inspection of the papers cited in the 69 reviews that used existing tools revealed that 10 of the tools (cited in 13 of the reviews) were not in fact quality assessment tools, but were methodological papers, textbooks or manuals on how to conduct reviews[118–123] or design primary studies.[50,124,125] Nineteen reviews used a variety of versions of the 'levels of evidence' framework originally developed by Sackett[126] and did not assess quality beyond identifying the study design.

Amongst the remaining 37 reviews, the most frequently used tools were originally developed for the assessment of RCTs (*Figure 6*): the Chalmers scale,[51] or modifications of it[76,77,92,96] (six reviews); the Maastricht Criteria List[69,70] (three reviews); and the Jadad scale[46] (three reviews). Other commonly used tools were ones developed for use in public health[127] (used in three reviews) and a tool developed by Oakley and colleagues[128] developed for and used by the same authors for reviews of the health promotion literature (three reviews).

Of the 28 reviews that modified existing tools, only one[26] reported details about how the tool (originally published by Downs and Black[85]) had been modified and also attempted to examine the validity of the tool. Twelve of these reviews reported inter-rater reliability, all producing kappa coefficients of >0.70 or percentage agreement of at least 80%.

In the review by MacLehose and colleagues,[26] the items considered to be the most important and unambiguous 'quality items' were identified by three authors and were categorised into one of four dimensions: quality of reporting, internal validity, susceptibility to confounding and external validity. The internal consistency of the individual items contributing to the first three of these dimensions was described using Cronbach's α. The resulting low α values indicated that individual items were poorly correlated with the sum of scores for the items, suggesting that the items were not assessing a homogeneous aspect of quality. An attempt was also made to validate the four quality dimensions and the sum of the four dimensions by investigating predictors of quality scores. Study design, intervention and the interaction of study design and intervention were entered into the regression model. For total quality scores, both cohort and case–control studies had significantly poorer quality than RCTs. There was no independent effect of intervention and no interaction of intervention and study design. Case–control studies scored more highly than cohort studies.

For the 71 tools developed by review authors, no details of the tool development were provided and none justified the choice of quality items. Twenty-one reviews examined inter-rater reliability and found similar levels of agreement to those discussed previously. In addition, two further reviews[93,129] addressed the validity of their tools.

TABLE 12 *Quality assessment in systematic reviews: coverage of internal validity*

	No. with ⩾1 pre-specified items	Proportion of reviews specifying items (%) (*n* = 124)	Proportion of reviews claiming QA (%) (*n* = 168)	Proportion of reviews including NRS (%) (*n*= 511)
Coverage of individual domains				
Creation of groups	103	83	61	20
Blinding	78	63	46	15
Soundness of information	35	28	21	7
Follow-up	89	72	53	17
Comparability	68	55	41	13
Outcome	46	37	27	9
Summary of domains covered				
5 or 6 domains	34	27	20	7
All 6 domains	11	9	7	2
Summary of core items covered				
Number of core items met:				
1	62	50	37	12
2	19	15	11	4
3	5	4	3	1
4	0	0	0	0
Number meeting each core item:				
5.3. How allocation occurred	24	19	14	5
5.4. Any attempt to balance groups by design	17	14	10	3
9.2. Identification of prognostic factors	12	10	7	2
9.3. Case-mix adjustment	33	27	20	7
Reviews using 'best' tools:				
⩾5 domains and ⩾3 core items	3	2	2	1

QA, quality assessment; NRS, non-randomised study.

One[93] looked at criterion validity but did not present the results of the comparison. The other[129] used the judges panel technique to examine face and content validity. The results of this were not reported in detail, only that the preliminary list of over 200 constructs was reduced to 71 critical factors.

Twelve reviews in the sample (eight of the modified tools and four developed by review authors) did not report the quality items assessed by the authors.

Content of assessment tools

The quality criteria included in the assessment tools used could be identified for 124 reviews and these were examined in more detail (*Table 12*). The majority of these reviews (83%) included at least one item relating to the 'creation of groups' domain; 72% considered 'follow-up' and looked at 63% 'blinding'; and 55% included items relating to the 'comparability of groups' domain. Less than half of the 124 reviews considered 'analysis of outcome' or 'soundness of information'. Only 34 reviews contained items in at least five out of six internal validity domains and 11 contained items in all six domains.

Fewer than 21% of all reviews that included non-randomised studies addressed any one of the six internal validity domains (see final column of *Table 12*), 5% assessed five internal validity domains and 2% assessed six domains. Looking more closely at the four quality items that particularly distinguish non-randomised studies from RCTs, only 62 reviews (12%) assessed at least one of the four core items, 19 of which (4%) assessed two of the items and only five (1%) looked at three of the four. No reviews assessed all four items and 449 (88%) reviews addressed none of the four. Of the four items, reviews were most likely to consider whether the study had conducted any case-mix adjustment (33 reviews, 7% of the sample), how the allocation had occurred (24 reviews, 4.7%), whether any matching had been used to balance groups (17 reviews, 3%) and whether all important prognostic factors had been identified (12 reviews, 2%). Only three reviews[109,112,130] used one of what we judged

TABLE 13 *Reporting of quality assessment results*

	No. of reviews	No. of studies per review			Reviews claiming QA (%) (*n* = 168)
		Median	Mean	Range	
QA results per item	50	14	25	5–211	30
Overall quality score per study	25	18	30	5–107	15
No. of studies meeting each criterion	16	67	66	13–177	10
Narrative discussion	21	21	34	4–224	13
Average quality score across studies	12	30	41	9–106	7
Levels of evidence	9	24	23	8–39	5
Other	8	45	59	10–194	5
Not reported	27	46	76	9–474	16

QA, quality assessment.

TABLE 14 *Method of incorporating quality assessment*

Synthesis method	Narrative reviews *n* (%)	Meta-analyses *n* (%)	All reviews using QA *n* (%)
Narrative	76 (84)	11 (14)	87 (52)
Quantitative	4 (4)	57 (74)	61 (36)
Not reported	11 (12)	9 (12)	20 (12)
Total	91 (100)	77 (100)	168 (100)

QA, quality assessment.

to be the 'best' tools in our sample,[85,109] that is including items in at least five out of six internal validity domains and three out of four core items.

Reporting of study quality in systematic reviews

The various methods used to report the results of the quality assessment are given in *Table 13*. These varied from a detailed listing of the quality assessment results by item (30%) to reporting only the average quality score across all of the included studies (7%). In 27 cases (16%), the quality assessment results were not reported at all.

Use of study quality in systematic reviews

The means by which the results of the quality assessment were considered in the study syntheses are reported in *Table 14*, split by whether the review used only a narrative synthesis or included a meta-analysis. Overall, over half of the reviews (87/168) discussed the results of the quality assessment narratively. Generally, the authors provided a broad statement of the overall quality of included studies (often inferring that the studies were of too poor quality to allow firm conclusions to be drawn), although some attempted a narrative comparison of the results of better studies compared with weaker studies. A further 12% of all reviews did not incorporate the results of the quality assessment into the review synthesis.

Around three-quarters of the meta-analyses in the sample (57/77) attempted to incorporate the results of the quality assessment in a quantitative way, although the implications of study quality were not always discussed (Appendix 7). Half (28/57) investigated the impact of individual quality components, 13 looked for any association between quality score and effect size and the remainder used a variety of techniques, including quality weighting or quality thresholds.

Discussion

We found that 67% of systematic reviews that included non-randomised studies either did not conduct, or failed to report, any form of quality assessment. The remainder used a variety of quality assessment tools, including several that were explicitly developed for the assessment of RCTs. In a substantial proportion of cases (14%), review authors developed the tools themselves but gave very limited details regarding the method of development.

Our analysis of the content of the quality assessment tools used revealed that the majority of

the 168 reviews that claimed to have undertaken quality assessment did not comprehensively assess the internal validity of the included studies. Only 34 reviews (20%) included at least one quality item in five out of six internal validity domains and only 62 (37%) assessed at least one of the four key areas that distinguish non-randomised from randomised studies. Only three reviews (2%) used quality assessment tools that we judged to pay sufficient attention to key issues of selection bias for non-randomised studies.

Most of the reviews that assessed study quality reported the results in some form, although in less than one-third of cases were the results reported per study. This seems to be partially related to the number of studies per review – those with fewer studies were more able to present detailed results. A minority of reviews that assessed quality (12%) did not go on to consider the quality assessment results in the study syntheses; however, most provided a narrative discussion of study quality and its implications.

Those reviews that attempted to incorporate study quality into a quantitative synthesis did so in a variety of ways. Most included a wide variety of study designs, but the numbers of primary studies included were not large enough to allow the degree of bias introduced by variations in quality to be clearly identified; the impact of quality was confounded by differences in study design.

Our review has revealed that the conduct of systematic reviews that include non-randomised studies with respect to quality assessment is as poor as, if not worse than, that found by Moher and colleagues in 1999 for meta-analyses of RCTs.[131] Moher and colleagues also found that trial quality was not assessed in most meta-analyses (48% compared with 67% of reviews in our study) and that when it is assessed, the assessments are obtained with non-validated tools.

The infrequency of adequate quality assessment may in part occur owing to the lack of empirical evidence and controversy concerning the biases thought to act in non-randomised studies, and confusion both about what quality items to assess and what quality assessment tool to use. This is perhaps supported by our finding that those reviews that included both RCTs and non-randomised studies were much more likely to have conducted quality assessment than those that did not include RCTs or also included uncontrolled studies.

However, the absence of agreed criteria for assessing quality has not stopped reviews of non-randomised studies of healthcare interventions being carried out and their results being used to inform treatment and policy decisions. Given the clear evidence that inadequate randomisation procedures severely bias the results of RCTs, it seems reasonable to predict that non-random methods of allocation are equally if not more open to selection bias than concealed allocation. Where randomised evidence is unavailable, the potential for bias and resulting uncertainty inherent in estimates based on non-randomised evidence should be strongly emphasised and evaluated through quality assessment.

A particular strength of our review was the availability of a large number of systematic reviews for assessment provided via the DARE database. This database is fed by monthly, and in some cases weekly, extensive literature searches of a wide range of databases [such as *Current Contents*, MEDLINE and Cumulative Index of Nursing and Allied Health Literature (CINAHL)] published since 1994. Systematic reviews have to meet a certain standard of methodological quality before being included on the database. This in turn means that the reviews in our sample are of higher quality than many that are published, so that the situation in practice may be even worse than we have demonstrated here. In the past, the process of assessing reviews for inclusion in DARE and the subsequent writing of structured abstracts for each review has also meant that there was some time lag before reviews were loaded on to the database. The majority of reviews in our sample were published prior to 1999 and it is possible that reviewers are now more likely to conduct and report quality assessment of non-randomised studies than they have been in the past.

Nevertheless, reviewers who include non-randomised studies in their systematic reviews should be aware of the fundamental need to address the potential for selection bias in these studies (and also all of the other quality issues that affect all study designs), and should consider the impact of these biases in the synthesis of studies and, in turn, in their conclusions. In turn, users of systematic reviews should be careful not to over-interpret the results of reviews that included non-randomised studies. If biases have not been assessed then conclusions may be invalid or unjustified.

Chapter 6

Empirical estimates of bias associated with non-random allocation

Introduction

In Chapters 4 and 5, many aspects of study quality that may be related to bias in non-randomised studies were identified. Two quality items were emphasised that are specific to non-randomised studies rather than RCTs: bias introduced by non-random allocation and the use of case-mix adjustment methods. In this chapter, we report an empirical investigation into the impact of non-random allocation. In Chapter 7 we report a similar investigation into the benefits of case-mix adjustment.

Eight reviews were described in Chapter 3 that attempted to evaluate the importance of randomisation. Our appraisal, however, identified problems with their methodology, especially with regard to control for confounders (meta-confounding) and their inability to identify biases other than those acting in a systematic manner. For our evaluation, we have chosen a different methodology – resampling – to overcome these issues. By selectively sampling treated and control participants from a completed large randomised trial it is possible to make comparisons between groups of individuals who did and did not receive the intervention, but where the treatment allocation was based not on randomisation but on a non-random mechanism. Effectively we are answering the question, 'what results would the trialists have obtained if they had not compared their treated participants with random controls but with controls selected using a non-random process, all other aspects of study design being kept the same?'.

Two classical non-randomised designs can be constructed from the data of a multicentre trial: first, a concurrently controlled cohort study in which treated patients from one centre are compared with an untreated control group from a different centre recruited during the same time period; and second, a historically controlled study where the untreated controls are selected from the same centre but from a period before the treated patients were recruited. The results from such studies can be compared directly with the results of the RCT in a single centre. Repeating the comparisons using replacement sampling methodology with data from many centres can generate distributions of results for randomised and non-randomised studies. These distributions can be compared to describe the impact of both systematic and unpredictable components of bias associated with the allocation mechanism. The comparisons between study designs are based on the same patient data, and hence across repeated samples will be unconfounded.

We report the results of such resampling exercises for two large multicentre trials: the International Stroke Trial (IST)[132] and the European Carotid Surgery Trial (ECST).[133] The IST compared medical treatments in stroke patients to prevent long-term death or disability. The ECST was a trial of a surgical treatment in patients at risk of stroke to prevent stroke or death. Full descriptions are given in Appendix 8.

Note that because of the adaptations made to the trials for the purposes of the empirical investigation (such as grouping of centres, omission of data from centres and periods, and selection of particular time-points for follow-up), the results reported here differ from those of the original publications.[132,133] Notably, the full report of the ECST trial reports treatment benefit in participants with high-grade stenosis, whereas the simplified data set that we consider presents a finding of an average harmful effect. The original publications should be referred to for the appropriate analyses of these trials.[132,133]

Methods

Generation of non-randomised and randomised studies

The resampling was undertaken in a similar way for both the IST and ECST. First, participants were grouped into **regions** as described in Appendix 8. Secondly, participants were classified according to whether they were **early** or **late** recruits to the trial, depending on whether they were recruited before or after a time-point midway

through the recruitment of the trial. *Figure 7* illustrates this structure for the region-based analysis of the IST trial. Participants are classified according to the treatment they received (*T* or *C*), whether they were recruited in the first (subscript *B* for **before**) or second (subscript *A* for **after**) half of the trial, and according to the region number (subscripts *1–14*).

Historically controlled studies were constructed within each region by sampling participants who were recruited to the control group during the first half of the trial and participants who were recruited to the treatment group during the second half of the trial. For example, in North Italy, control participants were selected from cell C_{B1} and treated participants from cell T_{A1}.

Concurrently controlled studies were constructed by selecting a region from which treated participants were drawn and choosing control participants from another region chosen at random. Participants were sampled regardless of the recruitment period. For example, a concurrently controlled comparison could be generated between treated patients in Scotland (sampling from cells T_{B4} and T_{A4}) and control patients in Spain (sampling from C_{B12} and C_{A12}). It was expected that no average bias would be observed in large samples of concurrently controlled studies generated in this manner as there would be corresponding studies comparing treated patients from Spain with control patients from Scotland.

The IST recruited sufficient participants for samples of 100 treated and 100 controls to be sampled from each region for each design. For the ECST only samples of 40 treated and 40 controls were drawn as sample sizes in regions were lower. In all instances random samples were drawn using replacement sampling techniques.[134] This partial bootstrap process ensures that the variability between samples is as close as possible to that which would have been obtained if sampling from an infinite (or very large) population.

For the purpose of comparing distributions of results, randomised trials of the same size as the non-randomised studies were constructed using the same resampling methodology. RCTs for comparison with historically controlled studies were constructed by comparing treated patients sampled from the second half of the trial (the same treated sample as selected for the HCT) with control patients from the same region also sampled from the second half of the trial. For example, a North Italy RCT for comparison with a North Italy HCT was constructed by sampling treated participants from T_{A1} and control participants from C_{A1}. RCTs for comparison with concurrently controlled studies were constructed by comparing treated and control patients sampled from the same region regardless of the recruitment period. For example, a North Italy RCT for comparison with concurrently controlled studies was constructed by sampling treated participants from T_{B1} and T_{A1} and control participants from C_{B1} and C_{A1} (*Figure 7*).

Historically controlled studies were constructed in each of the 14 regions of the IST analysis and the eight regions of the ECST analysis. The whole resampling process was repeated 1000 times in both studies, generating results of 14,000 IST and 8000 ECST historically controlled studies, and the same number of comparable RCTs.

Concurrently controlled studies were constructed for each of the 14 regions of the IST analysis and the eight regions of the ECST analysis. In addition, in an attempt to limit some sources of variability, concurrently controlled studies were constructed using IST data from 10 cities in the UK (see Appendix 8). All resampling analyses were repeated 1000 times, generating 14,000 IST and 8000 ECST concurrently controlled region-based comparisons and 10,000 IST concurrently controlled UK city-based comparisons, all with the same number of comparable randomised trials.

Statistical analyses

The focus of the statistical analysis was on describing and comparing the location (average) and spread (variability) of the treatment effects observed in studies of different designs. Treatment effects were calculated using standard methods commonly reported in medical journals, and interpreted according to their statistical significance assessed at the 5% level using a two-sided test.

Observed outcomes of the resampled trial participants for each non-randomised and randomised study were tabulated, and the treatment effect expressed as an odds ratio (OR) with 95% confidence interval. For both IST and ECST, OR < 1 indicate that the experimental treatment (IST, aspirin; ECST, carotid endarterectomy) had better outcomes than controls. The effect was deemed to be statistically significant if the confidence interval excluded the null hypothesis of no treatment effect (OR = 1).

Region	Trial allocation	Recruitment period[a] 1st half	2nd half
Northern Italy	Aspirin	T_{B1}	T_{A1}
	Avoid aspirin	C_{B1}	C_{A1}
Central Italy	Aspirin	T_{B2}	T_{A2}
	Avoid aspirin	C_{B2}	C_{A2}
Southern Italy	Aspirin	T_{B3}	T_{A3}
	Avoid aspirin	C_{B3}	C_{A3}
Scotland	Aspirin	T_{B4}	T_{A4}
	Avoid aspirin	C_{B4}	C_{A4}
Northern England and Wales	Aspirin	T_{B5}	T_{A5}
	Avoid aspirin	C_{B5}	C_{A5}
Southern England	Aspirin	T_{B6}	T_{A6}
	Avoid aspirin	C_{B6}	C_{A6}
Australia	Aspirin	T_{B7}	T_{A7}
	Avoid aspirin	C_{B7}	C_{A7}
The Netherlands	Aspirin	T_{B8}	T_{A8}
	Avoid aspirin	C_{B8}	C_{A8}
New Zealand	Aspirin	T_{B9}	T_{A9}
	Avoid aspirin	C_{B9}	C_{A9}
Norway	Aspirin	T_{B10}	T_{A10}
	Avoid aspirin	C_{B10}	C_{A10}
Poland	Aspirin	T_{B11}	T_{A11}
	Avoid aspirin	C_{B11}	C_{A11}
Spain	Aspirin	T_{B12}	T_{A12}
	Avoid aspirin	C_{B12}	C_{A12}
Sweden	Aspirin	T_{B13}	T_{A13}
	Avoid aspirin	C_{B13}	C_{A13}
Switzerland	Aspirin	T_{B14}	T_{A14}
	Avoid aspirin	C_{B14}	C_{A14}

[a]The IST recruitment period was split at 15 January 1995

FIGURE 7 *Resampling structure for IST used to generate non-randomised and randomised studies*

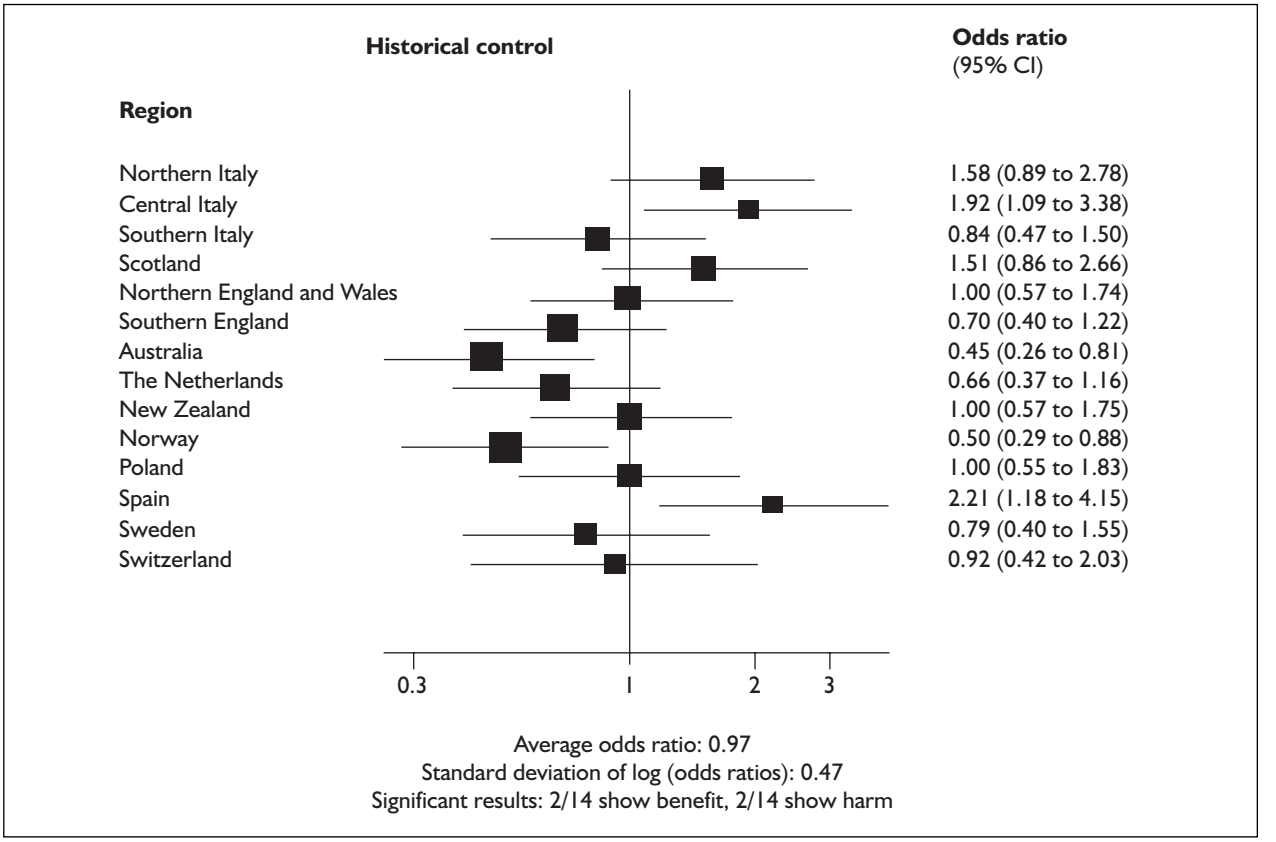

FIGURE 8 *Illustrative results from one resampling iteration of historically controlled studies resampled from the IST. CI, confidence interval*

Distributions of results for each design were first plotted using dotplots. Location was described by calculating the 'average' OR as the exponential of the mean of the log OR. The ratio of estimates of average ORs between the RCTs and non-randomised studies provided an estimate of systematic bias. *Figure 8* shows the results of the first 14 historically controlled studies generated from the IST data set, with an average log OR of –0.03 and standard deviation (SD) of 0.47. The estimates of SD were averaged across the 1000 sample repeats to summarise the spread of results for each study design. As the sample size and average event rate are the same for randomised and non-randomised studies, differences in spread can be wholly attributed to unpredictablity in the size and direction of the bias introduced through non-random allocation. The difference in spread was expressed as the ratio of the SD of log OR for non-randomised studies compared with that for RCTs. This ratio estimates how many times more variable results from non-randomised studies are than results of RCTs.

To investigate the likely conclusions of the studies, the percentage reaching the conventional 5% level of statistical significance was also tabulated. Results showing significant benefit were reported separately from results showing significant harm.

Results

Concurrently controlled studies compared with randomised controlled trials

The overall results of the resampled concurrent controlled studies and RCTs are reported in *Table 15*. As expected, the concurrent controlled studies from the IST showed no average systematic bias (the average OR, of concurrently controlled studies and RCTs were very similar), but increased variation (the variability of the results of concurrently controlled studies were higher than those of RCTs). Owing to this increased variability, more concurrently controlled studies were statistically significant than were RCTs. Concurrent controlled studies from the ECST showed only a small increase in variability and no systematic bias.

Results from the IST region-based analysis

The results of the 14,000 concurrently controlled studies generated from the IST data set are plotted in *Figure 9*, together with results of 14,000

TABLE 15 *Comparisons of results of RCTs with results of concurrently controlled non-randomised studies based on analyses of 1000 sets of resampled studies*

			Percentage of studies with statistically significant results showing beneficial or harmful effects of the intervention ($p < 0.05$)		
	Average OR	**SD of log OR**	**Beneficial**	**Harmful**	**Total**
IST – 14 regions					
RCTs	0.91	0.34	7	2	9
Concurrent controls	0.91	0.85	29	21	50
Ratio of SDs		*2.5-fold*			
IST – 10 UK cities					
RCTs	1.01	0.49	7	6	13
Concurrent controls	1.02	0.90	22	23	45
Ratio of SDs		*1.8-fold*			
ECST – 8 regions					
RCTs	1.08	0.68	3	5	9
Concurrent controls	1.08	0.69	3	6	9
Ratio of SDs		*1.01-fold*			

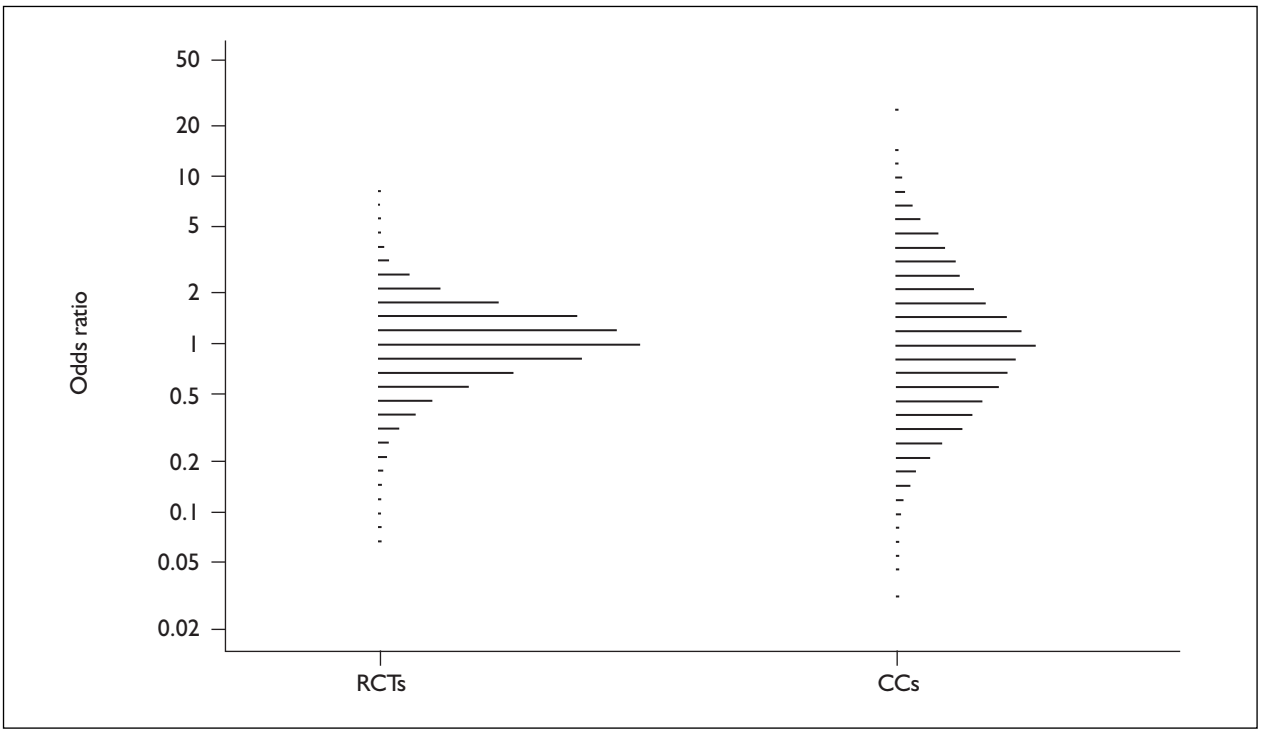

FIGURE 9 *Comparison of distributions of results of 14,000 concurrently controlled studies (CCs) and 14,000 RCTs resampled from 14 regions within the IST*

RCTs generated from the same data. Although there was no difference in average results, the increase in variability of results in concurrently controlled trials is clear. The SD of the distribution was 2.5 times that of the RCTs (*Table 15*). Some 50% of the concurrently controlled studies gave statistically significant results compared with 9% of the RCTs, with 29% showing significant benefit of aspirin and 21% showing significant harm.

Results from the IST UK city-based analysis

The results of the 10,000 concurrently controlled studies generated from the UK IST data are plotted in *Figure 10*, together with results from the

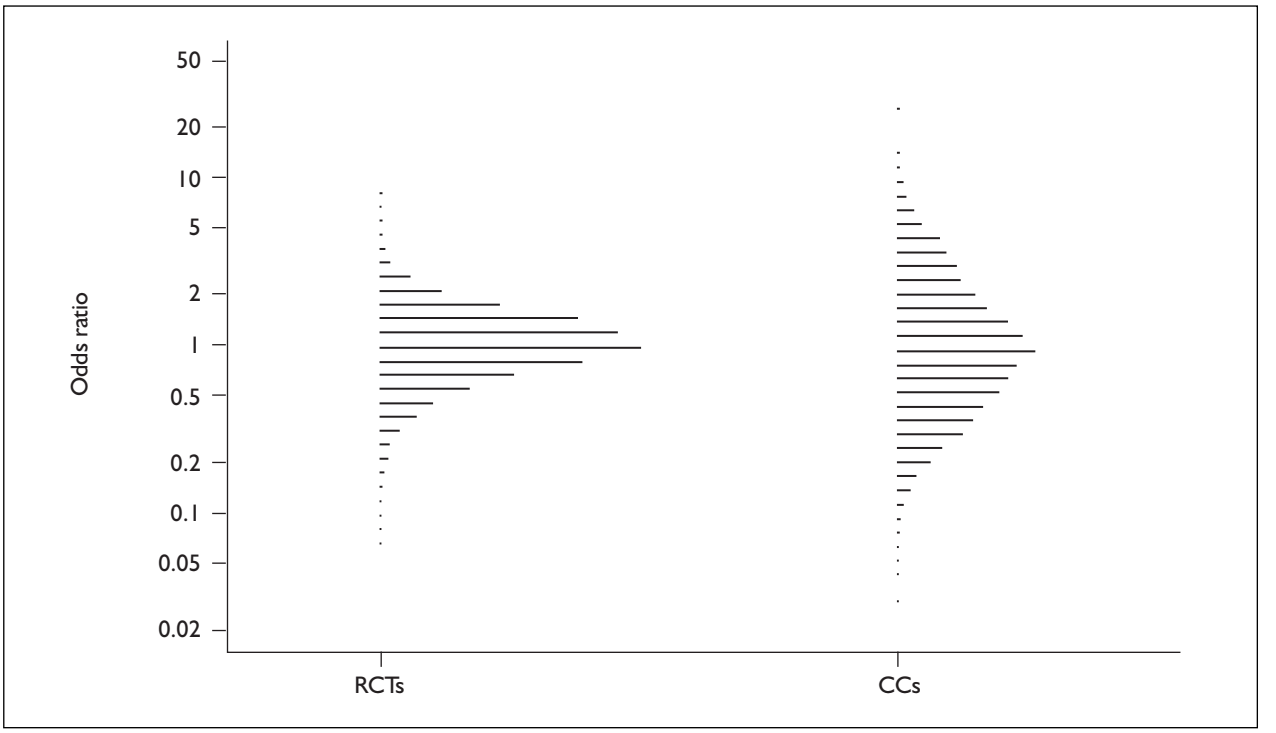

FIGURE 10 *Comparison of distribution of results of 10,000 concurrently controlled studies and 10,000 RCTs resampled from 10 UK cities within the IST*

comparable RCTs. Again, the increase in variability of results for concurrently controlled studies was evident, the distribution of results being 1.8 times wider than for RCTs, slightly less than for the regional IST comparisons (*Table 15*). Some 45% of concurrently controlled results were statistically significant compared with 13% of RCTs, the significant results being evenly distributed between benefit (22%) and harm (23%).

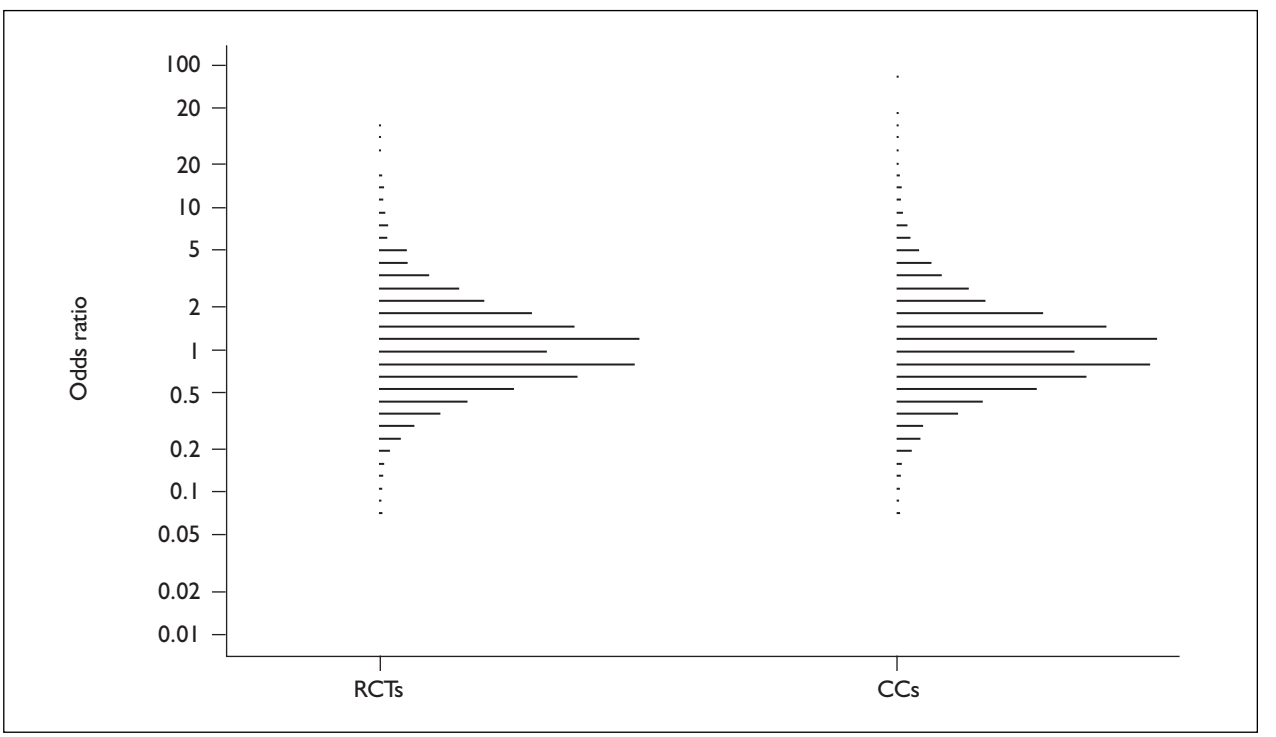

FIGURE 11 *Comparison of distribution of results of 8000 concurrently controlled studies and 8000 RCTs resampled from eight regions within the ECST*

Results from the ECST region-based analysis

Figure 11 displays results of 8000 concurrently controlled comparisons and 8000 RCTs generated from the ECST data set. The pattern of results of concurrent controlled studies are similar to those of RCTs. The concurrently controlled comparisons were not more variable than the RCTs; their distribution was only 1.01 times as wide as the RCTs (*Table 15*); 9% of both concurrently controlled studies and RCTs were statistically significant.

Historically controlled studies compared with randomised controlled trials

Table 16 shows results of comparisons of historically controlled studies and RCTs sampled from the IST and ECST. The data from the historically controlled studies generated from the two trials display different patterns.

Results from the IST region-based analysis

In the overall results of the historically controlled studies from the IST there was no evidence of systematic bias. However, there was an increase in unpredictability in study results; the distribution of historically controlled results was 20% wider than that for the RCTs (*Table 16*). This increase is discernible in *Figure 12*, where the results of the 14,000 historically controlled studies and 14,000 comparable RCTs are displayed. The increased unpredictability increased the percentage of studies deemed statistically significant from 11 to 20%, with increases in the number of findings of both statistically significant benefit and harm.

Table 17 presents the same results broken down according to region. The results disaggregated at this level appeared rather different. For each region there was little evidence of an increase in unpredictability (the SDs of the historically controlled studies were all within 7% of the SD of the RCTs). However, there was evidence of systematic bias within many of the regions, although the magnitude and direction of the bias vary. For example, in Scotland the average OR for historically controlled studies was 1.23 compared with an average OR of RCTs of 0.78. In contrast to Scotland, in Sweden the average OR from historically controlled studies was 0.44 compared with 0.94 from the RCTs, 81% of historically controlled studies concluding statistically significant benefit. When aggregated across the regions, these varying systematic biases on average cancel out and manifest as an increase in unpredictability.

Results from the ECST region-based analysis

The overall results of the historical controlled studies sampled from the ECST analyses (*Table 16*) showed a large systematic bias, but with little increase in unpredictability. Whereas the average OR in the 8000 RCTs was 1.23, the average OR of the 8000 historically controlled studies was 1.06. This difference is noticeable in the comparison of the distribution of the results of the studies in *Figure 13*. The results of the two study designs also show similar levels of variability. The systematic bias impacted on statistical significance by shifting the distribution of results in one direction: 6% of the historically controlled studies concluding significant benefit from carotid surgery compared with 4% of the RCTs, whereas 10% of the historically controlled studies concluded significant harm compared with 12% of the RCTs.

TABLE 16 *Comparisons of results of RCTs with results of historically controlled non-randomised studies based on analyses of 1000 sets of resampled studies*

			Percentage of studies with statistically significant results showing beneficial or harmful effects of the intervention ($p < 0.05$)		
	Average OR	**SD of log OR**	**Beneficial**	**Harmful**	**Total**
IST – 14 regions					
RCTs	0.89	0.35	9	2	11
Historical controls	0.88	0.44	16	4	20
Ratio of SDs		*1.2-fold*			
ECST – 8 regions[a]					
RCTs	1.23	0.83	4	12	16
Historical controls	1.06	0.85	6	10	16
Ratio of SDs		*1.03-fold*			

[a] All patients entering the trial after 1990 were excluded when the protocol's inclusion criteria were changed.

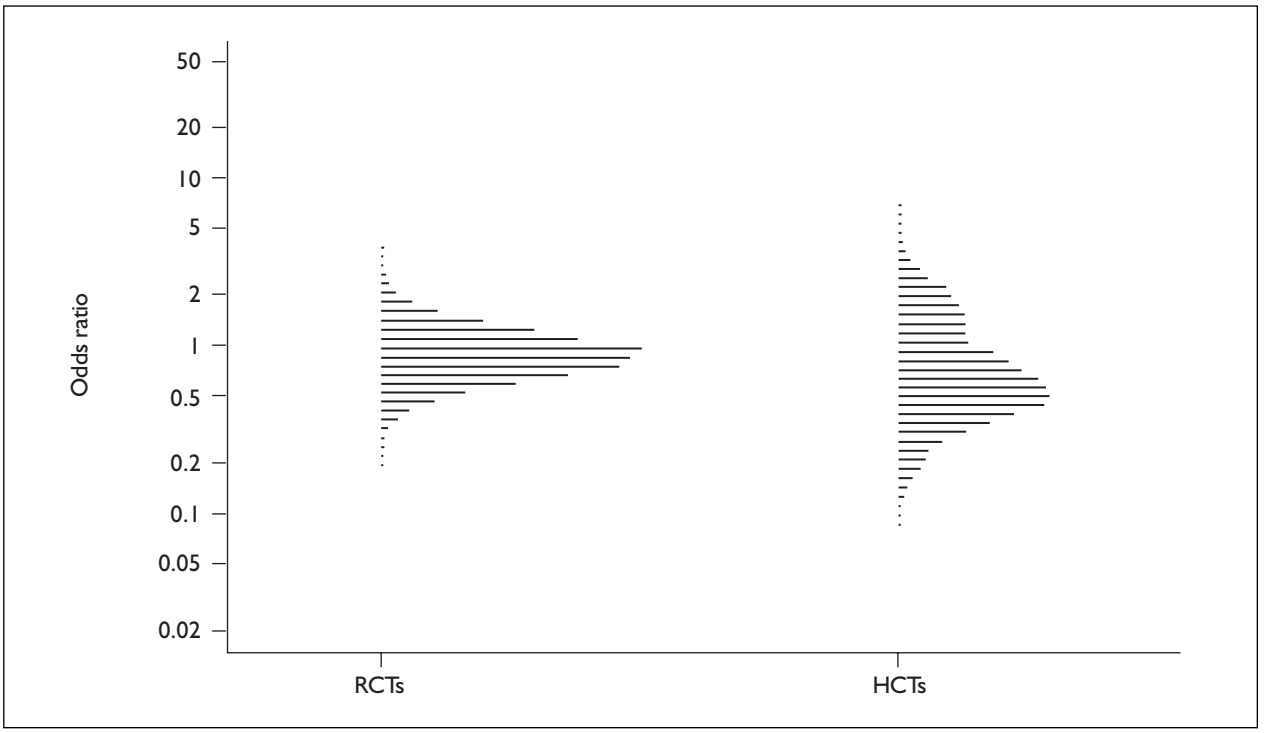

FIGURE 12 *Comparison of distribution of results of 14,000 historically controlled studies and 14,000 RCTs resampled from 14 regions within the IST*

TABLE 17 *Comparisons of results of RCTs with results of historically controlled non-randomised studies based on 1000 sets of resampled studies analysed by region (IST)*

	Average OR			SD log OR			Percentage of studies with statistically significant results ($p < 0.05$): benefit/harm	
Region	**RCT**	**HCT**	**Ratio**	**RCT**	**HCT**	**Ratio**	**RCT**	**HCT**
South Italy	0.56	1.00	1.79	0.29	0.30	1.03	49/0	3/3
Scotland	0.78	1.23	1.58	0.35	0.33	0.93	12/0	0/9
North Italy	0.94	1.15	1.22	0.29	0.29	0.99	4/2	1/8
New Zealand	0.81	0.97	1.20	0.29	0.30	1.03	12/0	4/2
Switzerland	0.97	1.07	1.10	0.30	0.30	1.00	3/2	1/4
Australia	1.30	1.39	1.07	0.28	0.30	1.07	0/13	0/23
Mid-Italy	0.89	0.92	1.03	0.31	0.30	0.98	6/1	5/1
The Netherlands	0.88	0.85	0.97	0.28	0.29	1.03	6/1	8/1
Northern England	0.95	0.92	0.97	0.36	0.35	0.98	4/2	4/2
Spain	1.00	0.93	0.93	0.27	0.28	1.05	2/2	5/1
Southern England	1.01	0.87	0.86	0.37	0.36	0.99	2/2	6/1
Poland	0.91	0.70	0.77	0.28	0.28	1.01	6/1	27/0
Norway	0.78	0.48	0.62	0.28	0.28	0.99	14/1	73/0
Sweden	0.94	0.44	0.47	0.32	0.29	0.93	4/2	81/0
Total	0.89	0.88	0.99	0.35	0.44	1.23	9/2	16/4

Table 18 displays the same results according to the region. These results demonstrate a broadly similar pattern to the overall findings: there was no increase in variability of historically controlled studies compared with RCTs and a systematic bias in six of the eight regions in favour of carotid surgery. There was variability in the average magnitude of the bias by region (Region 1 showing the largest overestimate of harm and Region 8 showing the largest overestimate of benefit). There was notable heterogeneity in the results of the RCTs between

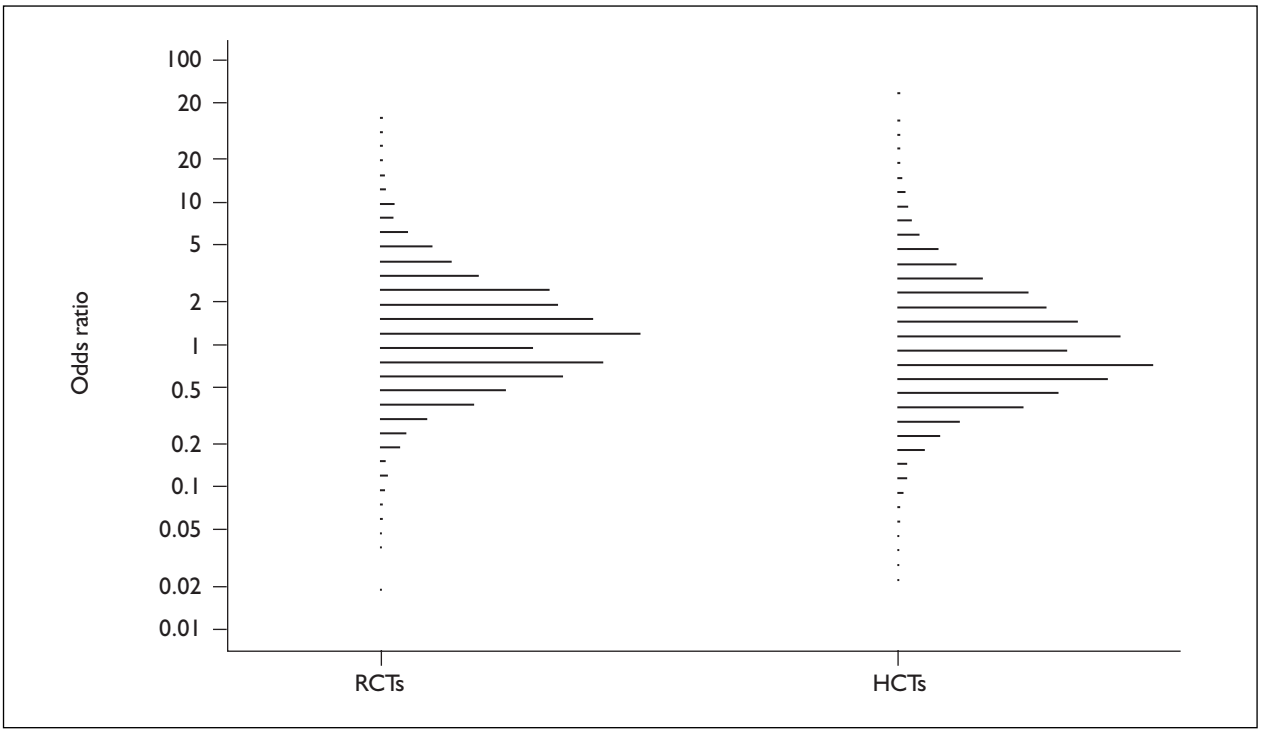

FIGURE 13 *Comparison of distribution of 8000 historically controlled studies and 8000 RCTs resampled from eight regions within the ECST*

TABLE 18 *Comparisons of results of RCTs with results of historically controlled non-randomised studies based 1000 sets of resampled studies analysed by region (ECST)*

	Average OR			SD log OR			Percentage of studies with statistically significant results ($p < 0.05$): benefit/harm	
Region	**RCT**	**HCT**	**Ratio**	**RCT**	**HCT**	**Ratio**	**RCT**	**HCT**
Region 1	1.53	3.24	2.12	0.52	0.60	1.15	0/14	0/54
Region 2	0.78	0.79	1.01	0.57	0.56	0.98	7/1	6/1
Region 3	0.71	0.64	0.90	0.58	0.55	0.95	7/0	11/0
Region 4	0.52	0.42	0.81	0.68	0.68	1.00	13/0	25/0
Region 5	2.68	2.04	0.76	0.72	0.67	0.93	0/24	0/16
Region 6	1.20	0.87	0.73	0.54	0.52	0.96	1/4	4/1
Region 7	1.28	0.88	0.69	0.59	0.53	0.90	0/6	4/2
Region 8	2.85	1.47	0.52	0.57	0.50	0.88	0/47	0/10
Total	1.23	1.06	0.86	0.83	0.85	1.02	4/12	6/10

these regions, but very similar heterogeneity was seen between the results of the historically controlled trials.

