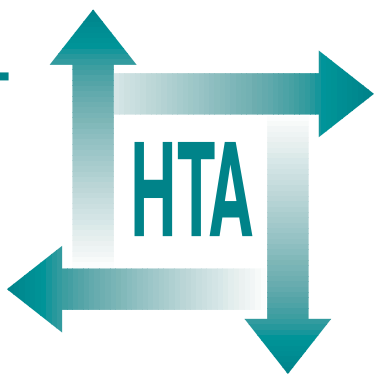
Health Technology Assessment 2000; Vol. 4: No. 34

Methodology

A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies

RR MacLehose
BC Reeves
IM Harvey
TA Sheldon
IT Russell
AMS Black

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.

A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies

RR MacLehose[1]
BC Reeves[2*]
IM Harvey[3]
TA Sheldon[4]
IT Russell[5]
AMS Black[6]

[1] South Essex Health Authority, UK
[2] Clinical Effectiveness Unit, Royal College of Surgeons and Health Services Research Unit, London School of Hygiene and Tropical Medicine, UK
[3] School of Health Policy and Practice, University of East Anglia, UK
[4] York Health Policy Group, University of York, UK
[5] Department of Health Sciences, University of York, UK
[6] Sir Humphry Davy Department of Anaesthesia, University of Bristol, Bristol Royal Infirmary, UK

[*] Corresponding author

Competing interests: none declared

Published December 2000

This report should be referenced as follows:

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).

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

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.

Although the National Coordinating Centre for Health Technology Assessment (NCCHTA) commissions research on behalf of the Methodology Programme, it is the Methodology Group that now considers and advises the Methodology Programme Director on the best research projects to pursue.

The research reported in this monograph was funded as project number 93/45/02.

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

M

Contents

List of abbreviations

List of abbreviations

CAGB	coronary artery bypass graft*
CCS	comprehensive cohort study
CI	confidence interval
COF	control group outcome frequency
COLA	change-to-open-label
CONSORT	Consolidation of Standards for Reporting Trials
CPTT	clinician-preferred-treatment trial
DARE	Database of Abstracts of Reviews of Effectiveness
df	degrees of freedom*
DOM	Diagnostich Onderzoek Mammacarcinom
DRCD	double randomised consent design
ECMO	extracorporeal membrane oxygenation
EXV	external validity
FAS	folic acid supplementation
ICC	control group only available for QEO element of a strategy 1 comparison
ICI	intervention group only available for QEO element of a strategy 1 comparison
ICIC	intervention and control groups available for both RCT and QEO elements of a strategy 1 comparison
IOF	intervention outcome frequency
ITT	intention-to-treat
IVB	internal validity – bias
IVC	internal validity – confounding
MI	myocardial infarction*
MRC	Medical Research Council
MSBC	mammographic screening for breast cancer
NTD	neural tube defect
PPT	patient-preference trial
PRIT	placebo run-in trial
QEO	quasi-experimental or observational
RCT	randomised controlled trial
RD	risk difference
RDT	randomised discontinuation trial
REP	quality of reporting
RPWD	randomised play-the-winner design
RR	relative risk
SOLVD	studies of left ventricular dysfunction
SRCD	single randomised consent design
TSTD	two-stage trial design
TURP	transurethral resection of the prostate

* Used only in tables and figures

Executive summary

Background

There is controversy about the value of evidence about the effectiveness of healthcare interventions from non-randomised study designs. Advocates for quasi-experimental and observational (QEO) studies argue that evidence from randomised controlled trials (RCTs) is often difficult or impossible to obtain, or is inadequate to answer the question of interest. Advocates for RCTs point out that QEO studies are more susceptible to bias and refer to published comparisons that suggest QEO estimates tend to find a greater benefit than RCT estimates. However, comparisons from the literature are often cited selectively, may be unsystematic and may have failed to distinguish between different explanations for any discrepancies observed.

Objectives

The aim was to investigate the association between methodological quality and the magnitude of estimates of effectiveness by comparing systematically estimates of effectiveness derived from RCTs and QEO studies. Quantifying any such association should help healthcare decision-makers to judge the strength of evidence from non-randomised studies. Two strategies were used to minimise the influence of differences in external validity between RCTs and QEO studies:

- a comparison of the RCT and QEO study estimates of effectiveness of any intervention, where both estimates were reported in a single paper
- a comparison of the RCT and QEO study estimates of effectiveness for specified interventions, where the estimates were reported in different papers.

The authors also sought to identify study designs that have been proposed to address one or more of the problems often found with conventional RCTs.

Methods

Data sources

Relevant literature was identified from:

- the Cochrane Library, MEDLINE, EMBASE, DARE, and the Science Citation Index
- references of relevant papers already identified
- experts.

Electronic searches were very difficult to design and yielded few papers for the first strategy and when identifying study designs.

Choice of interventions to review for strategies 1 and 2

For strategy 1, any intervention was eligible. For strategy 2, interventions for which the population, intervention and outcome investigated were anticipated to be homogeneous across studies were selected for review:

- mammographic screening (MSBC) of women to reduce mortality from breast cancer
- folic acid supplementation (FAS) to prevent neural tube defects in women trying to conceive.

Data extraction and quality assessment

Data were extracted by the first author and checked by the second author. Disagreements were negotiated with reference to the paper concerned.

For strategy 1, study quality was scored using a checklist to assess whether the RCT and QEO study estimates were derived from the same populations, whether the assessment of outcomes was 'blinded', and the extent to which the QEO study estimate took account of possible confounding. For strategy 2, a more detailed instrument was used to assess study quality on four dimensions: the quality of reporting, the generalisability of the results, and the extent to which estimates of effectiveness may have been subject to bias or confounding. All quality assessments were carried out by three people.

Data synthesis and analysis

For strategy 1, pairs of comparisons between RCT and QEO study estimates were classified as high or low quality. Seven indices of the size of discrepancies between estimates of effect size and outcome frequency were calculated, where possible, for each comparison. Distributions of the size and direction of discrepancies were compared for high- and low-quality comparisons.

For strategy 2, three analyses were carried out:

- Attributes of the instrument were described by κ statistics, percentage agreement, and Cronbach's α values.
- Regression analyses were used to investigate variations in study quality.
- Meta-regression was used to investigate associations between study attributes and the size of estimates of effect for each intervention separately; the attributes considered included study design, study quality and sources of heterogeneity of the intervention and population between studies.

Results

Strategy 1

Fourteen papers were identified, yielding 38 comparisons between RCT and QEO study estimates; 25 were classified as low and 13 as high quality. Discrepancies between RCT and QEO study estimates of effect size and outcome frequency for intervention and control groups were smaller for high- than low-quality comparisons. For high-quality comparisons, no tendency was observed for QEO study estimates of effect size to be more extreme than RCT ones, but this tendency was seen with low-quality comparisons.

Strategy 2

Thirty-four papers were identified, 17 evaluating MSBC and 17 FAS; eight and four papers, respectively, were individually or cluster assigned RCTs, five and six were non-randomised trials or cohort studies, and three and six were matched or unmatched case–control studies. Two studies, one of MSBC and one of FAS, used some other study design.

κ statistics for most items were < 0.4, although the percentage agreement usually exceeded 60%. Cronbach α values for different aspects of quality were < 0.5, suggesting that the instrument had limited ability to differentiate aspects of quality.

Regression analyses showed that both cohort and case–control studies had lower total quality scores than RCTs; cohort studies also had significantly lower scores than case–control studies. The latter, counter-intuitive finding may reflect a general tendency for quasi-experimental studies (which must use cohort designs) to have lower quality than observational studies.

Meta-regression of study attributes against relative risk estimates showed no association between effect size and study quality. Estimates from RCTs and cohort studies were not significantly different, but case–control studies gave significantly different estimates for both MSBC (greater benefit) and FAS (less benefit).

Identification of study designs

Ten study designs were identified; four, which include elements of both RCT and QEO study methods, were classified as hybrids and six, which adhere to the principle of randomisation but include some modification, were classified as RCT variants. Apart from the two-stage trial design, hybrid designs assume that non-randomised estimates are unbiased and that discrepancies between RCT and non-randomised estimates reflect the factors of interest (e.g. treatment preference). The majority of RCT variants have been designed to overcome the problems of non-compliance and patient drop-out; these designs therefore promote measures of efficacy as opposed to effectiveness. Three other types of variant were identified, namely response adaptive, randomised consent designs and change-to-open-label.