Discussion

Our evaluations have detected the bias in non-randomised studies using either concurrent or historical controls. Compared with the use of randomised controls, we have shown that use of non-randomised controls increases the probability of statistically significant study findings (and hence the likely conclusions), in addition to altering the estimates of treatment effects. The bias observed in the non-randomised studies acted in two ways, which we refer to as systematic and unpredictable components of bias.

Systematic bias

Systematic bias acts consistently in a given direction leading to differences in average overall

results between non-randomised and randomised studies. The bias was observed for the ECST historically controlled comparisons leading to overestimates of the benefit of carotid surgery, both for individual regions and when the results were aggregated across regions. This pattern is consistent with the conclusions of Sacks and colleagues,[27] who noted in their review of historically controlled studies for six medical interventions that "biases in patient selection may irretrievably weight the outcome of historically controlled studies in favour of new therapies".

Systematic biases were also noted in some of the historically controlled studies in the individual regions in the IST analysis, but here they were seen to vary in direction and magnitude, sometimes overestimating benefit and sometimes overestimating harm.

Systematic bias in historically controlled studies arises from there being time trends in the average outcomes of participants in a study, regardless of which treatment they receive. Details of the outcomes and characteristics of the participants in the ECST are presented in *Tables 41* and *42* in Appendix 8. For five regions there was a reduction in the adverse event rate of between 1 and 7% (averaged across both treatment and control) between the trial periods, whereas for three regions there was an increase of between 1 and 14%. The change was statistically significant ($p < 0.01$) in one region.

How do such trends arise? There are a limited number of options: they must arise through variation over time in the case-mix, and hence prognosis, of participants recruited to the trial (as proposed by Sacks and colleagues[27]), through differences in other healthcare interventions that the participants receive or through changing assessments of outcome. These variations may themselves be haphazard or due to systematic mechanisms (such as changes in patient referral and recruitment or in patient management). Some of these potential causes may be measured, such as baseline risk factors, but many may go unnoticed and are not assessed.

Tables 39 and *42* in Appendix 8 show summaries of the distribution of important baseline risk factors for IST and ECST, respectively. For both trials there were differences in the risk factors of participants between the first and second halves of the trial, although the patterns of these differences were not consistent between regions, and it is not immediately obvious how they relate to differences in outcome. It seems likely that the differences occur in part due to unmeasured changes within the trials, but that there may also be different mechanisms causing systematic bias in different regions.

Why should there be a time trend in outcome in the ECST? Patients were only entered into the trial when an investigator judged that in the case of the individual patient there was uncertainty as to whether surgery would be beneficial or harmful. One possibility is therefore that throughout the very long recruitment period (12.5 years) investigators joined or left the trial who had systematically different opinions on who was suitable for randomisation. Six of the eight regions showed significant reductions ($p < 0.05$) in the proportion of patients recruited with <50% stenosis between the two periods considered. The time trend may therefore relate to the enthusiasm of the investigators for the surgical intervention: those joining the trial early may have been more likely to recruit all patients whereas those joining the trial later may have been selecting only patients with higher degree stenosis.

A second explanation of the time trend relates to changes in the intervention. Increasing delays between the index event [usually a transient ischaemic attack (TIA)] and obtaining surgery may decrease the value of surgery. If referral and investigation processes became faster and more efficient as the trial progressed, the benefits of surgery would become greater.

Additionally, stenosis was found to be an important variable upon which treatment decisions have become based. Early results from ECST and a similar trial [North American Symptomatic Carotid Endarterectomy Trial (NASCET)][135] were published indicating a clear benefit of endarterectomy in patients with high-grade stenosis. This led to changes in the inclusion criteria of the ECST from 1990. All the historical comparisons reported here are based on participants recruited before this protocol change was made, so it is not in any way responsible for the observed historical trend.

For IST the time trends in outcome are all smaller and are probably due to haphazard patterns in case-mix associated with referral and recruitment to the trial.

Unpredictability in bias

When bias acts unpredictably, it will sometimes lead to an overestimation and sometimes to underestimation of an effect. Although these biases may on average 'cancel out' across a set of studies such that no difference is observed in average ORs, the biases will still affect the results of individual studies. The presence of systematic bias may therefore be missed if the comparison of results is restricted to a comparison of average values, as was done in five of the eight previous reviews summarised in Chapter 3.[25–28,32]

Unpredictable over- and underestimation will increase the variability (or heterogeneity) of the results of a set of studies. In the concurrent comparisons such an increase in variability (measured by the standard deviation) was observed for the IST (*Table 15*), even though the average treatment effects in the concurrently controlled and randomised studies were the same. A similar pattern was observed for historically controlled studies generated from the IST when the haphazard within-region time trends were aggregated in the overall analysis (*Table 16*).

How do these biases occur, and how do they differ from the variability seen between RCTs? Variability always occurs between the results of multiple RCTs. The principal reason is the 'play of chance' or sampling variation. A treatment effect observed in a particular RCT is unlikely to be the precise effect of the intervention. For example, randomly dividing the study sample into two does not guarantee that the groups are identical in all respects, and the differences that do exist in case-mix will lead to either under- or overestimates of the treatment effect in an individual trial. We do not normally talk about these differences as **biases**, but rather as **uncertainties**. We know the distribution with which under-and overestimates arise in RCTs, enabling us to draw correct inferences within specified degrees of certainty. We cannot identify whether a particular trial is affected by such bias, but we can calculate bounds within which we are reasonably sure possible bias is encompassed, which we term confidence intervals. Importantly, we know that the possible differences between the groups due to sampling variation (and hence confidence intervals) reduce with increasing sample size.

The extra variability we see in the non-randomised studies arises in a similar but more troubling manner. Rather than randomly dividing a single group of individuals, we start with two different groups of individuals. We therefore start with differences between the groups in measurable and unmeasurable factors. These potentially include differences in case-mix, additional treatments and methods of assessment of outcome. Importantly, in addition to not being able to identify all these differences, we may not know in which way many of the factors act, so that there is overall uncertainty as to whether they will cause under- or overestimates of the treatment effect. Sampling from these populations introduces the same sampling variation as in the RCT. While we can estimate the impact of the sampling variation (and calculate standard confidence intervals), there is no mathematical way of knowing how pre-existing differences between the groups behave. It is therefore not possible to include an allowance in the confidence interval for a single study that accounts for the extra uncertainty introduced through unsystematic bias. As we cannot mathematically allow for this variation when drawing conclusions, it is appropriate to call such extra variation 'bias' even though it is 'uncertain'. In contrast to sampling variation, the extra uncertainty is independent of sample size as it is a feature of the pre-existing differences between the two populations from which the samples were drawn.

Our resampling studies provide a unique opportunity to calculate the distribution of this extra uncertainty for the specific situations studied in the IST and ECST by calculating the increase in variance seen with non-randomised concurrently controlled studies compared with RCTs. This computation is possible as we ensured that for each study the RCTs are the same size as the concurrent comparisons, such that the differences in variability cannot be explained by differences in sampling variability. The results of these computations are given in *Table 19*. The extra variance in log OR was 0.61 for regional IST comparisons, 0.57 for UK city IST comparisons and 0.01 for regional ECST comparisons. Given these estimates, it is possible to calculate new adapted confidence intervals for these studies that allow for these potential uncertain biases in addition to sampling variation. They are expressed in *Table 19* as multiplicative increases in the width of the standard confidence intervals. As sampling variability decreases with increasing sample size but the unsystematic bias remains constant, the ratio of the extra allowance in the width of the confidence interval due to unsystematic bias increases with sample size. The ratios presented in *Table 19* reveal that standard confidence intervals for many non-randomised

TABLE 19 *Impact of observed increased variability with sample size*

	IST – 14 regions	IST – 10 UK cities	ECST – 8 regions
Observed ratio of SDs for concurrent controls	2.5	1.8	1.01
Increase in variance in log OR attributable to non-random allocation	0.607	0.570	0.014
Total sample size	**Multipliers to confidence interval width to give correct coverage**		
100	1.9	1.5	1.0
200	2.5	1.8	1.0
500	3.8	2.6	1.1
1000	5.2	3.5	1.2
2000	7.3	4.8	1.3
5000	12	7.5	1.7
10000	16	11	2.2
20000	23	15	2.9
50000	36	24	4.4

studies may be an order of magnitude too narrow to describe correctly the true uncertainty in their results, but that there are differences in the adjustments that are needed in different situations. For example, the confidence interval calculated from a concurrently controlled study of 1000 participants may be five times too narrow to describe the true uncertainty for regional IST-type comparisons, three times too narrow to describe the true uncertainty in UK city IST-type comparisons, but only 20% too narrow for regional ECST-type comparisons. For sample sizes of 10,000 the confidence intervals are estimated to be more than 10 times too narrow for the IST situations and half the width needed for the ECST situation. Of course, in practice one would not know to what extent the standard confidence interval under-represented the true uncertainty.

Generalisability and limitations of the findings

The value of these findings and estimates depends on the generalisability of the results obtained from the IST and ECST and the degree to which the slightly artificial methodology and samples used in these evaluations are representative of the reality of non-randomised studies.

Generalisability

The IST and ECST were chosen for this investigation as (a) they were large trials, (b) they had an outcome which was not rare, (c) they were multicentre trials and (d) the trialists were willing to provide reduced and anonymised data sets suitable for our analyses. Other than the fact that both trials relate to stroke medicine, the trials differ considerably. One is a trial of pharmacological agents (aspirin and heparin) whereas the other is a trial of a surgical procedure (carotid endarterectomy). The treatment in one is acute, being given immediately after the patients have suffered a severe stroke, whereas in the other it is preventive, being given to high-risk patients. It is difficult to argue that these trials can be regarded as representative and therefore that the results are generalisable. However, their results should be regarded as being **indicative** of the biases associated with the use of non-random controls. Ideally these resampling study methods should be repeated in more trials. In the case of this project, the time required to generate the resampling studies and the difficulty in obtaining data sets from multicentre clinical trials prevented additional evaluations being undertaken.

It is important also to consider whether the time trend observed in the ECST is likely to be typical of those that may be observed in other areas of healthcare – especially as it is in agreement with the trends observed by Sacks and colleagues in their review across six clinical contexts.[27] The trend is one of patient outcomes improving over time. It is consistent with a general pattern of average outcomes improving with progress in medical care, which may apply across all medical specialities. However, this argument assumes that the case-mix of patients being treated is stable, which may not be the case. In some circumstances changes in case-mix over time, for good reason, may lead to apparent increases in adverse outcomes. For example, if medical information leads to knowledge that the treatment is not suited to patients at low risk, then a change to excluding lower risk patients from receiving that treatment

may lead to increases in average event rates. Historically controlled studies undertaken in such a situation may be prone to underestimating the benefits of treatment and may even falsely conclude that treatment does more harm than good.

The lack of systematic bias with the use of geographical controls is based on the presumption that geographical differences act in a haphazard manner, and are as likely to lead to overestimates of treatment effects as to underestimates. The random manner in which concurrent control groups were selected in the resampling exercise ensured that across a large number of studies these differences would be seen to balance each other out, albeit possibly increasing unpredictability. This result does not indicate that geographically controlled studies are unbiased. In reality, a single comparison between two areas is likely to be biased, as are meta-analyses of several studies, although the direction in which the bias acts may be unknown. In addition, if an investigator chose a geographical control group with knowledge of the likely differences in case-mix, it would be possible for the selection to be manipulated (consciously or subconsciously) in such a way as to introduce a particular bias, akin to the bias observed in RCTs when treatment allocation is not concealed.[20]

Similarly, we should consider whether the mechanisms leading to unpredictability in bias, especially in studies generated from the IST, are likely to apply widely across different clinical areas. *Tables 38* and *40* in Appendix 8 show that the case-mix of patients recruited to the IST varied between locations, both internationally and between cities in the UK. These haphazard differences, together with differences in other unknown risk factors and aspects of patient management and outcome assessment, will have caused the unpredictability in the bias that was observed. Evidence is available in all areas of medicine that such differences exist, and therefore it seems reasonable to conclude that the unpredictable behaviour in biases will be observed elsewhere, although the degree of unpredictability may vary.

Limitations of the resampling methodology

The resampling method used participants recruited to a randomised controlled trial to generate non-randomised studies. This, of course, is not what happens in reality, but there are reasons to believe that our approach is more likely to have led to underestimates than overestimates of bias.

The degree of bias in a non-randomised study depends on the similarity of the two groups from which treated participants and controls are drawn. Sampling these groups from the same randomised trial is likely to have reduced such differences for the following reasons:

1. All participants included in the RCT will have been judged to have been suitable for either treatment. In a non-randomised study participants who are suitable for only one of the two treatments may have been recruited to that arm: there is usually no formal assessment that they would have been considered suitable for the alternative. This difference will nearly always act to increase differences in outcome between the groups.
2. The RCT was conducted according to a protocol, describing methods for recruiting, assessing, treating and evaluating the patients. This will have reduced the variability within the trial. Although some non-randomised studies are organised using a protocol, many are not.
3. All participants included in the trial were recruited prospectively. In non-randomised studies, especially those using historical controls, participants are likely to have been 'retrospectively' included in the study, potentially introducing additional bias.

On balance, it could be argued that using randomly chosen international comparisons for selection of concurrent controls may be regarded as rather artificial and likely to have increased differences between groups. In reality, a concurrent control group in a non-randomised study would be selected to minimise likely differences between groups, and such **long-distance** geographical comparisons would probably be avoided. It is perhaps more realistic to focus on the magnitude of the biases observed in the concurrent comparisons generated from the UK cities in the IST as being more representative of what might occur in reality. The unsystematic bias seen here was less than that observed in international comparisons, but still large enough to lead many studies falsely to obtain significant findings of both benefit and harm.

Importantly, we have concentrated on only one aspect of quality in non-randomised studies: there are other biases to which they are susceptible in the same way as are RCTs.

Conclusions

Individual non-randomised studies may have seriously biased results when either historical or concurrent controls are used. Biases may be large enough to mask both small and moderate treatment effects, to the extent that non-randomised studies may observe statistically significant effects acting in the wrong direction.

While the use of historical controls may frequently lead to overestimates of effectiveness of the experimental treatment, this depends on the underlying time trend of improving outcomes in the patients being studied. Such a trend may not always apply, however, especially when the case-mix of those being considered for treatment also changes over time.

Geographical variations in the provision of care and in the case-mix of patients being considered for treatment will lead to bias in studies using geographical concurrent controls. Differences in case-mix between treatment and control groups are likely to be haphazard, such that the size and magnitude of the biases may be unpredictable.

The magnitude and nature of biases may differ considerably between clinical situations.

While results of RCTs and non-randomised studies may appear to agree on average across many studies, this does not indicate that non-randomised studies are reliable, as individually they are affected by additional unpredictability. Failure to recognise this aspect may have led previous reviews falsely to underestimate the bias in non-randomised studies when they compared average results.

Chapter 7

Empirical evaluation of the ability of case-mix adjustment methodologies to control for selection bias

Introduction

Case-mix (or risk) adjustment methods are used widely throughout healthcare research to enable comparisons to be made between groups that are not directly comparable owing to differences in prognostic (or **confounding**) factors. For example, comparisons between outcomes from different hospitals are often confounded by the severity of the patients that they treat, and it is recommended that outcomes should be compared only when differences in case-mix (severity) are adjusted for.[136,137] The philosophy of case-mix adjustment is to 'level the playing field' during analysis, enabling a comparison of like-with-like to be made at the point of analysis, even if comparable groups could not be generated by the study design. Case-mix adjustment is an attempt to achieve by analysis what could not be done (or was not done) by the design.

Consideration of case-mix is routinely recommended in the analysis of non-randomised studies. Guides for assessing the validity of non-randomised studies for both therapy and harm recommend readers to assess first whether 'investigators demonstrate similarity in all known determinants of outcome' and, if not, whether they 'adjust for differences in analysis'.[138,139] Many epidemiological and biostatistical texts advocate adjustment for differences in baseline covariates when they are observed, both in non-randomised controlled studies and even in RCTs.[140–142]

We will focus on comparisons between two groups, which we refer to as treatment (or experimental) and control. In principle, many of the methods discussed extend to multiple armed trials. Four approaches to dealing with differences in case-mix in non-randomised studies are commonly encountered in the medical literature:

1. **Comparison of baseline characteristics**. The baseline clinical and demographic characteristics of the two groups are compared to ascertain whether differences in case-mix exist and hence determine the certainty with which the observed difference can be attributed to the intervention and not to confounding factors. This method does not attempt to adjust for differences in case-mix but simply to discover whether there is evidence of confounding and make inferences accordingly. The presence of baseline differences is usually determined by tests of statistical significance, although such tests do not relate directly to comparability.[143]
2. **Standardised or stratified analyses**. Study participants are divided into groups (strata) with the same characteristics. Stratified analyses work by making 'like-with-like' comparisons within each of these groups. The overall treatment effect is calculated by computing a weighted average of the within-strata estimates of the treatment effect. The most popular method is that of Mantel and Haenszel.[144] Stratification is best used when there are only one or two baseline characteristics, and is frequently used in epidemiological research to standardise for differences in age and/or sex.
3. **Multiple regression**. Regression models (**linear** regression if the outcome is continuous, **logistic** regression if it is binary and **Cox models** if censoring occurs) estimate how each prognostic factor relates to outcome. When the comparison between the groups is made, adjustments are added to or subtracted from the estimated treatment effect to account for the impact of differences in each of the baseline covariates according to their estimated relationship with outcome. The two stages of the process happen simultaneously, so that the results depend on the correlations between baseline characteristics and treatment allocation. Stepwise regression procedures are commonly used to identify covariates that are significantly related to outcome, and work in a sequential manner such that adjustments are only made for the subset of covariates thought to matter.
4. **Propensity score methods**. Propensity scores are the least familiar method. Whereas

regression models described in (3) model how covariates relate to outcome, propensity score methods model how the same covariates relate to treatment allocation.[145–147] The principle is that the propensity score summarises the manner in which baseline characteristics are associated with treatment allocation, so that selection bias is removed when comparisons are made between groups with similar propensity scores.[146] The method involves calculation for each individual in the dataset the propensity probability that estimates their chance of receiving the experimental intervention from their baseline characteristics. In an RCT, where there should be no relationship between baseline characteristics and treatment assignment, this probability will be the same for each participant (e.g. 0.5 if the groups are of equal size). In non-randomised studies, it is likely that treatment assignment does depend on baseline covariates, and that propensity scores will vary between individuals. For example, patients with more severe disease may be more likely to receive one treatment than the other. Propensity scores in such a situation would relate to disease severity, and the average propensity score in the experimental group will differ from the average in the control group. Estimation of propensity scores is typically undertaken using multiple logistic regression models, the outcome variable being treatment allocation.

What evidence is there that these methods actually adjust for selection bias in non-randomised studies? Although there is plenty of literature demonstrating that case-mix adjustment can change estimates of treatment effects, none of the texts that we have consulted cite any empirical evidence demonstrating that case-mix adjustment, on average, reduces bias. Broader searches of the methodological literature related to case-mix likewise did not identify supporting empirical evidence.

To the contrary, there are hints in the literature that case-mix adjustment methods may not adequately perform the task to which they are applied. Of the eight reviews in Chapter 3, two compared adjusted and unadjusted estimates of treatment effects from non-randomised studies with the results of similar RCTs, and noted that there was little evidence that adjustment consistently moved the estimates of treatment effects from non-randomised studies towards those of RCTs.[25,27]

In an attempt to provide empirical evidence of the value of case-mix adjustment, we have used the same non-randomised studies generated by resampling participants from the IST[132] and ECST[133] data sets (described in Chapter 6) to evaluate the performance of eight different case-mix adjustment methods in controlling for selection bias. As explained previously, resampling of data from the trials was used to generate randomised and non-randomised studies of different designs in such a way that the differences between their results with respect to location and spread could only be attributed to selection bias. Case-mix adjustment methods were then applied to each resampled non-randomised study making use of the available baseline data (different in the two trials – see *Table 20*). The distribution of results of the adjusted non-randomised studies was then compared with both the distribution of unadjusted results and the distribution of the results of the RCTs, to see whether there was any evidence that adjustment had reduced or removed the selection bias.

In our investigations we evaluate the ability of case-mix adjustment to control for three different types of bias:

1. Naturally occurring systematic bias, where results are consistently either all overestimates or all underestimates of the treatment owing to some naturally occurring unknown allocation mechanism (as was observed in the ECST historical comparisons in Chapter 6).
2. Naturally occurring unpredictable bias, where results are biased variably with different magnitudes in different directions leading to both underestimates and overestimates of treatment effects (as was observed in the IST concurrent comparisons in Chapter 6).
3. Bias arising where allocation to treatment is probabilistically linked to a prognostic variable, such that participants are more likely or less likely to receive treatment according to their observed characteristics. This in fact is a model of practice in much of medicine, where treatment decisions are made according to the conditions and characteristics that each patient displays (allocation by indication). Although we have no realistic way of mimicking such a mechanism for the two trial data sets, we generate two artificial scenarios when allocation related (a) to the value of a single prognostic covariate, and, more realistically, (b) to an unknown function of a set of prognostic covariates.

TABLE 20 *Baseline covariates used in case-mix adjustment models for the IST and ECST*

IST	ECST
Binary covariates	
Sex (male/female)	Sex (male/female)
Symptoms noted on waking	Residual neurological signs
Consciousness	Previous MI
Atrial fibrillation	Angina
	Current prophylactic aspirin use
Continuous covariates	
Age	Age
Delay to presentation	Degree of stenosis
Systolic blood pressure	
Unordered categorical variables	
Infarct visible on CT scan	
Type of stroke	
Ordered categorical variables	
Neurological deficit score (7 categories)	Presenting stroke (4 categories)

The investigative method provides an opportunity to make comparisons between different case-mix adjustment strategies: matching baseline groups, stratification, regression and propensity score methods.

Methods

Generation of samples

The principles of our resampling methodology were discussed in detail in Chapter 6. Studies of fixed sample sizes with randomised and non-randomised designs (described below) were generated for each region in each of the IST and ECST data sets by selectively sampling participants, the whole process being repeated 1000 times. In each trial baseline data on important prognostic variables had been recorded for each participant at the point of recruitment. These variables (we will refer to them as **covariates**) were used in the analyses to adjust for differences in case-mix. Details of the covariates available for each study are given in *Table 20* and Appendix 8.

Samples with 'naturally' occurring biases

Historically controlled and concurrently (non-randomised) controlled studies were generated from the IST and ECST data sets as described in Chapter 6. We thus obtained results for:

1. 14,000 **historically controlled studies** based on the 14 international regions from the IST and 14,000 corresponding RCTs, all of sample size 200 (100 per arm)
2. 14,000 **concurrently controlled studies** based on the 14 international regions from the IST and 14,000 corresponding RCTs, all of sample size 200 (100 per arm)
3. 10,000 **concurrently controlled studies** based on the 10 UK cities within the IST and 10,000 corresponding RCTs, all of sample size 200 (100 per arm)
4. 8000 **historically controlled studies** based on the eight international regions from the ECST and 8000 corresponding RCTs, all of sample size 80 (40 per arm)
5. 8000 **concurrently controlled studies** based on the eight international regions from the ECST and 8000 corresponding RCTs, all of sample size 80 (40 per arm).

Different case-mix adjustment methods were applied individually to each of these 54,000 non-randomised studies, and the results were compared with the results from the corresponding 54,000 RCTs.

Samples with bias related to 'known' differences in case-mix (allocation by indication)

In addition to the standard historically and concurrently controlled designs, for the purposes of evaluating the performance of case-mix adjustment we have included two further designs in which bias relates to known relationships with prognostic variables. This has been done for two reasons. First, we wished to evaluate how well case-mix methods work in situations when we have direct knowledge of the bias-inducing mechanism that they are trying to correct. Second, we wished to mimic crudely clinical database-type studies, in

Region	Trial allocation	Number of neurological deficits						
		0	1	2	3	4	5	6
Northern Italy	Aspirin	$T_{0,1}$	$T_{1,1}$	$T_{2,1}$	$T_{3,1}$	$T_{4,1}$	$T_{5,1}$	$T_{6,1}$
	Avoid aspirin	$C_{0,1}$	$C_{1,1}$	$C_{2,1}$	$C_{3,1}$	$C_{4,1}$	$C_{5,1}$	$C_{6,1}$
Central Italy	Aspirin	$T_{0,2}$	$T_{1,2}$	$T_{2,2}$	$T_{3,2}$	$T_{4,2}$	$T_{5,2}$	$T_{6,2}$
	Avoid aspirin	$C_{0,2}$	$C_{1,2}$	$C_{2,2}$	$C_{3,2}$	$C_{4,2}$	$C_{5,2}$	$C_{6,2}$
Southern Italy	Aspirin	$T_{0,3}$	$T_{1,3}$	$T_{2,3}$	$T_{3,3}$	$T_{4,3}$	$T_{5,3}$	$T_{6,3}$
	Avoid aspirin	$C_{0,3}$	$C_{1,3}$	$C_{2,3}$	$C_{3,3}$	$C_{4,3}$	$C_{5,3}$	$C_{6,3}$
Scotland	Aspirin	$T_{0,4}$	$T_{1,4}$	$T_{2,4}$	$T_{3,4}$	$T_{4,4}$	$T_{5,4}$	$T_{6,4}$
	Avoid aspirin	$C_{0,4}$	$C_{1,4}$	$C_{2,4}$	$C_{3,4}$	$C_{4,4}$	$C_{5,4}$	$C_{6,4}$
Northern England and Wales	Aspirin	$T_{0.5}$	$T_{1.5}$	$T_{2.5}$	$T_{3.5}$	$T_{4.5}$	$T_{5.5}$	$T_{6.5}$
	Avoid aspirin	$C_{0,5}$	$C_{1.5}$	$C_{2.5}$	$C_{3.5}$	$C_{4.5}$	$C_{5.5}$	$C_{6.5}$
Southern England	Aspirin	$T_{0,6}$	$T_{1,6}$	$T_{2,6}$	$T_{3,6}$	$T_{4,6}$	$T_{5,6}$	$T_{6,6}$
	Avoid aspirin	$C_{0,6}$	$C_{1,6}$	$C_{2,6}$	$C_{3,6}$	$C_{4,6}$	$C_{5,6}$	$C_{6,6}$
Australia	Aspirin	$T_{0,7}$	$T_{1,7}$	$T_{2,7}$	$T_{3,7}$	$T_{4,7}$	$T_{5,7}$	$T_{6,7}$
	Avoid aspirin	$C_{0,7}$	$C_{1,7}$	$C_{2,7}$	$C_{3,7}$	$C_{4,7}$	$C_{5,7}$	$C_{6,7}$
The Netherlands	Aspirin	$T_{0,8}$	$T_{1,8}$	$T_{2,8}$	$T_{3,8}$	$T_{4,8}$	$T_{5,8}$	$T_{6,8}$
	Avoid aspirin	$C_{0,8}$	$C_{1,8}$	$C_{2,8}$	$C_{3,8}$	$C_{4,8}$	$C_{5,8}$	$C_{6,8}$
New Zealand	Aspirin	$T_{0,9}$	$T_{1,9}$	$T_{2,9}$	$T_{3,9}$	$T_{4,9}$	$T_{5,9}$	$T_{6,9}$
	Avoid aspirin	$C_{0,9}$	$C_{1,9}$	$C_{2,9}$	$C_{3,9}$	$C_{4,9}$	$C_{5,9}$	$C_{6,9}$
Norway	Aspirin	$T_{0,10}$	$T_{1,10}$	$T_{2,10}$	$T_{3,10}$	$T_{4,10}$	$T_{5,10}$	$T_{6,10}$
	Avoid aspirin	$C_{0,10}$	$C_{1,10}$	$C_{2,10}$	$C_{3,10}$	$C_{4,10}$	$C_{5,10}$	$C_{6,10}$
Poland	Aspirin	$T_{0,11}$	$T_{1,11}$	$T_{2,11}$	$T_{3,11}$	$T_{4,11}$	$T_{5,11}$	$T_{6,11}$
	Avoid aspirin	$C_{0,11}$	$C_{1,11}$	$C_{2,11}$	$C_{3,11}$	$C_{4,11}$	$C_{5,11}$	$C_{6,11}$
Spain	Aspirin	$T_{0,12}$	$T_{1,12}$	$T_{2,12}$	$T_{3,12}$	$T_{4,12}$	$T_{5,12}$	$T_{6,12}$
	Avoid aspirin	$C_{0,12}$	$C_{1,12}$	$C_{2,12}$	$C_{3,12}$	$C_{4,12}$	$C_{5,12}$	$C_{6,12}$
Sweden	Aspirin	$T_{0,13}$	$T_{1,13}$	$T_{2,13}$	$T_{3,13}$	$T_{4,13}$	$T_{5,13}$	$T_{6,13}$
	Avoid aspirin	$C_{0,13}$	$C_{1,13}$	$C_{2,13}$	$C_{3,13}$	$C_{4,13}$	$C_{5,13}$	$C_{6,13}$
Switzerland	Aspirin	$T_{0,14}$	$T_{1,14}$	$T_{2,14}$	$T_{3,14}$	$T_{4,14}$	$T_{5,14}$	$T_{6,14}$
	Avoid aspirin	$C_{0,14}$	$C_{1,14}$	$C_{2,14}$	$C_{3,14}$	$C_{4,14}$	$C_{5,14}$	$C_{6,14}$

FIGURE 14 *Resampling structure for IST based on number of neurological deficits*

TABLE 21 *Sampling probabilities use to generate non-randomised designs mimicking 'allocation by indication'*

(a) Sampling probabilities by number of neurological deficits, IST

	0	1	2	3	4	5	6
Aspirin	1.0	0.9	0.8	0.6	0.4	0.2	0.1
Avoid aspirin	0.1	0.2	0.4	0.6	0.8	0.9	1.0

(b) Sampling probabilities by consciousness on presentation, IST

	Fully alert	Drowsy or unconscious
Aspirin	1.0	0.5
Avoid aspirin	0.5	1.0

(c) Sampling probabilities by outcome, IST

	Adverse outcome	No adverse outcome
Experimental	0.75	1.0
Control	1.0	0.75

which patients have been allocated interventions based on their characteristics. These additional analyses were carried out on the IST data only.

Non-randomised studies mimicking 'allocation by indication' were generated by identifying strong prognostic covariates in the IST and biasing allocation according to values of these covariates. These studies were based on a simplistic form of allocation by indication, in which the decision to treat depends on a single observed covariate. Two different covariates were considered separately: the number of neurological deficits and the level of consciousness at admission. The prognostic value of these covariates is summarised in Appendix 8.

To generate bias, participants within each region were divided into groups according to their values of the covariate and the treatment that they received. For example, *Figure 14* shows the categorisation of participants in the IST according to whether they receive aspirin (T) or avoid aspirin (C), the number of neurological deficits with which they present (0–6, first subscript) and the region within which they were recruited (1–14, second subscript). Participants were resampled from this structure with sampling probabilities determined according to their allocation and covariate value. For both covariates sampling probabilities were chosen such that the treated group was weighted towards participants with good prognosis, whereas the control group was weighted towards participants with poor prognosis, leading to an overestimate of the benefit of experimental treatment. The sampling probabilities used with the two covariates are given in *Table 21(a)* and *(b)*. The same sampling probabilities were used in all regions.

In practice, allocation by indication is more complicated than the single covariate allocation described above – it depends on some unknown combination of multiple pieces of covariate information that are predictive of prognosis, and is likely to vary both between individual clinicians and over time. We wished to include studies with such a complex bias in our evaluation to investigate how well case-mix adjustment methods might work in more typical database-type analyses. Research into medical decision making has characterised some decision processes using regression techniques to provide models explaining how covariate information is combined to make treatment decisions. Although it would be possible to generate biased allocations using such a decision model, we did not use this approach as it is too simplistic and artificial to test properly the ability of case-mix adjustment methods to adjust for selection bias. Such analyses would overestimate the ability of case-mix adjustment methods to adjust for bias due to the circularity of one regression model (the case-mix adjustment method) being used to estimate parameters of an underlying regression model (the allocation process). The many treatment decisions that did not fit with the underlying regression model would not be accounted for.

The second approach that we used is based on the observation that when decisions are made in relation to prognosis, those decisions naturally

relate to patient outcome. To generate a bias related to prognosis, therefore, we can take advantage of the fact that the outcomes in these patients are already known, thus skipping the process of selecting a prognostic model to be used to represent treatment decisions. We have therefore generated a prognostic bias mimicking 'allocation by indication' by stratifying patients in each region according to outcome and treatment, and differentially sampling patients with good and bad outcomes in the two treatment arms. The sampling probabilities used are given in *Table 21(c)*. Probabilities were again chosen to lead to overestimation of treatment benefit, and the same sampling probabilities were used in all samples. Correcting for this selection process is likely to be more testing than correcting for real allocation by indication as it is based on information not directly available to the treatment allocator, who will only have available signs, symptoms, history and results of investigations on which to base treatment.

The resampling methods were repeated 1000 times as for the historically and concurrently controlled studies (Chapter 6). Thus, for the IST 14,000 non-randomised studies (1000 for each region) were generated with bias related to neurological deficit score, 14,000 with bias related to level of consciousness and 14,000 with bias related to outcome, all of sample size 200. The results of these studies were compared with those of 14,000 RCTs of the same sample size sampled from the same data. The case-mix adjustment methods were applied to each of these 42,000 non-randomised studies.

Case-mix adjustment methods

Eight case-mix adjustment strategies were investigated: matching baseline groups, stratification, three variants of regression models and three propensity score methods.

Exclusion of non-matching groups

The first analysis investigated the degree to which the absence of differences in case-mix was related to the absence of selection bias. The significance of baseline differences for the 10 prognostic variables for IST and eight prognostic variables for ECST that are listed in *Table 20* was tested using *t*-tests (continuous variables), chi-squared tests (binary and unordered categorical variables), and chi-squared tests for trend (ordered categorical variables) using the STATA commands `ttest`, `tabulate` and `mhodds`, respectively. Studies were then classified by the number of covariates with significant ($p < 0.05$) baseline differences. Although we do not imagine that researchers would desist from analysing or publishing results of non-randomised studies with significant differences in one or more baseline covariates, this analysis allows us to investigate the advice given to readers to assess whether 'investigators demonstrate similarity in all known determinants of outcome'.[138,139]

Stratification on a single factor

The Mantel–Haenszel stratified method[144] was applied to all non-randomised studies for both trials. Stratification was undertaken according to the most strongly prognostic factor for each of the two trial data sets: the neurological deficit score was used for IST (seven categories) and the degree of stenosis for ECST (four categories). The STATA `mhodds` command was used to obtain an overall OR together with 95% confidence interval (calculated using the standard equation proposed by Robins and colleagues[148]).

Multiple logistic regression models

Multiple logistic regression models were fitted to the data from each non-randomised study for both trials. Three variants of the method were applied. The first included all prognostic variables listed in *Table 20*, and is termed the 'full model'. For both the IST and ECST studies a variable was also included indicating treatment group, the value of which yielded the estimate of log OR (and 95% confidence interval) of the adjusted treatment effect. Analysis was done using the STATA `logit` command.

Two stepwise alternatives to the full model were considered, both using the STATA `sw logit` command. Both used backward selection techniques removing non-significant variables from the full model one at a time. The models differed in the significance level used to decide whether to remove a variable. The first method removed variables with p-values >0.15. The second method used a stricter p-to-remove value of 0.05, requiring greater evidence of a relationship between a covariate and outcome for the covariate to be included in the final model. In both situations the variable for treatment group was forced to remain in the model throughout the stepwise procedure, its value again yielding the estimate of log OR (and 95% confidence interval) for the adjusted treatment effect.

The logistic regression models estimated only the main effects associated with the covariates in the model (as is typical of logistic regression analyses published in the medical literature). No attempt

was made to investigate interactions, or to investigate non-linear models of continuous covariates. Although these could be regarded as important oversimplifications, the size of the data set (n = 200 for IST, n = 80 for ECST) means that such investigations would be underpowered.

Propensity score methods

As noted above, an individual's propensity score is their probability of being in the treatment group given their baseline data. Once the score has been calculated for all individuals in the data set, it can be used in the analysis in one of three ways. First, participants in the experimental and control groups can be selected to generate groups with similar distributions of propensity score by matching each individual in the experimental group with an individual in the control group who shares the same (or a very similar) propensity score. Participants for whom no match can be found are excluded from the analysis. Second, a stratified analysis can be undertaken, the participants being divided into, say, quintiles according to their propensity score, and comparisons made within each of the five groups. As with conventional stratified analysis, the overall treatment effect is estimated by calculating a weighted average of the within strata estimates of the treatment effect. Third, the propensity score can be used as a predictor of outcome in a logistic regression model. The model will estimate the relationship between propensity score and outcome, and make a suitable adjustment to the estimate of treatment effect according to the difference in mean propensity score between treated and control groups.

The first stage of the propensity score method involved calculating propensity probabilities for all members of each data set, using a logistic regression model with 'treatment' as the outcome variable. The standard sets of 10 covariates for the IST and eight covariates for the ECST (as listed in *Table 20*) were included in this model, with no interaction terms being considered. The `logistic` command in STATA was used to fit the model and obtain estimated propensity scores for each participant. The propensity score was then used in the three different ways: matching, stratification and regression.

Matching on propensity scores involved pairing each treated participant with a control participant who has an identical or very similar propensity score. Our definition of 'very similar' was Rosenbaum and Rubin's suggestion of being within 0.25 SD of the logit of the propensity score.[145] This approach takes into account the observed variability of the computed propensity scores in each study. The difference d in logit propensity scores was calculated for all possible pairs of treated and control participants. The pairings were sorted in ascending order according to d, and pairs where d was greater than our matching criterion were dropped. Then the first matched pair (t_1, c_1) was selected, and all other pairs including either of t_1 and c_1 dropped. The process was repeated through the data set, selecting the next available matched pair (t_i, c_i), and dropping all other pairs including either of t_i or c_i. Participants in both groups who were not matched were excluded from the analysis. The individual pairing was not exploited in the analysis: the OR was computed as the ratio of the odds of the outcome in the treated group to the odds in the control group in the standard way, with a standard error for log OR calculated using the large sample estimate.

The stratified analysis used the Mantel–Haenszel method to estimate a common OR across the participants grouped according to their propensity score. In each analysis five groups were used, defined by calculating the quintiles of the observed propensity scores (pooled across treatment and control groups). Calculations were performed using the STATA `mhodds` command.

The regression analysis estimated the treatment effect adjusted for differences in propensity score by fitting a logistic regression model to the outcome data with just two covariates, the treatment group and the log OR of the propensity score. The model was fitted using the STATA `logistic` command.

Statistical analyses

As with the analysis of bias related to study designs (Chapter 6), the focus of the statistical analysis was on describing and comparing the location (average) and spread (variability) of the treatment effects. Distributions were considered first for the RCTs and the non-randomised studies without adjustment, and second for the non-randomised studies with each of the eight case-mix strategies described above. In all analyses treatment effects were expressed as log OR with 95% confidence intervals, and interpreted according to their statistical significance assessed at the 5% level using a two-sided test.

Distributions of results were considered graphically using dotplots, and through statistics summarising location [the 'average OR', computed

TABLE 22 *Comparison of concurrently and historically controlled studies with results of RCTs resampled from 14 regions within the IST analysed according to the number of significant differences in baseline covariates*

Study design	Maximum no. of significant baseline differences ($p < 0.05$) for match[a]	Percentage of studies meeting criterion	Average OR	Ratio of SD of log OR	Percentage of studies with statistically significant results ($p < 0.05$)		
					Benefit	Harm	Total
RCT			0.89		9	2	11
Non-randomised historical control	Any number	100	0.88	1.3	15	4	19
	0 or 1 or 2	79	0.89	1.3	14	4	18
	0 or 1	55	0.89	1.2	12	3	15
	0	23	0.92	1.2	12	3	15
RCT			0.91		7	2	9
Non-randomised concurrent control	Any number	100	0.91	2.6	29	22	51
	0 or 1 or 2	20	0.93	2.8	27	19	46
	0 or 1	6	0.91	2.1	26	16	42
	0	1	0.86	1.5	20	7	27

[a] Baseline differences were considered for 10 variables: time since symptoms, consciousness, age, sex, presence of atrial fibrillation, stroke noticed on waking, systolic blood pressure, infarct visible on CT scan, type of stroke and a score of eight neurological deficits.

TABLE 23 *Comparison of concurrently and historically controlled studies with results of RCTs resampled from eight regions within the ECST analysed according to the number of significant differences in baseline covariates*

Study design	Maximum no. of significant baseline differences ($p < 0.05$) for match[a]	Percentage of studies meeting criterion	Average OR	Ratio of SD of log OR	Percentage of studies with statistically significant results ($p < 0.05$)		
					Benefit	Harm	Total
RCT			1.23		4	12	16
Non-randomised historical control	Any number	100	1.06	1.0	6	10	16
	0 or 1 or 2	86	1.01	1.0	7	9	16
	0 or 1	57	0.93	1.0	8	7	15
	0	18	0.81	0.9	9	3	12
RCT			1.08		3	5	9
Non-randomised concurrent control	Any number	100	1.08	1.0	3	6	9
	0 or 1 or 2	69	1.06	1.0	3	5	8
	0 or 1	40	1.04	1.0	3	4	7
	0	12	1.17	0.9	3	4	7

[a] Baseline differences were considered for eight variables: severity of presenting stroke, degree of stenosis, age, prophylactic aspirin use, angina, previous myocardial infarction, residual neurological signs and sex.

as the exponential of the mean of the log OR] and spread [the standard deviation of the observed log OR]. As in Chapter 6, ratios of the the average ORs of the RCTs and the average ORs of the non-randomised studies (adjusted or unadjusted) quantify systematic bias, while ratios of their SDs indicate unpredictability in the bias. The likely conclusions of the analyses were investigated by considering the percentage of studies for each analysis which reported statistically significant ($p < 0.05$) results, separately in the direction of harm and of benefit.

Results

Exclusion of non-matching groups

Table 22 presents the results for analyses of the IST according to the number of covariates that differed significantly between treatment and control groups. *Table 23* presents comparable analyses for the ECST.

For the IST, 23% of the historically controlled studies appeared to have comparable groups in that they had no statistically significant differences in baseline covariates (*Table 22*). There did appear to be a reduction in both systematic and unpredicatable dimensions of bias as the number of differing covariates reduced. However, among studies with no significantly unbalanced covariates the results were still more variable than those from the corresponding RCTs, and there was still an excess of statistically significant results.

Table 23 shows that 18% of the historically controlled non-randomised studies generated from the ECST had no statistically significant differences in baseline covariates. A reverse trend was noted with study results becoming more biased with fewer significant differences in baseline, contrary to the hypothesis that comparability at baseline predicts reliable results.

Only 1% of concurrently controlled studies from the IST had no significant differences at baseline (*Table 22*), and reductions in excess variability were less marked than for historically controlled studies. Further, the absence of statistically significant differences did not guarantee comparability, with spuriously statistically significant treatment effects still being common when treated and control groups appeared to match.

As shown in *Table 23*, 12% of the concurrently controlled non-randomised studies generated from the ECST had no statistically significant differences in baseline covariates, but again systematic bias was actually higher in this subgroup of apparently comparable studies than in the studies which had significant differences in at least one baseline covariates.

Adjustment for 'naturally' occurring biases

Systematic bias originating from historically controlled studies

The clearest example of systematic bias was observed in historically controlled studies in the ECST data, where the average OR estimate was 1.06 compared with 1.23 in the RCTs, grossly underestimating the harmfulness of treatment (see Chapter 6). The seven case-mix adjustment methods were applied to each of the historically controlled studies to investigate the degree to which the case-mix methods could adjust for this bias. The results are presented in *Table 24* and *Figure 15*. The adjusted results from six of the seven methods were on average more biased than the unadjusted results: only the full logistic regression model appeared to reduce bias. All methods also increased variability (unpredictability in bias), the greatest increase being with the full logistic regression. This increase in the width of distributions of results is discernible in *Figure 15*.