Conclusions

The findings of strategy 1 suggest that QEO study estimates of effectiveness may be valid if important confounding factors are controlled for. The small size of discrepancies for high-quality comparisons also implies that psychological factors (e.g. treatment preferences or willingness to be randomised) had a negligible effect on outcome. However, the authors caution against generalising their findings to other contexts, for three main reasons:

- Few papers were reviewed, and the findings may depend on the specific interventions evaluated.
- Most high-quality comparisons studied RCT and QEO study populations that met the same eligibility criteria, which may have reduced the importance of controlling for confounding.
- The literature reviewed is likely to have been subject to some form of publication bias. Authors of papers appeared to have strong *a priori* views about the usefulness of evidence from QEO studies, and the findings of papers appeared to support these views.

Strategy 2 found no association between study quality and effect size for either intervention, after taking account of study design. The lack of

association between quality and effect size could have arisen for a variety of reasons, the most likely being that study quality is not associated with relative risk in a predictable way or that the instrument failed to characterise methodological quality adequately.

There are several possible reasons for the finding that effect size estimates for case–control studies were significantly different from those for RCTs and cohort studies. The inconsistency of the direction of the discrepancy suggests that the direction is unpredictable and may be intervention specific. Case–control estimates of effectiveness should therefore be interpreted with extreme caution.

Several study designs were identified, which had been proposed to overcome a range of problems experienced with conventional RCTs, although the reported advantages were rarely substantiated. Discrepancies between RCT and QEO study estimates should not be attributed to factors such as patient preferences by default, since there may be residual confounding. Randomising patients prior to obtaining consent can cause as many problems as it solves, but may be useful when patients have a strong preference for an intervention. Other RCT variants may have a role when the aim is to measure efficacy.

The primary aim of quantifying any association between methodological quality and effect size was thwarted by several obstacles. For objective 1, the authors were unable to draw strong conclusions because of the paucity of evidence, and the potentially unrepresentative nature of the evidence they reviewed. For objective 2, the authors were unable adequately to distinguish, and measure, the variations in different aspects of quality between studies. The authors' recommendations relate directly to these obstacles.

Recommendations

- Most quasi-experiments reviewed were of poor quality. Quasi-experimental designs should not be rejected on the basis of this evidence.
- Standards for reporting of quasi-experimental and observational studies should be introduced. Enforcement of such standards, in the long term, might be expected to improve the standard of the research as well as reporting.
- More direct evidence about the comparability of findings from RCTs and QEO studies is needed. The comprehensive cohort study is probably the best study design for obtaining such evidence. Studies should be carried out in areas where RCTs are the preferred design, and in areas where RCTs are problematic, to assess the generalisability of evidence about the validity of QEO studies.
- There is a need to develop methods for identifying studies that provide a direct comparison of estimates from randomised and non-randomised data. A register should be established and studies entered into the register as they are identified. There is also a need for innovative search strategies to be developed.

Developing an instrument to characterise the quality of different studies is an urgent priority. The instrument must be able to assess all aspects of study design that may influence effect size. Separate instruments may be required for different study designs.

Chapter 1
Introduction

The use of RCTs and QEO studies to measure effectiveness

There is a long-standing debate about the advantages and disadvantages of different research designs for assessing the effectiveness of healthcare interventions. This review arose from a perception that this debate has become polarised (see appendix 1), with strong advocates for randomised controlled trials (RCTs) on the one hand[1–7] and quasi-experimental or observational (QEO) research designs on the other.[8–12]

RCTs are widely perceived as the gold standard research design for evaluating effectiveness[1–7,13–15] because they minimise confounding of the intervention of interest by differences in known and unknown prognostic factors between groups. This advantage can be compromised if allocation is not truly random or if randomisation is inadequately concealed.[16] Other advantages include approximately balanced treatment allocation within subgroups which have differential outcomes, best achieved by stratified randomisation, and a clearly defined 'time zero'[4] (see chapter 7).

These strengths are particularly important when evaluating effectiveness because:

- small effects of comparable size to those arising from bias and confounding may be clinically important[2,17]
- quantifying the effect of an intervention accurately is very important, since all interventions have financial costs, and many have side-effects or complications, as well as benefits; the decision to adopt an intervention often depends on weighing up the relative magnitudes of the benefits and the complications and costs, and so accurate estimates are important.