The use of adjustment methods inflates the standard error of the estimate of a treatment effect. The use of propensity score matching led in addition to a reduction in sample size (as 45% of participants were discarded), further reducing power. Thus, although many of the adjusted estimates were on average more biased than the unadjusted estimates, only logistic regression methods had markedly increased spurious significance rates.

Systematic bias was also observed in the historically controlled studies generated from the IST, although it differed between regions. Overall a small systematic bias was noted in the aggregated results (*Table 25*). Logistic regression modelling showed a similar pattern of behaviour as for the ECST comparison, increasing both average bias and the variability of results. Propensity score methods slightly overadjusted for the bias, but gave results closest to the RCT results. Spurious statistical significance rates decreased slightly with logistic regression and were completely removed by propensity score methods.

The case-mix adjusted results for the historically controlled studies from the IST analysis are

TABLE 24 *Comparison of methods of case-mix adjustment applied to results of historically[a] controlled studies with results of RCTs resampled from eight regions within the ECST*

	Average OR	Variability of results		Percentage of studies with statistically significant results ($p < 0.05$)		
		SD of log OR	Ratio with RCT	Benefit	Harm	Total
RCTs	1.23	0.83		4	12	16
Historically controlled studies						
Unadjusted	1.06	0.85	1.03	6	10	16
Stratification	0.99	0.93	1.12	8	10	18
Logistic regression						
Full model[b]	1.17	1.96	2.37	12	14	26
Stepwise $p_r = 0.05$[c]	1.04	1.16	1.40	11	14	25
Stepwise $p_r = 0.15$[d]	1.01	1.32	1.59	13	15	27
Propensity score						
Matched[e]	0.96	1.09	1.32	5	10	16
Stratified	1.01	1.12	1.35	12	10	22
Regression	1.00	1.11	1.34	12	10	21

[a] Data excluded after protocol change in 1990.
[b] Full model includes eight covariates.
[c] Mean number of covariates included: 2.3.
[d] Mean number of covariates included: 3.5.
[e] Mean number of patients matched: 44/80.

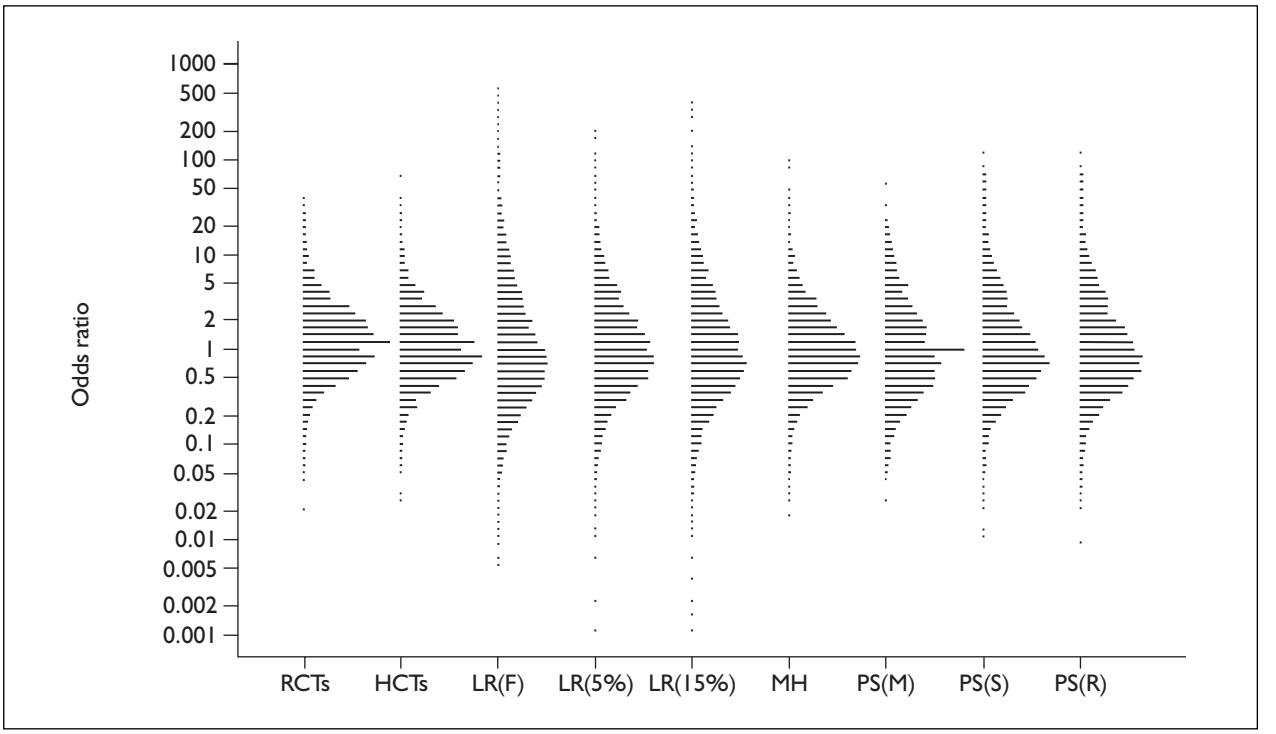

FIGURE 15 *Comparison of methods of case-mix adjustment applied to results of historically controlled studies resampled from eight regions within the ECST. RCT: results from corresponding randomised controlled trials; HCTs: unadjusted historical controls without adjustment, LR(F): adjustment using full logistic regression analysis; LR(5%): adjustment with stepwise logistic regression with $p_r = 0.05$; LR(15%): adjustment with stepwise logistic regression with $p_r = 0.15$; MH: adjustment by Mantel–Haenszel stratification; PS(M): adjustment by matching on propensity score; PS(S): adjustment by stratification on propensity score; PS(R): regression adjustment based on propensity score.*

TABLE 25 *Comparison of methods of case-mix adjustment applied to results of historically controlled studies with results of RCTs resampled from 14 regions within the IST*

	Average OR	Variability of results		Percentage of studies with statistically significant results ($p < 0.05$)		
		SD of log OR	Ratio with RCT	Benefit	Harm	Total
RCTs	0.89	0.35		9	2	11
Historically controlled studies						
Unadjusted	0.88	0.44	1.23	16	4	20
Stratification	0.88	0.53	1.51	17	5	22
Logistic regression						
Full model[a]	0.85	0.56	1.60	13	3	16
Stepwise $p_r = 0.05$[b]	0.84	0.52	1.49	15	3	18
Stepwise $p_r = 0.15$[c]	0.85	0.53	1.51	14	3	17
Propensity score						
Matched[d]	0.91	0.43	1.23	7	2	9
Stratified	0.90	0.40	1.14	9	2	11
Regression	0.91	0.39	1.11	9	2	11

[a] Full model includes 10 covariates.
[b] Mean number of covariates included: 4.5.
[c] Mean number of covariates included: 5.7.
[d] Mean number of patients matched: 132 out of 200.

considered by region in the last six columns of *Table 26*. Stratification reduced systematic bias in only two of the 14 regions, logistic regression adjustments reduced bias in six regions, propensity score methods reduced bias in five of these six, and in two additional regions. Overall, bias introduced by use of an historical control group was consistently reduced by case-mix adjustment methods in less than half of the regions.

Where biases were increased by adjustment, the direction of the increase was unpredictable. In Scotland the historical control result (OR = 1.23) suggested the treatment to be harmful, in contrast to a beneficial result observed in the RCTs (OR = 0.78). Logistic regression further increased this bias (OR = 1.40). In Norway the opposite pattern was seen, with adjustment by logistic regression (OR = 0.41) increasing the overestimate of treatment benefit in the historical controls (OR = 0.48) compared with RCTs (OR = 0.78). However, in some regions, such as The Netherlands, adjustment moved the historical control estimates (OR = 0.85) from a value which was lower than the RCTs (OR = 0.88) to a higher value (OR = 1.14), changing a relatively correct estimate of the benefit of the intervention to a biased estimate suggestive of harm. Similar, but less extreme, changes occurred with propensity score methods.

Unpredictability in bias originating from concurrently controlled studies

Unpredictability in bias was observed most clearly in the IST concurrently controlled comparisons. The ability of the seven case-mix adjustment methods to correct these biases is summarised in *Tables 27* and *28* for regional comparisons in the IST and UK city comparisons in the IST. Regional comparisons in the ECST are given in *Table 29* for completeness. The results for the studies demonstrating the largest unpredictable biases, the regional IST comparison (see Chapter 6), are also shown in *Figure 16*.

As with the historically controlled studies, logistic regression increased the variability of results for all three situations, the increased spread of results being evident in *Figure 16*. Use of the full logistic model (including all covariates) increased spread more than use of stepwise models. Again, there was little corresponding increase in spurious significance levels as the power of the analyses was reduced. Propensity score methods slightly reduced unpredictable bias and spurious significance rates for two of the three situations, while stratification made little difference. Although there was no evidence of a systematic bias in the unadjusted results, both logistic regression and propensity score methods introduced small systematic biases in most

TABLE 26 *Comparison of methods of case-mix adjustment applied to results of historically controlled studies with results of RCTs resampled from 14 regions within the IST, analysed by region*

	RCTs	Historically controlled studies							
		Unadjusted		Logistic regression[a]		Stratified		Propensity score[b]	
	Average OR	Average OR	% over or under estimate	Average OR	% over or under estimate	Average OR	% over or under estimate	Average OR	% over or under estimate
Scotland	0.78	1.23	+58	1.40	+79	1.48	+90	1.30	+67
Southern Italy	0.56	1.00	+79	0.98	+75	0.89	+59	0.97	+73
Northern Italy	0.94	1.15	+22	1.06	+13	1.20	+28	1.05	+12
New Zealand	0.81	0.97	+20	0.94	+16	1.09	+35	1.08	+33
Australia	1.30	1.39	+7	1.09	–16	1.49	+15	1.08	–17
Switzerland	0.97	1.07	+10	1.05	+8	1.13	+16	1.04	+7
Central Italy	0.89	0.92	+3	0.85	–4	0.89	0	0.87	–2
The Netherlands	0.88	0.85	–3	1.14	+30	1.11	+26	1.10	+25
Northern England	0.95	0.92	–3	0.79	–17	1.03	+8	0.86	–9
Spain	1.00	0.93	–7	0.59	–41	0.57	–43	0.75	–25
Southern England	1.01	0.87	–14	0.72	–29	0.85	–16	0.81	–20
Poland	0.91	0.70	–23	1.08	+19	0.69	–24	1.06	+16
Norway	0.78	0.48	–38	0.41	–47	0.47	–40	0.51	–35
Sweden	0.94	0.44	–53	0.47	–50	0.37	–61	0.58	–38

[a] Logistic regression was performed using the full model with 10 covariates.
[b] Propensity scores were used in a regression adjustment.

TABLE 27 *Comparison of methods of case-mix adjustment applied to results of concurrently controlled studies with results of RCTs resampled from 14 regions within the IST*

	Average OR	Variability of results		Percentage of studies with statistically significant results ($p < 0.05$)		
		SD of log OR	Ratio with RCT	Benefit	Harm	Total
RCTs	0.91	0.34		7	2	9
Concurrently controlled studies						
Unadjusted	0.91	0.85	2.51	29	21	50
Stratification	0.92	0.89	2.70	27	21	48
Logistic regression						
Full model[a]	0.92	1.12	3.39	23	18	41
Stepwise $p_r = 0.05$[b]	0.91	1.01	3.06	26	20	46
Stepwise $p_r = 0.15$[c]	0.91	1.03	3.12	26	20	46
Propensity score						
Matched[d]	0.95	0.79	2.39	15	12	27
Stratified	0.94	0.76	2.30	19	15	34
Regression	0.95	0.75	2.27	19	15	34

[a] Full model includes 10 covariates.
[b] Mean number of covariates included: 4.6.
[c] Mean number of covariates included: 5.8.
[d] Mean number of patients matched: 101 out of 200.

TABLE 28 *Comparison of methods of case-mix adjustment applied to results of concurrently controlled studies with results of RCTs resampled from 10 UK cities within the IST*

	Average OR	Variability of results		Percentage of studies with statistically significant results ($p < 0.05$)		
		SD of log OR	Ratio with RCT	Benefit	Harm	Total
RCT	1.01	0.49		7	6	13
Concurrently controlled studies						
Unadjusted	1.02	0.90	1.84	22	23	45
Stratification	0.99	0.91	1.86	22	21	43
Logistic regression						
Full model[a]	1.05	0.96	1.96	16	17	33
Stepwise $p_r = 0.05$[b]	1.03	0.89	1.82	18	20	38
Stepwise $p_r = 0.15$[c]	1.04	0.89	1.82	18	20	38
Propensity score						
Matched[d]	1.05	0.84	1.71	11	13	24
Stratified	1.05	0.79	1.61	14	16	30
Regression	1.03	0.74	1.51	13	15	28

[a] Full model includes 10 covariates.
[b] Mean number of covariates included: 3.4.
[c] Mean number of covariates included: 4.4.
[d] Mean number of patients matched: 109 out of 200.

situations, and a large bias for concurrently controlled studies from the ECST.

Adjustment for bias related to 'known' differences in case-mix

In this section we summarise the results for the three sets of studies constructed to mimic studies where allocation occurred by indication.

Adjusting for bias due to a known and measured covariate

The results of case-mix adjustment for the IST studies with bias relating to neurological deficit are given in *Table 30*. In these analyses systematic bias was introduced according to values of a single covariate (neurological deficit). The same covariate was then included in the models, and

TABLE 29 *Comparison of methods of case-mix adjustment applied to results of concurrently controlled studies with results of RCTs resampled from eight regions within the ECST*

	Average OR	Variability of results		Percentage of studies with statistically significant results ($p < 0.05$)		
		SD of log OR	Ratio with RCT	Benefit	Harm	Total
RCT	1.08	0.69		3	5	9
Concurrently controlled studies						
Unadjusted	1.08	0.69	1.01	3	6	9
Stratification	1.09	0.74	1.08	3	6	9
Logistic regression						
Full model[a]	1.20	1.43	2.10	4	6	10
Stepwise $p_r = 0.05$[b]	1.06	0.82	1.21	5	7	12
Stepwise $p_r = 0.15$[c]	1.07	0.91	1.34	5	8	13
Propensity score						
Matched[d]	1.05	0.87	1.27	2	6	8
Stratified	1.06	0.82	1.20	5	3	8
Regression	1.06	0.80	1.17	5	3	8

[a] Full model includes eight covariates.
[b] Mean number of covariates included: 1.90.
[c] Mean number of covariates included: 3.06.
[d] Mean number of patients matched: 40/80.

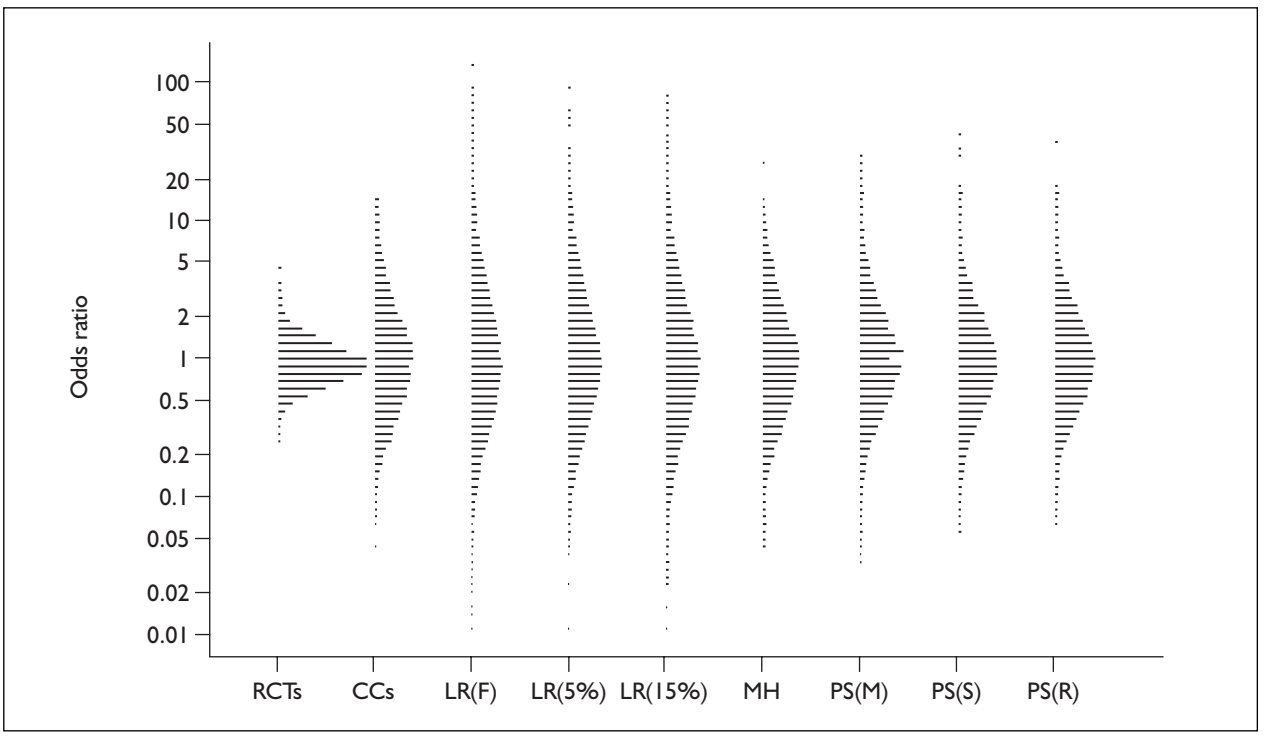

FIGURE 16 *Comparison of methods of case-mix adjustment applied to results of concurrently controlled studies resampled from 14 regions within the IST. RCT: results from corresponding randomised controlled trials; CCs: unadjusted concurrent controls without adjustment, LR(F): adjustment using full logistic regression analysis; LR(5%): adjustment with stepwise logistic regression with $p_r = 0.05$; LR(15%): adjustment with stepwise logistic regression with $p_r = 0.15$; MH: adjustment by Mantel–Haenszel stratification; PS(M): adjustment by matching on propensity score; PS(S): adjustment by stratification on propensity score; PS(R): regression adjustment based on propensity score.*

TABLE 30 *Case-mix adjustment with bias caused by a known and measured single covariate [neurological deficit score (IST)]*

	Average OR	Variability of results		Percentage of studies with statistically significant results ($p < 0.05$)		
		SD of log OR	Ratio with RCT	Benefit	Harm	Total
RCT	0.91	0.33		7	2	9
Studies where treatment is related to condition						
Unadjusted	0.53	0.36	1.09	53	0	53
Stratification	0.93	0.41	1.24	7	3	10
Logistic regression						
Full model[a]	0.89	0.55	1.67	8	3	11
Stepwise $p_r = 0.05$[b]	0.87	0.51	1.55	12	3	15
Stepwise $p_r = 0.15$[c]	0.88	0.52	1.58	10	3	13
Propensity score						
Matched[d]	0.93	0.45	1.36	4	2	6
Stratified	0.90	0.40	1.21	6	2	8
Regression	0.92	0.39	1.18	5	2	7

[a] Full model includes 10 covariates (including neurological deficit score).
[b] Mean number of covariates included: 4.6.
[c] Mean number of covariates included: 5.9.
[d] Mean number of patients matched: 127 out of 200.

TABLE 31 *Case-mix adjustment with bias caused by a known but unmeasured single covariate [consciousness level (IST)]*

	Average OR	Variability of results		Percentage of studies with statistically significant results ($p < 0.05$)		
		SD of log OR	Ratio with RCT	Benefit	Harm	Total
RCT	0.91	0.33		7	2	9
Studies where treatment is related to condition						
Unadjusted	0.64	0.34	1.03	34	0	34
Stratification	0.70	0.38	1.15	23	0	23
Logistic regression						
Full model[a]	0.71	0.48	1.45	17	1	18
Stepwise $p_r = 0.05$[b]	0.69	0.43	1.30	21	1	22
Stepwise $p_r = 0.15$[c]	0.69	0.44	1.33	20	1	21
Propensity score						
Matched[d]	0.79	0.38	1.15	9	0	9
Stratified	0.77	0.34	1.03	12	0	12
Regression	0.78	0.34	1.03	11	0	11

[a] Full model includes 10 covariates (excludes consciousness).
[b] Mean number of covariates included: 4.0.
[c] Mean number of covariates included: 5.3.
[d] Mean number of patients matched: 134 out of 200.

used as the variable for stratification. The bias was appropriately adjusted for by all seven methods. Adjustments made by stratification gave results closest to those of the RCTs.

Adjusting for bias due to a known but unmeasured covariate

Table 31 reports results of case-mix adjustment for the bias in the IST trial introduced when treatment allocation was linked to the level of consciousness at admission. However, in this situation, the adjustments were made without the covariate indicating level of consciousness being included in the models. As before, stratification was undertaken using the neurological deficit score, but the regression and propensity score models were based on the remaining nine out of 10 covariates.

TABLE 32 *Case-mix adjustment with bias caused by multiple covariates, some measured some unmeasured, with unknown mechanism (based on observed outcomes in the IST)*

	Average OR	Variability of results		Percentage of studies with statistically significant results ($p < 0.05$)		
		SD of log OR	Ratio with RCT	Benefit	Harm	Total
RCT	0.91	0.33		7	2	9
Studies where treatment is related to condition						
Unadjusted	0.51	0.34	1.03	60	0	60
Stratification	0.51	0.38	1.15	54	0	54
Logistic regression						
Full model[a]	0.45	0.50	1.52	50	0	50
Stepwise $p_r = 0.05$[b]	0.47	0.44	1.33	51	0	51
Stepwise $p_r = 0.15$[c]	0.47	0.46	1.39	51	0	51
Propensity score						
Matched[d]	0.58	0.37	1.12	31	0	31
Stratified	0.57	0.34	1.03	39	0	39
Regression	0.57	0.33	1.00	39	0	39

[a] Full model includes 10 covariates.
[b] Mean number of covariates included: 4.5.
[c] Mean number of covariates included: 5.8.
[d] Mean number of patients matched: 137 out of 200.

All methods adjusted the crude estimate of the treatment effect (OR = 0.64) in the direction of the result of the RCTs (OR = 0.91), thus removing some of the selection bias. However, stratification and logistic regression (LR) removed only a small fraction of the bias (OR for LR full model = 0.71, OR for stratification = 0.70), propensity score (PS) methods did somewhat better (OR for a matched PS of 0.79), but remained substantially biased and hence gave far too many statistically significant results. While the selection mechanism was not designed to introduce an unpredictable bias, the results from the logistic regression model were much more variable than the unadjusted results.

Adjusting for bias due to unknown multiple covariates

Selection according to outcome, as anticipated, introduced strong biases into the data. For the IST, the non-randomised unadjusted results (OR = 0.51) significantly overestimated treatment efficacy compared with the RCT results (OR = 0.91) (*Table 32*).

Stratification failed to adjust for the bias at all, with the results being identical with the unadjusted results. Significance rates decreased slightly.

Adjustment using logistic regression increased bias. For the IST the unadjusted average OR of 0.51 decreased to 0.45 (full model). Variability of results also increased, the distribution of adjusted results being 1.48 times the variability of unadjusted results for the IST. Significance rates decreased slightly.

PS methods slightly reduced bias in the IST. The variability of results increased, but not as much as for logistic regression results, whilst significance rates decreased.

Discussion

> "The first experience with multivariate analysis is apt to leave the impression that a miracle in the technology of data analysis has been revealed; the method permits control for confounding and evaluation of interactions for a host of variables with great statistical efficiency. Even better, a computer does all the arithmetic and neatly prints out the results. The heady experience of commanding a computer to accomplish all these analytic goals and the simply gathering and publishing the sophisticated 'output' with barely a pause for retyping is undeniably alluring. However useful it may be, multivariate analysis is not a panacea. The extent to which this process represents improved efficiency rather than just bias depends on the adequacy of the assumptions built into the mathematical model."
>
> From Rothman[149]

The results of our investigations can be summarised by the following four key results, all of

which raise concerns about the performance of case-mix adjustment methods:

1. Comparisons between non-randomised groups that appear comparable in terms of case-mix are often biased, sometimes more than for non-randomised groups that do not appear comparable.
2. Case-mix adjustment methods rarely adequately adjust for differences in case-mix.
3. Logistic regression always increases variability in study results.
4. All adjustment methods can on occasion increase systematic bias.

The first and second observations are not surprising, and have been discussed before.[150] The third observation has been demonstrated theoretically always to be the case,[151] although we suspect that this result is not well known. The fourth observation is contrary to most beliefs about case-mix adjustment methods, and has only become detectable through the unique resampling design of our investigation.

As Rothman described above, statistical risk adjustment methods are alluring, appearing to provide a simple solution to many of the inadequacies of the design and execution of non-randomised studies. This message has been widely disseminated throughout the medical research community, leading to their routine use in epidemiology and health services research.[136] However, the validity of a risk-adjustment model depends on fulfilling a demanding set of assumptions. Below we consider the assumptions that may be most critical.

Why adjustment methods might not work?

Omitted covariates

Risk adjustment models can only adjust for differences in variables that have been observed and measured. As seen with the 'allocation by indication' mechanisms based on neurological deficits (IST) (*Table 30*), if all variables linked to the allocation mechanism have been measured and observed, then adequate adjustment may be made. However, in most situations we do not know the variables upon which allocation is based. There may be important prognostic factors that the investigators do not know about or have not measured which are unbalanced between groups and responsible for differences in outcome. For example, when allocation may be influenced by level of consciousness (IST) (*Table 31*) but the consciousness variable was not included in the model, inadequate adjustment for bias was made. If the missing covariates affecting allocation are correlated with the observed covariates, some degree of adjustment is likely to be observed (as was seen for degree of consciousness), but it is unlikely to be adequate unless the correlations are very strong. Many texts refer to the unadjusted effect as 'residual confounding' or 'hidden bias'. Rosenbaum proposed a strategy for investigating the robustness of an observational finding of hidden bias based on sensitivity analyses which determine the size of covariate association required to nullify an observed treatment effect.[152] In some situations this approach could help to decide whether hidden bias could fully explain an observed effect, but such an assessment retains a degree of subjective judgement.

Our first and second results could be explained by the common situation of treatment assignment depending on unmeasured covariates. However, missing covariates cannot explain the third and fourth results.

Misspecified continuous covariates and omitted interactions

Two forms of misspecification can occur in case-mix adjustment analyses. The first is when a continuous variable is categorised, or a categorical variable is regrouped into a smaller number of categories. Cochran[153] presented analytical investigations that showed that for a continuous covariate related to outcome with a monotonic trend, dichotomisation would leave 36% of the variability caused by the relationship unexplained (subject to some distributional assumptions). His results suggest that five categories are needed to explain successfully at least 90% of the variability. Brenner and Blettner[154] extended this work to consider the efficiencies of different approaches of modelling various monotonic trends using multiple categorisations and linear terms.

However, Brenner[155] also showed that if the covariate does not have a monotonic effect, correct categorisation of the data can be crucial to obtaining a sensible result. *Table 33* presents hypothetical data taken from Brenner's paper where there is a U-shaped relationship between the covariate and the outcome. Adjusting for the covariate classified in three categories (*Table 33a*) makes an appropriate adjustment, the observed treatment effect being OR = 1 in all categories. However, if the risk factor is dichotomised different results are obtained depending on where the dichotomisation is made: in *Table 33(b)* the

TABLE 33 *Hypothetical example demonstrating the potential impact of crudely classifying a covariate*

(a) Variable classified in three categories

	CV = 1		
	Dead	**Alive**	**Total**
T	20	10	30
C	160	80	240
Total	180	90	270
	OR = 1		

	CV = 2		
	Dead	**Alive**	**Total**
T	20	110	130
C	20	110	130
Total	40	220	260
	OR = 1		

	CV = 3		
	Dead	**Alive**	**Total**
T	16	8	24
C	2	1	3
Total	18	9	27
	OR = 1		

(b) Variable dichotomised at CV = 1

	CV = 1		
	Dead	**Alive**	**Total**
T	20	10	30
C	160	80	240
Total	180	90	270
	OR = 1		

	CV > 1		
	Dead	**Alive**	**Total**
T	36	118	154
C	22	111	133
Total	58	229	287
	OR = 1.54		

(c) Variable dichotomised at CV = 2

	CV < 2		
	Dead	**Alive**	**Total**
T	40	120	160
C	180	190	370
Total	220	310	530
	OR = 0.35		

	CV = 3		
	Dead	**Alive**	**Total**
T	16	8	24
C	2	1	3
Total	18	9	27
	OR = 1		

Adapted from Brenner, 1997[155]. C, control group; CV, covariate; T, treatment group.

stratified estimate of the OR is >1 and in *Table 33(c)* the stratified estimate of the OR is >1.

Misspecification of relationships for continuous or ordinal covariates could therefore also (partly) explain the lack of adjustment we observed (Result 2). It is also possible that misspecification of a U-shaped relationship as a dichotomy or a linear trend could explain the increase in bias observed with adjustment (Result 4). However, it seems unlikely that this is the full explanation of these results as there are few true continuous or ordinal covariates in the analyses, and U-shaped relationships are rare.

The second type of misspecification is the omission of interactions between covariates, or between covariates and allocation. If covariates do interact, omitting the interaction term from a case-mix adjustment model will lead to a reduction in the potential adjustment that can be made, hence increasing residual confounding is another possible explanation of our second result.

Interactions between covariates and the treatment allocation are more complicated, and imply that the treatment effect should not be summarised as a single value but as a set of values dependent on the covariate. We have not considered this level of complexity for either the results of RCTs or non-randomised studies (despite this being likely for the ECST, where degree of stenosis is known to relate to benefit of carotid endarterectomy). When these interactions are omitted, the effect of treatment is summarised as a single 'average' value, the average depending on the distribution of covariates of participants included in the study.

As the omission of interactions applied to both the RCTs and the non-randomised studies, it is unlikely to be the key explanation of the results that we observed.

Multicollinearity among covariates

If a model includes two or more covariates which are strongly correlated with each other, it is difficult to disentangle their effects. In this situation, it has often been noted that the estimated regression coefficients may change drastically according to what other variables are included in the model, which is disconcerting. In addition, the estimates of effect are likely to be made with low precision. Such multicollinearity among covariates may not be too much of a problem in a non-randomised study unless they cause the adjustment procedures to 'explode', producing infeasibly small or large estimates of parameters or standard errors. However, multicollinearity between a covariate and the treatment allocation will make it impossible to separate the independent effects of the covariate and of treatment on outcome.[156] The allocation relationships in non-randomised studies are rarely strong enough for this to be a major concern, and this is unlikely to be an explanation for the results that we observed.

Misclassification and measurement error

We have already noted that in order to be able to adjust for a confounding factor, it must be measured and included in the adjustment model. To do this it is important (a) that the real confounding factor is used, and not a surrogate or proxy variable, and (b) that it is measured without error. Covariate misclassification through the use of poor proxies, measurement error and within participant instability in covariates (e.g. because of circadian rhythms) all lead to underestimation of the effect of each covariate on outcome. This phenomenon, often known as regression dilution or attenuation,[157,158] is demonstrated using hypothetical data in *Table 34*. The first row presents data from a non-randomised study stratified according to the true values of two covariates, where the underlying treatment effect is OR = 0.5. Owing to the confounding effect of the first covariate, the observed result is biased: OR = 1.28 (*Table 34a*). Adjusting for the first covariate corrects for this bias (OR = 0.50). However, if the first covariate is assessed with measurement error, we could observe the categorisation observed in *Table 34(b)*. Here, adjustment for the same covariate has much less effect, the adjusted estimate being OR = 1.17. The impact of misclassification in this example is large, but is generated by a typical degree of misclassification corresponding to a kappa coefficient of 0.4, which is routinely interpreted as showing 'fair to moderate' agreement.

Methods have been developed to correct for regression dilution and are widely used throughout epidemiology.[158–160] However, they require the degree of misclassification to be estimated, either within the study or in additional reliability investigations, and their ability to correct for misclassification depends on the precision with which these estimates are made. When there are multiple sources of misclassification, correction for measurement error becomes complex.[160,161]

Where confounders cannot be corrected for misclassification, it is often assumed that adjustment using the covariate will adjust only

TABLE 34 Hypothetical example demonstrating the potential impact of adjusting for covariates when misclassifications are correlated

Observed unstratified results

	Dead	Alive	Total
T	165	201	366
C	143	223	366
Total	308	424	732

Unadjusted OR = 1.28

(a) Results stratified according to underlying distribution of covariates (unobserved)[a]

CV1 = 0, CV2 = 0	Dead	Alive	Total
T	6	24	30
C	100	200	300
Total	106	224	330

OR = 0.5

CV1 = 1, CV2 = 0	Dead	Alive	Total
T	30	30	60
C	40	20	60
Total	70	50	120

OR = 0.5

CV1 = 0, CV2 = 1	Dead	Alive	Total
T	6	24	30
C	1	2	3
Total	7	26	33

OR = 0.5

CV1 = 1, CV2 = 1	Dead	Alive	Total
T	123	123	246
C	2	1	3
Total	125	124	249

OR = 0.5

True underlying OR = 0.50, OR adjusted only for CV1 (observed without misclassification) = 0.50

(b) Results stratified according to covariates with misclassification[b] in CV1[a]

CV1 = 0, CV2 = 0	Dead	Alive	Total
T	13	26	39
C	82	146	228
Total	95	172	267

OR = 0.89

CV1 = 1, CV2 = 0	Dead	Alive	Total
T	23	28	51
C	58	74	132
Total	81	102	183

OR = 1.05

CV1 = 0, CV2 = 1	Dead	Alive	Total
T	41	54	95
C	1	2	3
Total	42	56	98

OR = 1.52

CV1 = 1, CV2 = 1	Dead	Alive	Total
T	88	93	181
C	2	1	3
Total	90	94	184

OR = 0.47

OR adjusted only for CV1 (observed with misclassification) = 1.17

(c) Results stratified according to covariates with additional misclassification[c] in CV2 for those misclassified on CV1[a]

CV1 = 0, CV2 = 0	Dead	Alive	Total
T	50	63	113
C	83	146	229
Total	133	209	342

OR = 1.40

CV1 = 1, CV2 = 0	Dead	Alive	Total
T	21	21	42
C	28	14	42
Total	49	35	84

OR = 0.5

CV1 = 0, CV2 = 1	Dead	Alive	Total
T	4	17	21
C	1	1	2
Total	5	18	23

OR = 0.24

CV1 = 1, CV2 = 1	Dead	Alive	Total
T	90	100	190
C	31	62	93
Total	121	162	283

OR = 1.80

OR additionally adjusted for CV2 (observed with correlated misclassification) = 1.32

[a] Adjusted OR obtained from Mantel–Haenszel estimators.
[b] Misclassification in CV1: 30% of CV1=0 misclassified as CV1=1, 30% of CV1=1 misclassified as CV1=0.
[c] Misclassification in CV2: CV2 always classified as CV2=1 if CV1 misclassified as CV1=1; CV2 always classified as CV2=0 if CV1 misclassified as CV1=0.
CV, covariate; CV1, first covariate; CV2, second covariate.

partially for the effect of the covariate. The logical corollary of this assumption is that it is always worthwhile to adjust for a covariate, however poorly it has been measured. However, this notion in turn depends critically on the assumption that misclassification in the covariate is itself unbiased.[157] Greenland and Robins[162] discussed scenarios for case-control studies where this assumption is relaxed. They considered differential misclassification in disease, exposure and a single covariate, and showed that in some misclassification scenarios it can actually be detrimental to adjust for the covariate – the unadjusted estimate may be closer to the underlying effect than the adjusted estimate. In these situations it is not a matter of adjustment being inefficient but rather that adjustment leads to totally incorrect conclusions. Such an error will be magnified by the additional credence often given to an adjusted analysis when reporting study results.

Greenland and Robins' scenarios hint at a mechanism that might explain our troubling fourth key result. However, their scenarios do not translate directly to consideration of non-randomised intervention studies: they considered measurement error in the exposure (treatment) and case–control status (outcome). They also only assessed the impact of measurement error in a single covariate, whereas case-mix adjustment involves many covariates, all of which could be subject to misclassification. We have extended the hypothetical example in *Table 34* to include stratification for a second covariate (CV2). In *Table 34(c)* we have introduced misclassification in CV2 but only for those misclassified for CV1. We observe that the OR adjusted for observed CV1 and CV2 (OR = 1.32) is now further from the underlying true value (OR = 0.5) than the unadjusted estimate (OR = 1.28). Hence such a mechanism could explain our fourth observation, that adjustment in some situations can increase bias in estimates of treatment effects.

If misclassification was solely a matter of measurement error, such correlated misclassifications would be rare. However, misclassification includes all processes that lead us to observe a value that differs from the true value of the confounder. For example, consider a study in which the true confounder is a person's average blood pressure. The blood pressure reading used in the analysis could be misclassified owing to measurement error, but it could also be misclassified owing to natural within-person variation in blood pressure related to the time of day when the measurement was made, or to a short period of stress. If a second variable was included in the analysis that was also influenced by circadian rhythm or stress, misclassification in the second variable will be correlated with misclassification in blood pressure.

If case-mix adjustment models are being influenced by correlated misclassification, it could be hypothesised that including more variables in an adjustment model will increase the chances of correlated misclassifications, and hence increase expected bias. Comparison of the three different logistic models across *Tables 24*, *25*, *27–29* and *32* potentially supports this hypothesis: in nearly all cases the logistic regression model including the most covariates (the 'full model') was more biased than the logistic regression model with the fewest covariates (backward stepwise with p-to-remove of 0.05).

Misclassification and measurement error in covariates has been cited as the possible root of many controversial associations detected through observational studies.[163] We are correspondingly concerned that measurement and misclassification errors may be largely responsible for our fourth result.

Differences between unconditional and conditional estimates

Covariates which are prognostic but balanced across groups are rarely routinely adjusted for in analyses of randomised controlled trials. Even when adjustments are made, the adjusted estimates are often ignored when results of trials are included in meta-analyses. We have followed this standard approach in calculating the results of the RCTs that we created from the IST and ECST data sets. These unadjusted estimates are known as unconditional or population average results.

Estimates of treatment effects where prognostic variables have been adjusted for are known as conditional estimates, being conditional on knowledge of the covariates included in the analysis. Our comparison of results of RCTs with adjusted results of non-randomised studies is therefore comparing unconditional estimates of treatment effects from RCTs with conditional estimates of treatment effects from non-randomised studies.

When using ORs, it has been noted that adjustment for prognostic covariates leads to differences between unconditional and conditional estimates of treatment effects, even for covariates that are balanced across treatment groups.[164,165]

TABLE 35 *Hypothetical example demonstrating the potential impact of not adjusting for a balanced prognostic covariate in an RCT*

Crude analysis

	Dead	Alive	Total
T	140	60	200
C	60	140	200

Unconditional estimate of OR = 5.4

Adjusted analysis

CV = 0

	Dead	Alive	Total
T	90	10	100
C	50	50	100

Stratum-specific estimate of OR = 9

CV = 1

	Dead	Alive	Total
T	50	50	100
C	10	90	100

Stratum-specific estimate of OR = 9

Conditional estimate of OR = 9

Adapted from Gail, 1984.[164]

Consider the trial in *Table 35*. The unconditional estimate of the treatment effect is OR = 5.4. The lower half of the table shows the results of the same trial stratified by a prognostic covariate that is perfectly balanced across treatment groups. The estimate of the treatment effect in each strata is OR = 9. Thus the estimate of the treatment effect conditional on knowledge of the covariate is OR = 9. It can be deduced that if there were a further balanced prognostic covariate to adjust for, the result would change further, always moving further from the null effect value of OR = 1.[164]

This conditional result would be obtained through adjustment using both logistic regression and stratification. However, the propensity score for participants in the hypothetical trial is 0.5 regardless of their covariate value, and therefore the estimates of the treatment effect using propensity score methods will be OR = 5.4 – the unconditional estimate. Propensity scores methods only make adjustments for covariates that are not balanced across treatment groups.

Hence the difference between unconditional and conditional results is one possible explanation of the differences observed between RCT results and the results of the logistic regression adjusted analyses of non-randomised studies, and also between the results of adjustment using logistic regression and adjustment using propensity score methods.

Comparison of methods

Stratification

Stratification is best used to adjust for a single covariate. When stratification is used for several covariates, the strata become numerous and so small in size that many of the cells contain only treated participants or control participants, or participants all of whom have the same outcome state. In these situations the strata do not contribute to the analysis, and the data from those participants are effectively discarded. Even so, when bias relates to a single ordinal covariate, stratification can yield the best adjustment (as was seen in *Table 30*) as stratification estimates a separate parameter for each category, avoiding specifying a trend across categories to be either linear or monotonic. As selection bias rarely relates to a single variable, stratification will either be an inefficient (if multiple covariates are stratified) or an inadequate (if only one covariate is stratified) method for adjusting for differences in case-mix in non-randomised studies.

Logistic regression

In clinical trials, covariate adjustment is often recommended as a method of improving the precision of an estimate of treatment effect, even if there is no overt imbalance between the groups. This result, however, is particular to the use of linear regression and continuous outcome measures. Robinson and Jewell have shown that logistic regression always leads to a loss of

precision.[151] Their theoretical finding explains the increased variability of adjusted results that we observed with all applications of logistic regression, which we have interpreted as increased unpredictability in bias. However, unlike the increased variability observed with historically and concurrently controlled non-randomised studies in Chapter 6, the standard errors of the adjusted estimates are also inflated, such that the extra increased variability does not further increase spurious statistical significance rates.

One dilemma in all regression models is the process by which covariates are selected for adjustment. Many texts discuss the importance of combining clinical judgement and empirical methods to ensure that the models select and code variables in ways that have clinical face validity. There are three strategies that are commonly used in health care research to achieve this, described below.

Recently there has been a trend to include all scientifically relevant variables in the model, irrespective of their contribution to the model.[166] The rationale for this approach is to provide as complete control of confounding as possible within the given data set. This idea is based on the fact that it is possible for individual variables not to exhibit strong confounding, but when taken collectively considerable confounding can be present in the data. One major problem with this approach is that the model may be overfitted and produce numerically unstable estimates. However, as we have observed, a more important problem may be the increased risk of including covariates with correlated misclassification errors.

The stepwise approaches to selecting covariates are often criticised for using statistical significance to assess the adequacy of a model rather than judging the need to control for specific factors on the basis of the extent of confounding involved, and in using sequential statistical testing, known to lead to bias.[167] Research based on simulations has found that stepwise selection strategies which use higher p-values (0.15–0.20) are more likely to correctly select confounding factors than those which use a p-value of 0.05.[168,169] In our evaluations, little practical difference was observed between these two stepwise strategies.

A pragmatic strategy for deciding which estimates to adjust for involves undertaking unadjusted and adjusted analyses and using the results of the adjusted analysis when they differ from those of the unadjusted analysis. This is based on an argument that if the adjustment for a covariate does not alter the treatment effect the covariate is unlikely to be important.[141] An extension of this argument is used to determine when all necessary confounders have been included in the model, suggesting that confounders should keep being added to a model so long as the adjusted effect keeps changing (e.g. by at least 10%). The assumed rationale for this strategy sometimes misleads analysts to reach the unjustified conclusion that when estimates become stable all important confounders have been adjusted for, such that the adjusted estimate of the treatment effect is unbiased. We did not attempt to automate this variable selection approach in our evaluations.

Propensity score methods

Propensity score methods are not widely used in healthcare research, and are difficult to undertake owing to the lack of suitable software routines. However, there may be benefits of the propensity score approach over traditional approaches in making adjustments in non-randomised studies. Whilst Rosenbaum and Rubin showed that for bias introduced through a single covariate the propensity score approach is equivalent to direct adjustment through the covariate,[146] our analyses have shown that when there are multiple covariates the propensity score method may in fact be superior as it does not increase variability in the estimates. In addition, propensity score methods give unconditional (or population average) estimates of treatment effects, which are more comparable to typical analyses of RCTs. Simulation studies have also shown that propensity scores are less biased than direct adjustment methods when the relationship of covariates is misspecified.[170]

The impact of misclassification and measurement error on propensity score methods appears not to have been studied. It is unclear whether these problems can explain the occasional overcorrection of propensity score methods that we observed. Also, our implementation of the propensity score method did not include interaction terms in the estimation of propensity scores, as is sometimes recommended.[147] It would be interesting to evaluate whether including additional terms would have improved the performance of the model.

Conclusions

The problems of underadjustment for confounding are well recognised. However, in a non-randomised study it is not possible to assess

directly the likely degree of residual confounding that may be present, and therefore we cannot gauge how biased adjusted results may still be. By comparing with results based on randomisation, our investigations suggest that the degree of underadjustment may be large. Indeed, our results may in fact be overoptimistic, as the covariate data used were recorded in a standard way according to trial protocols, and were complete for all participants. In many non-randomised studies measurement methods are not standardised. Also, covariate data are incomplete (especially in retrospective studies), leading to bias if the observations are not missing at random.

Our two greatest concerns are the potential increase in bias that could occur as a result of the existence of correlated misclassification of covariates, and the differences between conditional and unconditional estimates. Correlated misclassification is a problem inherent to the data, and cannot be adjusted for. It is very difficult to know the degree of misclassification and error in a variable, and impossible to know whether the variable being used is the 'true' confounder or just a proxy. These findings question the appropriateness of the strategy of including data on all available potential confounders when adjusting for case-mix, which has been the starting point of many risk-adjustment methods used throughout healthcare.

However, the same findings could be explained by the peculiar differences between unconditional and conditional estimates of treatment effects observed when results are expressed as ORs, although this mechanism only applied to estimates obtained from logistic regression and stratification methods.

The finding of high levels of residual confounding and the detrimental effect of adjustment were seen in both historically controlled studies, known to be prone to systematic bias, and in concurrently controlled studies, more prone to unpredictability in bias. The relationships were also noted in studies mimicking allocation by indication.

It is important to find out whether such destructive relationships between covariates are common. We have examined data from only two clinical situations, but in both we observed results that undermine the use of case-mix adjustment. Also in the IST, case-mix adjustment was found to be detrimental in eight of the 14 regions.

There appears to be a small potential benefit of using propensity score methods over logistic regression for case-mix adjustment in terms of the consistency of estimates of treatment effects. While logistic regression always increased the range of observed treatment effects, propensity score methods did not. This finding may indicate a greater role for propensity score methods in healthcare research, although in the particular applications investigated neither approach performed adequately.

For those critically appraising non-randomised studies, the recommendation to assess whether "investigators demonstrate similarity in all known determinants of outcome"[138,139] has not been universally supported by our empirical investigations. The second recommendation, to assess whether they "adjust for these differences in analysis" is also not supported empirically. Our analyses suggest that there are considerable complexities in assessing whether a case-mix adjustment analysis will increase or decrease bias.

These findings may have a major impact on the certainty which we assign to many effects in healthcare which have been made on the basis of using risk adjustment methods.

Chapter 8

Discussion and conclusions

Chapters 3–7 have reported results from five separate evaluations concerning non-randomised studies. The results have been discussed in detail in each chapter. We summarise their main findings below.

Summary of key findings

Our review of previous empirical investigations of the importance of randomisation (Chapter 3) identified eight studies that fulfilled our inclusion criteria. Each investigation reported multiple comparisons of results of randomised and non-randomised studies. Although there was overlap in the comparisons included in these reviews, they reached different conclusions concerning the likely validity of non-randomised data, mainly reflecting weaknesses in the meta-epidemiological methodology that they all used, most notably that it was not able to account for confounding factors in the comparisons between randomised and non-randomised studies, nor to detect anything other than systematic bias.

We identified 194 tools that could be used to assess the quality of non-randomised studies (Chapter 4). Overall the tools were poorly developed: the majority did not provide a means of assessing the internal validity of non-randomised studies and almost no attention was paid to the principles of scale development and evaluation. However, 14 tools were identified that included items related to each of our pre-specified core internal validity criteria, which related to assessment of allocation method, attempts to achieve comparability by design, identification of important prognostic factors and adjustment of differences in case-mix. Six of the 14 tools were considered potentially suitable for use as quality assessment tools in systematic reviews, but all require some modification to meet all of our pre-specified criteria.

Of 511 systematic reviews we identified that included non-randomised studies, only 169 (33%) assessed study quality, and only 46% of these reported the results of the quality assessment for each study (Chapter 5). This is lower than the rate of quality assessment in systematic reviews of randomised controlled trials.[131] Among those that did assess study quality, a wide variety of quality assessment tools were used, some of which were designed only for use in evaluating RCTs, and many were designed by the review authors themselves. Most reviews (88%) did not assess key quality criteria of particular importance for the assessment of non-randomised studies. Sixty-nine reviews (41%) investigated the impact of quality on study results in a quantitative manner. The results of these analyses showed no consistent pattern in the way that study quality relates to treatment effects, and were confounded by the inclusion of a variety of study designs and studies of variable quality.

A unique 'resampling' method was used to generate multiple unconfounded comparisons between RCTs and historically controlled and concurrently controlled studies (Chapter 6). These empirical investigations identified two characteristics of the bias introduced by using non-random allocation. First, the use of historical controls can lead to systematic over- or underestimations of treatment effects, the direction of the bias depending on time trends in the case-mix of participants recruited to the study. In the studies used for the analyses, these time trends varied between study regions, and were therefore difficult to predict. Second, the results of both study designs varied beyond what was expected from chance. In a very large sample of studies the biases causing the increased unpredictability on average cancelled each other out, but in individual studies the bias could be fairly large, and could act in either direction. These biases again relate to differences in case-mix, but the differences are neither systematic nor predictable.

Four commonly used methods of dealing with variations in case-mix were identified: (i) discarding comparisons between groups which differ in their baseline characteristics, (ii) regression modelling, (iii) propensity score methods and (iv) stratified analyses (Chapter 7). The methods were applied to the historically and concurrently controlled studies generated in Chapter 6, and also to studies designed to mimic 'allocation by indication'. None of the methods successfully removed bias in

historical and concurrent cohort studies. Logistic regression in fact tended to increase the bias. Propensity score methods performed slightly better than other methods, but did not yield satisfactory adjustments in most situations. Detailed investigation revealed that adequate adjustment for selection bias could only be made when selection depends on a single prognostic factor that is measured and included in the adjustment model. Although apparent under-adjustment could be explained by omission of important confounding factors, the observation that adjustment could also increase bias required different explanations. Of possible explanations identified, we considered the most likely to be the difference between unconditional and conditional estimates of ORs and the inclusion of confounders in the adjustment models that have correlated mismeasurement or misclassification errors, and the differences between conditional and unconditional estimates.

Discussion

> "Scientific evidence is commonly and properly greeted with objections, scepticism and doubt. Responsible scientists are responsibly sceptical. We look for failures of observation, gaps in reasoning and alternative explanations. This scepticism is itself scrutinised. Scepticism must itself be justified, defended. One needs 'grounds for doubt'."
>
> From Rosenbaum[145]

Non-randomised studies are widely used throughout healthcare to evaluate the intended effects of healthcare interventions. They have been included in many systematic reviews and, on occasion, used as the sole basis for healthcare decisions and policy. While they are widely perceived to have greater 'external validity' than RCTs,[25] their internal validity is questionable, principally owing to problems of selection bias, although they often have weaknesses in other areas. These weaknesses lead many to be sceptical about the validity of their results. In this project we have attempted to evaluate the degree to which this scepticism is justified, through reviews of existing evidence and through new empirical investigations.

Two studies recently published in the *New England Journal of Medicine* (*NEJM*)[32,33] challenged the existing evidence base that has accumulated about the validity of non-randomised studies for assessing the intended effects of healthcare interventions. These two studies and, to a large extent, the research of the five other groups reviewed in Chapter 3, attempted to answer the simple question, 'are non-randomised studies biased?'. Between them, these reviews have accumulated many instances where the results of randomised and non-randomised studies of the same intervention are on average the same, but also many examples where they differ. Although not the conclusion of all the authors, the one conclusion that fits with all previous research and is supported by our new empirical investigations (Chapter 6) is that non-randomised studies are **sometimes** but **not always** biased. Having established that bias is a likely possibility, it is of more importance to understand (a) what causes bias in non-randomised studies, (b) how often and how badly biased are the results from non-randomised studies, (c) whether the presence of bias can in any way be predicted and (d) what the implications are for those producing, using and reviewing non-randomised studies.

What are the causes of bias in non-randomised studies?

The principal difference between randomised and non-randomised studies can be characterised as a difference in their susceptibility to selection bias, that is, a bias which acts such that participants selected to receive one intervention are in some way different from those selected to receive the alternative intervention. Concealed randomisation specifically removes the possibility of selection bias, the only differences between the outcome of different groups being attributable to chance or to the intervention, all else being equal. In non-randomised studies, allocation to groups depends on other factors, sometimes known, sometimes unknown. When these factors are also related to outcome, bias will be introduced.

There are other issues that commonly lead to bias in non-randomised studies. For example, numbers of exclusions in non-randomised studies are frequently unclear, treatment and outcome assessment are rarely conducted according to standardised protocols and outcomes may not be assessed blind. None of these issues are insurmountable in a prospective non-randomised study, but while the threat of increased selection bias can be reduced, it can never be removed.

How often and how biased are non-randomised studies?

Our resampling studies have demonstrated that the biases associated with historically and concurrently controlled non-randomised studies are large enough to impact on the conclusions of a systematic review. Some 50% of the concurrently controlled non-randomised studies sampled from the IST met standard criteria for statistical

significance compared with only 9% of comparable RCTs. Taking a significant result as leading to a conclusion of effectiveness, selection bias in these concurrently controlled studies was strong enough for 41% to reach unjustified conclusions. Notably, 21% of these studies wrongly concluded that aspirin was harmful. Similarly, 40% of the ECST historically controlled studies were statistically significant compared with 15% of RCTs, implying that the use of historical comparisons introduced bias severe enough to lead to unjustified conclusions in favour of surgery in 25% of the studies. This high frequency of 'conclusion-changing' bias raises concern.

However, are these frequencies and magnitudes of bias likely to be typical? In many clinical situations the variation in case-mix is large, as, for example, in the treatment of those who have suffered a stroke, who, as seen in the IST trial, vary in their levels of consciousness, paralysis and many other prognostic factors. Here the possibility of producing groups with large differences in case-mix (resulting in serious bias) is high, such that the results of non-randomised comparisons are unlikely to be reliable. In other situations there is little prognostic information available at the time of treatment allocation, so that the possibility of introducing case-mix related biases might be lower. For example, in childhood vaccination trials in the developed world there is little possibility of predicting the occurrence of infectious disease at the time of vaccine administration. It could therefore be hypothesised that non-randomised studies in such contexts would be less likely to be biased. These issues require further research.

Can we predict bias?

One issue that is clearly important when trying to predict bias is the quality of the study design. Whilst there is hardly any empirical evidence about the importance of particular dimensions of study quality for non-randomised studies, the importance of evaluating the quality of RCTs for the likelihood of performance, attrition and detection bias has been well established, and there is good empirical evidence of bias associated with certain methodological features.[23] Despite the lack of empirical evidence specific to non-randomised studies, it is likely that they are equally if not more susceptible to these biases than RCTs.

It has been suggested that to assess the likelihood of selection bias in a non-randomised study, it is necessary to know the methods by which participants were allocated to intervention groups.[18] Specifically, in order to assess the degree of selection bias we need to know the factors influencing the allocation, the manner in which those factors are related to outcome and how the factors were dealt with in the analysis. These requirements underlie our choice of the four core items we propose for inclusion in quality assessment tools for non-randomised studies (Chapter 4), although we found that they were rarely included in the tools that we assessed.

In classical non-randomised designs, allocation is often made according to centre (concurrent controlled cohort studies) or period (historically controlled cohort studies). However, knowing that there is a difference in location or era of the treated and control samples is inadequate to judge comparability: it is necessary to know the manner in which these differences lead to specific differences in case-mix.

The situation is more complex when interventions being compared have been used across centres during the same period, where allocation is by indication or patient preference. Then it is not usually possible to describe fully the factors influencing allocation, nor is it likely that the allocation mechanism will be consistent across patients and clinicians. In addition, patients may be included in a study who are suitable for only one of the two treatments. These factors all mean that allocation may be strongly linked to prognostic factors, such that the bias related to differences in case-mix will be large.[3]

Theoretically, the second stage in assessing the likelihood of bias in a non-randomised study is to consider whether differences in confounding factors are comparable between the groups and, if not, whether they have been adjusted for in the analysis. Many investigators routinely claim effects as 'adjusted' or 'independent' or 'unbiased' when they have used statistical methods to correct for differences in case-mix. However, their trust in the observed comparability of case-mix and the ability of statistical adjustment methods to adjust properly for differences in case-mix is unfounded. The unique design of our investigation has allowed the performance of these case-mix adjustment methods to be assessed by comparing unadjusted and adjusted results with those from randomised comparisons. We have only undertaken these evaluations in two data sets and we found consistent results indicating that case-mix adjustment cannot reliably compensate for biases in selection, and in fact it may introduce additional bias.

For example, the biases observed in our empirical investigations based on cohort designs could not be removed using case-mix adjustment methods, even though we could describe the process by which allocation occurred (Chapter 7). This was despite having at our disposal complete data on several highly prognostic variables. Our principal failure resulted from being unable to describe how allocation was linked to prognostic factors in a manner that could be adjusted for in a statistical analysis. Moses[18] stated that there are three requirements for successful case-mix adjustment: (1) knowledge of which variables must be taken into account, (2) measuring those variables on each participant and (3) using those measurements appropriately to adjust the treatment comparison. He suggested that researchers are likely to fail at all three. Although it is well known that failure to identify all relevant variables will lead to under-adjustment (residual confounding), our evaluations have suggested that failing to measure the correct variables properly and failing to model their effects appropriately can lead to spurious conclusions.

The use of case-mix adjustment analyses should therefore not be regarded as a guarantee that a study is unbiased, or that the observed differences in case-mix have been adjusted for. Investigators using case-mix adjustment methodologies, especially logistic regression models, should draw conclusions cautiously, and probably regard adjusted analyses as exploratory rather than confirmatory. Our observations require further assessment and confirmation from investigations in additional data sets to explore further the mechanisms that are causing these failures. Nonetheless, they raise serious concern about the routine promotion and use of case-mix adjustment methods for the analysis of non-randomised studies. We do not propose that case-mix adjustment be abandoned, but recommend that greater scepticism be applied in the interpretation of results.

What are the implications for those producing, reviewing and using non-randomised studies?

An investigator planning to undertake a non-randomised study should first make certain that an RCT cannot be undertaken.[18] The advantages of concealed randomised allocation are so great compared with the inadequacies that we have quantified for non-randomised studies that it should be with the greatest reluctance that an investigator concludes that randomisation is not possible. The ability to eradicate bias at the design stage is crucial to establishing the validity of a study. In particular, investigators should not assume that statistical methods can be used reliably to compensate for biases introduced through suboptimal allocation methods.

A prospective non-randomised study should be undertaken according to a protocol that is carefully followed to ensure consistent inclusion criteria, that all relevant factors are measured accurately for each participant and that participants are all monitored in a standard manner and blinded to treatment if possible. In some situations it may even be possible to match prospectively treated and control patients on important prognostic factors.[171] Byar pointed out that the one piece of information required for successful adjustment is knowledge of the reasons why each patient received their particular treatment.[16] This information is usually not recorded in non-randomised studies, and how it should be measured is a real challenge worthy of research.[18] As a minimum, recording details of reasons for allocations of particular interventions should allow the subset of patients to be identified who are not considered suitable for both treatments, and should not be included in analyses. In addition, all prognostic variables should be measured in such a way as to allow measurement error, misclassification and within-person variability to be assessed. Some authors have attempted to acquire reliable risk assessments by retrospective case-note review,[172] but obtaining these data prospectively has many attractions. Retrospective studies cannot use consistent inclusion criteria, or ensure that data are complete and consistently recorded. Phillips and Davey Smith have pointed out that it is probably more worthwhile to put effort into undertaking a small observational study to a high standard than in obtaining poor quality and inadequate data on a large number of participants.[173] However, despite all these efforts, the possibility of residual confounding will not be removed.

Several authors have suggested additional strategies that investigators might consider as a way of identifying hidden bias. Rosenbaum proposed the routine use of extra control groups and the inclusion of additional outcomes that are known not to be altered by the treatment as further checks on comparability.[145,152] Several authors have emphasised the use of sensitivity analyses to quantify the strength of confounding in a missing

covariate that would be needed to nullify an observed effect.[145,174] However, none of these methods can correct for bias.

Investigators undertaking systematic reviews of effectiveness should include the results of non-randomised studies with discretion. If results from good-quality RCTs are available, there seems to be no justification for additionally considering non-randomised studies. If non-randomised studies are to be included in a review, their quality needs to be carefully assessed to evaluate the likelihood of bias. Quality assessment should pay particular attention to the description of the allocation mechanisms and the demonstration of there being comparability in all important prognostic factors at baseline. Although we have not identified a quality assessment tool that met all our requirements, the six best tools could all be adapted for use in a systematic review. The conclusions of a review should take into account the extra uncertainty associated with the results of non-randomised studies.

The results of non-randomised studies should be treated with a healthy degree of scepticism. Healthcare decision-makers should be cautious not to over-interpret results from non-randomised studies. Importantly, checking that treated and control groups appear comparable does not guarantee freedom from bias, and it should never be assumed that case-mix adjustment methods can fully correct for observed differences between groups. The uncertainty in the result of a non-randomised study is not properly summarised in the confidence interval for the overall effect: our analyses have shown that the true uncertainty in the results of non-randomised studies may be 10 times greater.

Conclusions

Non-randomised studies are sometimes but not always biased:

- The results of non-randomised studies can differ from the results of RCTs of the same intervention.
- All other issues remaining equal, lack of randomisation introduces bias into the assessment of treatment effects. The bias may have two components. It may be systematic and appear on average to act in a particular direction if the non-random allocation mechanism leads to a consistent difference in case-mix. However, if the allocation mechanism can lead to haphazard differences in case-mix, the bias can act in either direction, increasing uncertainty in outcome in ways that cannot be predicted. The extent of systematic bias and increased uncertainty varies according to the type of non-randomised comparison and clinical context.
- Meta-epidemiological techniques tend not to provide useful information on the sources of and degrees of bias in non-randomised studies owing to the existence of meta-confounding and lack of systematic or predictable bias in the results of non-randomised studies.

Statistical methods of analysis cannot properly correct for inadequacies of study design:

- Case-mix comparability and standard methods of case-mix adjustment do not guarantee the removal of bias. Residual confounding may be high even when good prognostic data are available, and in some situations adjusted results may appear more biased than unadjusted results.

Systematic reviews of effectiveness often do not adequately assess the quality of non-randomised studies:

- Quality assessment has not routinely been undertaken in systematic reviews of effectiveness that include non-randomised studies. When study quality has been investigated, there is variability in the tools and quality criteria that have been used and there has been no consistent pattern between quality and review findings. Not all reviews that have assessed quality have considered the findings when synthesising study results and drawing conclusions.
- Although many quality assessment tools exist and have been used for appraising non-randomised studies, few are suitable for this task as they omit key domains of quality assessment (assessment of allocation mechanisms, attempts to achieve case-mix comparability by design, identification of confounding factors, use of case-mix adjustment methods). Few were developed using appropriate methodological procedures. Fourteen tools were identified which appear to have reasonable coverage of core domains of internal validity, six of which were considered potentially suitable for use in systematic reviews of non-randomised studies. All six would require modification to cover adequately the key issues in non-randomised studies (of identifying prognostic factors and accounting for them in the design and analysis).

Non-randomised studies provide a poor basis for treatment or health policy decisions:

- The inability of case-mix adjustment methods to compensate for selection bias and our inability to identify non-randomised studies which are free of selection bias indicate that non-randomised studies should only be used in situations where RCTs cannot be undertaken.
- Healthcare policies based upon non-randomised studies or systematic reviews of non-randomised studies may need re-evaluation if the uncertainty in the true evidence base was not fully appreciated when the decisions were made.

Recommendations for further research

1. The resampling methodology that we have employed in this project should be applied in other clinical areas where suitable RCTs exist. These evaluations should consider (a) the distribution of biases associated with non-randomised allocation, (b) whether non-randomised studies with similar baseline characteristics are less biased and (c) the performance of case-mix adjustment methods. It would be valuable to study different contexts to evaluate the degree to which bias is related to the amount of prognostic information known at allocation.
2. Efforts should be focused on the development of a new quality assessment tool for non-randomised studies or the refinement and development of existing tools. Appropriate methodological procedures of tool development should be employed and key indicators of internal validity covered. These indicators include both those for which empirical evidence is available from work on RCTs and those supported by our empirical investigations. The latter should include the method used to allocate participants to groups; specification of the factors that influenced these allocations; the way in which these factors are thought to relate to outcome; and appropriate adjustment in the analysis. In the meantime, systematic reviewers should be strongly encouraged to use and adapt those tools that do cover key quality issues.
3. Research should be undertaken to develop methods of measuring and characterising reasons for treatment choices in patient preference and allocation by indication studies, and evaluations undertaken to assess whether recording such information allows effective adjustment for selection bias.
4. Empirical work is needed to investigate how quality assessments of non-randomised studies should be incorporated in the synthesis of studies in a systematic review and to study the implications of individual quality features for the interpretation of review results.
5. Reasons for the failure of case-mix adjustment methods should be further investigated, including assessment of the generalisability of our results to risk assessments and epidemiological studies where they are frequently utilised. The impact of differences between unconditional and conditional estimates of treatment effects should be assessed.
6. Guidelines should be produced to advise investigators on the best ways of undertaking prospective non-randomised studies to minimise bias.
7. The role of propensity scoring in adjusting for selection bias should be further evaluated, and computer macros made available for its application.

Recommendations for those producing and using health technology assessments

1. Systematic reviewers and those conducting health technology assessments should be strongly encouraged to base any estimates of effectiveness or cost-effectiveness on evidence from RCTs. Where such evidence is unavailable, the uncertainty inherent in estimates based on non-randomised evidence should be strongly emphasised.
2. Decision-makers should review healthcare policies based on the results of non-randomised studies to assess whether the inherent uncertainty in the evidence base was properly appreciated when the policy decisions were made.
3. Agencies funding primary research should fund non-randomised studies only when they are convinced that a randomised study is not feasible.

Acknowledgements

We would like to thank the following people for their help with this project: Peter Sandercock and Peter Rothwell for organising access to the IST and ECST data sets respectively; Jane Harrison for carrying out the extensive set of searches and Kath Wright for help with search queries; Alison Brown and Vanda Castle for secretarial support. We would also like to acknowledge the helpful comments of the members of our expert panel: Professor Brian Haynes, Professor Nick Black and Dr Barnaby Reeves.

References

1. D'Agostino RB, Kwan H. Measuring effectiveness: what to expect without a randomized control group. *Med Care* 1995;**33**:95–105.
2. Abel U, Koch A. The role of randomization in clinical studies: myths and beliefs. *J Clin Epidemiol* 1999;**52**:487–97.
3. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Stat Med* 1984;**3**:361–73.
4. Chalmers I. Assembling comparison groups to assess the effects of health care. *J R Soc Med* 1997;**90**:379–86.
5. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;**312**:1215–18.
6. Altman DG, Schulz KF. Statistics notes – concealing treatment allocation in randomised trials. *BMJ* 2001;**323**:446–7.
7. Feinstein AR, Horwitz RI. Problems in the "evidence" of "evidence-based medicine". *Am J Med* 1997;**103**:529–35.
8. Bradley C. Designing medical and educational intervention studies: a review of some alternatives to conventional randomized controlled trials. *Diabetes Care* 1993;**16**:509–18.
9. Torgerson D, Klaber-Moffett J, Russell I. Patient preferences in randomised trials: threat or opportunity. *J Health Serv Res Policy* 1996;**1**:194–7.
10. Bradford Hill A. A short textbook of medical statistics. 10th ed. London: Hodder & Stoughton, 1977.
11. Miettinen OS. Theoretical epidemiology: principles of occurrence research in medicine. New York: Wiley, 1985.
12. Greenland S, Neutra R. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;**9**:361–7.
13. Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. *Am J Epidemiol* 1999;**149**:981–3.
14. Walker AM. Confounding by indication. *Epidemiology* 1996;**7**:335–6.
15. Horwitz RI, Feinstein AR. The application of therapeutic trial principles to improve the design of epidemiologic research. *J Chron Dis* 1981;**34**:575–83.
16. Byar DP. Problems with using observational data bases to compare treatments. *Stat Med* 1991;**10**:663–6.
17. Horwitz RI, Viscoli CM, Clemens JD, Sadock RT. Developing improved observational methods for evaluating therapeutic effectiveness. *Am J Med* 1990;**89**:630–8.
18. Moses LE. Measuring effects without randomized trials? Options, problems, challenges. *Med Care* 1995;**33**:AS8–AS14.
19. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, *et al*. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;**134**:663–94.
20. 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.
21. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, *et al*. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;**352**:609–13.
22. Sterne JAC, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;**21**:1513–24.
23. Jüni P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. In Egger M, Davey Smith G, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2001. pp. 87–108.
24. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.
25. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;**2**(13).
26. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AMS. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;**4**(34).

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

28. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;**317**:1185–90.

29. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials (Cochrane Methodology Review). In The Cochrane Library, Issue 4. Oxford: Update Software; 2002.

30. Chalmers TC, Matta RJ, Smith HJ. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1982;**72**:233–40.

31. Diehl LF, Perry DJ. A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid? *J Clin Oncol* 1986;**4**:1114–20.

32. Benson K, Hartz A. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342**:1878–86.

33. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**:1887–92.

34. Ioannidis JPA, Haidich A, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, *et al*. Comparison of evidence of treatment effects in randomized and non randomized studies. *JAMA* 2001;**286**:821–30.

35. Lipsey MW, Wilson DB. The efficacy of psychological educational, and behavioural treatments: confirmation from meta-analysis. *Am Psychol* 1993;**48**:1181–209.

36. Wilson DB, Lipsey MW. The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychol Methods* 2001;**6**:413–29.

37. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: surgical. *Stat Med* 1989;**8**:455–66.

38. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: medical. *Stat Med* 1989;**8**:441–54.

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

40. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *N Engl J Med* 2000;**342**:1907–9.

41. Kunz R, Khan KS, Nuemayer H, Sacks HS, Liu P-Y, Anderson G, *et al*. Observational studies and randomized trials. *N Engl J Med* 2000; **343**:1194–7.

42. Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Educ Prev* 1997; **9**:15–21.

43. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2001;**135**:982–9.

44. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995;**16**:62–73.

45. Cooper H. The integrative research review: a systematic approach. Newbury Park, CA: Sage; 1984.

46. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;**17**:1–12.

47. Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, Fergusson D, *et al*. Assessing the quality of randomized trials: reliability of the Jadad scale. *Control Clin Trials* 1999;**20**:448–52.

48. Wortman PM. Judging research quality. In Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994. pp. 97–109.

49. Campbell DT, Stanley JC, editors. Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally; 1966.

50. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Chicago, IL: Rand McNally; 1979.

51. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, *et al*. A method for assessing the quality of a randomized control trial. *Control Clin Trials* 1981;**2**:31–49.

52. Bangert-Drowns RL, Wells-Parker E, Chevillard I. Assessing the methodological quality of research in narrative reviews and meta-analyses. In Bryant KJ, Windle M, editors. The science of prevention: methodological advances from alcohol and substance abuse research. Washington, DC: American Psychological Association; 1997. pp. 405–29.

53. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. *Int J Technol Assess Health Care* 1996;**12**:195–208.

54. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;**323**:42.

55. Schulz KF, Chalmers I, Altman DG, Grimes DA, Dore CJ. The methodologic quality of randomization as assessed from reports of trials in specialist and general medical journals. *Online J Curr Clin Trials* 1995;Aug 26:Doc No 197.

56. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;**309**:1358–61.

57. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 1995.

58. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;**140**:290–6.

59. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;**282**:1054–60.

60. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. *Health Technol Assess* 1999; **3**(9).

61. Morris JJ, Smith R, Glower DD, Muhlbaier LH, Reves JG, Wechsler AS, *et al*. Clinical evaluation of single versus multiple mammary artery bypass. *Circulation* 1990;**82**:214–23.

62. Nichol K, Baken L, Wuorenma J, Nelson A. The health and economic benefits associated with pneumococcal vaccination of elderly persons with chronic lung disease. *Arch Intern Med* 1999;**159**:2437–42.

63. Walker JD, Dodds RA, Murrells TJ, Bending JJ, Mattock MB, Keen H. Restriction of dietary protein and progression of renal failure in diabetic nephropathy. *Lancet* 1989;**ii**:1411–14.

64. Critical Appraisal Skills Programme. 12 questions to help you make sense of a cohort study. Oxford: Critical Appraisal Skills Programme; 1999.

65. Thomas H. Quality assessment tool for quantitative studies. Effective Public Health Practice Project. McMaster University, Toronto.

66. Wells G, Shay B. Data extraction for non-randomised systematic reviews. University of Ottawa, Ottawa.

67. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Knipschild PG. Balneotherapy and quality assessment: interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. *J Clin Epidemiol* 1998; **51**:335–41.

68. ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria-based meta-analysis. *J Clin Epidemiol* 1990;**43**:1191–9.

69. Kleijnen J, Knipschild P, Ter Riet G. Clinical trials of homeopathy. *BMJ* 1991;**302**:316–32.

70. Koes BW, Assendelft WJ, van der Heijden GJ, Bouter LM, Knipschild PG. Spinal manipulation and mobilisation for back and neck pain: a blinded review. *BMJ* 1991;**303**:1298–303.

71. van der Windt DA, van der Heijden GJ, Scholten RJ, Koes BW, Bouter LM. The efficacy of non-steroidal anti-inflammatory drugs for shoulder complaints. *J Clin Epidemiol* 1995;**48**:691–704.

72. de Vet HC, de Bie RA, van der Heijden GJ, Verhagen AP, Sijpkes P, Knipschild PG. Systematic reviews on the basis of methodologic criteria. *Physiotherapy* 1997;**83**:284–9.

73. Bours GJ, Ketelaars CA, Frederiks CM, Abu Saad HH, Wouters EF. The effects of aftercare on chronic patients and frail elderly patients when discharged from hospital: a systematic review. *J Adv Nurs* 1998;**27**:1076–86.

74. Kleijnen J, Riet ter G, Knipschild P. Vitamine C en verkoudheid. Overzicht van een megadosis literatuur. *Ned Tijdschr Geneeskd* 1989;**133**:1535.

75. Riet ter G, Kleijnen J, Knipschild P. De meta-analyse als review-methode [De effectiviteit van acupunctuur]. *Huisarts en Wetenschap* 1989;**32**: 176–81.

76. Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized control trials in dental research 1. Methods. *J Periodontal Res* 1986; **21**:305–14.

77. Nicolucci A, Grilli R, Alexanian A, Apolone G, Torri V, Liberati A. Quality, evolution, and clinical implications of randomized, controlled trials on the treatment of lung cancer. *JAMA* 1989; **262**:2101–7.

78. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-based Medicine Working Group. *JAMA* 1994;**271**:59–63.

79. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-based Medicine Working Group. *JAMA* 1993;**270**:2598–601.

80. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre. How to read clinical journals. V. To distinguish useful from useless or even harmful therapy. *CMAJ* 1981;**124**:1156–62.

81. Anders JF, Jacobson RM, Poland GA, Jacobsen SJ, Wollan PC. Secondary failure rates of measles vaccines: a meta-analysis of published studies. *Pediatr Infect Dis J* 1996;**15**:62–6.

82. Baker DW, Jones R, Hodges J, Massie BM, Konstam MA, Rose EA. Management of heart failure. III. The role of revascularization in the treatment of patients with moderate or severe left ventricular systolic dysfunction. *JAMA* 1994;**272**:1528–34.

83. Cameron I, Crotty M, Currie C, Finnegan T, Gillespie L, Gillespie W, *et al*. Geriatric rehabilitation following fractures in older people: a systematic review. *Health Technol Assess* 2000;**4**(2).

84. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA* 1994;**272**:101–4.

85. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Commun Health* 1998;**52**:377–84.

86. Zaza S, Wright-de Aguero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, *et al*. Data collection instrument and procedure for systematic reviews in the 'Guide to Community Preventive Services'. *Am J Prev Med* 2000;**18**:44–74.

87. Miller WR, Brown JM, Simpson TL. What works? A methodological analysis of the alcohol treatment outcome literature. In Hester RK, Miller WR, editors. Handbook of alcoholism treatment approaches: effective alternatives. Boston, MA: Allyn & Bacon; 1995. pp. 12–44.

88. Bass JL, Christoffel KK, Widome M, Boyle W, Scheidt P, Stanwick R, *et al*. Childhood injury prevention counseling in primary care settings: a critical review of the literature. *Pediatrics* 1993; **92**:544–50.

89. Duffy ME. A research appraisal checklist for evaluating nursing research reports. *Nurs Health Care* 1985;**6**:538–47.

90. Shay CB, Zimmermann WS. The factorial validity of a rating scale for the evaluation of research articles. *Educ Psychol Meas* 1972;**32**:453–7.

91. Wells-Parker E, Bangert-Drowns R. Meta-analysis of research on DUI remedial interventions. *Drugs Alcohol Driving* 1990;**6**:147–60.

92. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta analysis. *J Clin Epidemiol* 1992;**45**:255–65.

93. Melchart D, Linde K, Worku F, Bauer R, Wagner H. Immunomodulation with Echinacea – a systematic review of controlled clinical trials. *Phytomedicine* 1994;**1**:245–54.

94. Standards of Reporting Trials Group. A proposal for structured reporting of randomised controlled trials. *JAMA* 1994;**272**:1926–31.

95. Linde K, Scholz M, Ramirez G, Clausius N, Melchart D, Jonas WB. Impact of study quality on outcome in placebo-controlled trials of homeopathy. *J Clin Epidemiol* 1999;**52**:631–6.

96. de Oliveira IR, Dardennes RM, Amorim ES, Diquet B, de Sena EP, Moreira EC, *et al*. Is there a relationship between antipsychotic blood levels and their clinical efficacy? An analysis of studies design and methodology. *Fundam Clin Pharmacol* 1995;**9**:488–502.

97. Coleridge Smith P. The management of chronic venous disorders of the leg: an evidence-based report of an International Task Force. *Phlebology* 1999;**14**:3–19.

98. Dawson-Saunders B, Trapp R. Reading the medical literature. Basic and clinical biostatistics. Norwalk, CT: Appleton & Lange; 1990. pp. 264–76.

99. DuRant RH. Checklist for the evaluation of research articles. *J Adolesc Health* 1994;**15**:4–8.

100. Friedenreich CM. Methods for pooled analyses of epidemiologic studies. *Epidemiology* 1993; **4**:295–302.

101. Gordis L, Kleinman JC, Klerman LV, Mullen PD, Paneth N. Criteria for evaluating evidence regarding the effectiveness of prenatal interventions. In Merkatz IR, Thompson JE, editors. New perspectives on prenatal care. New York: Elsevier; 1990. pp. 31–8.

102. Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol* 1996;**49**:749–54.

103. Wilson A, Henry DA. Meta-analysis. Part 2. Assessing the quality of published meta-analyses. *Med J Aust* 1992;**156**:173–87.

104. Bracken MB. Reporting observational studies. *Br J Obstet Gynaecol* 1989;**96**:383–8.

105. Spitzer WO, Lawrence V, Dales R, Hill G, Archer MC, Clarck P, *et al*. Links between passive smoking and disease; a best-evidence synthesis. *Clin Invest Med* 1990;**13**:17–42.

106. Talley NJ, Nyren O, Drossman DA. The irritable bowel syndrome: toward optimal design of controlled treatment trials. *Gastroenterology Int* 1993;**6**:189–211.

107. Fowkes FG, Fulton PM. Critical appraisal of published research. Introductory guidelines. *BMJ* 1991;**302**:1136–40.

108. Weintraub M. How to critically assess clinical drug trials. *Drug Ther* 1982;**12**:131–48.

109. Cowley DE. Prostheses for primary total hip replacement. A critical appraisal of the literature. *Int J Technol Assess Health Care* 1995;**11**:770–8.

110. Vickers A. Critical appraisal: how to read a clinical research paper. *Complement Ther Med* 1995;**3**:158–66.

111. Reisch J, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. *Pediatrics* 1989;**84**:815–27.

112. Downs, Black N. Systematic review of the literature on the effectiveness of surgery for stress incontinence for women. London: London School of Hygiene and Tropical Medicine; 1996.

113. Taggart DP, D'Amico R, Altman DG. Effect of arterial revascularisation on survival: a systematic review of studies comparing bilateral and single internal mammary arteries. *Lancet* 2001; **358**:870–5.

114. Tyson JE, Furzan JA, Reisch JS, Mize SG. An evaluation of the quality of therapeutic studies in perinatal medicine. *Obstet Gynecol* 1983;**62**:99–102.

115. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet* 2000;**356**:1228–31.

116. Khan KS, Daya S, Jadad A. The importance of quality of primary studies in producing unbiased systematic reviews. *Arch Intern Med* 1996;**156**:661–6.

117. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, *et al*. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;**283**:2008–12.

118. Clarke M, Oxman AD, editors. Cochrane Reviewers Handbook 4.1.5 [updated April 2002]. In The Cochrane Library, Issue 2. Oxford: Update Software; 2002.

119. NHS Centre for Reviews and Dissemination. Undertaking systematic reviews of research on effectiveness. CRD Report No. 4. York: University of York, 1996.

120. L'Abbe KA, Detsky AS, O'Rourke KF. Meta-analysis in clinical research. *Ann Intern Med* 1987; **107**:224–33.

121. Schmid JE, Koch GC, LaVange LM. An overview of statistical issues and methods of meta-analysis. *J Biopharm Stat* 1991;**1**:103–20.

122. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol* 1995; **48**:9–18.

123. Ganong LH. Integrative reviews of nursing research. *Res Nurs Health* 1987;**10**:1–11.

124. Gartland JJ. Orthopaedic clinical research: deficiencies in experimental design and determinations of outcome. *J Bone Joint Surg Am* 1988;**70**:309–15.

125. Smith TMF, Sugden RA. Sampling and assignment mechanisms in experiments, surveys and observational studies. *Int Stat Rev* 1988;**56**:165–80.

126. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1989;**95**:2s–4s.

127. Gyorkos TW, Tannenbaum TN, Abrahamowicz M, Oxman AD, Scott EA, Millson ME, *et al*. An approach to the development of practice guidelines for community health interventions. *Can J Public Health* 1994;**85**:s8–s13.

128. Oakley A, Fullerton D, Holland J, Arnold S, Hickey D, Kelley P, *et al*. SSRU database project: reviews of effectiveness: HIV prevention and sexual health education interventions. London: University of London, Social Science Research Unit; 1994.

129. Wells-Parker E, Bangert-Drowns R, McMillen R, Williams M. Final results from a meta-analysis of remedial interventions with drink/drive offenders. *Addiction* 1995;**90**:907–26.

130. Cheek CM, Black NA, Devlin HB, Kingsnorth AN, Taylor RS, Watkin DF. Groin hernia surgery: a systematic review. *Ann R Coll Surg Engl* 1998; **80**:S1–S80.

131. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, *et al*. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess* 1999;**3**(12).

132. International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with ischaemic stroke. *Lancet* 1997;**349**:1569–81.

133. European Carotid Surgery Trialists' Collaborative Group. Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European Cartoid Surgery Trial (ECST). *Lancet* 1998;**351**:1379–87.

134. Efron B, Tibshirani RJ. An introduction to the bootstrap. London: Chapman & Hall; 1993.

135. North American Symptomatic Carotid Endarterectomy Trial (NASCET) investigators. Clinical alert: benefit of carotid endarterectomy for patients with high-grade stenosis of the internal carotid artery. National Institute of Neurological Disorders and Stroke, Stroke and Trauma Division. *Stroke* 1991;**22**:816–17.

136. Iezzoni LI. Risk adjustment for measurement healthcare outcomes. 2nd ed. Chicago, IL: Health Administration Press; 1997.

137. Sowden AJ, Deeks J, Sheldon TA. Volume and outcome in coronary artery bypass graft surgery: true relationship or artefact? *BMJ* 1995;**311**:151–5.

138. Naylor CD, Guyatt GH. Therapy and harm – outcomes of health services. In Guyatt GH, Rennie D, editors. Users' guides to the medical literature – a manual for evidence-based clinical practice. Chicago, IL: American Medical Association Press; 2002. pp. 233–46.

139. Levine M, Haslam D, Walter S, Cumming R, Lee H, Haines T, *et al*. Harm. In Guyatt GH, Rennie D, editors. Users' guides to the medical literature – a manual for evidence-based clinical

practice. Chicago, IL: American Medical Association Press; 2002. pp. 84–100.

140. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991.

141. Breslow NE, Day NE, editors. Statistical methods in cancer research. Analysis of case control studies. Lyon: IARC Scientific Publications; 1980.

142. Breslow NE, Day NE. Statistical methods in cancer research. Design and analysis of cohort studies. Lyon: IARC Scientific Publications; 1987.

143. Altman DG. Adjustment for covariate imbalance. In Armitage P, Colton T, editors. Encyclopedia of biostatistics. Chichester: Wiley; 1998. pp. 1005–7.

144. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;**22**:719–48.

145. Rosenbaum PR. Observational studies. New York: Springer; 1995.

146. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrica* 1983;**70**:41–55.

147. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**:2265–81.

148. Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel–Haenszel odds ratio. *Am J Epidemiol* 1986;**124**:719–23.

149. Rothman KJ. Modern epidemiology. Boston, MA: Little Brown; 1986.

150. Leon DA. Failed or misleading adjustment for confounding. *Lancet* 1993;**342**:479–81.

151. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic-regression models. *Int Stat Rev* 1991;**59**:227–40.

152. Rosenbaum PR. Discussing hidden bias in observational studies. *Ann Intern Med* 1991; **115**:901–5.

153. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;**24**:295–313.

154. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. *Epidemiology* 1997;**8**:429–34.

155. Brenner H. A potential pitfall in control of covariates in epidemiologic studies. *Epidemiology* 1997;**9**:68–71.

156. Davey Smith G, Phillips A. Declaring independence: why we should be cautious. *J Epidemiol Community Health* 1990;**44**:257–8.

157. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980;**112**:564–9.

158. Caroll RJ. Covariance analysis in generalized linear measurement error models. *Stat Med* 1989;**8**:1075–93.

159. Savitz DA, Baron AE. Estimating and correcting for confounder misclassification. *Am J Epidemiol* 1989;**129**:1062–71.

160. Knuiman MW, Divitini ML, Buzas JS, Fitzgerald PEB. Adjustment for regression dilution in epidemiological regression analyses. *Ann Epidemiol* 1998;**8**:56–63.

161. Rosner B, Spiegelman D, Willet WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol* 1990;**132**:734–45.

162. Greenland S, Robins JM. Confounding and misclassification. *Am J Epidemiol* 1985; **122**:495–506.

163. Prentice RL. Design issues in cohort studies. *Stat Methods Med Res* 1995;**4**:273–92.

164. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;**71**:431–44.

165. Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Stat Med* 2002;**21**:2899–908.

166. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley; 1989.

167. Bancroft TA. Analysis and inference for incompletely specified models involving the use of preliminary tests of significance. *Biometrics* 1964;**20**:247–439.

168. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989;**129**:125–37.

169. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol* 1993;**138**:923–36.

170. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993;**49**:1231–6.

171. Charpentier PA, Bogardus ST, Inouye SK. An algorithm for prospective individual matching in a non-randomized clinical trial. *J Clin Epidemiol* 2001;**54**:1166–73.

172. Johnston SC. Identifying confounding by indication through blinded prospective review. *Am J Epidemiol* 2001;**154**:276–84.

173. Phillips AN, Davey Smith G. The design of prospective epidemiological studies: more subjects or better measurements? *J Clin Epidemiol* 1993; **46**:1203–11.

174. Psaty BM, Koepsell TD, Lin D, Weiss NS, Siscovick DS, Rosendaal FR, *et al*. Assessment and control for confounding by indication in observational studies. *J Am Geriatr Soc* 1999; **47**:749–54.

175. Khan KS, ter Riet, G, Glanville J, Sowden AJ, Kleijnen J. Undertaking systematic reviews of research on effectiveness. CRD Report No. 4. Update. York: University of York; 2001.

176. Audet N, Gagnon R, Ladouceur R, Marcil M. How effective is the teaching of critical analysis of scientific publications? Review of studies and their methodological quality. *CMAJ* 1993;**148**:945–52.

177. Campos-Outcalt D, Senf J, Watkins AJ, Bastacky S. The effects of medical school curricula, faculty role models, and biomedical research support on choice of generalist physician careers: a review and quality assessment of the literature. *Acad Med* 1995;**70**:611–19.

178. Carter J, Verhoef MJ. Efficacy of self-help and alternative treatments of premenstrual syndrome. *Womens Health Issues* 1994;**4**:130–7.

179. Cochrane Musculoskeletal Injuries Group. Assessment of methodological quality of included trials. In The Cochrane Library, Issue 4. Oxford: Update Software; 2002.

180. Cuddy PG, Elenbaas RM, Elenbaas JK. Evaluating the medical literature. Part I: abstract, introduction, methods. *Ann Emerg Med* 1983;**12**:549–55.

181. Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. *BMJ* 1986;**292**:810–2.

182. Glantz JC, McNanley TJ. Active management of labor: a meta-analysis of cesarean delivery rates for dystocia in nulliparas. *Obstet Gynecol Surv* 1997; **52**:497–505.

183. Greenhalgh T. Assessing the methodological quality of published papers. *BMJ* 1997;**315**:305–8.

184. Gurman A, Kniskern D. Research on marital and family therapy: progress, perspective and prospect. In Garfield S, Bergin A, editors. Handbook of psychotherapy and behavior change: an empirical analysis. New York: Wiley; 1978. pp. 817–901.

185. Cole MG, Primeau F, McCusker J. Effectiveness of interventions to prevent delirium in hospitalized patients: A systematic review. *CMAJ* 1996; **155**:1263–8.

186. Kay EJ, Locker D. Is dental health education effective? A systematic review of current evidence. *Community Dent Oral Epidemiol* 1996;**24**:231–5.

187. Kreulen CM, Creugers NH, Meijering AC. Meta-analysis of anterior veneer restorations in clinical studies. *J Dent* 1998;**26**:345–53.

188. Kwakkel G, Wagenaar RC, Koelman TW, Lankhorst GJ, Koetsier JC. Effects of intensity of rehabilitation after stroke. A research synthesis. *Stroke* 1997;**28**:1550–6.

189. Lee TM, Chan CC, Paterson JG, Janzen HL, Blashko CA. Spectral properties of phototherapy for seasonal affective disorder: a meta-analysis. *Acta Psychiatr Scand* 1997;**96**:117–21.

190. Levine J. Trial Assessment Procedure Scale (TAPS). Bethesda, MD: Department of Health and Human Services, Public Health Service, Alcohol, Drug Abuse and Mental Health Administration, National Institute of Mental Health; 1980.

191. MacMillan HL, MacMillan JH, Offord DR, Griffith DL, MacMillan A. Primary prevention of child physical abuse and neglect: a critical review. Part 1. *J Child Psychol Psychiatry* 1994;**35**:835–56.

192. Maziak DE, Meade MO, Todd TR. The timing of tracheotomy: a systematic review. *Chest* 1998; **114**:605–9.

193. Meijman F, Melker de RA. The extent of inter- and intra-reviewer agreement on the classification and assessment of designs of single-practice research. *Fam Pract* 1995;**12**:93–7.

194. Barley ZA. Assessment of quality of studies for inclusion in meta-analyses. Doctoral dissertation, University of Colorado; 1989. *Dissertation Abstracts International* 1989;**50A**:575.

195. Moncrieff J, Drummond DC. The quality of alcohol treatment research: an examination of influential controlled trials and development of a quality rating system. *Addiction* 1998;**93**:811–23.

196. Morley JA, Finney JW, Monahan SC, Floyd AS. Alcoholism treatment outcome studies, 1980–1992: methodological characteristics and quality. *Addict Behav* 1996;**21**:429–43.

197. Mulrow CD, Lichtenstein MJ. Blood glucose and diabetic retinopathy: a critical appraisal of new evidence. *J Gen Intern Med* 1986;**1**:73–7.

198. Salisbury C. What is the impact of different models of care on patients' quality of life, psychological well-being or motivation? Appropriate and cost effective models of service delivery in palliative care: report March 1996–July 1997. Bristol: Division of Primary Health Care, University of Bristol; 1997.

199. Schechter MT, Leblanc FE, Lawrence VA. Critical appraisal of published research. In Mulder D, McPeek B, Troidl H, Spitzer W, McKneally M, Weschler A, editors. Principles and practice of research: strategy for surgical investigators. 2nd ed. New York: Springer; 1991. pp. 81–7.

200. Sheldon TA, Song F, Davey Smith G. Critical appraisal of the medical literature: how to assess whether health-care interventions do more harm than good. In Drummond MF, Maynard A, Wells N, editors. Purchasing and providing cost

effective health care. London: Churchill Livingstone; 1993. pp. 31–48.

201. Vickers AJ. Can acupuncture have specific effects on health? A systematic review of acupuncture antiemesis trials. *J R Soc Med* 1996;**89**:303–11.

202. Wingood GM, DiClemente RJ. HIV sexual risk reduction interventions for women: a review. *Am J Prev Med* 1996;**12**:209–17.

203. Wright J, Dye R. Systematic review on obstructive sleep apnoea: its effect on health and benefit of treatment. Leeds: University of Leeds; 1995. pp. 1–60.

204. Hayes C, Antzcak-Bouckoms A, Burdick E. Quality assessment and meta-analysis of systemic tetracycline use in chronic adult periodontitis. *J Clin Periodontol* 1992;**19**:164–8.

205. Norman GR, Shannon SI. Effectiveness of instruction in critical appraisal (evidence-based medicine) skills: a critical appraisal. *CMAJ* 1998; **158**:177–81.

206. Avis M. Reading research critically. II. An introduction to appraisal: assessing the evidence. *J Clin Nurs* 1994;**3**:271–7.

207. Bailar JC, Louis TA, Lavori PW, Polansky M. Studies without internal controls. *N Engl J Med* 1984;**311**:156–62.

208. Bauman LJ, Drotar D, Leventhal JM, Perrin EC, Pless IB. A review of psychosocial interventions for children with chronic health conditions. *Pediatrics* 1997;**100**:244–51.

209. Berard A, Bravo G, Gauthier P. Meta-analysis of the effectiveness of physical activity for the prevention of bone loss in postmenopausal women. *Osteoporos Int* 1997;**7**:331–7.

210. Beyer WE, Palache AM, Osterhaus AD. Comparison of serology and reactogenicity between influenza subunit vaccines and whole virus or split vaccines. A review and meta-analysis of the literature. *Clin Drug Invest* 1998;**15**:1–12.

211. Bland JM, Jones DR, Bennett S, Cook DG, Haines AP, MacFarlane AJ. Is the clinical trial evidence about new drugs statistically adequate? *Br J Clin Pharmacol* 1985;**19**:155–60.

212. Boers M, Ramsden M. Longacting drug combinations in rheumatoid arthritis: a formal overview. *J Rheumatol* 1991;**18**:316–24.

213. Borghouts JAJ, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. *Pain* 1998;**77**:1–13.

214. Brown SA. Measurement of quality of primary studies for meta-analysis. *Nurs Res* 1991;**40**:352–5.

215. Brown SA. Studies of educational interventions and outcomes in diabetic adults: a meta-analysis revisited. *Patient Educ Couns* 1990;**16**:189–215.

216. Brown SA, Grimes DE. A meta-analysis of nurse practitioners and nurse midwives in primary care. *Nurs Res* 1995;**44**:332–9.

217. Brown SA, Upchurch S, Anding R, Winter M, Ramirez G. Promoting weight loss in type II diabetes. *Diabetes Care* 1996;**19**:613–24.

218. Callahan CM, Dittus RS, Katz BP. Oral corticosteroid therapy for patients with stable chronic obstructive pulmonary disease: a meta-analysis. *Ann Intern Med* 1991;**114**:216–23.

219. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *CMAJ* 1979;**121**:1193–254.

220. Canadian Task Force on the Periodic Health Examination. Methodology. In Canadian Task Force on the Periodic Health Examination. The Canadian Guide to Clinical Preventive Health Care. Ottawa: Health Canada; 1994. p. xxxvii.

221. Watt D, Verma S, Flynn L. Wellness programs: a review of the evidence. *CMAJ* 1998;**158**:224–30.

222. Catre MG. Anticoagulation in spinal surgery. A critical review of the literature. *Can J Surg* 1997;**40**:413–19.

223. Loblaw DA, Laperriere NJ. Emergency treatment of malignant extradural spinal cord compression: an evidence-based guideline. *J Clin Oncol* 1998; **16**:1613–24.

224. Richard CS, McLeod RS. Follow-up of patients after resection for colorectal cancer: a position paper of the Canadian Society of Surgical Oncology and the Canadian Society of Colon and Rectal Surgeons. *Can J Surg* 1997;**40**:90–100.

225. Griffiths AM, Sherman PM. Colonoscopic surveillance for cancer in ulcerative colitis: a critical review. *J Pediatr Gastroenterol Nutr* 1997; **24**:202–10.

226. Cao WC, Van der Ploeg CP, Plaisier AP, van der Sluijs IJ, Habbema JD. Ivermectin for the chemotherapy of bancroftian filariasis: a meta-analysis of the effect of single treatment. *Trop Med Int Health* 1997;**2**:393–403.

227. Carlson M, Fanchiang SP, Zemke R, Clark F. A meta-analysis of the effectiveness of occupational therapy for older persons. *Am J Occup Ther* 1996; **50**:89–98.

228. Ball C, Sackett D, Phillips B, Haynes B, Straus S. Oxford Centre for Evidence-based Medicine Levels of Evidence (May 2001). http://minerva.minervation.com/cebm/docs/levels.html#levels. 2002.

229. Gam AN, Johannsen F. Ultrasound therapy in musculoskeletal disorders: a meta-analysis. *Pain* 1995;**63**:85–91.

230. Cheatham ML, Chapman WC, Key SP, Sawyers JL. A meta-analysis of selective versus routine

nasogastric decompression after elective laparotomy. *Ann Surg* 1995;**221**:469–76.

231. Ciliska D, Hayward S, Thomas H, Mitchell A, Dobbins M, Underwood J, *et al*. A systematic overview of the effectiveness of home visiting as a delivery strategy for public health nursing interventions. *Can J Public Health* 1996;**87**:193–8.

232. Ploeg J, Ciliska D, Dobbins M, Hayward S, Thomas H, Underwood J. A systematic overview of adolescent suicide prevention programs. *Can J Public Health* 1996;**87**:319–24.

233. Cochrane Effective Practice and Organisation of Care Group. Assessment of methodological quality. The Cochrane Library, Issue 4. Oxford: Update Software; 2002.

234. Oakley A, Fullerton D, Holland J. Behavioural interventions for HIV/AIDS prevention. *AIDS* 1995;**9**:479–86.

235. Towler B, Irwig L, Glasziou P, Kewenter J, Weller D, Silagy C. A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, Hemoccult. *BMJ* 1998; **317**:559–65.

236. Cohen JE, Goel V, Frank JW, Bombardier C, Peloso P, Guillemin F. Group education interventions for people with low back pain: an overview of the literature. *Spine* 1994;**19**:1214–22.

237. Colditz GA, Brewer TF, Berkey CS, Wilson ME, Burdick E, Fineberg HV, *et al*. Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *JAMA* 1994; **271**:698–702.

238. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1992; **102**:305s–11s.

239. Schmidt-Nowara W, Lowe A, Wiegland L, Cartwright R, Perez-Guerra F, Menn S. Oral appliances for the treatment of snoring and obstructive sleep apnoea. *Sleep* 1995;**18**:501–10.

240. Chaulk CP, Kazandjian VA. Directly observed therapy for treatment completion of pulmonary tuberculosis: Consensus Statement of the Public Health Tuberculosis Guidelines Panel. *JAMA* 1998;**279**:943–8.

241. Mayo-Smith MF. Pharmacological management of alcohol withdrawal. A meta-analysis and evidence-based practice guideline. American Society of Addiction Medicine Working Group on Pharmacological Management of Alcohol Withdrawal. *JAMA* 1997;**278**:144–51.

242. Cox GL, Merkel WT. A qualitative review of psychosocial treatments for bulimia. *J Nerv Ment Dis* 1989;**177**:77–84.

243. Cuijpers P. Psychological outreach programmes for the depressed elderly: a meta-analysis of effects and dropout. *Int J Geriatr Psychiatry* 1998;**13**:41–8.

244. de Craen AJ, Di Giulio G, Lampe Schoenmaeckers JE, Kessels AG, Kleijnen J. Analgesic efficacy and safety of paracetamol–codeine combinations versus paracetamol alone: a systematic review. *BMJ* 1996; **313**:321–5.

245. de Kruif YP, van Wegen EE. Pelvic floor muscle exercise therapy with myofeedback for women with stress urinary incontinence: a meta-analysis. *Physiotherapy* 1996;**82**:107–13.

246. de Oliveira IR, de Sena EP, Pereira EL, Miranda AM, de Oliveira NF, Ribeiro MG, *et al*. Haloperidol blood levels and clinical outcome: a meta-analysis of studies relevant to testing the therapeutic window hypothesis. *J Clin Pharm Ther* 1996;**21**:229–36.

247. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med* 1982;**306**:1332–7.

248. Curatolo M, Petersen-Felix S, Scaramozzino P, Zbinden AM. Epidural fentanyl, adrenaline and clonidine as adjuvants to local anaesthetics for surgical analgesia: meta-analyses of analgesia and side-effects. *Acta Anaesthesiol Scand* 1998;**42**:910–20.

249. Aronson R, Offman HJ, Joffe RT, Naylor CD. Triiodothyronine augmentation in the treatment of refractory depression. A meta-analysis. *Arch Gen Psychiatry* 1996;**53**:842–8.

250. Devine EC. Effects of psychoeducational care for adult surgical patients: a meta-analysis of 191 studies. *Patient Educ Couns* 1992;**19**:127–42.

251. Devine EC, Cook TD. A meta-analytic analysis of effects of psycho-educational interventions on length of post-surgical hospital stay. *Nurs Res* 1983;**32**:267–74.

252. Devine EC, Reifschneider E. A meta-analysis of the effects of psychoeducational care in adults with hypertension. *Nurs Res* 1995;**44**:237–45.

253. Devine EC, Westlake SK. The effects of psychoeducational care provided to adults with cancer: meta-analysis of 116 studies. *Oncol Nurs Forum* 1995;**22**:1369–81.

254. Brown SA. Effects of educational interventions in diabetes care: a meta-analysis of findings. *Nurs Res* 1988;**37**:223–30.

255. Durlak JA, Wells AM. Primary prevention mental health programs: the future is exciting. *Am J Community Psychol* 1997;**25**:233–43.

256. Elvik R. The safety value of guardrails and crash cushions: a meta-analysis of evidence from evaluation studies. *Accid Anal Prev* 1995;**27**:523–49.

257. Esdaile JM, Horwitz RI. Observational studies of cause-effect relationships: an analysis of methodological problems as illustrated by the

conflicting data for the role of oral contraceptives in the etiology of rheumatoid arthritis. *J Chron Dis* 1986;**39**:841–52.

258. Farquhar WW, Krumboltz JD. A checklist for evaluating experimental research in psychology and education. *J Educ Res* 1959;**52**:353–4.

259. Fox JH. Criteria of good research. *Phi Delta Kappan* 1958;**39**:284–6.

260. Collingsworth S, Gould D, Wainwright SP. Patient self-administration of medication: a review of the literature. *Int J Nurs Stud* 1997;**34**:256–69.

261. Gansevoort RT, Sluiter WJ, Hemmelder MH, de Zeeuw D, de Jong PE. Antiproteinuric effect of blood-pressure-lowering agents: a meta-analysis of comparative trials. *Nephrol Dial Transplant* 1995; **10**:1963–74.

262. Garber BG, Hebert PC, Yelle JD, Hodder RV, McGowan J. Adult respiratory distress syndrome: a systematic overview of incidence and risk factors. *Crit Care Med* 1996;**24**:687–95.

263. Garg A, Shiau J, Guyatt G. The ineffectiveness of immunosuppressive therapy in lymphocytic myocarditis: an overview. *Ann Intern Med* 1998; **128**:317–22.

264. Callahan CM, Drake BG, Heck DA, Dittus RS. Patient outcomes following unicompartmental or bicompartmental knee arthroplasty: a meta-analysis. *J Arthroplasty* 1995;**10**:141–50.

265. Gifford RH, Feinstein AR. A critique of methodology in studies of anticoagulant therapy for acute myocardial infarction. *N Engl J Med* 1969;**280**:351–7.

266. Good M. Effects of relaxation and music on postoperative pain: a review. *J Adv Nurs* 1996; **24**:905–14.

267. Fiscella K. Does prenatal care improve birth outcomes? A critical review. *Obstet Gynecol* 1995; **85**:468–79.

268. Gray MA, McPherson IG. Agoraphobia: a critical review of methodology in behavioural treatment research. *Curr Psychol Rev* 1982;**2**:19–45.

269. Gyorkos TW, Tannenbaum TN, Abrahamowicz M, Bedard L, Carsley J, Franco ED, *et al*. Evaluation of the effectiveness of immunization delivery methods. *Can J Public Health* 1994;**85**:S14–30.

270. Riben PD, Mathias RG, Campbell E, Wiens M. The evaluation of the effectiveness of routine restaurant inspections and education of food handlers: critical appraisal of the literature. *Can J Public Health* 1994;**85**:S56–60.

271. Oxman AD, Scott EA, Sellors JW, Clarke JH, Millson ME, Rasooly I, *et al*. Partner notification for sexually transmitted diseases: an overview of the evidence. *Can J Public Health* 1994;**85**:S41–7.

272. Hancock L, Sanson Fisher RW, Redman S, Burton R, Burton L, Butler J, *et al*. Community action for health promotion: a review of methods and outcomes 1990–1995. *Am J Prev Med* 1997; **13**:229–39.

273. Haynes RB, Sackett DL, Taylor DW, Hackett BC, Luterbach E, Cloak JR. An annotated bibliography (including notes on methodologic standards for compliance research). In Sackett DL, Haynes RB, editors. Compliance with therapeutic regimens. Baltimore, MD: Johns Hopkins University Press; 1976. pp. 193–7.

274. Mullen PD, Mains DA, Velez R. A meta-analysis of controlled trials of cardiac patient education. *Patient Educ Couns* 1992;**19**:143–56.

275. Mullen PD, Green LW, Persinger G. Clinical trials of patient education for chronic conditions: a comparative meta-analysis of intervention types. *Prev Med* 1985;**14**:753–81.

276. Heacock H, Koehoorn M, Tan J. Applying epidemiological principles to ergonomics: a checklist for incorporating sound design and interpretation of studies. *Appl Ergon* 1997;**28**:165–72.

277. Hedges LV. Issues in meta-analysis. *Review of Research in Education* 1986;**13**:353–98.

278. Helfenstein U, Steiner M. Fluoride varnishes (Duraphat): a meta-analysis. *Comm Dent Oral Epidemiol* 1994;**22**:1–5.

279. Heneghan AM, Horwitz SM, Leventhal JM. Evaluating intensive family preservation programs: a methodological review. *Pediatrics* 1996; **97**:535–42.

280. Hill MJ, Blane HT. Evaluation of psychotherapy with alcoholics: a critical review. *Q J Stud Alcohol* 1967;**28**:76–104.

281. Hines DC, Goldziehr JW. Clinical investigation: a guide to its evaluation. *Am J Obstet Gynecol* 1969; **105**:450–87.

282. Hoag M, Burlingame G. Evaluating the effectiveness of child and adolescent group treatment: a meta-analytic review. *J Clin Child Psychol* 1997;**26**:234–46.

283. Hoffman RM, Wheeler KJ, Deyo RA. Surgery for herniated lumbar discs: a literature synthesis. *J Gen Intern Med* 1993;**8**:487–96.

284. Howell WH, McNamara DJ, Tosca MA, Smith BT, Gaines JA. Plasma lipid and lipoprotein responses to dietary fat and cholesterol: a meta-analysis. *Am J Clin Nutr* 1997;**65**:1747–64.

285. Jabbour M, Osmond MH, Klassen TP. Life support courses: are they effective? *Ann Emerg Med* 1996; **28**:690–8.

286. Lucassen PL, Assendelft WJ, Gubbels JW, van Eijk JT, van Geldrop WJ, Neven AK. Effectiveness of

treatments for infantile colic: systematic review. *BMJ* 1998;**316**:1563–9.

287. Ernst E, Pittler MH. The effectiveness of acupuncture in treating acute dental pain: a systematic review. *Br Dent J* 1998;**184**:443–7.

288. Saint S, Elmore JG, Sullivan SD, Emerson SS, Koepsell TD. The efficacy of silver alloy-coated urinary catheters in preventing urinary tract infection: a meta-analysis. *Am J Med* 1998; **105**:236–41.

289. Johnston ME, Langton KB, Haynes RB, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. A critical appraisal of research. *Ann Intern Med* 1994;**120**:135–42.

290. Sullivan F, Mitchell E. Has general practitioner computing made a difference to patient care? A systematic review of published reports. *BMJ* 1995;**311**:848–52.

291. Kasiske BL, Ma JZ, Kalil RS, Louis TA. Effects of antihypertensive therapy on serum lipids. *Ann Intern Med* 1995;**122**:133–41.

292. Kasiske BL, Lakatua JD, Ma JZ, Louis TA. A meta-analysis of the effects of dietary protein restriction on the rate of decline in renal function. *Am J Kidney Dis* 1998;**31**:954–61.

293. Kasiske BL, Heimduthoy K, Ma JZ. Elective cyclosporine withdrawal after renal transplantation – a meta-analysis. *JAMA* 1993;**269**:395–400.

294. King PR. Methodological guidelines for reading drug evaluation research. *Can J Psychiatry* 1984;**29**:575–82.

295. King PR. Appraising the quality of drug-evaluation research: I. A method of meta-analysis for acute treatment medications. *Can J Psychiatry* 1990;**35**:316–9.

296. King PR. Appraising the quality of drug-evaluation research: II. A methodological review procedure for prophylactic treatment trials. *J Clin Psychopharmacol* 1990;**10**:22–6.

297. Kingery WS. A critical review of controlled clinical trials for peripheral neuropathic pain and complex regional pain syndromes. *Pain* 1997;**73**:123–39.

298. Kinney MR, Burfitt SN, Stullenbarger E, Rees B, DeBolt MR. Quality of life in cardiac patient research: a meta analysis. *Nurs Res* 1996;**45**:173–80.

299. Klassen T, Verhey FR, Sneijders GH, Rozendaal N, De Vet HC, Van Praag HM. Treatment of depression in Parkinson's disease: a meta-analysis. *J Neuropsychiatry Clin Neurosci* 1995;**7**:281–6.

300. Kleijnen J, Knipschild P. Mistletoe treatment for cancer. Review of controlled trials in humans. *Phytomedicine* 1994;**1**:255–60.

301. Kleijnen J, Knipschild P. Hyperbaric oxygen for multiple sclerosis. Review of controlled trials. *Acta Neurol Scand* 1995;**91**:330–4.

302. Ernst E, Barnes J. Are homeopathic remedies effective for delayed-onset muscle soreness? A systematic review of placebo-controlled trials. *Perfusion* 1998;**11**:4–8.

303. Lijesen GKS, Theeuwen I, Assendelft WJJ, Van der Wal G. The effect of human chorionic gonadotropin (hCG) in the treatment of obesity by means of the simeons therapy: a criteria-based meta-analysis. *Br J Clin Pharmacol* 1995;**40**:237–43.

304. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Knipschild PG. Taking baths: the efficacy of balneotherapy in patients with arthritis. A systematic review. *J Rheumatol* 1997;**24**:1964–71.

305. Kristensen TS. Cardiovascular diseases and the work environment. A critical review of the epidemiologic literature on nonchemical factors. *Scand J Work Environ Health* 1989;**15**:165–79.

306. Krause N, Dasinger LK, Neuhauser F. Modified work and return to work: a review of the literature. *J Occup Rehab* 1998;**8**:113–39.

307. Krogh CL. A checklist system for critical review of medical literature. *Med Educ* 1985;**19**:392–5.

308. Kulik CL, Kulik JA, Bangert-Drowns RL. Effectiveness of mastery learning programs: a meta-analysis. *Rev Educ Res* 1990;**60**:265–99.

309. Detsky AS, Baker JP, O'Rourke K, Goel V. Perioperative parenteral nutrition: a meta-analysis. *Ann Intern Med* 1987;**107**:195–203.

310. Labrecque M. Evaluation critique des publications médicales: mise à jour pratique. *Can Fam Physician* 1989;**35**:792–5.

311. Leboeuf C. A review of data reports published in the *Journal of Manipulative and Physiological Therapeutics* from 1986 to 1988. *J Manipulative Physiol Ther* 1990;**13**:89–95.

312. Lee C-M, Picard M, Blain M-D. A methodological and substantive review of intervention outcome studies for families undergoing divorce. *J Fam Psychol* 1994;**8**:3–15.

313. Cooper BC, Mullins PR, Jones MR, Lang SD. Clinical efficacy of roxithromycin in the treatment of adults with upper and lower respiratory tract infection due to *Haemophilus influenzae*. A meta-analysis of 12 clinical studies. *Drug Invest* 1994; **7**:299–314.

314. Linde K, Melchart D. Randomized controlled trials of individualized homeopathy: a state-of-the-art review. *J Altern Complement Med* 1998; **4**:371–88.

315. Lionel ND, Herxheimer A. Assessing reports of therapeutic trials. *BMJ* 1970;**3**:637–40.

316. Littenberg B, Weinstein LP, McCarren M, Mead T, Swiontkowski MF, Rudicel SA, *et al*. Closed fractures of the tibial shaft. A meta-analysis of three methods of treatment. *J Bone Joint Surg Am* 1998;**80**:174–83.

317. Longnecker MP, Berlin JA, Orza MJ, Chalmers TC. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 1988; **260**:652–6.

318. Luborsky L, Singer B, Luborsky L. Comparative studies of psychotherapies. Is it time that "everyone has won and all must have prizes?" *Arch Gen Psychiatry* 1975;**32**:995–1008.

319. Lyons LC, Woods PJ. The efficacy of rational-emotive therapy: a quantitative review of the outcome research. *Clin Psychol Rev* 1991; **11**:357–69.

320. Mahon WA, Daniel EE. A method for the assessment of reports of drug trials. *CMAJ* 1964; **90**:565–9.

321. Margetts BM, Thompson RL, Key T, Duffy S, Nelson M, Bingham S, *et al*. Development of a scoring system to judge the scientific quality of information from case-control and cohort studies of nutrition and disease. *Nutr Cancer* 1995; **24**:231–9.

322. Margolis LH. A critical review of studies of newborn discharge timing. *Clin Pediatr Phila* 1995; **34**:626–34.

323. Marrs RW. A meta-analysis of bibliotherapy studies. *Am J Community Psychol* 1995;**23**:843–70.

324. Massy ZA, Ma JZ, Louis TA, Kasiske BL. Lipid-lowering therapy in patients with renal disease. *Kidney Int* 1995;**48**:188–98.

325. McAweeney MJ, Tate DG, McAweeney W. Psychosocial interventions in the rehabilitation of people with spinal cord injury: a comprehensive methodological inquiry. *SCI Psychosocial Process* 1997;**10**:58–66.

326. Flint AJ. Effects of respite care on patients with dementia and their caregivers. *Int Psychogeriatr* 1995;**7**:505–17.

327. Cole M-G. The impact of geriatric medical services on mental state. *Int Psychogeriatr* 1993;**5**:91–101.

328. McCusker J, Cole M, Keller E, Bellavance F, Berard A. Effectiveness of treatments of depression in older ambulatory patients. *Arch Intern Med* 1998;**158**:705–12.

329. Meinert CL. Essential design features of a controlled clinical trial. In Clinical trials: design, conduct and analysis. Oxford: Oxford University Press; 1986. pp. 65–70.

330. Morin CM, Culbert JP, Schwartz SM. Nonpharmacological interventions for insomnia: a meta-analysis of treatment efficacy. *Am J Psychiatry* 1994;**151**:1172–80.

331. Morris SB, Turner CW, Szykula SA. Psychotherapy outcome research: an application of a new method for evaluating research methodology. *Psychotherapy* 1988;**25**:18–26.

332. Murtagh DR, Greenwood KM. Identifying effective psychological treatments for insomnia: a meta-analysis. *J Consult Clin Psychol* 1995;**63**:79–89.

333. Naldi L, Parazzini F, Cainelli T. Role of topical immunotherapy in the treatment of alopecia areata. Quality analysis of articles published between January 1977 and January 1988 about three treatments. Reading Group. *J Am Acad Dermatol* 1990;**22**:654–6.

334. Naylor CD, Detsky AS, O' Rourke K, Fonberg E. Does treatment with essential amino acids and hypertonic glucose improve survival in acute renal failure? A meta-analysis. *Ren Fail* 1987;**10**:141–52.

335. Camma C, Almasio P, Craxi A. Interferon as treatment for acute hepatitis C. A meta-analysis. *Dig Dis Sci* 1996;**41**:1248–55.

336. Nielsen ME, Reilly PL. A guide to understanding and evaluating research articles. *Gifted Child Q* 1985;**29**:90–2.

337. Nyberg G. Assessment of papers of clinical trials. *Med J Aust* 1974;**2**:381.

338. EPI Centre. Review of effectiveness of sexual health promotion interventions for young people. London: Social Science Research Unit, University of London Institute of Education; 1996.

339. Oakley A, Fullerton D. Sexual health education interventions for young people: a methodological review. *BMJ* 1995;**310**:158–62.

340. Foxcroft DR, Lister-Sharp D, Lowe G. Alcohol misuse prevention for young people: a systematic review reveals methodological concerns and lack of reliable evidence of effectiveness. *Addiction* 1997; **92**:531–7.

341. Ogilvie-Harris DJ, Gilbart M. Treatment modalities for soft tissue injuries of the ankle: a critical review. *Clin J Sport Med* 1995;**5**:175–86.

342. Osberg TM. Teaching psychotherapy outcome research methodology using a research-based checklist. *Teach Psychol* 1997;**24**:271–4.

343. Oxman A. No magic bullets: a systematic review of 102 trials of interventions to help health care professionals deliver services more effectively or efficiently. London: North East Thames Regional Health Authority; 1994.

344. Powe NR, Tielsch JM, Schein OD, Luthra R, Steinberg EP. Rigor of research methods in studies of the effectiveness and safety of cataract extraction with intraocular lens implantation.

Cataract Patient Outcome Research Team. *Arch Ophthalmol* 1994;**112**:228–38.

345. Puig Barbera J, Marquez Calderon S. Effectiveness of influenza vaccine in the elderly. A critical review of the bibliography. *Med Clin (Barc)* 1995; **105**:645–8.

346. Rey JM, Walter G. Half a century of ECT use in young people. *Am J Psychiatry* 1997;**154**:595–602.

347. Ried LD, Carter KA, Ellsworth A. Acetazolamide or dexamethasone for prevention of acute mountain sickness: a meta-analysis. *J Wilderness Med* 1994;**5**:34–48.

348. Rowe DE, Bernstein SM, Riddick MF, Adler F, Emans JB, Gardner Bonneau D. A meta-analysis of the efficacy of non-operative treatments for idiopathic scoliosis. *J Bone Joint Surg Am* 1997; **79**:664–74.

349. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston, MA: Little Brown; 1985.

350. Heyland DK, Cook DJ, Guyatt GH. Enteral nutrition in the critically ill patient: a critical review of the evidence. *Intensive Care Med* 1993; **19**:435–42.

351. Ciliska D, Kelly C, Petrov N, Chalmers J. A review of weight loss interventions for obese people with non-insulin-dependent diabetes mellitus. *Can J Diabetes Care* 1995;**19**:10–15.

352. Choi HW, Breman JG, Teutsch SM, Liu S, Hightower AW, Sexton JD. The effectiveness of insecticide-impregnated bed nets in reducing cases of malaria infection: a meta-analysis of published results. *Am J Trop Med Hyg* 1995;**52**:377–82.

353. Sellers DE, Crawford SL, Bullock K, McKinlay JB. Understanding the variability in the effectiveness of community heart health programs: a meta-analysis. *Soc Sci Med* 1997;**44**:1325–39.

354. Selley S, Donovan J, Faulkner A, Coast J, Gillatt D. Diagnosis, management and screening of early localised prostate cancer. *Health Technol Assess* 1997;**1**(2).

355. Shaver JP, Curtis CK, Jesunathadas J, Strong CJ. The modification of attitudes towards persons with disabilities: is there a best way? *Int J Special Educ* 1989;**4**:33–57.

356. Simons MP, Kleijnen J, van Geldere D, Hoitsma HF, Obertop H. Role of the Shouldice technique in inguinal hernia repair: a systematic review of controlled trials and a meta-analysis. *Br J Surg* 1996;**83**:734–8.

357. Slavin RE. Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educ Researcher* 1986;**15**:5–11.

358. Schoemaker C. Does early intervention improve the prognosis in anorexia nervosa? A systematic review of the treatment-outcome literature. *Int J Eat Disord* 1997;**21**:1–15.

359. Smeenk FW, van Haastregt JC, de Witte LP, Crebolder HM. Effectiveness of home care programmes for patients with incurable cancer on their quality of life and time spent in hospital: systematic review. *BMJ* 1998;**316**:1939–44.

360. Smith MC, Stullenbarger E. A prototype for integrative review and meta-analysis of nursing research. *J Adv Nurs* 1991;**16**:1271–83.

361. Smith NL, Glass GV. Meta-analysis of psychotherapy outcome studies. *Am Psychol* 1977; **32**:752–60.

362. Burckhardt CS. The effect of therapy on the mental health of the elderly. *Res Nurs Health* 1987; **10**:277–85.

363. Krywanio ML. Meta-analysis of physiological outcomes of hospital-based infant intervention programs. *Nurs Res* 1994;**43**:133–7.

364. Smith NL, Glass GV, Miller TI. Benefits of psychotherapy. Baltimore, MD: Johns Hopkins University Press; 1980.

365. Solomon MJ, McLeod RS. Clinical studies in surgical journals – have we improved? *Dis Colon Rectum* 1993;**36**:43–8.

366. Solomon DH, Bates DW, Panush RS, Katz JN. Costs, outcomes, and patient satisfaction by provider type for patients with rheumatic and musculoskeletal conditions: a critical review of the literature and proposed methodologic standards. *Ann Intern Med* 1997;**127**:52–60.

367. Stevens A, Raftery J. The stimulus for needs assessment: reforming health services. In Stevens A, Raftery J, editors. Health care needs assessment. Oxford: Radcliffe Medical Press; 1994. pp. 11–30.

368. Dodhia H, Miller E. Review of the evidence for the use of erythromycin in the management of persons exposed to pertussis. *Epidemiol Infect* 1998;**120**:143–9.

369. Stock S. Workplace ergonomic factors and the development of musculoskeletal disorders of the neck and upper limbs: a meta-analysis. *Am J Ind Med* 1991;**19**:87–107.

370. Suydam MN. An instrument for evaluating experimental educational research reports. *J Educ Res* 1968;**61**:200–3.

371. Floyd M, Scogin F. Effects of memory training on the subjective memory functioning and mental health of older adults: a meta-analysis. *Psychol Aging* 1997;**12**:150–61.

372. Tangkanakul C, Counsell CE, Warlow CP. Local versus general anaesthesia in carotid endarterectomy: a systematic review of the evidence. *Eur J Vasc Endovasc Surg* 1997;**13**:491–9.

373. Theis JG, Selby P, Ikizler Y, Koren G. Current management of the neonatal abstinence syndrome: a critical analysis of the evidence. *Biol Neonate* 1997;**71**:345–56.

374. Thomas JP, Lawrence TS. Common deficiencies of NIDRR research applications. *Am J Phys Med Rehabil* 1991;**70**:161–4.

375. Thomas J, DeHueck A, Kleiner M, Newton J, Crowe J, Mahler S. To vibrate or not to vibrate: usefulness of the mechanical vibrator for clearing bronchial secretions. *Physiother Can* 1995;**47**:120–5.

376. Tuckman BW. A proposal for improving the quality of published research. *Educ Researcher* 1990;**19**:2–25.

377. US Preventive Services Task Force. Task force ratings. In US Preventive Services Task Force, editor. Guide to clinical preventive services. Baltimore, MD: Williams & Wilkins; 1996. pp. 861–85.

378. US Preventive Services Task Force. Task force ratings. In US Preventive Services Task Force, editor. Guide to Clinical Preventive Services. Baltimore, MD: Williams & Wilkins; 1989. pp. 387–98.

379. Robert G, Stevens A. Should general practitioners refer patients directly to physical therapists? *Br J Gen Pract* 1997;**47**:314–18.

380. Long T, Soderstrom E. A critical appraisal of positioning infants in the neonatal intensive care unit. *Phys Occup Ther Pediatr* 1995;**15**:17–31.

381. Witlin AG, Sibai BM. Magnesium sulfate therapy in preeclampsia and eclampsia. *Obstet Gynecol* 1998;**92**:883–9.

382. Grullon KE, Grimes DA. The safety of early postpartum discharge: a review and critique. *Obstet Gynecol* 1997;**90**:860–5.

383. Grimes DA. Medical abortion in early pregnancy: a review of the evidence. *Obstet Gynecol* 1997; **89**:790–6.

384. van Balkom AJ, Bakker A, Spinhoven P, Blaauw BM, Smeenk S, Ruesink B. A meta-analysis of the treatment of panic disorder with or without agoraphobia: a comparison of psychopharmacological, cognitive-behavioral, and combination treatments. *J Nerv Ment Dis* 1997; **185**:510–16.

385. van Dalen DB. A research checklist in education. *Educ Admin Supervision* 1958;**44**:174–81.

386. Vickers AJ. Hypnotherapy for irritable bowel syndrome. A report commissioned by North East Thames Regional Health Authority. London: Research Council for Complementary Medicine; 1994.

387. Ward M, Fetler M. What guidelines should be followed in critically evaluating research reports. *Nurs Res* 1979;**28**:120–5.

388. Weiler JM. Medical modifiers of sports injuries. The use of non-steroidal anti-inflammatory drugs in sports soft tissue injury. *Clinc Sports Med* 1992; **11**:625–44.

389. Wolf RM. Judging educational research based on experiments and surveys. Fundamentals of educational planning series No. 45. Paris: United Nations Educational, Scientific and Cultural Organization; 1993.

390. Woolf SH, Battista RN, Andeson GM, Logan AG, Wang E, Canadian Task Force on the Periodic Health Examination. Assessing the clinical effectiveness of preventive manoeuvres: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol* 1990; **43**:891–905.

391. NHS Centre for Reviews and Dissemination. Review of the research on the effectiveness of health service interventions to reduce variations in health. CRD Report 3. York: University of York, 1995. pp. 1–157.

392. Zola P, Volpe T, Castelli G, Sismondi P, Nicolucci A, Parazzini F, *et al*. Is the published literature a reliable guide for deciding between alternative treatments for patients with early cervical cancer? *Int J Radiat Oncol Biol Phys* 1989;**16**:785–97.

393. ACCP/AACVPR Pulmonary Rehabilitation Guidelines Panel. Pulmonary rehabilitation: joint ACCP/AACVPR evidence-based guidelines. *Chest* 1997;**112**:1363–96.

394. Eastwood AE, Sheldon TA. Organisation of asthma care: what differences does it make? A systematic review of the literature. *Qual Health Care* 1996; **5**:134–43.

395. Hearn J, Higginson IJ. Do specialist palliative care teams improve outcomes for cancer patients? *Palliat Med* 1998;**12**:317–32.

396. MacMillan HL, MacMillan JH, Offord DR, Griffith DL, MacMillan A. Primary prevention of child sexual abuse and neglect: a critical review. Part 2. *J Child Psychol Psychiatry* 1994;**35**:857–76.

397. Mathews F. Antioxidant nutrients in pregnancy: a systematic review of the literature. *Nutr Res Rev* 1996;**9**:175–95.

398. Parker J, Bisits A, Proietto AM. A systematic review of single-dose intramuscular methotrexate for the treatment of ectopic pregnancy. *Aust NZ J Obstet Gynaecol* 1998;**38**:145–50.

399. Powe NR, Schein OD, Gieser SC, Tielsch JM, Luthra R, Javitt J, *et al*. Synthesis of the literature on visual acuity and complications following cataract extraction with intraocular lens implantation. Cataract Patient Outcome Research Team. *Arch Ophthalmol* 1994;**112**: 239–52.

400. Snowdon SK, Stewart-Brown SL. Preschool vision screening: results of a systematic review. CRD Report 9. York: University of York, 1997. pp. 1–114.

401. Talley NJ, Owen BK, Boyce P, Paterson K. Psychological treatment for irritable bowel syndrome: a critique of controlled treatment trials. *Am J Gastroenterol* 1996;**91**:277–83.

402. Villalobos T, Arango C, Kubilis P, Rathore M. Antibiotic prophylaxis after basilar skull fractures: a meta-analysis. *Clin Infect Dis* 1998;**27**:364–9.

403. Poynard T. Evaluation de la qualite methodologique des essais therapeutiques randomises. *Presse Med* 1988;**17**:315–18.

404. Fiske DW, Hunt HF, Luborsky L, Orne MT *et al*. Planning of research on effectiveness of psychotherapy. *American Psychologist* 1970;**25**:727–37.

405. Bergin AE, Lambert MJ. The evaluation of therapeutic outcomes. In Garfield SL, Bergin AE, editors. Handbook of psychotherapy and behavior change. New York: John Wiley; 1978. pp. 139–89.

406. Jonsson C, Martens S, Sjoqvist F. [Attitude toward clinical tests of drugs – especially psychopharmaca]. *Nordisk Psykiatrisk Tidsskrift*. 1969;**23**:281–9.

407. Evans M, Pollock AV. A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infections. *Br J Surg* 1985;**72**:256–60.

Appendix 1

Search strategy for Chapters 3–5

MEDLINE (1966–99) via Ovid

(check adj list$).tw.
check-list$.tw.
checklist$.tw.
1 or 2 or 3
methodolog$.tw.
(((check adj list$) or check-list$ or checklist$) adj5 methodolog$).tw.
quality.tw.
validity.tw.
validat$.tw.
assess$.tw.
evaluat$.tw.
7 or 8 or 9 or 10 or 11
study.tw.
studies.tw.
paper$.tw.
article$.tw.
literature.tw.
report$.tw.
research.tw.
13 or 14 or 15 or 16 or 17 or 18 or 19
(((check adj list$) or check-list$ or checklist$) adj5 (quality or validity or validat$ or assess$ or evaluat$) adj5 (study or studies or paper$ or article$ or literature or report$ or research)).tw.
bias$.tw.
exp "bias (epidemiology)"/
22 or 23
4 and 24
6 or 21 or 25
((scale or scales) adj5 methodolog$).tw.
((scale or scales) adj3 (quality or validity or validat$ or assess$ or evaluat$) adj3 (study or studies or paper$ or article$ or literature or report or reports or research)).tw.
(inter adj rater adj reliability).tw.
(interrater adj reliability).tw.
(test adj retest adj reliability).tw.
(testretest adj reliability).tw.
(testre adj test adj reliability).tw.
(test adj re adj test adj reliability).tw.
(scales or scale).tw.
29 or 30 or 31 or 32 or 33 or 34
4 and 36
((scales or scale) adj3 ((inter adj rater adj reliability) or (interrater adj reliability) or (test adj retest adj reliability) or (testretest adj reliability) or (testre adj test adj reliability) or (test adj re adj test adj reliability))).tw.
27 or 28 or 38
evaluat$.ti.
confounding.tw.
(critical$ adj apprais$).tw.
40 or 41 or 42
4 and 43
case-control studies/
(observational adj (study or studies or data)).tw.
(non-random$ or nonrandom$).tw.
(natural adj experiment$).tw.
(quasi adj experiment$).tw.
quasiexperiment$.tw.
(non adj experiment$).tw.
nonexperiment$.tw.
intervention studies/
cohort studies/
45 or 46 or 47 or 48 or 49 or 50 or 51 or 52 or 53 or 54
4 and 55
46 or 47 or 48 or 49 or 50 or 51 or 52
7 or 8 or 9
(((observational adj (study or studies or data)) or (non-random$ or nonrandom$) or (natural adj experiment$) or (quasi adj experiment$) or quasiexperiment$ or (non adj experiment$) or nonexperiment$) adj5 (quality or validity or validat$)).tw.
(assess$ or judg$ or measur$ or analy$ or evaluat$).ti.
quality.ti.
(critical$ adj apprais$).ti.
(research or literature or report$ or paper$ or study or studies or article$).ti.
((assess$ or judg$ or measur$ or analy$ or evaluat$) adj3 quality adj3 (research or literature or report$ or paper$ or study or studies or article$)).ti.
(critical$ adj apprais$ adj3 (research or literature or report$ or paper$ or study or studies or article$)).ti.
data interpretation, statistical/
23 and 66
55 and 67
37 or 39 or 44 or 56 or 59 or 64 or 65 or 68
((scales or scale) adj3 bias$).tw.
(methodolog$ adj5 quality).ti.
(bias$ adj3 ((observational adj (study or studies or

data)) or (non-random$ or nonrandom$) or (natural adj experiment$) or (quasi adj experiment$) or quasiexperiment$ or (non adj experiment$) or nonexperiment$)).tw.
26 or 69 or 70 or 71 or 72
(nonrandom$ adj allocation).tw.
(non-random$ adj allocation).tw.
(historical adj control$).tw.
exp clinical trials/
random allocation/
research design/
(gold adj standard).tw.
(randomi#ed adj controlled adj trial$).tw.
research design/
55 or 77 or 78 or 79
76 and 83
80 and 81
78 and 82
74 or 75 or 84 or 85 or 86
non random$.tw.
nonrandom$.tw.
allocat$.tw.
((non random$ or nonrandom$) adj5 allocat$).tw.
88 or 89 or 90 or 91
73 or 87 or 92

Appendix 2

Data extraction forms

Author Year EN DARE

Published as:

Source

Topic Area

Tool purpose:

Type of studies Design1 Design2

Quality issues Quality1 Quality2

Type of tool Specify modified tool

Separate questions by study design

If Scale,

defines high quality threshold ☐

weighting

Type/no of items Total Generic Topic-specific

Was choice of items justified? Quality items reported?

How were items generated?

Was scale validity assessed?

Validity1

Validity2

Validity3

Validity4

Was scale reliability assessed?

Reliability1

Reliability2

Reliability3

Reliability4

Reliability5

Time to complete Average time taken

Tested on sample

OR

Describe method

Describe studies

Describe topic area

What items are included?

Comments and/or useful references

Checklist references EN no

Other references EN No

Has been used by other reviews

Author: Accession No: Endnote No:

Journal: Year:

Did the review use validity assessment? ☐ Were results clearly reported by study design? ☐

Did the review consider CMA? ☐ Was the VA/CMA considered in the study synthesis? ☐

Type of review: Intervention type:

Aim of intervention: Disease category:

Literature search

Database1: Other search1:

Database2: Other search2:

Database3: Other search3:

Database4: Other search4:

Database5: Other search5:

Database6: Other search6:

Inclusion criteria

Inclusion criteria1: No of studies considered:

Inclusion criteria2: No of studies excluded:

Languages included:

Type/number of studies

Total no of studies: Total no of patients:

No of RCTs: No patients:

No of non-RCTs: No patients:

List type of non-RCTs:

No of non-comp: No patients:

List type of non-comp:

Quality assessment

Was VA carried out? Same elig criteria?

Was independent VA carried out? Similar study dates?

Was same tool used for all designs?

Type of tool: Specify tool:

Tool method Study designs:

Tool described?:

Tool details:

VA results reported Specify other:

Endnote No:

Study synthesis

Was VA considered in study synthesis? ☐

If Yes, how?

Weighted by sample size? ☐ ☐ Extract results?

If quantitative

Quality threshold defined ☐

Define qual threshold:

Describe VA method:

Summarise results

Results

Primary outcome

What was the overall conclusion of the review?

1:

2:

3:

4:

5:

6:

Describe quantitative results: Statistical measure: Specify other

Comparison	Point estimate	95% CI Lower	Upper	P value
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0

	0	0	0	0

Additional outcomes reported ☐

Comments

Author: Accession No: Endnote No:

Journal: Year:

Did the review use validity assessment? ☐ Were results clearly reported by study design? ☐

Did the review consider CMA? ☐ Was the VA/CMA considered in the study synthesis? ☐

Type of review: Intervention type:

Aim of intervention: Disease category:

Literature search

Database1: Other search1:

Database2: Other search2:

Database3: Other search3:

Database4: Other search4:

Database5: Other search5:

Database6: Other search6:

Inclusion criteria

Inclusion criteria1: No of studies considered:

Inclusion criteria2: No of studies excluded:

Languages included:

Type/number of studies

Total no of studies: Total no of patients:

No of RCTs: No patients:

No of non-RCTs: No patients:

List type of non-RCTs:

No of non-comp: No patients:

List type of non-comp:

Quality assessment

Was VA carried out? Same elig criteria?

Was independent VA carried out? Similar study dates?

Was same tool used for all designs?

Type of tool: Specify tool:

Tool method Study designs:

Tool described?:

Tool details:

VA results reported Specify other:

Endnote No

Was CMA considered at study level? ☐ ☐ CMA quant stats

If yes, how?

Method described Not reported

Variables described

Impact on review

Was CMA considered at review level? ☐

If yes, what method was used?

Which variables were examined?

Impact on review

Results

Primary outcome

What was the overall conclusion of the review?

1:
2:
3:
4:
5:
6:

Describe quantitive results: Statistical measure Specify other

Comparison	Point estimate	95% CI Lower	Upper	P value

Additional outcomes reported? ☐

Comments

Appendix 3

Details of identified quality assessment tools

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Anders, 1996[81]	n		7	Coh	s e	n	y	6	na	na	na	na	na	n	n	
Antczak, 1986[76]	m	Chalmers, 1981[51]	7	RCTs	s u	n	n	30	na	na	na	na	na	n	y	204
Audet, 1993[176]	m	Poynard, 1988[403]	7	RCTs, NRS	s e	n	n	17	na	0.70^{K} ($p < 0.01$)	na	na	na	n	y	205
Avis, 1994[206]	m	Fowkes, 1991[107]	3	Any	c	n	n	24	na	na	na	na	na	n	n	
Bailar, 1984[207]	n		3	Weak/no internal controls	c	n	n	5	na	na	na	na	na	n	n	
Baker, 1994[82]	n		7	Coh	c	n	n	NR	na	na	na	na	na	n	n	
Bass, 1993[88]	n		7	RCTs, NRS; multiple time series and descriptive studies	s u	n	n	16	na	na	na	na	na	n	n	
Baumann, 1997[208]	n		7	RCTs, CTs	c	n	n	20	na	na	na	na	na	n	n	
Berard, 1997[209]	m	Chalmers, 1981[51]	7	RCTs, prospective CTs	s	n	n	NR	na	na	na	na	na	n	n	
Beyer, 1998[210]	n		7	RCTs, NRS	c	n	n	4	na	na	na	na	na	n	n	
Bland, 1985[211]	m	Gifford, 1969[265]	2	CCTs	c	n	n	18	na	Perfect/good except item 7 0.38^{K}, item 1 0.07^{K}, item 6 -0.19^{K}	na	na	na	n	n	
Boers, 1991[212]	m	Sackett, 1985[349]	7	RCTs, Coh	c	n	n	21	na	na	na	na	na	n	n	
Borghouts, 1998[213]	n		7	RCTs, Coh; longitudinal studies; FU; prospective studies; retrospective studies; CCS	s e	n	n	13	na	na	na	na	na	n	n	
Bours, 1998[73]	m	van der Windt, 1995[71]	7	RCTs, CCTs	s u	n	y	18	na	84%[a]	na	na	na	n	n	
Bracken, 1989[104]	n		6	Obs	c	n	n	36	na	na	na	na	na	n	n	
Brown, 1991[214]	m	Haynes, 1976[273]	7	RCTs, NRS; Pre/Post	s u	n	n	6	na	89%[a]	na	na	na	n	n	215

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Brown, 1995[216]	n		7	RCTs, NRS	c	n	n	5	na	89%[a]	na	na	na	n	n	
Brown, 1996[217]	m	Brown, 1991[214]	7	RCTs, NRS; Pre/Post	s	n	n	6	na	94.5%[a]	na	na	na	n	n	
Callahan, 1991[218]	n		7	RCTs, NRS	c	n	n	9	na	na	na	na	na	n	n	
Cameron, 2000[83]	n		7	RCTs, NRS	c	n	n	36	na	na	na	na	na	n	n	
Cameron, 2000[83]	n		7	RCTs, NRS; Coh; CCS	s u	n	n	9	na	na	na	na	na	n	n	
Campos-Outcalt, 1995[177]	n		7	RCTs, Quasi-Exp; Coh; CS; CCS; correlational studies	s e	n	n	7	na	na	na	na	na	n	n	
Canadian Task Force, 1997,[219] 1994[220]	L															221–225
Cao, 1997[226]	m	Chalmers, 1981[51]	7	RCTs, NRS	s	n	n	NR	na	na	na	na	na	n	n	
Carlson, 1996[227]	n		7	RCTs, NRS	c	n	n	7	na	na	na	na	na	n	n	
Carter, 1994[178]	m	Chalmers, 1981[51]	7	RCTs, NRS	s u	n	n	12	na	na	na	na	na	n	n	
CASP, 1999[64]	n		3	Coh	c	n	n	10	na	na	na	na	na	n	n	
CEBM, 2002[228]	L	Can Task Force, 1979[219]	4													
Chalmers, 1981[51]	n		3	RCT	s u	n	n	36	na	na	na	na	na	n	y	229
Cheatham, 1995[230]	m	Chalmers, 1981[51]	7	RCTs, NRS	s u	n	n	NR	na	na	na	na	na	n	n	
Cho & Bero, 1994[84]	m	Spitzer, 1990[105]	3	Exp, Obs	s e	n	n	31	FC, Cr	0.64[W]; 0.89[r] (CI: 0.73 to 0.96)	na	na	na	y	n	
Ciliska, 1996[231]	n		7	RCTs, Coh, CCS	c	n	n	6	na	0.80[K]	na	na	na	n	n	232
Cochrane EPOC Group, 2002[233]	n		7	RCTs, CCTs	c	n	n	7	na	na	na	na	na	n	y	EPOC reviews

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Cochrane Handbook, 2002[118]	X	Not a QA tool														234, 235
Cochrane Musculoskeletal Injuries Group, 2002[179]	m	Chalmers, 1981[51]	7	RCTs, CCTs	s e	n	n	12	na	na	na	na	na	n	y	Cochrane reviews
Cohen, 1994[236]	n		7	RCTs, NRS	s e	n	n	5	na	0.707[K], $p = 0.0005$	na	na	na	n	n	
Colditz, 1994[237]	n		7	RCTs, CCS	s	n	y	6	na	na	na	na	na	n	n	
Coleridge Smith, 1999[97]	n		7	CCTs, Coh, CCS, CS	c	n	y	49	na	C/M: 0.57[K] (CI: 0.49 to 0.65) C/C: 0.34[K] (CI: 0.19 to 0.49) M/M: 0.94[K] (CI: 0.71 to 1.0)	na	na	na	n	n	
Cook, 1979[50]	X	Not a QA tool														189
Cook, 1992[238]	L		4												y	239–241
Cowley, 1995[109]	n		7	RCTs, NRS; Obs studies without controls	c	n	n	12–17	na	na	na	na	na	n	n	
Cox, 1989[242]	n		7	RCTs, NRS; Obs	s u	n	n	14	na	0.98[K]	na	na	na	n	n	
Cuddy, 1983[180]	n		3	Exp	c	n	n	17	na	na	na	na	na	n	n	
Cuijpers, 1998[243]	n		7	RCTs, Pre/Post	c	n	n	NS	na	na	na	na	na	n	n	
Dawson-Saunders, 1990[98]	n		3	CCTs, Coh, CCS	c	n	y	49	na	na	na	na	na	n	n	
de Craen, 1996[244]	m	ter Riet, 1990[68]	7	RCTs, NRS	s u	n	n	13	na	na	na	na	na	n	n	
de Kruif, 1996[245]	n		7	RCTs, NRS; Coh	c	n	n	8	na	na	na	na	na	n	n	
de Oliveira, 1995[96]	m	Chalmers, 1981[51]	7	Not clear, CCTs	s u	n	n	30	na	Unfamiliar 0.75[r], familiar 0.87[r]	na	na	na	n	y	246

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
de Vet, 1997[72]	m	ter Riet, 1990[68]	7	RCTs, NRS	s e	n	n	15	na	0.77[r] (0.64–0.89)[67]	na	na	na	n	n	
DerSimonian, 1982[247]	n		6	CCTs	c	n	n	11	na	9 items > 80%[a]	na	na	na	n	n	
Detsky, 1992[92]	m	Chalmers, 1981[51]	7	RCTs	c	n	n	14	na	0.53[r248]	na	na	na	n	y	248–249
Devine, 1992[250]	n		7	Not clear	c	n	n	3?	na	89%[a], 0.79[K]	na	na	na	n	n	
Devine, 1983[251]	n		7	RCTs, NRS	c	n	n	6	na	92%[a]	na	na	na	n	n	
Devine, 1995[252]	n		7	RCTs, Quasi-Exp; Pre/Post	c	n	y	NR	na	87–100%[a]	na	na	na	n	n	
Devine, 1995[253]	n		7	RCTs, NRS	c	n	n	NR	na	na	na	na	na	n	n	
Downs, 1998[85]	n		3	RCTs, NRS	s u	n	n	27	FC, Cr	0.75[r]	na	0.88[r] @ 2 wks 0.89[Kr]	na	y	y	112, 130
Duffy, 1985[89]	n		3	Any	s e	n	n	51	na	0.89[254]	0.94 at 2 months	na	0.91[CA]	n	y	254
DuRant, 1994[99]	n		3	Exp, Quasi-Exp, Survey, CS	c	n	y	103	na	na	na	na	na	n	n	
Durlak, 1997[255]	n		7	RCTs, CTs	c	n	n	6	na	0.85[a] (range 0.75–0.99)	na	na	na	n	n	
Elvik, 1995[256]	n		7	NRS	c	n	n	NR	na	na	na	na	na	n	n	
Esdaile, 1986[257]	n		3	Coh, CCS	c	n	n	6	na	na	na	na	na	n	n	
Farquhar, 1959[258]	n		3	Exp,	c	n	n	18	na	na	na	na	na	n	n	
Fowkes, 1991[107]	n		3	CCTs, Coh, CCS, CS	c	n	n	23	na	na	na	na	na	n	y	97
Fox, 1958[259]	n		3/1	NS	c	n	n	7	na	na	na	na	na	n	n	
Friedenreich, 1993[100]	n		7	Coh, CCS	c	n	y	29	na	na	na	na	na	n	n	

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Ganong, 1987[123]	X	Not a QA tool													y	260
Gansevoort, 1995[261]	n		7	RCTs, NRS	c	n	n	3	na	na	na	na	na	n	n	
Garber, 1996[262]	n		7	Any, CS; Coh; CCS; CCTs	s e	n	n	6	na	na	na	na	na	n	n	
Gardner, 1986[181]	n		2	RCTs, NRS	c	n	n	26	na	na	na	na	na	n	n	
Garg, 1998[263]	n		7	RCTs, Coh; CCS	c	n	n	5	na	na	na	na	na	n	n	
Gartland, 1988[124]	X	Not a QA tool													y	264
Gifford, 1969[265]	n		2	RCTs, NRS	c	n	n	8	na	na	na	na	na	n	n	
Glantz, 1997[182]	m	Chalmers, 1981[51]	7	RCTs, NRS	s u	n	n	20	na	na	na	na	na	n	n	
Good, 1996[266]	n		7	RCTs, NRS	c	n	n	7	na	na	na	na	na	n	n	
Gordis, 1990[101]	n		3	Trials, Coh, CCS, Time series, CS	c	n	y	17	na	na	na	na	na	n	y	267
Gray, 1982[268]	n		2	RCTs, NRS	s e	n	n	8	na	na	na	na	na	n	n	
Greenhalgh, 1997[183]	n		3	RCTs, NRS, Coh, CCS	c	n	n	11	na	na	na	na	na	n	n	
Gurman, 1978[184]	n		2	RCTs, NRS	s u	n	n	14	na	na	na	na	na	n	n	
Guyatt, 1993[79]	n		3	RCTs,	c	n	n	12	na	0.79[r 185]	na	na	na	n	y	185
Gyorkos, 1994[127]	n		7	CCTs, Coh, CCS, CS, Community trials, Obs community studies, descriptive studies	c	n	y	40	na	na	na	na	na	n	y	269–271
Hadorn, 1996[102]	m	US Task Force, 1989[378]	4	RCTs, Coh, CCS, CS and registries, case reports and expert opinion	c	n	y	41	na	na	na	na	na	n	n	
Hancock, 1990–95[272]	n		7	RCTs, studies with a control and treatment group	c	n	n	10	na	na	na	na	na	n	n	
Hayes, 1992[204]	X	Not a QA tool													n	

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Haynes, 1976[273]	n		2	RCTs, Quasi-Exp, Analytic, Descriptive	s u	n	n	6	na	nr	na	na	na	n	n	274, 275
Heacock, 1997[276]	n		1/3	CCTs	s u	n	n	14	na	Range 0.36, 1.00[K]; excellent for 11 questions (75–100), good for 2 items (53–64%) and poor for 1 item (36%)	na	na	na	n	n	
Hedges, 1986[277]	X	Not a QA tool													n	
Helfenstein, 1994[278]	n		7	RCTs, NRS	c	n	n	11	na	na	na	na	na	n	n	
Heneghan, 1996[279]	m	Chalmers, 1981[51]	7	RCTs, NRS	s e	n	n	15	na	91%[a]; 0.83[K]	na	na	na	n	n	
Hill, 1967[280]	n		2	RCTs, NRS; Coh	c	n	n	10	na	na	na	na	na	n	n	
Hines, 1969[281]	n		3	Clinical investigations	c	n	n	10	na	na	na	na	na	n	n	
Hoag, 1997[282]	n		7	RCTs, NRS	c	n	n	29	na	0.88[K]				n	n	
Hoffman, 1993[283]	n		7	RCTs, NRS; Coh	c	n	n	9	na	na	na	na	na	n	n	
Howell, 1997[284]	n		7	Not stated	c	n	y	5	na	na	na	na	na	n	n	
Jabbour, 1996[285]	n	See Klassen, 1995[299]	7	RCTs, NRS; Coh	s e	n	n	5	na	87%[a], 0.81[K], SE=0.10	na	na	na	n	n	
Jadad, 1996[46]	n		3/7	RCTs	s e	n	n	3	FC, Co	0.66[r] (CI: 0.53 to 0.79); 86%[a] (0.71[K])[286]	na	na	na	n	y	286–288
Johnston, 1994[289]	n		7	RCTs, NRS	s e	n	n	5	na	na	na	na	na	n	n	290
Kasiske, 1995[291]	n		7	RCTs, CCTs	s u	n	n	NR	na	na	na	na	na	n	n	
Kasiske, 1998[292]	n		7	RCTs, prospective control group study; B/A	c	n	n	5	na	na	na	na	na	n	n	
Kasiske, 1993[293]	n		7	RCTs, NRS	c	n	n	9	na	na	na	na	na	n	n	

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Kay, 1996[186]	n		7	RCTs, NRS	s e	n	n	20	na	na	na	na	na	n	n	
King, 1984[294]	n		3	Drug trials	c	n	n	12	na	na	na	na	na	n	n	
King, 1990[295]	m	King, 1984[294]	3	Drug trials	c	n	n	13	na	na	na	na	na	n	n	
King, 1990[296]	m	King, 1984[294]	3	Drug trials	c	n	n	17	na	na	na	na	na	n	n	
Kingery, 1997[297]	m	Koes, 1991[70]	7	Not clear, CCTs	s u	n	n	13	na	na	na	na	na	n	n	
Kinney, 1996[298]	m	Smith and Stullenbarger, 1991[360]	7	Exp, Quasi-exp	s e	n	n	21	na	0.79–0.94[W]	na	na	na	n	n	
Klassen, 1995[299]	m	ter Riet, 1990[68]	7	RCTs, NRS	s u	n	n	45	na	74%[a]				n	n	
Kleijnen, 1994[300]	m	Kleijnen, 1991[69]	7	RCTs, NRS	s e	n	n	10	na	na	na	na	na	n	n	
Kleijnen, 1995[301]	m	Kleijnen, 1991?[69]	7	RCTs, NRS	s e	n	n	10	na	na	na	na	na	n	n	
Kleijnen, 1991[69]	m	ter Riet, 1990?[68]	7	RCTs, NRS	s u	n	n	7	na	na	na	na	na	n	n	302
Koes, 1991[70]	m	ter Riet, 1990[68]	7	RCTs,	s u	n	n	17	na	80%[a]	na	na	na	n	y	303, 304
Kreulen, 1998[187]	m	Antzcak, 1986[76]	7	Not clear, prospective study; cross-CS; retrospective study	s u	n	n	17	na	0–1[K]; 0.95[r], $p < 0.001$	na	na	na	n	n	
Kristensen, 1989[305]	n		7	NRS	c	n	n	5	na	na	na	na	na	n	y	306
Krogh, 1985[307]	n		3	Any	c	n	n	11	na	na	na	na	na	n	n	
Kulik, 1990[308]	n		7	RCTs, NRS	c	n	y	15	na	na	na	na	na	n	n	
Kwakkel, 1997[188]	n		7	RCTs, CCTs	s e	n	n	16	na	0.86[K]	na	na	na	n	n	
L'Abbe, 1987[120]	X	No QA tool reported		(Allegedly based on Chalmers)											y	309, 402
Labrecque, 1989[310]	n		3	Intervention studies	c	n	n	6	na	na	na	na	na	n	n	
Leboeuf, 1990[311]	n		6/2	RCTs, NRS; Coh; surveys	c	n	y	10	na	na	na	na	na	n	n	

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Lee, 1997[189]	m	Cook and Campbell, 1979[50]	7	RCTs, between-subjects design; within-subjects design	s e	n	n	7	na	na	na	na	na	n	n	
Lee, 1994[312]	n		7	RCTs, NRS	s e	n	n	29	na	94%[a]	na	na	na	n	n	
Levine, 1980[190]	n		3	CCTs	s e	n	n	30	na	na	na	na	na	n	y	313
Linde, 1999[95]	n		2/7	RCTs, NRS	s e	n	n	7	Cr	na	na	na	na	n	y	314
Lionel, 1970[315]	n		3	Prospective therapeutic trials	c	n	n	42	na	na	na	na	na	n	n	
Littenberg, 1998[316]	n		7	RCTs, comparative studies; CS	s e	n	n	5	na	na	na	na	na	n	n	
Longnecker, 1988[317]	n		7	CCS, CCS	s u	n	y	15	na	na	na	na	na	n	n	
Luborsky, 1975[318]	m	Fiske, 1970[404]	2	RCTs, NRS	s e	n	n	12	na	0.84[r]	na	na	na	n	n	
Lyons, 1991[319]	m	Smith, 1980[364]	7	RCTs, NRS	c	n	n	28	na	na	na	na	na	n	n	
MacLehose, 2000[26]	m	Downs and Black, 1998[85]	7	RCTs, NRS; Coh; CCS	s	n	n	NR	FC, Co	> 60%[a]	na	na	na	n	n	
MacMillan, 1994[191]	m	Chalmers, 1981[51]	7	RCTs, NRS	s u	n	n	9	na	>90%[a]	na	na	na	n	n	
Mahon, 1964[320]	n		3	Drug trials	c	n	n	4	na	na	na	na	na	n	n	
Margetts, 1995[321]	n		3	Coh, CCS	s u	n	y	31	na	CCS 0.71[r]; Coh 0.92[r]	Similar	na	na	n	n	
Margolis, 1995[322]	n		7	RCTs, NRS	c	n	n	6? NR	na	na	na	na	na	n	n	
Marrs, 1995[323]	n		7	RCTs, NRS	c	n	n	22	na	Categorical variables: 0.891[K] (0.625–1); continuous variables: 0.933[K] (0.693–1)	na	na	na	n	n	
Massy, 1995[324]	n		7	RCTs, CTs; uncontrolled studies	s u	n	y	14	na	na	na	na	na	n	n	

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Maziak, 1998[192]	n		7	RCTs, NRS	s e	n	n	10	na	0.87[K]	na	na	na	n	n	
McAweeney, 1997[325]	n		7	RCTs, Quasi-Exp; correlational studies	s e	n	n	36	na	na	na	na	na	n	n	
McCusker, 1998[328]	m	Chalmers, 1981[51]	7	Comparative trials, longitudinal comparison of at least two groups	s e	n	n	6	na	0.84[r] (CI: 0.55 to 0.95)	na	na	na	n	n	
McMaster, 1981[80]	n		3	RCTs	c	n	n	6	na	na	na	na	na	n	y	326, 327
Meijman, 1995[193]	m	Fowkes, 1991[107]	2	Any	s e	n	n	11	na	Mediocre/ good	Mediocre/ good at 15 months	na	na	n	n	
Meinert, 1986[329]	n		2	CCTs	c	n	n	25	na	na	na	na	na	n	n	
Melchart, 1994[93]	n		7	RCTs, NRS	s e	n	n	15	Cr	82%[a]	na	na	na	n	n	
Miller, 1995[87]	n		7	RCTs, Matched groups	s u	n	y	12	na	80.7%[a]	na	na	na	n	y	
Moncrieff, 1998[195]	n		3	CCTs	s e	n	n	30	na	0.924[r] (CI: 0.833 to 0.966)	na	na	0.873[CA]	n	n	
Morin, 1994[330]	n		7	RCTs, NRS	s e	n	n	6	na	na	na	na	na	n	n	
Morley, 1996[196]	m	Chalmers, 1981[51]	2/6	RCTs, NRS	s u	n	n	30	na	Dichotomous variables 80%[a] (range 62–98%); continuous variables: 0.91[r] (range 0.76–0.97)	na	na	na	n	y	
Morris, 1988[331]	m	Bergin and Lambert,1978[405]	2	RCTs, NRS; Coh; CS	s e	n	n	23	na	0.619[K] (range 0.308–0.824; SD 0.136)	na	na	Design factors, 0.52[a]; measurement integrity, 0.64; treatment validity, 0.25	na	n	

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Mullen, 1992[274]	m	Haynes, 1976[273]	7	RCTs, NRS	c	n	n	NR	na	90.6%[a]	na	na	na	n	n	
Mulrow, 1986[197]	n		7	RCTs, NRS; Coh	c	n	n	13	na	94%[a]	na	na	na	n	n	
Murtagh, 1995[332]	n		7	Any	c	n	y	17	na	95%[a]	na	na	na	n	n	
Naldi, 1990[333]	m	Chalmers, 1981[51]	2	RCTs, NRS; Coh	c	n	n	8	na	na	na	na	na	n	n	
Naylor, 1987[334]	m	Chalmers, 1981[51]	7	RCTs, NRS	s	n	n	NR	na	na	na	na	na	n	n	
Newcastle-Ottawa[66]	n		7	Coh, CCS	s u	n	y	8	na	na	na	na	na	n	n	
Nicolucci, 1989[77]	m	Chalmers, 1981[51]	7	RCTs	s u	n	n	16	na	88%[a]	na	na	na	n	y	335
Nielsen, 1985[336]	n		3	Any	c	n	n	24	na	na	na	na	na	n	n	
Nyberg, 1974[337]	n	Jonsson, 1969[406]	1/3	CCTs	s e	n	n	11	na	na	na	na	na	n	n	
Oakley, 1994[128]	n		7	RCTs, CTs	c	n	n	8	na	na	na	na	na	n	y	338–340
Ogilvie-Harris, 1995[341]	m	Weiler, 1992[388]	7	RCTs, CCTs	s e	n	n	10	na	na	na	na	na	n	n	
Osberg, 1997[342]	n		2	Outcome research	c	n	n	8	FC	na	na	na	na	n	n	
Oxman, 1994[343]	n		7	RCTs, NRS	c	n	n	6	na	na	na	na	na	n	n	
Powe, 1994[344]	n		7	NRS,	s e	n	y	11	na	93.5[K]	na	na	na	n	n	
Puig Barbera, 1995[345]	n		7	RCTs, Coh	c	n	n	8	na	na	na	na	na	n	n	
Reisch, 1989[111]	n		1/3	Therapeutic studies	s e	n	n	57	na	0.99[P] for objective score; 0.71[P] for subjective score[114]	na	na	na	n	y	Cochrane IBD Group
Rey, 1997[346]	n		7	RCTs, Coh or case reports	s	n	n	5	na	na	na	na	na	n	n	
Ried, 1994[347]	m	Chalmers, 1981[51]	7	RCTs, CCTs	s e	n	n	9	na	na	na	na	na	n	n	
Rowe, 1997[348]	n		7	RCTs, NRS; uncontrolled	s	n	n	12	na	na	na	na	na	n	n	

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Sackett, 1989[126], 1985[349]	L		4												y	350, 351
Salisbury, 1997[198]	n		7	RCTs, descriptive studies; methodological studies	c	n	n	45	na	na	na	na	na	n	n	
Schechter, 1991[199]	n		3	CCT	c	n	n	30	na	na	na	na	na	n	n	
Schmid, 1991[121]	X	Not a QA tool													y	352
Sellers, 1997[353]	n		7	RCTs, NRS	c	n	n	13	na	na	na	na	na	n	n	
Selley, 1997[354]	n		7	NRS, Coh	s	n	n	9	na	na	na	na	na	n	n	
Shaver, 1989[355]	n		7	RCTs, Quasi-Exp, Pre/Post	c	n	n	162	na	94%[a]	na	na	na	n	n	
Shay, 1972[90]	n		3	Any	c	n	n	25	Fa	high	na	na	na	n	n	
Sheldon, 1993[200]	n		3	RCTs, Quasi-Exp, Coh, CCS, CS	c	n	n	36	na	na	na	na	na	n	n	
Simons, 1996[356]	n		7	RCTs, CCTs	s e	n	n	9	na	85%[a]	na	na	na	n	n	
Slavin, 1986[357]	X	Not a QA tool													y	358
Smeenk, 1998[359]	n		7	RCTs, CCTs	s u	n	n	34	na	90.8%[a]	na	na	na	n	n	
Smith, 1991[360]	n		7	Exp, Non-Exp	s e	n	n	22	na	na	na	na	na	n	y	
Smith, 1977[361]	X	Not a QA tool													y	362
Smith, 1988[125]	X	Not a QA tool													y	363
Smith, 1980[364]	n		7	RCTs, CCTs	c	n	n	21 NR	na	92%[a]	na	na	na	n	n	
Solomon, 1993[365]	m	Evans and Pollock, 1985[407]	2	RCTs, NRS; case studies	s	n	n	10	na	0.69[K]	0.67[K]	na	na	n	n	
Solomon, 1997[366]	n		7	RCTs, Coh; survey	c	n	y	11	na	na	na	na	na	n	n	

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Spitzer, 1990[105]	n		7	RCTs, Coh; CCS; CS; uncontrolled series; descriptive studies	c	n	n	20	na	na	na	na	na	n	y	97
Stevens, 1994[367]	L	US Task Force, 1989[378]	4												y	368
Stock, 1991[369]	n		7	NRS, CCS, cross-sectional studies; longitudinal Coh	c	n	n	7	na	na	na	na	na	n	n	
Suydam, 1968[370]	n		3	Exp	s e	n	n	9	na	Coefficients 0.91, 0.94; 0.78[r]: 371	Coefficients 0.57, 0.77	na	na	n	y	371
Talley, 1993[106]	n		7	RCTs, CTs	c	n	n	29	na	na	na	na	na	n	n	
Tangkanakul, 1997[372]	n	Could be from Cochrane?	7	RCTs, NRS	c	n	y	10	na	na	na	na	na	n	n	
ter Riet, 1990[68]	n		7	RCTs, NRS	s u	n	n	18	na	na	na	na	na	n	y	
Theis, 1997[373]	n		7	RCTs, comparison studies	c	n	n	5	na	na	na	na	na	n	n	
Thomas, 1991[374]	X	Not a QA tool													y	325
Thomas, 1995[375]	n		7	RCTs, NRS	s u	n	n	10	na	na	na	na	na	n	n	
Thomas[65]	n		7	CCTs, Coh, CCS, B/A, CS	c	n	n	21	na	na	na	na	na	n	n	
Tuckman, 1990[376]	n		6?	Quantitative research	s	n	n	30	na	na	na	na	na	n	n	
US Task Force, 1996,[377] 1989[378]	L															379–383
van Balkom, 1997[384]	m	Chalmers, 1981[51]	7	RCTs, NRS	s e	n	n	14? NR	na	na	na	na	na	n	n	
van Dalen, 1958[385]	n		1/6	Exp	c	n	y	113	na	na	na	na	na	n	n	
van der Windt, 1995[71]	m	Koes, 1991[70]	7	RCTs,	s u	n	n	17	na	90%[a]	na	na	na	n	y	73

continued

Author	Origin[a]	Modified tool	Tool purpose[b]	Study designs covered[c]	Type of tool[d] (weighting)	Items justified?	Items by design?	No. of items	Tool validity[e]	Inter-rater reliability[f]	Intra-rater reliability[f]	Test-retest[f]	Internal consistency[f]	Time to complete	Used in other reviews?	Review reference number
Vickers, 1995[110]	n		3	CCTs	c	n	n	21	na	na	na	na	na	n	n	
Vickers, 1994[386]	n		7	RCTs, Coh	c	n	n	NR	na	na	na	na	na	n	n	
Vickers, 1996[201]	n	See also Vickers, 1994[386]	7	RCTs, NRS	c	n	n	12	na	"good"	na	na	na	n	n	
Ward, 1979[387]	n		3	Any	c	n	n	53	na	na	na	na	na	n	n	
Weiler, 1992[388]	n		7	RCTs, CTs	c	n	n	14	na	na	na	na	na	n	y	
Weintraub, 1982[108]	m	Lionel, 1970[315]	1/2/3	CCTs	c	n	n	47	na	na	na	na	na	n	n	
Wells-Parker, 1995[129]	n		7	RCTs, NRS	s e	n	n	?	FC	grouping strategies domain: 0.95[r]	na	na	na	n	n	
Wilson, 1992[103]	n		7	RCTs, Coh, CCS	c	n	y	10	na	na	na	na	na	n	n	
Wingood, 1996[202]	n		7	RCTs, NRS	c	n	n	16	na	na	na	na	na	n	n	
Wolf, 1993[389]	n		5	Any	s e	n	n	31	na	na	na	na	na	n	n	
Woolf, 1990[390]	L	Can Task Force, 1979[219]	4												y	391
Wright, 1995[203]	n		7	RCTs, NRS	c	n	n	13?	na	na	na	na	na	n	n	
Zaza, 2000[86]	n		7	RCTs, NRS	c	n	n	22	na	na	na	na	na	n	n	
Zola, 1989[392]	m	Chalmers, 1981[51]	2	RCTs, NRS; Coh	c	n	n	11	na	na	na	na	na	n	n	

[a] Origin: n, new; m, modified; X, not a quality assessment tool; L, levels of evidence classification system.
[b] Tool purpose: 1, planning a study; 2, assessing a statistical/methodological analysis; 3, evaluating a study when considering the practical application of the intervention; 4, evaluating or grading study recommendations when producing guidelines; 5, peer reviewing; 6, reporting a study; 7, assessing studies included in a systematic review (see *Box 3*).
[c] Study designs covered: RCT, randomised controlled trial; CCT, controlled clinical trial; Coh, cohort; CCS, case–control; Exp, experimental; B/A, before and after; Pre/Post, pre/post; Obs, observational; CS, case series (or cross-sectional); FU, follow-up studies; NRS, non-randomised studies.
[d] Type of tool: c, checklist; s, scale; u, unequal weighting; e, equal weighting.
[e] Tool validity assessments: FC, face and content; Cr, criterion; Co, content; Fa, factorial.
[f] Tool reliability assessments: IRR, inter-rater reliability; IaRR, intra-rater reliability; IC, internal consistency; T-R, test-retest; K, kappa; W, Kendall's W; [a], % agreement; [r], intra-class correlation; KR, Kuder–Richardson formula 20; CA, Cronbach's alpha; P, Pearson correlation; [α], coefficient alpha; C, clinician; M, methodologist; SE, standard error. NB: low internal consistency suggests that the criteria are relatively easy to satisfy.
na, not assessed; NR, not reported.

Appendix 4

Detailed description of 'unselected' top 14 quality assessment tools

Bracken, 1989[104]

The types of questions which led to the selection of this tool according to our pre-selected **core** criteria were:

5.3 How allocation occurred
If historical rather than concurrent controls have been used, has it been explained why this does not introduce bias into the study?
5.4 Any attempt to balance groups by design
If study groups were 'matched', have the rationale and detailed criteria for matching, and success (i.e. number of cases not matched) been provided?
9.2 Identification of prognostic factors
Has the measurement of important confounding or effect-modifying variables been described so the reader can judge how they have been controlled?
9.3 Case-mix adjustment
Has an analysis of potential confounding or effect modification of the principal relations been presented?

This tool provides a list of 36 questions to guide the reporting of observational studies. The tool is split into four sections: introduction to the report (4 items); description of materials and methods (17 items); presentation of results (7 items); and study conclusions (8 items). The tool took 20–30 minutes to complete. It was clearly not designed for use in a systematic review and to our knowledge has not been used for that purpose. The majority of the questions could be answered yes/no, making it possible to compare responses across studies. However, the questions are not phrased in such a way that any real judgements of the methodological quality of a study could be made, as the questions listed above demonstrate. Furthermore, the study designs that the tool could cover were limited to case-control and cohort designs.

The Bracken tool was not judged to be suitable for use in a systematic review.

Critical Appraisal Skills Programme, 1999[64]

The types of questions which led to the selection of this tool according to our pre-selected **core** criteria were:

5.3 How allocation occurred
Item not covered
5.4 Any attempt to balance groups by design
Have the authors taken account of the confounding factors in the design and/or analysis?
9.2 Identification of prognostic factors
What factors could potentially be related to both the exposure and the outcome of interest (confounding)?
9.3 Case-mix adjustment
Same item as 5.4 above

This tool provides a list of 12 questions to aid the critical appraisal of cohort studies. The tool is split into three sections: 'are the results of the study valid?' (7 items); 'what are the results?' (2 items) and 'will the results help locally?' (2 items). The tool took 15–20 minutes to complete. It was not used in our sample of systematic reviews, and in its current format would be difficult to apply in such a context. The questions prompt thinking about quality but were felt to require subjective answers that are unlikely to be consistent within or between reviewers. The questions are followed by 'hints' to help the reader answer the overall questions, but in some cases it is difficult to work out how the responses to the hints translate into yes/no/can't tell. For example, if a reviewer answered 'yes' to two hints and 'no' to two hints for a single question, it is not clear what the overall answer to that question should be.

The CASP tool was not judged to be suitable for use in a systematic review in its current format.

DuRant, 1994[99]

The types of questions which led to the selection of this tool according to our pre-selected **core** criteria were:

5.3 How allocation occurred
How were subjects assigned to experimental groups?
5.4 Any attempt to balance groups by design
Item not covered
9.2 Identification of prognostic factors
As for 9.3 below
9.3 Case-mix adjustment
Were appropriate variables or factors controlled for or blocked during the analysis? Were other potentially confounding variables handled appropriately?

This tool provides a list of 103 questions to aid the evaluation of research articles. The tool covers several study designs: experimental or quasi-experimental, survey designs and cross-sectional studies, retrospective chart reviews and retrospective studies and case–control studies. Some of the sections are general across all designs: 'introduction' (8 items); 'methods and procedures' (3 items); 'results' (12 items); and 'discussion' (9 items). Additional sections relevant to non-randomised intervention studies are: 'experimental or quasi-experimental designs' (26 items) and 'statistical analysis for experimental designs' (5 items). The tool took 15–25 minutes to complete. Again, it was not designed for use in a systematic review and although it covers the majority of issues relating to internal validity, the tool does not force the reader to answer the questions in a systematic manner. It is also very long and includes several irrelevant items. This tool is really a critical appraisal tool designed to prompt thinking regarding quality and does not provide a means of comparing the quality of studies included in a review.

The DuRant tool was not judged to be suitable for use in a systematic review.

Fowkes and Fulton, 1991[107]

The types of questions which led to the selection of this tool according to our pre-selected **core** criteria were:

5.3 How allocation occurred
Item not covered
5.4 Any attempt to balance groups by design
Control group acceptable? Matching or randomisation
9.2 Identification of prognostic factors
Distorting influences? Confounding factors
9.3 Case-mix adjustment
Distorting influences? Distortion reduced by analysis

This tool provides a list of six questions as a checklist for appraising a medical article. The tool aims to cover several study designs including cross-sectional, cohort, controlled trials and case–control studies. Each of the six questions lists items to be scored 'major problem', minor problem' or 'no problem'. The tool took around 10 minutes to complete. This is another example of a checklist that was designed to prompt thinking about quality but does not permit the systematic assessment of quality across studies. Limited detail of what the items mean is provided in the actual checklists and the entire paper needs to be read to provide any guidance.

The Fowkes tool was not judged to be suitable for use in a systematic review.

Hadorn and colleagues, 1996[102]

The types of questions which led to the selection of this tool according to our pre-selected **core** criteria were:

5.3 How allocation occurred
For cohort studies, the study groups were not treated concurrently
5.4 Any attempt to balance groups by design
Item not covered
9.2 Identification of prognostic factors
Known prognostic factors for the outcome of interest or possible confounders were not measured at baseline
9.3 Case-mix adjustment
A significant difference was found in one or more baseline characteristics that are known prognostic factors or confounders, but no adjustments were made for this in the analysis

This tool was developed for the rating of evidence for clinical practice guidelines. It provides a list of eight quality assessment criteria; each criterion lists what the authors consider to be major and minor flaws. The criterion for allocation of patients to treatment groups has separate responses for RCTs and cohort or registry studies. The tool took 15–20 minutes to complete. Although the tool covered a number of validity issues, it was found to be fairly difficult to use in its published format and perhaps concentrated overly on pharmaceutical interventions.

The Hadorn tool was not judged to be suitable for use in a systematic review.

Spitzer and colleagues, 1990[105]

The types of questions which led to the selection of this tool according to our pre-selected **core** criteria were:

5.3 How allocation occurred
Suitable assembly of comparison group
5.4 Any attempt to balance groups by design
Known confounders accounted for by design
Any methods to attempt comparability between groups, other than randomisation
9.2 Identification of prognostic factors
Sample size enables adequate precision in secondary variables reported (confounding variables or incidental findings)
9.3 Case-mix adjustment
Known confounders accounted for by analysis

This tool was developed for use in an epidemiological systematic review. It contains 20 quality items and takes 10–20 minutes to complete. It was used as a quality assessment tool in one of our sample of systematic reviews.[97] It scores reasonably well on the required quality items and can be applied to different types of study design; however, some of the questions are rather subjective and/or ambiguous (for example, 'Selection bias accounted for') and some quality issues appear to be covered by more than one item. Very little guidance for completion of the tool is provided. Several response options are provided, but it is difficult to differentiate these from each other (e.g. options include Uncertain/Incomplete/Substandard, Don't know/Not reported, N/A and N/C).

Overall, the Spitzer tool was not judged to be suitable for use in a systematic review.

Vickers, 1995[110]

The types of questions which led to the selection of this tool according to our pre-selected **core** criteria were:

5.3 How allocation occurred
Item not covered
5.4 Any attempt to balance groups by design
Adequacy of baseline matching
9.2 Identification of prognostic factors
All relevant variables included in baseline matching
9.3 Case-mix adjustment
Stratification/multivariate analysis adequate and appropriate?

This checklist is a tool for critical appraisal, containing 21 items, several of which have further sub-questions. The tool takes up to 20 minutes to complete. It covers a lot of validity issues, but some of the wording is rather ambiguous. This tool was designed to aid critical appraisal and prompts thinking about quality rather than providing a tool for systematic quality assessment. It would be difficult to use within the context of a systematic review.

The Vickers tool was not judged suitable for use in a systematic review.

Weintraub, 1982[108]

The types of questions which led to the selection of this tool according to our pre-selected **core** criteria were:

5.3 How allocation occurred
Concurrent or historical controls
5.4 Any attempt to balance groups by design
Assignment of treatment: randomised, matched or stratification or minimisation
9.2 Identification of prognostic factors
Comparability of treatment groups on specific criteria
9.3 Case-mix adjustment
Item not covered

This tool was developed to evaluate critically reports of clinical drug trials and to aid in the writing of protocols and papers. It contains 47 items split into seven categories: population, treatments, experimental design details, compliance, data collection, control of bias and data analysis. The checklist takes 10–15 minutes to complete and is aimed at studies of any design that evaluate drug therapy. The checklist is not structured in such a way that responses to the questions are systematic and the items do not really force the reader to make any quality judgements. There is a final section where an overall assessment of study quality should be made; however, it is not clear how this assessment should follow from the preceding items.

The Weintraub tool is not a suitable quality assessment tool for use in systematic reviews in its present format.

Appendix 5

Coverage of core items by selected top tools

Cowley, 1995[109]

5.3 How allocation occurred
Method of assignment of patients to different prostheses described
5.4 Any attempt to balance groups by design
Patient groups matched for diagnoses, age, and illness grade or indicators of activity level, sex, and/or weight, or effect of any differences evaluated in valid statistical analysis
9.2 Identification of prognostic factors
Item not covered
9.3 Case-mix adjustment
Covered by question under item 5.4 above

Downs and Black, 1998[85]

5.3 How allocation occurred
Were the patients in different intervention groups, or were the cases and controls recruited from the same population?
Were study subjects in different intervention groups, or were the cases and controls recruited over the same period of time?
5.4 Any attempt to balance groups by design
Item not covered
9.2 Identification of prognostic factors
Are the distributions of principal confounders in each group of subjects to be compared clearly described?
9.3 Case-mix adjustment
Was there adequate adjustment for confounding in the analyses from which the main findings were drawn?

Newcastle–Ottawa tool[66]

5.3 How allocation occurred
Item not covered
5.4 Any attempt to balance groups by design
Comparability of cohorts on the basis of design or analysis
9.2 Identification of prognostic factors
Study controls for most important prognostic factors
9.3 Case-mix adjustment
Comparability of cohorts on the basis of design or analysis

Reisch and colleagues, 1989[111]

5.3 How allocation occurred
Several items pertaining to allocation were included, such as, 'historical controls' or 'convenience (subjects selected for availability)' and 'other non-randomised method'
5.4 Any attempt to balance groups by design
Use of either prognostic stratification prior to study entry or retrospective stratification during data analyses
9.2 Identification of prognostic factors
Item not covered
9.3 Case-mix adjustment
Use of either prognostic stratification prior to study entry or retrospective stratification during data analyses

Thomas[65]

5.3 How allocation occurred
Item not covered
5.4 Any attempt to balance groups by design
Indicate the percentage of relevant confounders that were controlled in the design
9.2 Identification of prognostic factors
Were there important differences between groups prior to the intervention (a list of important confounding factors was provided)?
9.3 Case-mix adjustment
Indicate the percentage of relevant confounders that were controlled in the analysis

Zaza and colleagues, 2000[86]

5.3 How allocation occurred
Item not covered

5.4 Any attempt to balance groups by design
Considering the study design, were appropriate methods for controlling confounding variables and limiting potential biases used?

9.2 Identification of prognostic factors
Did the authors identify and discuss potential biases or unmeasured/contextual confounders, etc.?

9.3 Case-mix adjustment
As 5.4 above

Appendix 6

Details of review methods

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
ACCP/AACVPR, 1997[393]	Nar	Rehab	Pulmonary disease	76	49	?	0	LoE	US Task Force[377,378]	Nar
Anders, 1996[81]	MA	Vacc	Measles	10	0	10?	0	S	Author's own	Mean score
Aronson, 1996[249]	MA	Pharm	Depression	8	4	4	0	S	Detsky, 1992[92]	Itemised; Nar
Audet, 1993[176]	Nar	Educ	Medical students	10	0	10	0	S	Modified Poynard, 1988[403]	Itemised; OS
Baker, 1994[82]	Nar	Surgery	Coronary disease	38	0	7	31	C	Author's own	Itemised
Bass, 1993[88]	Nar	Behav	Childhood injury	20	7	10	3	S	Author's own	OS; ranked by score
Bauman, 1997[208]	Nar	Psych	Chronic conditions	15	11	4	0	C	Author's own	Itemised (most items)
Berard, 1997[209]	MA	Physio	Osteoporosis	18	5	13	0	S	Modified Chalmers, 1981[51]	Mean score and range
Beyer, 1998[210]	MA	Vacc	Influenza	24	22	2	0	S	Author's own	Itemised
Boers, 1991[212]	Nar	Pharm	Rheumatoid arthritis	7	6	1	0	C	Modified Sackett, 1985[349]	Itemised
Bours, 1998[73]	Nar	Org	Aftercare of elderly	17	6	11	0	S	Modified van der Windt, 1995[71]	Itemised and OS
Brown, 1988[254]	MA	Educ	Diabetes	47	NS	NS	0	C	Duffy, 1985[89]	Mean score; no. with 13 threats to validity presented in Brown, 1990[215]
Brown, 1990[215]	MA	Educ	Diabetes	82	NS	NS	0	S	Brown, 1991[214]	Not reported
Brown, 1995[216]	MA	Org	Primary care	53	14	39	0	C	Author's own	Not reported
Brown, 1996[217]	MA	Mixed	Diabetes	89	?	?	0	S	Modified Brown, 1991[214]	Mean score and range
Burckhardt, 1987[362]	MA	Psych	Elderly	41	23	18	0	C	Smith and Glass, 1977[361]	Nar
Callahan, 1991[218]	MA	Pharm	Stable chronic obstructive pulmonary disease	14	11	4	0	C	Author's own	Itemised
Callahan, 1995[264]	MA	Surg	Knee arthroplasty	64	0	64?	0	NR	Gartland 1988[124]	Not reported
Cameron, 2000[83]	MA	Rehab	Fractures in older people	41	14	27	0	S	Author's own	OS
Camma, 1996[335]	MA	Pharm	Hepatitis C	9	5	4	0	S	Nicolucci, 1989[77]	Mean score
Cao, 1997[226]	MA	Pharm	Bancroftian filariasis	15	9	6	0	S	Modified Chalmers, 1981[51]	Not reported
Carlson, 1996[227]	MA	Occ ther	Well-being of older persons	14	6	8	0	C	Author's own	Nar
Carter, 1994[178]	Nar	Alt med	Pre-menstrual syndrome	14	?	?	0	S	Modified Chalmers, 1981[51]	OS

continued

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
Catre, 1997[222]	Nar	Surg	Spinal surgery	15	1	6	8	C	Canadian Task Force 1994[220]	Categorised by level of evidence
Chaulk, 1998[240]	Nar	Other	Pulmonary tuberculosis	27	5	3	22	NR	Cook, 1992[238]	Level of evidence
Cheatham, 1995[230]	MA	Surg	Nasogastric decompress	26	15	2	9	S	Modified Chalmers, 1981[51]	OS
Cheek, 1998[130]	Nar	Surg	Groin hernia	71	45	26	0	S	Downs, 1996[112]	Itemised
Choi, 1995[352]	MA	Pharm	Bed nets for malaria	10	7	3	0	S	Schmid, 1991[121]	Not reported
Ciliska, 1995[351]	Nar	Pharm	Obesity in NIDDM[e]	10	8	2	0	C	Sackett and Haynes, 1985[349]	Classed strong/ moderate/weak
Ciliska, 1996[231]	Nar	Org	Home visiting	77	NS	NS	0	C	Author's own	Itemised
Cohen, 1994[236]	Nar	Educ	Low back pain	13	9	4	0	S	Author's own	No. scoring <5, 5–7 and 9.5 reported
Colditz, 1994[237]	MA	Vacc	Tuberculosis	25	7	8	10	S	Author's own	Not reported
Cole, 1993[327]	Nar	Org	Geriatric patients	10	7	3	0	C	McMaster, 1981[80]	Not reported
Cole, 1996[185]	Nar	Behav	Delirium	10	3	7	0	S	Guyatt, 1993[79], 1994[78]	Itemised
Coleridge Smith, 1999[97]	Nar	Mixed	Wound care	263	NS	NS	0	C	Modified various	Not reported
Collingsworth, 1997[260]	Nar	Org	Medication admin.	11	1	5	5	C	Ganong, 1987[123]	Nar
Cooper, 1994[313]	MA	Pharm	Respiratory tract infection	12	?	12	0	S	Levine, 1980[190]	OS, then categorised A–C
Cowley, 1995[109]	Nar	Other	Total hip replacement	81	8	17	56	C	Author's own	% meeting each criterion
Cox, 1989[242]	Nar	Psych	Bulimia	32	NS	10	22	S	Modified Gurman, 1978[184]	OS & no. meeting each criterion
Cuijpers, 1998[243]	MA	Psych	Depression in the elderly	14	12	2	0	C	Author's own	Itemised
Curatolo, 1998[248]	MA	Pharm	Intraoperative analgesia	26	25	1	0	S	Detsky, 1992[92]	Mean score
de Craen, 1996[244]	MA	Pharm	Pain control	24	22	2	0	S	Modified ter Riet, 1990[68]	Itemised; OS
de Kruif, 1996[245]	Nar	Physio	Stress urinary incontinence	10	4	2	4	C	Author's own	Nar

continued

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
De Oliveira, 1996[246]	MA	Pharm	Schizophrenia	18	?	?	0	S	de Oliveira, 1995[96]	NR; all scored above a certain threshold
Detsky, 1987[309]	MA	Diet	Surgical patients	18	11	7	0	S	L'Abbe, 1987[120]	OS
Devine, 1992[250]	MA	Psych	Surgical patients	173	119	54	0	C	Author's own	Not reported
Devine, 1983[251]	MA	Psych	Postsurgical patients	49	37	12	0	C	Author's own	No. meeting certain criteria
Devine, 1995[252]	MA	Psych-ed	Hypertension	88	51	37	0	NR	Author's own	No. using each method of assignment
Devine, 1995[253]	MA	Psych	Cancer	98	67	31	0	C	Author's own	Not reported
Dodhia, 1998[368]	Nar	Pharm	Pertussis transmission	13	2	4	7	LoE	Stevens and Raftery, 1994[367]	Not reported
Downs, 1996[112]	Nar	Surg	Stress incontinence	76	11	65	0	S	Downs, 1998[85]	Itemised
Durlak, 1997[255]	MA	Prim prev	Mental health	177	108	69	0	C	Author's own	No. meeting each criteria reported
Eastwood, 1996[394]	Nar	Org	Asthma care	27	9	4	14	C	CRD guidelines[119]	Nar
EPI Centre, 1996[338]	Nar	H promo	Sexual health	110	34	NS	NS	C	Oakley, 1994[128]	No. each criteria reported
Ernst, 1998[287]	Nar	Alt med	Dental pain	16	11	5	0	S	Jadad, 1996[46]	Itemised; OS
Ernst, 1998[302]	Nar	Alt med	Muscle soreness	8	3	5	0	S	Kleijnen, 1991[69]	OS
Fiscella, 1995[267]	Nar	Other	Pregnancy	50	11		12	C	Gordis, 1990[101]	Nar
Flint, 1995[326]	Nar	Respite care	Dementia	4	2	2	0	C	McMaster, 1981[80]	Nar
Floyd, 1997[371]	MA	Psych	Mental health	25	?	?	0	S	Suydam, 1968[370]	Not reported
Foxcroft, 1997[340]	Nar	Psych	Alcohol misuse	33	?	?	0	C	Oakley, 1994[128]	% meeting each criterion reported
Gam, 1995[229]	MA	Physio	Musculoskeletal disorders	22	9	13	0	C	Chalmers, 1981[51]	% meeting each criterion reported
Gansevoort, 1995[261]	MA	Pharm	Hypertension	41	20	21	0	C	Author's own	Itemised
Garg, 1998[263]	Nar	Pharm	Myocarditis	7	3	4	0	C	Author's own	Itemised
Glantz, 1997[182]	MA	Other	Obstetrics	5	2	3	0	S	Modified Chalmers, 1981[51]	Itemised; OS

continued

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
Good, 1996[266]	Nar	Psych	Postoperative pain	21	11	8	2	C	Author's own	Nar
Griffiths, 1997[225]	MA	Other	Ulcerative colitis	12	0	11	1	LoE	Canadian Task Force, 1994[220]	
Grimes, 1997[383]	Nar	Pharm	Pregnancy termination	27	7	3	17	LoE	US Task Force, 1996[377]	Categorised by level of evidence
Grullon, 1997[382]	Nar	Org	Obstetrics	18	5	10	13	LoE	US Task Force, 1996[377]	Nar
Gyorkos, 1994[269]	MA	Vacc	Various	54	11	43	0	C	Gyorkos, 1994[127]	OS; classed strong/ moderate/ weak
Hancock, 1990–1995[272]	Nar	H promo	N/A	13	3	9	1	C	Author's own	Itemised
Hayes, 1992[204]	Nar	Pharm	Periodontitis	13	9	4	0	S	Antczak, 1986[76]	% meeting each criterion reported
Hearn, 1998[395]	Nar	Pall care	Cancer	18	5	14	0	LoE	Cancer Guidance (US Task Force?[377])	Classified by level of evidence
Helfenstein, 1994[278]	MA	Dental	Caries prevention	8	8	4	0	C	Author's own	Itemised
Heneghan, 1996[279]	Nar	Other	Social work	10	5	5	0	S	Modified Chalmers, 1981[51]	Itemised; OS
Heyland, 1993[350]	Nar	Diet	Critically ill	39	?	?	?	LoE	Sackett, 1989[126]	Some categorisation by level of evidence
Hoag, 1997[282]	MA	Psych	Dysfunctional behaviour	56	NS	NS	0	C	Author's own	No. meeting each criteria reported
Hoffman, 1993[283]	Nar	Surg	Herniated lumber discs	81	2	17	62	C	Author's own	No. with each criterion reported; categorised by level of evidence
Howell, 1997[284]	MA	Behav/ Educ	Cholesterol reduction	224	?	?	0	C	Author's own	Nar; % low/medium/ high quality
Jabbour, 1996[285]	Nar	Educ	Life support	17	5	8	4	S	Author's own	OS
Johnston, 1994[289]	Nar	Org	Clinician performance	28	24	4	0	S	Author's own	Itemised; OS
Kasiske, 1995[291]	MA	Pharm	Hypertension	474	NS	NS	NS	Not reported	Author's own	Not reported
Kasiske, 1993[293]	MA	Pharm	Renal transplant	21	12	9	0	C	Author's own	Nar
Kasiske, 1998[292]	MA	Diet	Renal function	24	13	11	0	C	Author's own	Itemised

continued

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
Kay, 1996[186]	MA	Educ	Oral health	37	7	30	0	S	Author's own	Not reported
Kingery, 1997[297]	MA	Pharm	Pain syndromes	50	NS	NS	NS	S	Modified Koes, 1991[70]	OS
Kinney, 1996[298]	MA	Mixed	Cardiac patients	84	54	30	0	S	Modified Smith, 1991[360]	Mean/modal score and range
Klassen, 1995[299]	Nar	Pharm	Depression in Parkinson's disease	12	9	3	0	S	Modified ter Riet, 1990[68]	Itemised; OS
Kleijnen, 1995[301]	Nar	Other	Multiple sclerosis	14	11	3	0	S	Author's own	Itemised; OS
Kleijnen, 1994[300]	Nar	Alt med	Cancer	11	6	5	0	S	Author's own Kleijnen, 1991[69]	Itemised; OS
Kleijnen, 1991[69]	Nar	Alt med	Any	107	68	39	0	S	Author's own	OS; Itemised better quality ones
Krause, 1998[306]	Nar	Rehab	Occupational health	21	1	6	6	S	Kristensen, 1989[305]	Nar; OS
Kreulen, 1998[187]	MA	Dental	Veneer restorations	14			14	S	Modified Hayes, 1992[204]	OS; Mean/modal score per item
Krywanio, 1994[363]	MA	Other	Neonates in NICU	39	25	14	0	S	Smith, 1988[125]	OS; mean score and SD
Kulik, 1990[308]	MA	Educ	N/A	103*	28	75	0	C	Author's own	No. meeting each criterion reported
Kwakkel, 1997[188]	MA	Rehab	Stroke	9	8	1	0	S	Author's own	Itemised; Nar
Lee, 1997[189]	MA	Other	Seasonal affective disorder	40	?	?	0	S	Cook, 1979[50]	Not reported
Lee, 1994[312]	Nar	Psych	Divorce	15	11	4	0	S	Author's own	% meeting each criterion reported
Lijesen, 1995[303]	Nar	Pharm	Obesity	24	14	10	0	S	Koes, 1991[70]	Itemised; OS
Linde, 1998[314]	MA	Alt med	Various	32	23	9	0	S	Linde, 1999[95]	Itemised; OS
Littenberg, 1998[316]	MA	Surg	Tibial shaft fracture	19			13	S	Author's own	Mean/median scores and range
Loblaw, 1998[223]	Nar	Mixed	Spinal cord compression	24	?	?	?	LoE	Canadian Task Force, 1994[220]	Categorised by level of evidence
Long, 1995[380]	Nar	Other	Premature infants	31	11	20	0	LoE	US Task Force, 1989[378]	Categorised by level of evidence
Lucassen, 1998[286]	MA	Mixed	Infantile colic	27	NS	NS	0	S	Jadad, 1996[46]	OS

continued

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
Lyons, 1991[319]	MA	Behav	Psychological conditions	70	NS	NS	0	C	Modified Smith, 1980[364]	% of comparisons in each category
MacLehose, 2000[26]	MA	Screening/ other	Breast cancer/neural tube defects	34	12	11	11	S	Modified Downs, 1998[85]	Mean score per group of studies according to 4 dimensions of bias
MacMillan, 1994[396]	Nar	Psych	Child sexual abuse	19	14	5	0	S	Modified Chalmers, 1981[51]	Itemised
MacMillan, 1994[191]	Nar	Social work	Child phys abuse/neglect	11	6	5	0	S	Modified Chalmers, 1981[51]	Itemised
Margolis, 1995[322]	Nar	Org	Discharge of newborns	13	3	10	0	C	Author's own	Not reported
Marrs, 1995[323]	MA	Other	Psychological problems	70	59	11	0	C	Author's own	% meeting certain criteria reported
Massy, 1995[324]	MA	Pharm	Renal disease	154	NS	NS	NS	S	Author's own	Not reported
Mathews, 1996[397]	Nar	Diet	Pregnancy	27	8	19?	0	C	Can't tell	Nar
Mayo-Smith, 1997[241]	MA	Pharm	Alcohol withdrawal	65	NS	NS	0	C	Cook, 1992[238]	Not reported
Maziak, 1998[192]	Nar	Surg	Intensive care unit patients	5	3	2	0	S	Author's own	Itemised
McAweeney, 1997[325]	MA	Psych	Spinal cord injury	11			0	S	Author's own	Mean score per item
McCusker, 1998[328]	MA	Mixed	Depression in older people	37	NS	NS	0	S	Modified Chalmers, 1981[51]	Mean score and SD
Melchart, 1994[93]	Nar	Alt med	Immuno-modulation	26	16	10	0	S	Author's own and Detsky, 1992[92]	Itemised; OS
Miller, 1995[87]	Nar	Mixed	Alcoholism	211	159	52	0	S	Author's own	Itemised
Morin, 1994[330]	MA	Behav	Insomnia	59	NS	NS	0	S	Author's own	Not reported
Mullen, 1985[275]	MA	Educ	Chronic illness	70	43	27	0	S	Haynes, 1976[273]	OS
Mullen, 1992[274]	MA	Educ	Coronary heart disease	38	14	24	0	NR	Modified Haynes, 1976[273]	No. meeting certain criteria reported
Mulrow, 1986[197]	Nar	Other	Diabetic retinopathy	15	3	1	11	C	Author's own	Itemised
Murtagh, 1995[332]	MA	Psych	Insomnia	66	?	?	0	C	Author's own	Not reported
Naylor, 1987[334]	MA	Pharm	Renal failure	5	4	1	0	S	Modified Chalmers, 1981[51]	OS
NHS CRD, 1996[119]	Nar	Mixed	Reduction of injuries	49	22	27	0	C	CRD guidelines	Nar
NHS CRD, 1995[391]	Nar	Public health	N/A	94	38	56	0	C	Woolf, 1990[390]	Not reported

continued

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
Norman, 1998[205]	MA	Org		10	3	7	0	S	Audet, 1993[176]	Nar
Oakley, 1995[339]	Nar	Educ	Sexual health	65	20		0	C	Oakley, 1994[128]	% with certain quality attributes
Oakley,1994[128]	Nar	Educ	HIV/AIDS	68	28	40	0	C	Author's own	No. meeting certain criteria
Oakley, 1995[234]	Nar	Behav	HIV/AIDS	68	41	27	0	C	Cochrane, 1993[118]	% meeting each criterion reported
Ogilvie Harris, 1995[341]	Nar	Mixed	Soft tissue ankle injuries	84	58	14	12	S	Modified Weiler, 1992[388]	Itemised; OS
Oxman, 1994[343]	Nar	Org	Service delivery	102	NS	NS	0	C	Author's own	Not reported
Oxman, 1994[271]	Nar	H promo	STDs	13	6	7	0	S	Gyorkos, 1994[127]	OS
Parker, 1998[398]	MA	Pharm	Ectopic pregnancy	9	0	?	?	NR	Modified Chalmers, 1981[51]	Not reported
Ploeg, 1996[232]	Nar	Public health	Adolescent suicide	11	0	11	0	C	Ciliska, 1996[231]	Itemised
Powe, 1994[399]	MA	Surg	Cataract extraction	90	0	7	83	S	Author's own Powe, 1994[344]	OS
Puig Barbera, 1995[345]	MA	Vacc	Influenza	8	1		7	C	Author's own	Itemised
Rey, 1997[346]	Nar	Psych	Psychiatric disorders	60	0	1	59	S	Author's own	Nar
Riben, 1994[270]	Nar	Educ	Public health	13	NS	NS	0	S	Gyorkos, 1994[127]	OS
Richard, 1997[224]	Nar	Org	Colorectal cancer	17	4	5	8	LoE	Canadian Task Force, 1994[220]	Categorised by level of evidence
Ried, 1994[347]	MA	Pharm	Mountain sickness	20	?	?	0	S	Modified Chalmers, 1981[51]	OS
Robert, 1997[379]	Nar	Org	Physical therapy	8	1	5	2	LoE	US Task Force, 1989[378]	Classified by level of evidence
Rowe, 1997[348]	MA	Surg	Idiopathic sclerosis	20		1	0	S	Author's own	Not reported
Saint, 1998[288]	MA	Other	Urinary tract infection	8	6	2	0	C	Jadad, 1996[46]	Itemised
Salisbury, 1997[198]	Nar	Org	N/A	40	7	13	20	C	Author's own	Nar

continued

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
Schmidt-Nowara, 1995[239]	Nar	Other	Obstructive sleep apnoea	21	0	21	0	LoE	Cook, 1992[238]	Categorised by level of evidence
Schoemaker, 1997[358]	Nar	Psych	Anorexia nervosa	6		6?	0	C	Slavin, 1995[122]	Itemised
Sellers, 1997[353]	MA	Educ	Public health	7	0	7	0	C	Author's own	Itemised
Selley, 1997[354]	Nar	Surg	Prostate cancer	unclear			0	C	Author's own	Not reported
Simons, 1996[356]	MA	Surg	Inguinal hernia	11	8	3	0	S	Author's own	Itemised; OS
Smeenk, 1998[359]	Nar	Pall care	Cancer	8	5	3	0	S	Can't tell	Itemised
Snowdon, 1997[400]	Nar	Other	Pre-school vision	11	5	6	0	NR	Can't tell	Methodological shortcomings presented
Solomon, 1997[366]	Nar	Org	Rheumatic/ musculoskeletal conditions	16	1	12	3	C	Author's own	Itemised
Sullivan, 1995[290]	Nar	Org	GP computing	30	19	11	0	S	Johnston, 1994[289]	Itemised; OS
Talley, 1996[401]	Nar	Psych	Irritable bowel syndrome	14	13	1?	0	C	Author's own Talley, 1993[106]	Itemised
Tangkanakul, 1997[372]	MA	Pharm	Carotid end-arterectomy	20	17	3	0	C	Author's own	Nar
ter Riet, 1990[68]	Nar	Alt med	Chronic pain	51	40	11	0	S	Author's own	Itemised
Theis, 1997[373]	Nar	Pharm	Neonatal abstinence syndrome	14	8	6	0	C	Author's own	Nar
Thomas, 1995[375]	Nar	Physio	Pulmonary disease/CF	17	2	7	0	S	Author's own	OS
Towler, 1998[235]	MA	Other	Colorectal cancer	6	4	2	0	C	Cochrane Handbook, 1987[118]	Itemised
van Balkom, 1997[384]	MA	Mixed	Panic disorder	106	?	?	0	S	Modified Chalmers, 1981[51]	Mean score
Verhagen, 1997[304]	Nar	Alt med	Arthritis	14	7	7	0	S	Koes, 1991[70]	Itemised
Vickers, 1996[201]	Nar	Alt med	Nausea/vomiting	33	25?	8	0	C	Author's own	Itemised
Vickers, 1994[386]	Nar	Alt med	Irritable bowel syndrome	17	4	5	4	C	Author's own	Nar
Villalobos, 1998[402]	MA	Pharm	Meningitis	12	2	10	0	S	L'Abbe, 1987[120]	Not reported
Watt, 1998[221]	Nar	Behav	Not clear	11	7	4	0	NR	Canadian Task Force, 1994[220]	OS; level of evidence provided

continued

Author	Review type[a]	IV type[b]	Topic	Total	RCTs	NRS	Other	Type of tool[c]	QA Tool	QA results[d]
Wells-Parker, 1995[129]	MA	Behav	Drink/drive offenders	194	NS	NS	0	S	Author's own	No. scoring in certain ranges
Wingood, 1996[202]	Nar	Behav	HIV/AIDS	7	5	2	0	C	Modified Oakley, 1994[128]	Itemised; Nar
Witlin, 1998[381]	Nar	Pharm	Pre-eclampsia and Eclampsia	18	14	4?	0	LoE	US Task Force, 1989[378]	OS; level of evidence
Wright, 1995[203]	Nar	Other	Obstructive sleep apnoea	23	1	22	0	C	Author's own	Nar

[a] MA, meta-analysis; Nar, narrative.
[b] Rehab, rehabilitation; Vacc, vaccination; Diet, dietary; Pharm, pharmaceutical; Surg, surgery; Org, organisation; Educ, education; Mixed, several interventions; Psych, Psychiatry; Alt med, alternative medicine; Behav, behavioural therapy; Physio, physiotherapy; Psych-ed, psychoeducational; H promo, health promotion; Occ ther, occupational therapy; other, other types of intervention; Prim prev, primary prevention; Exerc, exercise; Pall care, palliative care.
[c] LoE, Levels of evidence; S, Scale; C, checklist; NR, not reported.
[d] Itemised, itemised per study; Nar, narrative description; OS, overall score.
[e] NIDDM = non-insulin-dependent diabetes mellitus.

Appendix 7

Results of identified systematic reviews

Review	Method of incorporating quality into synthesis	Results of quality investigation
ACCP/AACVPR, 1997[393]	*Qualitative*	Studies were classified according to level of evidence and recommendations were graded accordingly. Methodological issues were discussed narratively
Anders, 1996[81]	*Quantitative* Subgroup analysis according to size of effect; between group differences in mean quality scores investigated	The *t*-test of quality score differences was non-significant ($p = 0.25$). The authors conclude that they cannot reject the null hypothesis; no difference in the quality score between studies existed
Aronson, 1996[249]	*Quantitative* Sensitivity analysis sequentially removing groups of studies with shared methodological deficiencies	Pooling all studies indicated a strongly positive effect (RR 2.09, 95% CI: 1.31 to 3.32, $p = 0.002$). Eliminating one NRS with no control group and then one RCT with a treated control group lowered the relative response to treatment (RR 1.93, 95% CI: 1.13 to 3.29, $p = 0.02$). Further elimination of an RCT with possibly contaminated randomisation increased the relative response to treatment (RR 2.70, 95% CI: 1.89 to 3.86, $p = 0.006$), and eliminated the statistical heterogeneity of the results. The three NRS which remained in the analysis were HCTs and therefore likely to have biased results
Audet, 1993[176]	*Qualitative*	Study results were discussed in terms of methodological quality. Only two studies were of sufficiently high quality to demonstrate a positive effect from teaching
Baker, 1994[82]	*Quantitative* Forest plot used to present study results ranked by quality	Two higher quality studies indicated statistically significant reductions in mortality; of the five others, three had statistically significant reductions and two were non-significant
Bass, 1993[88]	*Qualitative*	Studies grouped by design, and quality score and overall rank presented: some NRS ranked more highly than RCTs. Found reasonably high-quality evidence to support the intervention
Bauman, 1997[208]	*Qualitative*	A statement was made regarding the overall quality of included studies. Quality stated to be poor, limiting the interpretation of results
Berard, 1997[209]	*Quantitative* Correlation analysis to investigate ES and quality score	The correlation between effect size and quality score ($r = -0.07$, $p = 0.19$) or sample size ($r = 0.04$, $p = 0.63$) were not found to be significant
Beyer, 1998[210]	*Quantitative* Subgroup analysis according to quality (high/medium/low)	When all studies were pooled, a small, positive, but non-significant effect from influenza vaccine was produced (RD –0.6, 95% CI: –3 to 1.9). The two poor-quality studies (both of the NRS) showed a much more strongly protective effect from the vaccine (RD –9.4, no CIs given), and the medium- and high-quality studies (all RCTs) indicated a much more conservative effect (RD –0.3 and 1.1, respectively, no CIs given)
Boers, 1991[212]	*Qualitative*	Methodological characteristics were discussed in detail and studies were categorised and discussed according to strength of evidence. Insufficient evidence was found to make any recommendations regarding drug combinations
Bours, 1998[73]	*Quantitative* Visual plot used to map relationship between study outcome and methodological score	Three studies with only positive results had lower quality scores; those with only negative results had similar scores to those with mixed (+ve/–ve) outcomes. Studies with highest scores demonstrated no effect. Five NRS scored more highly than any RCT

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Brown, 1988[254]	*Quantitative* 38 substantive and methodological characteristics were investigated in a correlational analysis	No significant correlations with unweighted mean effect sizes found. Experimental mortality (attrition) was the only variable significantly correlated with the overall weighted mean ES, $r = -0.52$, $p = 0.002$
Brown, 1990[215]	*Quantitative* Correlational analysis of ES and quality score. Subgroup analyses according to quality (high/low) and rigour of design for five outcome measures	Correlation analysis found no relation between attrition rate or rating of research quality and mean ES estimates. Lower-quality studies produced smaller ESs for three of the outcome measures and larger ESs for the other two. For four out of five outcomes the SDs associated with non-equivalent control groups were almost twice as large as for the other designs
Brown, 1995[216]	*Not considered*	
Brown, 1996[217]	*Quantitative* Correlation of quality index and effect sizes	None of the relationships were statistically significant. Strongest relationship was with ES for behavioural strategy ($r = -0.46$, $p = 0.07$, $n = 16$); the higher the quality the lower the ES for mean weight loss. Outcome effects also partitioned according to study design: all results were statistically significant for pre-/post-test studies (approx. twofold higher), but for experimental studies, only metabolic control variables were statistically significant. No data presented
Burckhardt, 1987[362]	*Quantitative* Subgroup analysis according to high/medium validity; reliability and reactivity of the outcome measure	High internal validity, high/adequate reliability of the outcome measure, and high reactivity of outcome measures demonstrated larger ES (0.60, 0.59, and 0.80 respectively compared with overall ES of 0.53). No information was provided regarding the statistical significance of these subgroup analyses. Authors comment that the significance of study design characteristics is unclear
Callahan, 1995[264]	*Quantitative* Correlation analysis used to determine if study characteristics were correlated with outcomes	No significant correlations were with publication year, mean follow-up, sample size, or % lost to FU.
Callahan, 1991[218]	*Quantitative* Only studies meeting all quality criteria included in main analysis; qualifying criteria were sequentially relaxed	Inclusion of lower quality studies marginally increased the overall effect and slightly narrowed the CIs. Results were not presented for low-quality studies only. Treatment ES was also examined visually in relation to sample size for studies meeting all quality criteria and the five studies meeting some criteria. The inclusion of lower quality studies did not significantly alter the appearance of the graph
Cameron, 2000[83]	*Qualitative*	Little methodological discussion on included studies; however authors conclude that "the available data have significant limitations in quantity and quality and allow only tentative conclusions", namely that there is limited evidence for the intervention in terms of hospital admission and residential status following discharge. Main analyses focus on comparing RCT and NRS results
Camma, 1996[335]	*Qualitative*	Methodological limitations of studies were discussed. Quality score was not used in the meta-analysis. Authors comment that quality was reasonably good.
Cao, 1997[226]	*Quantitative* Quality weighting	Weighting effect size by both quality score and sample size made essentially no difference to the overall results. Only sample size weighted results presented

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Carlson, 1996[227]	*Quantitative* Correlation analysis between ES and six potential moderator variables (year of publication, study design, residency of subjects, mean age, sample size and duration of treatment)	All correlations were non-significant. Less than 1% of the ES variability was associated with the quality of study design variable ($r^2 = 0.003$)
Carter, 1994[178]	*Qualitative*	Results used to indicate that existing research is methodologically weak
Catre, 1997[222]	*Not considered*	
Chaulk, 1998[240]	*Not considered*	
Cheatham, 1995[230]	*Quantitative* Subgroup analysis of high quality studies (15 RCTs and five case–control studies) alone for 15 outcome variables	In most cases, high-quality studies produced higher relative risk estimates, although the magnitude of the difference in effect varied. In three cases restriction to high quality studies changed the statistical significance of the results; in two cases a non-significant result became statistically significant, and in the other the pooled result became non-significant
Cheek, 1998[130]	*Qualitative*	Methodological limitations severely limited the conclusions that could be drawn from the literature, however, some recommendations were made
Choi, 1995[352]	*Quantitative* Subgroup analysis according to use of placebo-control group Presentation of results with and without weighting by quality score	Studies using a placebo control group generally led to higher pooled effects than those with a no treatment control group (incidence rate ratios 0.801 vs 0.626, respectively) Weighting by quality score produced a lower rate ratio compared with the unadjusted results; e.g. for placebo-controlled studies, the rate ratios fell from 0.801 (95% CI: 0.693 to 0.926) to 0.757 (95% CI: 0.611 to 0.937). For the analysis of permethrin-treated bed nets only ($n = 9$) the effect of using a placebo control group and of weighting by quality score were not as striking
Ciliska, 1996[231]	*Qualitative*	Quality (strong/weak, etc.) was discussed alongside the results of each study. Overall, despite limitations in the research, the authors conclude that sufficient high-quality studies are available to demonstrate a positive direction of effect
Ciliska, 1995[351]	*Qualitative*	Weak studies excluded from the review. Some methodological characteristics of included studies were discussed. Authors conclude that the evidence is equivocal
Cohen, 1994[236]	*Qualitative*	Only higher scoring studies were used to ascertain whether findings were related to characteristics of study participants, interventions or settings ($n = 6$). Little sound evdience was found to support the effect of group education on short-term outcomes (3-month FU), and no evidence of an effect at long-term FU (1 year)
Colditz, 1994[237]	*Quantitative* Regression analysis to examine quality score and ES	Quality explained 30% of the variation in BCG efficacy in the clinical trials; geographic altitude individually explained 41% of the variance. The authors could not determine whether these variables were surrogates for factors such as quality of FU or presence of non-tuberculous mycobacteria in the population. Quality explained 36% of the between-study heterogeneity in the case–control studies, and was the only variable to have any impact

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Cole, 1993[327]	*Qualitative*	Quality was not discussed in terms of the results of each study, but the authors do acknowledge that serious limitations in trial design and measures were present
Cole, 1996[185]	*Quantitative* Absolute risk reductions for each individual study were plotted against quality score on a forest plot	Quality score did not appear to be associated with the RD
Coleridge Smith, 1999[97]	*Qualitative*	Results of quality assessment were used to include/exclude studies. Only those with internal validity graded as strong or moderate were included in the review of treatments. Recommendations appear to be largely based on the strength of the evdience
Collingsworth, 1997[260]	*Qualitative*	Authors concluded that "despite their methodological shortcomings the conclusions from all 12 studies were broadly in agreement and substantiated the findings from the non-empirical work"
Cooper, 1994[313]	*Quantitative* Compared high and low quality	Clinical improvement was lower in higher quality studies, but the difference was not significant. In studies of 'very good' quality a 75% clinical improvement was observed, compared with 86% for 'good' studies and 83% for 'borderline' studies ($p = 0.17$)
Cowley, 1995[109]	*Qualitative*	Studies classified as A/B/C for each study design but only the results of 'A' category studies were discussed. One RCT, four NRS and 32 uncontrolled studies were given an A rating: key results were discussed. No conclusive evidence was identified
Cox, 1989[242]	*Qualitative*	Studies grouped by quality score (high/medium/low) and results discussed in terms of quality. Three high-quality studies were equivocal regarding treatments reviewed
Cuijpers, 1998[243]	*Quantitative* Use of control group investigated in subgroup analysis	Small non-significant difference in overall ES was observed between studies with and without a control group (1.14 and 1.06)
Curatolo, 1998[248]	*Not considered*	
de Craen, 1996[244]	*Quantitative* Quality weighting	Including the quality score as an additional weighting factor in the analysis did not change the results (data not presented)
de Kruif, 1996[245]	*Qualitative*	Results of QA were discussed with reviewers' conclusions. High-quality studies appeared to show positive effect and low-quality were equivalent. Identified trend of effectiveness of myofeedback not conclusive owing to methodological limitations
de Oliveira, 1996[246]	*Quantitative* Excluded lower quality studies	All studies bar one scored >0.6. When the lower-quality study was excluded from the meta-analysis the results did not change (no data presented)
Detsky, 1987[309]	*Quantitative* Correlation analysis	Quality scores were negatively correlated with the differences in complication (correlation coefficient –0.56, $p = 0.13$) and fatality rates (coefficient –0.51, $p = 0.031$) between groups, i.e. studies with better-quality scores showed a smaller trend in favour of intervention group

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Devine, 1992[250]	*Quantitative* Regression analysis to examine relationship between threats to validity (unpublished, high internal validity, little measurement subjectivity, placebo control) and estimates of effect	No relationship was found between validity and estimates of effect on length of stay, respiratory function, postsurgical pain and psychological distress
Devine, 1983[251]	*Quantitative* Subgroup analyses were conducted according to use of physician blinding, random assignment, high vs low internal validity and use of placebo control vs usual care control	Use of physician blinding to group assignment, higher internal validity and placebo-controlled studies demonstrated higher ES Random assignment produced lower ES than non-random assignment. The statistical significance of these differences was not provided The authors conclude that the intervention is effective and does not depend on quality factors inflating the ES
Devine, 1995[252]	*Quantitative* Multiple regression analysis between ES and three measures of quality (random assignment, repeated outcome assessment, post-test measurement outside the context of relaxation)	ES values for relaxation and blood pressure were negatively correlated with all three measures of quality
Devine, 1995[253]	*Quantitative* Weighted regression analysis for ES and threats to validity (unpublished, random assignment, placebo control)	No significant relationship found for any of the seven dependent variables. No data or results presented
Dodhia, 1998[368]	*Qualitative*	Some quality issues discussed narratively alongside individual study results. Overall the level of evidence provided was judged to be II-2, providing only weak evidence to support the intervention
Downs, 1996[112]	*Qualitative*	All results were reported individually per study and impact of methodological flaws on results/conclusions of studies was discussed. Methodological quality is poor and therefore the value of surgery and the relative effectiveness of different procedures is unclear
Durlak, 1997[255]	*Quantitative* Subgroup analysis for each quality item, including randomisation, placebo control, attrition <10%, FU data collected, multiple outcome measures used and normed outcome measures used	No significant effects found except for those relating to outcome measures used; normed outcome measures and single vs multiple outcome measures lead to lower ES
Eastwood, 1996[394]	*Qualitative*	Methodological comments provided in tables and text. Overall quality of the research is poor and no conclusive evidence was found on which to base any recommendations

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
EPI Centre, 1996[338]	*Qualitative*	All studies were discussed in terms of methodology, but only the results of the 21 methodologically sound studies were discussed in detail. There was disagreement between authors and reviewers regarding effectiveness of interventions in 24% of methodologically 'sound' studies
Ernst, 1998[287]	*Qualitative*	Authors focused only on highest quality studies (both scored 3). One found some increase in pain threshold and another found no effect. Authors' concluded that there is evidence of benefit
Ernst, 1998[302]	*Qualitative*	Three higher quality studies included (all RCTs); all reported no statistically significant differences; lower quality trials indicated some benefit from homeopathy
Fiscella, 1995[267]	*Qualitative*	Evidence was examined in regard to methodological criteria. Authors conclude that current evidence does not satisfy causal criteria necessary to establish that prenatal birth care definitely improves birth outcomes
Flint, 1995[326]	*Qualitative*	Some methodological details were provided in the discussion of each individual study. Author concludes that there is little evidence for the intervention but that this may be due to methodological limitations
Floyd, 1997[371]	*Quantitative* Correlation analysis to examine relationship between study and sample characteristics and outcome	Not reported in detail; only one variable (use of pretraining) stated to be significantly correlated with effect size
Foxcroft, 1997[340]	*Qualitative*	Quality issues discussed along with study results. Most had some methodological shortcomings. No one intervention could be recommended
Gam, 1995[229]	*Quantitative* Subgroup analysis to examine effect of blinding	Impact depended on type of ES used: when d/r was used there was no influence from blinding ($p = 0.78$), while the d/s showed significant influence ($p = 0.009$) (d/r and d/s stated to be standardised effect sizes)
Gansevoort, 1995[261]	*Quantitative* Regression analysis to examine patient and study characteristics and effect on primary outcome	Study design (randomised, double blind or cross-over) made no essential difference to the results. The most important variables were related to the intervention used or clinical characteristics of the population
Garg, 1998[263]	*Not considered*	
Glantz, 1997[182]	*Quantitative* Subgroup analysis of high quality studies only	Restricting the analysis to only high-quality studies (two RCTs and one NRS) made little difference to the summary effect: the OR increased from 0.61 (95% CI: 0.50 to 0.75) to 0.66 (95% CI: 0.54 to 0.81), indicating a lower reduction in the odds of caesarean due to active management of labour. The authors also presented quality-adjusted results for RCTs only (adjusted by a factor of 2), which further reduced the odds of caesarean: OR 0.70 (95% CI: 0.57 to 0.78). The authors recognise that the larger effect demonstrated when all studies were combined may be due to the fact all three NRS were HCTs.
Good, 1996[266]	*Qualitative*	The validity of the results was questioned owing to numerous methodological problems in the studies

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Griffiths, 1997[225]	*Qualitative*	Studies were categorised by level of evidence. The quality of positive European studies is superior to the negative US studies, so that 13 recommendations could be given
Grimes, 1997[383]	*Qualitative*	Studies classified by level of evidence and grade of recommendation. Methodological limitations were not discussed in detail
Grullon, 1997[382]	*Qualitative*	Studies were discussed according to level of evidence provided. Insufficient evidence was found either to support or to condemn the safety of early postpartum discharge
Gyorkos, 1994[269]	*Quantitative* Subgroup analysis according to internal validity rating (strong/moderate/weak)	For adult influenza vaccination strong/moderate quality comparisons demonstrated a lower rate difference (RD 17.9%, 95% CI: 16.7 to 19.0%) compared with weak quality comparisons (RD 20.0%, 95% CI: 18.4 to 21.6%). For adult pneumococcal immunisation programmes, the average increase in coverage in studies rated 'strong' (two RCTs and one NRS) was 34.6% (95% CI: –0.6 to 69.8%). Three RCTs and one cohort study of moderate internal validity had an average effect of 78% (95% CI: 72.3 to 83.8%). The five 'weak' studies (including three RCTs) had a pooled effect of 18.5% (95% CI: 17.7 to 19.3%). The authors made no further comment on these results and did not suggest any alternative explanations for these results
Hancock, 1990–95[272]	*Qualitative*	Methodological discussion of included studies – mainly relatively poor. The few methodologically adequate studies show little evidence of large benefits
Hayes, 1992[204]	*Qualitative*	Methodological issues were discussed, most studies received low scores. A trend towards effectiveness was indicated but the authors conclude that this is not clinically significant
Hearn, 1998[395]	*Not considered*	
Helfenstein, 1994[278]	*Quantitative* Correlation analysis between quality items and outcome (including study duration, professional cleaning, fluoride rinsing only in control group, randomisation, blinded outcome assessment, double-blinding, attrition)	The only significant explanatory variable was study duration (coefficient of determination $r^2 = 0.71$, $p = 0.002$)
Heneghan, 1996[279]	*Qualitative*	Only acceptable studies were included in the review. Methodological assessment used to describe limitations of the included studies, rather than to interpret results of the review
Heyland, 1993[350]	*Not considered*	
Hoag, 1997[282]	*Quantitative* Correlation analysis used to examine ES and client, treatment and methodological domains	Experimenter preference for the intervention (identified from introductory section of papers) produced significantly higher ES (0.72 vs 0.30, $p = 0.05$). Neither the content nor source of outcome measure were significantly related to improvement in group treatment
Hoffman, 1993[283]	*Qualitative*	Studies examined according to Sackett's levels of evidence. Both high- and low-quality trials showed positive effects from standard discectomy on reoperation rates. Methodological flaws limit the ability to estimate rates of successful outcomes

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Howell, 1997[284]	*Quantitative* Regression analysis to investigate design characteristics such as internal validity (high, medium or low), subject location, type of diet and the average of multiple observations	No association with outcome was found
Jabbour, 1996[285]	*Qualitative*	Only studies scoring 3/7 or higher were analysed in detail. Methodological flaws were discussed and in some cases the impact on the interpretation of results was discussed, e.g. quality did not appear to affect the impact on knowledge scores
Johnston, 1994[289]	*Qualitative*	Only seven studies scored highly in terms of quality; however, the authors concluded that several sound studies provided evidence for at least some of the outcomes evaluated
Kasiske, 1995[291]	*Quantitative* Quality weighting	A multiple linear regression model was used: few differences were found whether regression was weighted by study variance, quality weight or sample size or was unweighted (data not shown). Authors conclude that with respect to the regression weighting, the results appeared to be robust
Kasiske, 1998[292]	*Not considered*	
Kasiske, 1993[293]	*Quantitative* Subgroup analysis for each of the nine quality assessment criteria (including randomisation, drop-out rate, selection bias, discontinuation of treatment)	No significant effect was found (results not reported)
Kay, 1996[186]	*Qualitative*	Score 12 or more required for inclusion in review, 15 or more for inclusion in meta-analysis. Some discussion of methodological problems provided. Narrative review indicated positive effect on knowledge/attitudes. Only seven RCTs of high-enough quality to be included in meta-analysis
Kingery, 1997[297]	*Not considered*	
Kinney, 1996[298]	*Quantitative* Subgroup analysis according to quality score	Quality score was associated with ES: higher quality = larger effects. Larger effects also produced by smaller sample size and convenience sampling. Data not presented
Klassen, 1995[299]	*Qualitative*	Studies discussed in terms of methodological flaws; most scored poor to moderate. There is no empirical evidence available on which to base a treatment plan for depression in patients with Parkinson's disease
Kleijnen, 1994[300]	*Qualitative*	Methodological limitations and implications for study results were discussed. Most trials had shortcomings on five or more quality criteria, indicating a serious chance that the results are biased. The results of two 'highest' quality trials were discussed; however, several shortcomings made it difficult to draw conclusions
Kleijnen, 1995[301]	*Qualitative*	Results based on highest quality studies; eight higher quality trials indicated that hyperbaric oxygen is not efficacious in the treatment of multiple sclerosis

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Kleijnen, 1991[69]	*Qualitative*	Concentrated on results of trials with the best methodological quality. Evidence found to be largely positive, especially in trials of lower quality
Krause, 1998[306]	*Qualitative*	Only higher scoring studies were included. A 'best' estimate of effectiveness was derived from highest ranking studies, then a range of probable estimates of effectiveness was derived from the other studies. Authors conclude that there is a suggestion of effectiveness, but methodological limitations make stronger conclusions difficult
Kreulen, 1998[187]	*Qualitative*	Methodological characteristics were discussed narratively; authors conclude that studies were well designed but that methodological characteristics could be improved
Krywanio, 1994[363]	*Quantitative:various* Correlation analysis of ES and quality score. Weighting by sample size, quality score and mean birth weight in turn	No significant correlation between ES and quality score regardless of mean birthweight, i.e. the intervention groups in high-quality studies did not gain more/less than those in lower quality studies. In one mean birth weight subgroup, only weighting by quality score had an impact on ES [reduced from 1.54 (95% CI: 1.22 to 1.86) to 1.37]. Authors comment that this is indicative of poorer quality studies with larger ES. For the other subgroup, only weighting by sample size impacted on ES [reduced from 0.51 (95% CI: 0.39 to 0.63) to 0.42], indicating that studies in this group were of higher quality
Kulik, 1990[308]	*Quantitative* Subgroup analyses according to six quality criteria, further subdivided by teaching approach, i.e. 'Keller's Personalised System of Instruction' (largely self-directed learning), or Bloom's 'Learning for Mastery' (teacher-presented and controlled)	Random assignment and teacher effects (same or different teachers teaching experimental and control groups) made little difference to ES or SE. Use of national standardised tests produced lower ESs than locally developed tests. Subjectively scored criterion tests produced lower ES in the PSI group compared with objective, machine-scoreable exams, but made little difference to the LFM group. Where control groups received the same amount of feedback as the intervention groups, the effect size was found to be smaller than where they received standard feedback
Kwakkel, 1997[188]	*Qualitative*	Methodological limitations of studies were mentioned, and considered as major confounding factors in the conclusions
Lee, 1994[312]	*Qualitative*	Methodological characteristics discussed separately from results of individual studies; few high-quality studies were identified; however, some benefits from the interventions were reported
Lee, 1997[189]	*Not considered*	
Lijesen, 1995[303]	*Qualitative*	NRS all scored <50, so only RCTs discussed in detail. Higher proportion of NRS (7/10) had a positive result compared with the RCTs (2/14). Authors conclude no evidence of benefit. No attempt to do an meta-analysis, although it looks as though it would have been possible
Linde, 1998[314]	*Quantitative* Subgroup analysis according to quality (good quality; unlikely to have major flaws; obvious minor/moderate problems; major flaws). Only RCTs included in first two categories, quasi-randomised in lower quality categories	On the whole, the pooled rate ratio gradually increased as the quality of the studies worsened: rate ratio for all studies combined, 1.66 (95% CI: 1.20 to 2.28); for good-quality RCTs, 1.12 (95% CI: 0.87 to 1.44); for RCTs 'unlikely to have major flaws', 2.44 (95% CI: 1.30 to 4.59); for studies with 'obvious minor or moderate problems' (two RCTs and two quasi-RCTs), 2.02 (95% CI: 0.44 to 9.42); and for studies with major flaws, 3.89 (95% CI: 1.75 to 8.62). Authors' state that the categorisation of studies in this way did not always correlate with the quality score given; i.e. a study with a high score may have been categorised as lower quality and vice versa, owing to the subjective judgements involved

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Littenberg, 1998[316]	*Qualitative*	Quality and results of each study were discussed individually, starting with highest scoring studies, according to use of control group. Overall quality stated to be poor, but tentative conclusions were drawn from the results
Loblaw, 1998[223]	*Qualitative*	Summary of level of evidence available for each procedure/recommendation. Few high-quality papers were found
Long, 1995[380]	*Qualitative*	Studies discussed in terms of level of evidence and grading of recommendations. Authors conclude that a greater percentage of studies using indirect positioning as an intervention had stronger evidence to support the interventions than studies with direct positioning as the intervention. Sackett's framework should not be the sole method of treatment evaluation because the infant's medical status needs to be considered prior to implementing a treatment strategy
Lucassen, 1998[286]	*Quantitative* Correlation analysis between quality score and effect	Quality score and effect measure were not correlated ($r = -0.02$; $p = 0.92$)
Lyons, 1991[319]	*Quantitative* Subgroup analysis according to quality (high/medium/low)	Studies with high internal validity demonstrated larger pooled effect size (ES 1.138, SD 1.119) compared with medium (ES 0.818, SD 0.778) or low (ES 0.811, SD 0.670) validity studies. The low-validity comparison is significantly different from the medium and high comparisons ($p < 0.05$).
MacLehose, 2000[26]	*Quantitative* Correlation analysis to investigate impact of individual components of the quality score	Neither total quality nor any component of the quality score was significantly associated with effect size
MacMillan, 1994[396]	*Qualitative*	Some methodological limitations and their implications on study results were discussed narratively. The impact on the conclusions of the review is not clear
MacMillan, 1994[191]	*Qualitative*	Methodological limitations were discussed narratively. Study outcomes were also ranked by methodological score. There does not appear to be a great deal of difference in effect according to quality
Margolis, 1995[322]	*Qualitative*	Some narrative discussion of quality was provided. Methodological weaknesses make any conclusions regarding effectiveness impossible
Marrs, 1995[323]	*Quantitative* Subgroup analyses according to individual quality criteria	No-treatment control group led to higher ES than where placebo control used. Researcher contact with subjects had little impact on ES, and unpublished studies produced slightly higher ES than published, as did randomised compared with non-randomised studies. A significant correlation was also found between subject retention in study and ES ($r = 0.268$)
Massy, 1995[324]	*Quantitative* Regression model weighted by inverse variance and study quality. Results compared with weighting by inverse variance alone, sample size and unweighted models	In each case the results of the regression models were not substantively different (data not presented)
Mathews, 1996[397]	*Qualitative*	Studies discussed narratively according to grade (only grades A or B included). Methodological weaknesses make it difficult to assess the relative importance of each intervention

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Mayo-Smith, 1997[241]	*Not considered*	
Maziak, 1998[192]	*Qualitative*	Quality issues were discussed alongside the results of each study. A lack of rigorously controlled trials was identified and the reviewed evidence provides insufficient information to permit a reasonable conclusion to be made
McAweeney, 1997[325]	*Not considered*	
McCusker, 1998[328]	*Quantitative* Quality weighting	Results stated to be similar for quality-adjusted and unadjusted effect measures. Only quality-adjusted results presented
Melchart, 1994[93]	*Qualitative*	Methodological issues and results discussed. Reviewer's estimate of strength of evidence provided by each study was presented. Inconclusive or limited evidence of no efficacy was provided by 21 comparisons and limited or good evidence of efficacy was provided by 10 comparisons
Miller, 1995[87]	*Quantitative* Correlation analysis of quality score and 'outcome logic scores' (measure of the strength of support for treatment efficacy)	No significant relation between these codes and methodological quality score was found ($r = 0.06$). Quality score was also found to be unrelated to mean problem severity and to estimated cost of delivering the treatment
Morin, 1994[330]	*Quantitative* Correlations between outcome and quality of study design were calculated	Design quality was negatively correlated with ES for sleep onset latency ($r = –0.24$, df = 91, $p < 0.03$) only, indicating that lower quality studies produced larger ESs
Mullen, 1985[275]	*Quantitative* Regression analysis to examine effect of study design and type of comparison group	No impact on ES found
Mullen, 1992[274]	*Qualitative*	Some narrative discussion of methodological characteristics. Pre-/post-test studies were excluded from the meta-analysis owing to methodological limitations. Further implications of methodological flaws not discussed in any detail
Mulrow, 1986[197]	*Qualitative*	Some quality issues were discussed alongside study results. Authors conclude that methodological limitations do not allow many definitive conclusions but that some new and useful information was obtained
Murtagh, 1995[332]	*Quantitative* Design variables used as inputs into a regression to quantify impact on effect sizes (including source of sample, study design, use of objective validation)	Most effect were non-significant – only the source of participants, non-use of drugs and placebo treatment contributed significantly to the prediction of the dependent variable
Naylor, 1987[334]	*Quantitative* Sensitivity analysis conducted using weighting by quality score. Weighted and unweighted *p*-value for the difference between groups was provided	When all studies were combined (3 RCTs and 1 NRS), quality weighting abolished the significant treatment effect for both overall survival from acute episode and overall survival to discharge from hospital. (Inclusion of the NRS led to larger treatment effects). Authors conclude that despite the (unweighted) significant findings, there is insufficient evidence for the intervention

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
NHS CRD, 1996[119]	*Qualitative*	Only good-quality CTs and case series were presented in the tables. Some methodological comment also provided in the text. It was concluded that the quality of the research needs to be improved
NHS CRD, 1995[391]	*Not considered*	
Norman, 1998[205]	*Qualitative*	Three studies were rejected owing to low quality; the other seven studies were considered to have relatively minor methodological problems. The interventions were found to be effective at undergraduate level, but not at residency level
Oakley, 1995[339]	*Qualitative*	Only studies judged 'sound' were used to evaluate effectiveness. Seven sound studies were reviewed of which two showed short-term effects on reported sexual behaviour. No evidence to indicate that information provision leads to risk-taking behaviour, but some indication that it may encourage sexual experimentation
Oakley, 1994[128]	*Qualitative*	Characteristics of 'sound' studies were discussed in more detail. The effectiveness of interventions was examined for both sound and flawed studies. Overall, only a small proportion of interventions were found to have been evaluated in such a way that conclusions regarding effectiveness could be drawn
Oakley, 1995[234]	*Qualitative*	The conclusions from sound and flawed studies were compared; in general, a larger proportion of sound studies suggested interventions to be beneficial
Ogilvie Harris, 1995[341]	*Qualitative*	Validity assessment results were presented and discussed but not really used to interpret study results. For pharmacological studies around half of the studies were of reasonable quality, providing some evidence that the agents reviewed were better than placebo, but no evidence for one agent over any other. No quality considerations in discussion of other interventions. Studies were examined in regard to the number of statistically significant and non-significant conclusions presented, as there is no standardised method of scoring the severity of ankle injuries or the results of treatment
Oxman, 1994[343]	*Not considered*	
Oxman, 1994[271]	*Qualitative*	Quality described separately to results. The numbers of methodologically strong studies which provided data for each outcome were reported. Overall five studies were found to be methodologically strong. Conclusions are limited owing to paucity of well-designed studies
Parker, 1998[398]	*Not considered*	
Ploeg, 1996[232]	*Qualitative*	Quality features were not discussed in detail; however, authors conclude that the findings of the review should be considered in the light of serious methodological limitations of the studies
Powe, 1994[399]	*Quantitative*	No association between visual acuity outcomes and quality score.
	Quality score and ES	For three complications, higher-quality studies reported higher proportions of complications than studies of
	Compared high and low quality	lower quality ($p = 0.006, 0.02, 0.09$)
	Weighted by quality score	Quality score was used as a weight in the pooled analysis. Unweighted results were not presented
Puig Barbera, 1995[345]	*Quantitative* Excluded lower quality studies	Two studies excluded from meta-analysis for not meeting quality criteria; the impact of this on the results of the meta-analysis was not discussed

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Rey, 1997[346]	*Qualitative*	No conclusive results available because of poor study quality. Quality improved in the more recent studies. Authors conclude that electroconvulsive therapy in the young is similar in effectiveness and side-effects to its use in adults
Riben, 1994[270]	*Qualitative*	Effect of quality on results was discussed. For education: strong/moderate studies indicate no relationship between on-site education and sanitation level. For inspections: strong studies indicate evidence not conclusive
Richard, 1997[224]	*Qualitative*	Inconclusive evidence found to support/refute the value of FU surveillance programmes owing to lack of large and well-designed studies
Ried, 1994[347]	*Quantitative* Correlation analysis	Zero-order correlation (Pearson's product moment *r*) was not statistically significant for study ES and quality score (–0.12), indicating no effect of quality. However, power to detect a statistically significant correlation was low
Robert, 1997[379]	*Not considered*	
Rowe, 1997[348]	*Quantitative* Correlation analysis	No significant relationship between quality score and size of treatment effect was found (Pearson's correlation coefficient, $r = -0.22$, $p = 0.33$)
Saint, 1998[288]	*Not considered*	
Salisbury, 1997[198]	*Qualitative*	Comment on methodological quality provided in table along with individual study results. Authors conclude that there is evidence of benefit from the intervention but that this is based on studies that are methodologically weak
Schmidt-Nowara, 1995[239]	*Not considered*	
Schoemaker, 1997[358]	*Qualitative*	Only high-quality studies were considered in the review (6/33). Methodological shortcomings and lack of control for potential confounders meant that the prognostic value of early intervention could not be estimated
Sellers, 1997[353]	*Quantitative* Each coded study characteristic was investigated (including sample size, matching, response rate, covariate adjustment, study setting)	Use of matching led to larger ES for weight and systolic and diastolic blood pressure; use of covariate adjustment produced smaller ES for smoking and systolic blood pressure. Longer FU time was associated with larger ES for cholesterol, weight and CHD risk
Selley, 1997[354]	*Not considered*	
Simons, 1996[356]	*Quantitative*	Only high-quality studies included in the meta-analysis (all RCTs). Summary effect size was larger when all RCTs (including low-quality ones) were included

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Smeenk, 1998[359]	*Qualitative*	Narrative summary of study results and their quality scores was provided. RCTs and NRS appeared to demonstrate similar effects. Effectiveness of the intervention remains unclear.
Snowdon, 1997[400]	*Qualitative*	Quality issues discusses narratively. Authors conclude there is no evidence of benefit and that most of the studies were methodologically flawed
Solomon, 1997[366]	*Qualitative*	Quality issues discussed alongside results of each study. Some indication of differences between care providers was found, but important methodological limitations remained
Sullivan, 1995[290]	*Not considered*	
Talley, 1996[401]	*Qualitative*	The methodological quality of the studies was discussed in detail; only one study was deemed of acceptable quality; however, its results were not felt to be generalisable
Tangkanakul, 1997[372]	*Qualitative*	Methodological quality discussed separately from study results. Some indication of benefit for the intervention was found, but too few RCTs and too many problems with NRS to make strong recommendations
ter Riet, 1990[68]	*Qualitative*	Methodological issues were discussed and study results (positive or negative) were examined in relation to the quality of the studies. The authors conclude that no high-quality studies exist and no definitive conclusions can therefore be reached regarding efficacy.
Theis, 1997[373]	*Qualitative*	Some quality issues discussed. Methodological problems mean that available studies provide only very limited comparative data on the benefits of different treatment protocols
Thomas, 1995[375]	*Qualitative*	Comment was made on poor methodological quality of included studies, making definitive conclusions impossible
Towler, 1998[235]	*Qualitative*	Design issues discussed narratively. Quality of design generally high. Studies examined according to design
van Balkom, 1997[384]	*Quantitative* Correlation between quality score and effect size Multiple regression analysis using eight variables: % of drop-outs, male/female ratio, duration of illness, age at onset, age at start of trial, type of diagnosis, % of patients with panic disorder alone and the total quality score	Magnitude of effect size was not significantly associated with quality score. Multivariate analysis indicated that eight independent variables including quality score and confounding variables accounted for 15% of the variance for panic ($F = 0.9$; df = 8,42; $p = 0.53$), and 21% for agoraphobia ($F = 2.81$; df = 8,87; $p = 0.008$). In the Discussion section, the authors further state that total quality score was only weakly correlated with outcome
Verhagen, 1997[304]	*Qualitative*	The methodological quality and results of the studies were discussed separately. Authors conclude that a conclusion regarding efficacy cannot be drawn owing to poor methodological quality
Vickers, 1996[201]	*Qualitative*	Methodological quality of the studies was discussed in some detail. The authors conclude that the intervention is effective but acknowledge that methodological flaws may lead some readers to conclude that there is insufficient evidence

continued

Review	Method of incorporating quality into synthesis	Results of quality investigation
Vickers, 1994[386]	*Qualitative*	Some methodological items were discussed in the text and the impact of any limitations on the results of some of the individual studies were discussed. Most of the studies reviewed indicated favourable effects from hypnotherapy, despite the lack of good control groups
Villalobos, 1998[402]	*Quantitative* Sensitivity analyses according to prospective/ retrospective design (prospective studies may all be RCTs – not fully clear)	Prospective studies found no effect from intervention, whereas retrospective studies (and all studies combined) suggested that intervention was effective
Watt, 1998[221]	*Qualitative*	All studies supported the use of the intervention in question; results usually statistically or clinically significant, but inconclusive because of small sample size or methodological bias
Wells-Parker, 1995[129]	*Quantitative* Regression analysis used to investigate impact of methodological factors (including 'grouping' quality, source of ES estimate, purity or definition of outcome measure and sample size) Various quality score thresholds were used as inclusion criteria for the meta-analysis Visual plot	Methodological factors jointly accounted for significant variance ($R^2 = 14$; $p < 0.01$). Log (sample size) proved to be the strongest predictor of ES ($\beta = -0.27$; $t = 1.29$, $p = 0.21$) Lower quality studies demonstrated larger ES with larger SDs than high-quality studies [e.g. high-quality (score $\leqslant$5), ES 0.08, SD 0.13; low-quality (score $\geqslant$6), ES 0.25, SD 0.36]. Using 5.0 or 5.5 as the threshold for 'high quality' had little impact on the results The funnel plot indicated that poorer methodological quality was associated with greater variation in effect size
Wingood, 1996[202]	*Qualitative*	Study weaknesses discussed, but more detailed impact of this on results not made. All studies seem to produce positive effects on knowledge and/or behaviour. Authors suggest RCTs more likely to give positive effects, but little evidence of this, given only two NRS
Witlin, 1998[381]	*Qualitative*	Both level I and II evidence supported the use of magnesium sulphate for seizure prophylaxis in women with severe pre-eclampsia
Wright, 1995[203]	*Qualitative*	Results of individual studies were discussed with their design and quality. Overall studies were of insufficient quality to allow any strong conclusions to be made. Authors did not state that they used QA, but data extraction sheet covers major quality issues and is very similar to one by Sheldon. May be a modified version

ES, effect size; df, degrees of freedom; RD, risk difference.

Appendix 8

Descriptions of the IST and ECST

The International Stroke Trial (IST)[132]

Description of the trial

The IST investigated the safety and efficacy of aspirin and heparin on the outcome of ischaemic stroke. Between January 1991 and May 1996, 19,435 patients with suspected acute ischaemic stroke entering 467 hospitals in 36 countries were centrally randomised using a minimisation algorithm within 48 h of stroke onset. Using a 2×2 factorial design, half of the patients were allocated 'heparin' and half 'avoid heparin', whilst simultaneously half were allocated 'aspirin' and half 'avoid aspirin'. Aspirin was prescribed at 300 mg per day for 14 days (or the duration of hospital stay). At discharge, clinicians were encouraged to prescribe all patients long-term aspirin. Among patients allocated heparin, subcutaneous heparin was administered for 14 days at one of two randomly allocated doses: 5000 IU twice daily (low dose) and 12,500 IU twice daily (medium dose). All analyses presented below combine these two dose groups.

The primary outcomes were death within 14 days and death or dependency at 6 months. Outcome data were 99.99% complete for 14-day outcome and 99.2% complete for 6-month outcome. In terms of compliance, low-dose heparin was received throughout the scheduled treatment by 90% and medium-dose heparin by 88% of those allocated to it; 94% of those allocated to 'avoid heparin' did not receive it. Aspirin was taken throughout the scheduled treatment period by 92% allocated to receive it; 93% of those allocated to avoid it did not receive it.

In both aspirin- and heparin-allocated patients there were non-significantly fewer deaths within 14 days (9.0% heparin versus 9.3% no heparin; 9.0% aspirin versus 9.4% no aspirin). At 6 months there was a non-significant trend towards a smaller percentage of aspirin group being dead or dependent [62 versus 63%, $2p = 0.07$; a difference of 13 (SD 7) per 1000]. After adjustment for baseline stroke severity, the benefit from aspirin was significant [14 events prevented per 1000 patients (SD 6), $2p = 0.03$]. There was no interaction between aspirin and heparin in the main outcomes.

Baseline characteristics in the IST

The distribution of 15 baseline characteristics associated with prognosis following stroke is reported in Table 36. These data were collected during the randomisation process and are therefore available for every randomised patient. The ORs describe the relationship between a change in the baseline variable and the outcome of death or disability at 6 months.

The trial protocol required the use of a computed tomography (CT) scan to rule out intracranial haemorrhage. However, where obtaining a CT scan was likely to require a long delay, and the clinician regarded it as very likely that the stroke was ischaemic, a non-comatose patient could be randomised before the CT scan. *Table 36* shows that a high proportion of patients did not report CT scan results at baseline, and that the absence of a CT scan indicated poor prognosis. This observation arises through there being a relationship between stroke severity and the decision to obtain a CT scan.

For the purpose of our analyses, the last six variables were converted to a score (0–6) giving the number of presenting neurological characteristics. The distribution of this score is given in *Table 37*.

Deficit score (*Table 37*), stroke type and consciousness (*Table 36*) appear to have the largest ranges of event ranges across their categorisations, indicating that they are likely to be three of the most prognostic variables.

Adaptations made to the IST for the resampling project

For the resampling project, only the comparison of 'aspirin' against 'avoid aspirin' was considered for the outcome of death or disability at 6 months. To use the IST dataset for this project, the multicentre nature of the trial was exploited to construct a series of smaller randomised and non-randomised 'sub-studies'. Centres falling within geographical regions were grouped together such that sufficient participants were accrued in each

TABLE 36 *Dead or dependent at 6 months according to treatment and baseline characteristics (IST)*

	Aspirin (n = 9639) No. (% adverse outcomes)	No aspirin (n = 9646) No. (% adverse outcomes)	OR (95% CI)
Events at 6 months	6000 (62.2)	6125 (63.5)	0.95 (0.89 to 1.01)
Delay (h from symptoms)			
0–3	414 (70.5)	422 (70.6)	
4–6	1157 (66.1)	1148 (69.5)	
7–12	2005 (63.5)	2079 (63.3)	
13–24	2798 (60.8)	2730 (63.0)	
25–48	3265 (60.3)	3267 (61.0)	0.90 (0.84 to 0.96) per day
Sex			
F	4530 (69.5)	4436 (70.5)	1.00
M	5109 (55.8)	5210 (57.4)	0.56 (0.53 to 0.60)
Age (years)			
<50	491 (34.4)	477 (36.1)	
50–59	980 (43.4)	1022 (43.2)	
60–69	2235 (51.8)	2219 (53.6)	
70–79	3392 (64.7)	3369 (65.7)	
⩾80	2541 (80.9)	2559 (82.4)	1.74 (1.69 to 1.79) per 10 years
Systolic pressure			
<140	1789 (63.1)	1764 (65.2)	
140–159	2730 (61.9)	2628 (63.2)	
160–179	2463 (62.5)	2542 (61.8)	
⩾180	2657 (61.8)	2712 (64.3)	0.99 (0.98 to 1.00) per 10 mmHg
Symptoms noted on waking			
No	6842 (62.8)	6805 (63.7)	1.00
Yes	2797 (60.8)	2841 (63.0)	0.94 (0.87 to 1.01)
Consciousness			
Fully alert	7405 (54.2)	7404 (55.9)	1.00
Drowsy or unconscious	2234 (89.0)	2242 (88.5)	6.37 (5.77 to 7.02)
Type of stroke			
Lacunar	2308 (48.2)	2308 (48.3)	1.00
Partial anterior	3908 (60.1)	3939 (62.0)	1.69 (1.57 to 1.82)
Posterior circulation	1115 (49.3)	1103 (49.8)	1.05 (0.95 to 1.16)
Total anterior	2308 (86.2)	2296 (87.2)	6.99 (6.31 to 7.75)
Atrial fibrillation			
No	7537 (58.8)	7614 (60.6)	1.00
Yes	1611 (78.0)	1542 (79.8)	2.52 (2.30 to 2.76)
Infarct visible at CT scan			
No CT	3210 (68.8)	3163 (69.6)	1.00
CT and no infarct visible	3281 (52.5)	3265 (53.7)	0.50 (0.47 to 0.54)
CT and infarct visible	3148 (65.8)	3218 (67.5)	0.89 (0.83 to 0.96)
Face deficit			
No	2668 (51.1)	2633 (52.9)	1.00
Yes	6971 (66.5)	7013 (67.5)	1.87 (1.76 to 2.00)
Arm/hand deficit			
No	1394 (40.0)	1382 (44.0)	1.00
Yes	8245 (66.0)	8264 (66.8)	1.65 (1.58 to 1.72)
Leg/foot deficit			
No	2389 (42.2)	2338 (45.7)	1.00
Yes	7250 (68.9)	7308 (69.2)	1.42 (1.39 to 1.45)
Dysphasia			
No	5400 (56.8)	5424 (57.2)	1.00
Yes	4231 (69.3)	4222 (71.2)	1.15 (1.14 to 1.71)
Hemianopia			
No	8119 (59.5)	8098 (60.2)	1.00
Yes	1520 (77.0)	1548 (80.8)	1.20 (1.18 to 1.22)
Visual disorder			
No	8066 (58.6)	8064 (59.8)	1.00
Yes	1573 (81.1)	1582 (82.2)	1.20 (1.19 to 1.22)

TABLE 37 *Dead or dependent at 6 months according to neurological deficit score (IST)*

Deficit score	Aspirin (*n* = 7749) No. (% adverse outcomes)	No aspirin (*n* = 7753) No. (% adverse outcomes)	OR (95% CI)
0	267 (43.8)	246 (39.8)	
1	719 (41.6)	767 (38.5)	
2	1373 (53.0)	1421 (51.4)	
3	2298 (63.4)	2252 (63.4)	
4	2028 (78.8)	2020 (77.7)	
5	861 (88.6)	843 (88.6)	1.64 (1.60 to 1.68) per unit increase in score
6	203 (92.1)	204 (86.7)	

TABLE 38 *Distribution of baseline characteristics by country (IST)*

	Sample size	Adverse outcome (%)	Delay (h) (mean)	Female (%)	Age in (years) (mean)	Systolic blood pressure (mean)	Symptoms on waking (%)	Drowsy or unconscious (%)	Atrial fibrillation (%)	Deficit score (mean)	No CT scan (%)	Partial anterior + Total anterior strokes (%)
Australia	597	56	19	60	71	155	29	26	12	3.0	13	32 + 27 = 59
Northern Italy	1911	55	17	52	72	162	31	20	20	3.1	28	39 + 22 = 61
Central Italy	980	61	21	60	74	161	24	21	18	2.9	32	50 + 17 = 67
Southern Italy	546	60	16	64	70	154	27	20	19	3.2	31	46 + 18 = 64
The Netherlands	728	54	16	52	69	169	30	20	10	3.1	14	37 + 27 = 64
New Zealand	453	51	25	51	72	155	28	19	17	3.0	4	37 + 22 = 59
Norway	526	60	24	53	72	165	28	19	6	2.9	6	44 + 19 = 63
Poland	759	53	18	51	69	163	34	28	25	3.3	49	44 + 19 = 63
Spain	478	46	19	56	71	157	31	15	15	3.1	21	38 + 24 = 62
Sweden	636	42	24	55	75	167	30	15	15	2.5	2	50 + 16 = 66
Switzerland	1631	67	17	52	74	164	29	31	16	3.1	26	39 + 28 = 67
Scotland	1043	75	23	50	72	158	27	25	21	3.2	46	39 + 29 = 68
Northern England and Wales	2762	79	20	53	72	159	27	25	18	3.1	61	41 + 27 = 68
Southern England	2452	81	22	51	76	159	27	28	24	3.3	61	40 + 30 = 70

arm for the resampling to be undertaken. In addition, the sequential nature of the study recruitment was used to split participants into 'early recruits' and 'later recruits' to enable historically controlled studies to be constructed. Early recruits were those recruited up to and including the 15 January 1995; late recruits were those recruited after this date. This date was chosen to maximise the number of regions available for inclusion in the study and is close to the median recruitment date.

To be considered as a sub-study for the concurrent cohort design, we required that there were at least 100 participants within each arm of each trial. For the historical cohort design we required that when the arms were divided into early and late recruits, there were at least 100 early control recruits and 100 late treatment recruits.

Fourteen geographical regions were constructed on this basis for the main analysis for both concurrent cohort and historical cohort designs.

TABLE 39 *Distribution of baseline characteristics by country and recruitment period (IST)*

	Before (B) or after (A)	Sample size	Adverse outcome (%)	Delay (h) (mean)	Female (%)	Age (years) (mean)	Systolic blood pressure (mean)	Symptoms on waking (%)	Drowsy or unconscious (%)	Atrial fibrillation (%)	Deficit score (mean)	No CT scan (%)	Partial anterior + total anterior strokes (%)
Australia	B	331	55	20	60	71	155	29	28	14	2.9	12	28 + 27 = 54
	A	266	57	18	61	72	157	29	24	9	3.0	15	37 + 28 = 65
Northern Italy	B	1216	53	18	54	72	162	30	20	20	3.1	26	38 + 22 = 61
	A	695	57	17	49	73	162	32	18	19	3.1	31	41 + 21 = 62
Central Italy	B	710	60	21	60	74	160	24	20	19	2.9	33	50 + 17 = 66
	A	270	63	19	58	74	163	26	24	16	3.0	31	50 + 19 = 69
Southern Italy	B	241	56	16	62	70	155	22	20	18	3.0	29	46 + 15 = 62
	A	305	63	16	66	70	154	31	20	20	3.3	32	45 + 21 = 66
The Netherlands	B	521	55	17	53	69	168	30	21	10	3.2	15	36 + 29 = 65
	A	207	52	16	51	70	170	29	17	11	3.0	12	41 + 23 = 64
New Zealand	B	250	51	25	49	71	155	29	20	18	3.1	3	36 + 24 = 60
	A	203	51	26	53	73	156	27	18	15	2.9	4	39 + 19 = 58
Norway	B	316	65	24	52	73	165	28	21	7	2.8	9	43 + 18 = 62
	A	210	53	23	56	72	164	27	17	5	3.0	3	44 + 21 = 65
Poland	B	209	55	18	48	70	160	30	30	27	3.4	33	46 + 22 = 68
	A	550	53	19	52	69	164	36	28	24	3.2	55	43 + 18 = 61
Spain	B	237	45	19	58	71	155	32	16	14	2.9	21	40 + 23 = 63
	A	241	47	19	54	72	159	30	14	16	3.2	21	36 + 26 = 62
Sweden	B	409	46	24	56	75	166	30	16	16	2.5	1	48 + 17 = 66
	A	227	33	24	53	75	167	31	14	14	2.5	4	54 + 14 = 68
Switzerland	B	841	66	16	53	73	162	26	32	17	3.1	28	40 + 29 = 69
	A	790	68	18	50	75	166	33	30	15	3.1	24	39 + 27 = 66
Scotland	B	694	73	22	51	71	158	28	26	20	3.2	48	35 + 31 = 66
	A	349	78	25	49	74	158	26	24	23	3.1	40	46 + 25 = 71
Northern England and Wales	B	1562	80	21	52	72	158	28	25	19	3.2	58	41 + 28 = 69
	A	1200	78	19	53	72	160	25	25	17	3.0	65	42 + 26 = 68
Southern England	B	1384	82	21	51	76	159	25	28	23	3.3	62	40 + 30 = 70
	A	1068	80	24	50	76	157	29	28	24	3.3	60	41 + 30 = 71

TABLE 40 *Distribution of baseline characteristics in 10 UK cities combining results across 28 hospitals (IST)*

	Sample size	Adverse outcome (%)	Delay (h) (mean)	Female (%)	Age (years) (mean)	Systolic blood pressure (mean)	Symptoms on waking (%)	Drowsy or unconscious (%)	Atrial fibrillation (%)	Deficit score (mean)	No CT scan (%)	Partial anterior + total anterior strokes (%)
Belfast	234	85	26	47	76	151	30	38	22	3.1	28	40 + 29 = 70
Bishop Auckland	234	80	16	48	74	159	30	31	19	3.2	75	41 + 31 = 72
Blackburn	221	83	15	52	72	161	30	21	18	3.2	74	45 + 24 = 69
Edinburgh	329	57	22	49	69	159	29	20	20	3.0	34	36 + 27 = 63
Leeds	226	58	19	54	65	160	28	19	10	2.9	64	34 + 24 = 58
Liverpool	536	80	17	54	72	162	26	19	19	2.8	89	49 + 21 = 71
London	344	82	24	49	76	160	27	26	23	3.2	12	39 + 27 = 66
Newcastle	361	83	26	46	75	156	26	31	21	3.3	25	39 + 32 = 71
Nottingham	305	83	17	50	77	161	22	24	22	3.2	97	31 + 30 = 61
Sheffield	308	71	21	58	71	159	27	21	14	3.0	50	38 + 23 = 61

The distribution of the outcome of dead or dependent at 6 months and of the baseline characteristics in these sub-studies are described in *Table 38*. *Table 39* describes the baseline characteristics for the cohorts split according to 'early recruits' (before) and 'late recruits' (after). Differences are evident in both tables in the frequency of the outcome and for some of the baseline characteristics, notably the percentage with total anterior strokes, or admitted 'drowsy or unconscious'.

These 14 regions contain data from 15,502 (80%) of the randomised participants. The 3933 excluded participants were recruited in Argentina, Austria, Belgium, Brazil, Canada, Chile, Czech Republic, Denmark, Eire, Finland, France, Greece, Hong Kong, Hungary, India, Israel, Japan, Portugal, Romania, Singapore, Slovak Republic, Slovenia, South Africa, Sri Lanka, Turkey and the USA, but without sufficient numbers in any region to construct an additional centre.

A second set of 10 sub-studies was constructed from the UK data, by grouping hospitals within cities, including a total of 3098 participants (50% of those recruited in the UK). The distribution of baseline variables across the 10 cities is shown in *Table 40*. Data from these sub-studies were used only to construct concurrently controlled studies, insufficient data being available for the historically controlled comparisons.

The European Carotid Surgery Trial (ECST)[133]

Description of the trial

The ECST investigated the risks and benefits of carotid endarterectomy, primarily in terms of stroke prevention. Between October 1981 and March 1994, 3024 patients with recently symptomatic carotid stenosis entering 100 centres in 14 countries were randomised to carotid endarterectomy or control (avoid surgery as long as possible). Patients were eligible for inclusion if they had had at least one transient or mild symptomatic ischaemic vascular event within the last 6 months, and had some degree of carotid stenosis. Randomisation was performed centrally using a minimisation algorithm, 60% of patients being allocated to surgery and 40% to control.

Of the 1811 allocated surgery, 1745 (96%) had received it within 1 year. Of the 1213 allocated control, 42 (3%) received surgery within 1 year. The control treatment usually consisted of advice

TABLE 41 *Baseline characteristics according to country (ECST)*

	Sample size	Adverse outcome (%)	Female (%)	Age (years) (mean)	Previous major stroke (%)	Neurological signs (%)	Previous MI (%)	Angina (%)	Prophylactic aspirin use (%)	>50% stenosis (%)
Region 1	458	31	33	62	75	73	86	79	45	70
Region 2	271	43	22	66	65	69	92	85	75	48
Region 3	294	39	26	64	70	63	89	86	35	68
Region 4	264	27	21	63	67	65	91	92	45	50
Region 5	458	30	27	60	71	71	90	87	54	58
Region 6	308	46	35	61	79	74	82	77	14	81
Region 7	439	36	33	61	80	79	88	80	33	73
Region 8	341	51	28	62	77	72	86	79	39	73

MI, myocardial infarction.

TABLE 42 Baseline characteristics according to country and recruitment period (ECST)

	Before (B) or after (A)	Sample size	Adverse outcome (%)	Female (%)	Age (years) (mean)	Previous major stroke (%)	Neurological signs (%)	Previous MI (%)	Angina (%)	Prophylactic aspirin use (%)	>50% stenosis (%)
Region 1	B	131	45	36	60	75	31	10	15	41	54
	A	327	26	32	63	75	26	16	23	61	76
Region 2	B	149	54	23	66	69	32	7	15	21	39
	A	122	30	21	67	61	30	10	15	29	60
Region 3	B	119	49	29	63	76	34	10	8	71	60
	A	175	33	24	64	66	38	12	17	60	74
Region 4	B	128	33	21	61	66	38	11	4	57	43
	A	136	22	21	64	68	32	7	13	54	57
Region 5	B	243	39	26	59	70	30	13	13	38	48
	A	215	21	28	61	73	27	7	14	55	69
Region 6	B	136	59	35	60	80	28	17	23	82	79
	A	172	37	35	62	77	24	18	23	90	82
Region 7	B	201	55	35	61	81	18	14	23	56	70
	A	143	32	31	61	79	24	10	16	83	78
Region 8	B	231	53	25	62	73	34	13	22	48	73
	A	205	24	32	63	81	21	16	20	77	72

against smoking, treatment of raised blood pressure and antiplatelet drugs, although it was noted to vary between centres and over the years.

The primary outcome was death or major stroke. Overall there was a non-significant difference in outcome (37.0% of surgery-group patients versus 36.5% of control-group patients), although there was a relationship between benefit of surgery and degree of stenosis, surgery being of value when the degree of stenosis was >80%.

Baseline characteristics in the ECST

Eight variables for baseline characteristics associated with prognosis were included in the data set. Details of the values of these variables are not given in the report as they will soon be published in other research publications produced by the trial collaborative group. These characteristics were collected during the randomisation process, and were therefore available for every randomised patient.

Adaptations made to the ECST for the resampling project

To use the ECST dataset for this project, the multicentre nature of the trial was exploited to construct a series of smaller randomised and non-randomised studies, in the same way as for the IST. Centres falling within geographical regions were grouped together such that sufficient participants were accrued in each arm for the resampling to be undertaken. In addition, the sequential nature of the study recruitment was used to split participants into 'early recruits' and 'later recruits' to enable historically controlled studies to be constructed. For all regions other than Region 4, early recruits were those recruited up to and including 5 December 1987; late recruits were those recruited after this date. A cut-off date of 19 December 1988 was used for Region 4. These dates were chosen to maximise the number of regions available for inclusion in the study.

Sample sizes used in the resampling for the ECST study were smaller than those for the IST. To be considered as a sub-study for the concurrent cohort design, we required that there were at least 40 participants within each arm of each trial. For the historical cohort design we required that when the arms were divided into early and late recruits there were at least 40 early control recruits and 40 late treatment recruits.

Eight geographical regions were constructed on this basis for the main analysis for both concurrent cohort and historical cohort designs. As centres were smaller than for the IST, neighbouring countries were grouped where necessary. These eight regions recruited 2833 (94%) of trial participants. For reasons of future publications, the regions are not identifiable in the tables we present.

The ECST is one of two trials investigating the efficacy and safety of carotid endarterectomy. Both the ECST and the second trial, NASCET,[135] published early results during the recruitment period of the ECST, which led to the data monitoring committee of the ECST changing the inclusion criteria of the ECST in 1990. Cases recruited after this change were deleted from the data set where this change would be confounded with the generation of groups (e.g. for the historically controlled study analyses).

The distribution of the outcome of dead or major stroke, and of the baseline characteristics in these sub-studies are described in *Table 41*. *Table 42* describes the baseline characteristics for the cohorts split according to 'early recruits' (before) and 'late recruits' (after). Differences are evident in both tables in the frequency of the outcome and for some of the baseline characteristics, but are greatest in the historical comparisons. In these centres the event rates are lower in the second period in all eight centres, whereas the percentages with high degrees of stenosis increase with time in all but one centre. The pattern observed with degree of stenosis is likely to reflect a change in case-mix due to emerging evidence of the dangers of surgery for low-grade surgery during the trial's recruitment period.

Health Technology Assessment Programme

Prioritisation Strategy Group

Members

Chair,
Professor Kent Woods,
Director, NHS HTA Programme & Professor of Therapeutics, University of Leicester

Professor Bruce Campbell, Consultant Vascular & General Surgeon, Royal Devon & Exeter Hospital

Professor Shah Ebrahim, Professor in Epidemiology of Ageing, University of Bristol

Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Radcliffe Hospital, Oxford

Dr Ron Zimmern, Director, Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

HTA Commissioning Board

Members

Programme Director,
Professor Kent Woods, Director, NHS HTA Programme, Department of Medicine and Therapeutics, Leicester Royal Infirmary, Robert Kilpatrick Clinical Sciences Building, Leicester

Chair,
Professor Shah Ebrahim,
Professor in Epidemiology of Ageing, Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol

Deputy Chair,
Professor Jenny Hewison,
Professor of Health Care Psychology, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds School of Medicine, Leeds

Professor Douglas Altman, Professor of Statistics in Medicine, Centre for Statistics in Medicine, Oxford University, Institute of Health Sciences, Cancer Research UK Medical Statistics Group, Headington, Oxford

Professor John Bond, Professor of Health Services Research, Centre for Health Services Research, University of Newcastle, School of Health Sciences, Newcastle upon Tyne

Professor John Brazier, Director of Health Economics, Sheffield Health Economics Group, School of Health & Related Research, University of Sheffield, ScHARR Regent Court, Sheffield

Dr Andrew Briggs, Public Health Career Scientist, Health Economics Research Centre, University of Oxford, Institute of Health Sciences, Oxford

Dr Christine Clark, Medical Writer & Consultant Pharmacist, Cloudside, Rossendale, Lancs
and
Principal Research Fellow, Clinical Therapeutics in the School of Pharmacy, Bradford University, Bradford

Professor Nicky Cullum, Director of Centre for Evidence Based Nursing, Department of Health Sciences, University of York, Research Section, Seebohm Rowntree Building, Heslington, York

Dr Andrew Farmer, Senior Lecturer in General Practice, Department of Primary Health Care, University of Oxford, Institute of Health Sciences, Headington, Oxford

Professor Fiona J Gilbert, Professor of Radiology, Department of Radiology, University of Aberdeen, Lilian Sutton Building, Foresterhill, Aberdeen

Professor Adrian Grant, Director, Health Services Research Unit, University of Aberdeen, Drew Kay Wing, Polwarth Building, Foresterhill, Aberdeen

Professor Alastair Gray, Director, Health Economics Research Centre, University of Oxford, Institute of Health Sciences, Headington, Oxford

Professor Mark Haggard, Director, MRC ESS Team, CBU Elsworth House, Addenbrooke's Hospital, Cambridge

Professor F D Richard Hobbs, Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham, Primary Care and Clinical Sciences Building, Edgbaston, Birmingham

Professor Peter Jones, Head of Department, University Department of Psychiatry, University of Cambridge, Addenbrooke's Hospital, Cambridge

Professor Sallie Lamb, Research Professor in Physiotherapy/Co-Director, Interdisciplinary Research Centre in Health, Coventry University, Coventry

Dr Donna Lamping, Senior Lecturer, Health Services Research Unit, Public Health and Policy, London School of Hygiene and Tropical Medicine, London

Professor David Neal, Professor of Surgical Oncology, Oncology Centre, Addenbrooke's Hospital, Cambridge

Professor Tim Peters, Professor of Primary Care Health Services Research, Division of Primary Health Care, University of Bristol, Cotham House, Cotham Hill, Bristol

Professor Ian Roberts, Professor of Epidemiology & Public Health, Intervention Research Unit, London School of Hygiene and Tropical Medicine, London

Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh, Western General Hospital NHS Trust, Bramwell Dott Building, Edinburgh

Professor Martin Severs, Professor in Elderly Health Care, Portsmouth Institute of Medicine, Health & Social Care, St George's Building, Portsmouth

Dr Jonathan Shapiro, Senior Fellow, Health Services Management Centre, Park House, Birmingham

Current and past membership details of all HTA 'committees' are available from the HTA website (www.ncchta.org)

Diagnostic Technologies & Screening Panel

Members

Chair,
Dr Ron Zimmern, Director of the Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

Dr Paul Cockcroft, Consultant Medical Microbiologist/ Laboratory Director, Public Health Laboratory, St Mary's Hospital, Portsmouth

Professor Adrian K Dixon, Professor of Radiology, Addenbrooke's Hospital, Cambridge

Dr David Elliman, Consultant in Community Child Health, London

Dr Andrew Farmer, Senior Lecturer in General Practice, Institute of Health Sciences, University of Oxford

Dr Karen N Foster, Clinical Lecturer, Dept of General Practice & Primary Care, University of Aberdeen

Professor Jane Franklyn, Professor of Medicine, University of Birmingham

Professor Antony J Franks, Deputy Medical Director, The Leeds Teaching Hospitals NHS Trust

Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London

Dr Susanne M Ludgate, Medical Director, Medical Devices Agency, London

Dr William Rosenberg, Senior Lecturer and Consultant in Medicine, University of Southampton

Dr Susan Schonfield, CPHM Specialised Services Commissioning, Croydon Primary Care Trust

Dr Margaret Somerville, Director of Public Health, Teignbridge Primary Care Trust, Devon

Mr Tony Tester, Chief Officer, South Bedfordshire Community Health Council, Luton

Dr Andrew Walker, Senior Lecturer in Health Economics, University of Glasgow

Professor Martin J Whittle, Head of Division of Reproductive & Child Health, University of Birmingham

Dr Dennis Wright, Consultant Biochemist & Clinical Director, Pathology & The Kennedy Galton Centre, Northwick Park & St Mark's Hospitals, Harrow

Pharmaceuticals Panel

Members

Chair,
Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Oxford Radcliffe Hospital

Professor Tony Avery, Professor of Primary Health Care, University of Nottingham

Professor Iain T Cameron, Professor of Obstetrics & Gynaecology, University of Southampton

Mr Peter Cardy, Chief Executive, Macmillan Cancer Relief, London

Dr Christopher Cates, GP and Cochrane Editor, Bushey Health Centre, Bushey, Herts.

Mr Charles Dobson, Special Projects Adviser, Department of Health

Dr Robin Ferner, Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Dr Karen A Fitzgerald, Pharmaceutical Adviser, Bro Taf Health Authority, Cardiff

Professor Alastair Gray, Professor of Health Economics, Institute of Health Sciences, University of Oxford

Mrs Sharon Hart, Managing Editor, *Drug & Therapeutics Bulletin*, London

Dr Christine Hine, Consultant in Public Health Medicine, Bristol South & West Primary Care Trust

Professor Robert Peveler, Professor of Liaison Psychiatry, Royal South Hants Hospital, Southampton

Dr Frances Rotblat, CPMP Delegate, Medicines Control Agency, London

Mrs Katrina Simister, New Products Manager, National Prescribing Centre, Liverpool

Dr Ken Stein, Senior Lecturer in Public Health, University of Exeter

Professor Terence Stephenson, Professor of Child Health, University of Nottingham

Dr Richard Tiner, Medical Director, Association of the British Pharmaceutical Industry, London

Professor Dame Jenifer Wilson-Barnett, Head of Florence Nightingale School of Nursing & Midwifery, King's College, London

Therapeutic Procedures Panel

Members

Chair,
Professor Bruce Campbell,
Consultant Vascular and General Surgeon, Royal Devon & Exeter Hospital

Dr Mahmood Adil, Head of Clinical Support & Health Protection, Directorate of Health and Social Care (North), Department of Health, Manchester

Professor John Bond, Head of Centre for Health Services Research, University of Newcastle upon Tyne

Mr Michael Clancy, Consultant in A & E Medicine, Southampton General Hospital

Dr Carl E Counsell, Senior Lecturer in Neurology, University of Aberdeen

Dr Keith Dodd, Consultant Paediatrician, Derbyshire Children's Hospital, Derby

Professor Gene Feder, Professor of Primary Care R&D, Barts & the London, Queen Mary's School of Medicine and Dentistry, University of London

Ms Bec Hanley, Freelance Consumer Advocate, Hurstpierpoint, West Sussex

Professor Alan Horwich, Director of Clinical R&D, The Institute of Cancer Research, London

Dr Phillip Leech, Principal Medical Officer for Primary Care, Department of Health, London

Mr George Levvy, Chief Executive, Motor Neurone Disease Association, Northampton

Professor James Lindesay, Professor of Psychiatry for the Elderly, University of Leicester

Dr Mike McGovern, Senior Medical Officer, Heart Team, Department of Health, London

Dr John C Pounsford, Consultant Physician, North Bristol NHS Trust

Professor Mark Sculpher, Professor of Health Economics, Institute for Research in the Social Services, University of York

Dr L David Smith, Consultant Cardiologist, Royal Devon & Exeter Hospital

Professor Norman Waugh, Professor of Public Health, University of Aberdeen

Expert Advisory Network

Members

Mr Gordon Aylward,
Chief Executive,
Association of British Health-Care Industries, London

Ms Judith Brodie,
Head of Cancer Support Service, Cancer BACUP, London

Mr Shaun Brogan,
Chief Executive, Ridgeway Primary Care Group, Aylesbury, Bucks

Ms Tracy Bury,
Project Manager, World Confederation for Physical Therapy, London

Mr John A Cairns,
Professor of Health Economics, Health Economics Research Unit, University of Aberdeen

Professor Howard Stephen Cuckle,
Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Professor Nicky Cullum,
Director of Centre for Evidence Based Nursing, University of York

Dr Katherine Darton,
Information Unit, MIND – The Mental Health Charity, London

Professor Carol Dezateux,
Professor of Paediatric Epidemiology, London

Professor Martin Eccles,
Professor of Clinical Effectiveness, Centre for Health Services Research, University of Newcastle upon Tyne

Professor Pam Enderby,
Professor of Community Rehabilitation, Institute of General Practice and Primary Care, University of Sheffield

Mr Leonard R Fenwick,
Chief Executive, Newcastle upon Tyne Hospitals NHS Trust

Professor David Field,
Professor of Neonatal Medicine, Child Health, The Leicester Royal Infirmary NHS Trust

Mrs Gillian Fletcher,
Antenatal Teacher & Tutor and President, National Childbirth Trust, Henfield, West Sussex

Ms Grace Gibbs,
Deputy Chief Executive, Director for Nursing, Midwifery & Clinical Support Servs., West Middlesex University Hospital, Isleworth, Middlesex

Dr Neville Goodman,
Consultant Anaesthetist, Southmead Hospital, Bristol

Professor Robert E Hawkins,
CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor F D Richard Hobbs,
Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham

Professor Allen Hutchinson,
Director of Public Health & Deputy Dean of ScHARR, Department of Public Health, University of Sheffield

Professor Rajan Madhok,
Medical Director & Director of Public Health, Directorate of Clinical Strategy & Public Health, North & East Yorkshire & Northern Lincolnshire Health Authority, York

Professor David Mant,
Professor of General Practice, Department of Primary Care, University of Oxford

Professor Alexander Markham,
Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Dr Chris McCall,
General Practitioner, The Hadleigh Practice, Castle Mullen, Dorset

Professor Alistair McGuire,
Professor of Health Economics, London School of Economics

Dr Peter Moore,
Freelance Science Writer, Ashtead, Surrey

Dr Andrew Mortimore,
Consultant in Public Health Medicine, Southampton City Primary Care Trust

Dr Sue Moss,
Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton, Surrey

Professor Jon Nicholl,
Director of Medical Care Research Unit, School of Health and Related Research, University of Sheffield

Mrs Julietta Patnick,
National Co-ordinator, NHS Cancer Screening Programmes, Sheffield

Professor Chris Price,
Visiting Chair – Oxford, Clinical Research, Bayer Diagnostics Europe, Cirencester

Ms Marianne Rigge,
Director, College of Health, London

Professor Sarah Stewart-Brown,
Director HSRU/Honorary Consultant in PH Medicine, Department of Public Health, University of Oxford

Professor Ala Szczepura,
Professor of Health Service Research, Centre for Health Services Studies, University of Warwick

Dr Ross Taylor,
Senior Lecturer, Department of General Practice and Primary Care, University of Aberdeen

Mrs Joan Webster,
Consumer member, HTA – Expert Advisory Network

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278