The potential importance of small effects and the need to quantify effect sizes accurately have led to some researchers adopting an extreme position about the value of non-randomised (i.e. QEO) methods. For example:

> Observational methods provide no useful means of assessing the value of a therapy.
>
> (Doll,[18] page 313)

> [Non-randomised designs] cannot discriminate reliably between moderate differences and negligible differences in outcome, and the mistaken clinical conclusions that they engender could well result in the under-treatment, over-treatment or other mistreatment of millions of future patients.
>
> (Peto and co-workers,[17] page 24)

However, there are also proponents of using QEO research designs to evaluate effectiveness, because there are perceived to be many circumstances in which RCTs may be unnecessary, inappropriate, impossible or inadequate[9,12] (see *Box 1*). Black[12] and others[8] contend that the polarity of the debate about the value of QEO methods for health technology assessment arises because QEO study designs are seen as alternatives to experimentation, rather than as "a set of complementary approaches".[12] QEO studies can provide estimates of the effectiveness of interventions when RCTs are not possible,

BOX 1 Reasons given by Black[12] for using observational studies to evaluate the effectiveness of healthcare

Experimentation may be:

Unnecessary
- because the effect of an intervention is dramatic

Inappropriate
- due to the large sample sizes required to measure rare adverse outcomes or to evaluate interventions to prevent rare outcomes
- due to the duration of follow-up required to measure long-term outcomes
- because the act of randomly allocating participants may reduce the effectiveness of the intervention

Impossible
- due to the reluctance of clinicians and others to participate
- due to ethical objections
- due to political and legal obstacles
- due to contamination
- due to a lack of resources for health technology assessment

Inadequate
- due to the poor external validity of an RCT (the patients, health carers and interventions studied may not be representative)

and can help to interpret the findings of RCTs, for example the extent to which they can be generalised to patients not included in the original RCTs. Sackett and Wennberg[19] have argued that the complementary nature of different study designs reflects the different types of research questions that they are best suited to address.

The internal validity of non-experimental approaches must always be suspect, since it is impossible to be certain that all important confounding factors have been identified and adequately controlled for.[4,7,20,21] QEO studies also offer less opportunity to control for biases. Although outcome assessment can often be blinded, healthcare providers and study participants are usually aware of treatment allocations. The extent of the distrust of evidence derived from these non-experimental approaches is illustrated by the above quotations, although Black comments that:

> it is unclear how serious and how insurmountable a methodological problem the threat to internal validity is in practice.
>
> (Black,[12] page 1218)

RCTs also have disadvantages. Although they have good internal validity, they are usually expensive to set up and conduct and may raise ethical problems. There are also doubts about the generalisability of the findings from RCTs, since individuals who take part in RCTs are often highly selected.[12,22–24]

Differences of opinion about the value of QEO study designs for evaluating effectiveness may also depend on the type of intervention being evaluated. RCTs are relatively easy to carry out when evaluating pharmacological interventions, whereas the constraints described by Black[12] are more likely to arise when evaluating surgical interventions, alternative methods of healthcare delivery, or 'educational' or health promotion interventions. For example, surgeons may not be in 'equipoise'[25,26] or may be unable to carry out both the procedures being compared. Surgeons who are truly equivocal about alternative procedures, and who are prepared and able to carry them out, are unlikely to be representative of most surgeons operating on patients who have the problem of interest; randomising patients to specialist centres using one or other procedure is likely to be impracticable. For educational and health promotion interventions that are extremely unlikely to carry any risk and where the research question is about weighing up potential benefit against cost, interventions can be difficult to 'blind', and patients may be unwilling to be deprived of a potentially valuable intervention and can experience strong placebo effects. Contamination can occur if researchers randomise individuals in such circumstances, and randomising by cluster is usually logistically complex.[27]

The criticisms that are levelled at RCTs and QEO studies both predict that estimates of effect size are likely to differ for the two types of design, but for different reasons. On the one hand, if QEO studies have dubious internal validity, their results might be expected to differ from those of RCTs because of biases. On the other hand, if RCTs study highly selected populations, or are carried out in atypical settings, their results may not generalise to more inclusive populations and more typical settings usually studied by QEO study designs.

In order to use the results of RCTs and QEO studies appropriately to guide healthcare decision-making, questions about both the internal and external validity of QEO studies need to be answered:

Internal validity:

- To what extent can effect size estimates derived from QEO studies be trusted?
- What attributes promote the internal validity of QEO studies?
- Can one quantify additional uncertainty, over and above the statistical imprecision of effect size estimates, or consistent biases that should be considered when interpreting evidence from QEO studies?

External validity:

- What attributes of studies, other than differences in internal validity, cause RCTs and QEO studies to yield discrepant effect size estimates?
- Do some of these attributes have a consistent influence (direction and strength of effect) on effect size that might help users to interpret evidence from RCTs?

This review focuses primarily on the first set of questions (see chapter 2). The second set of question has been dealt with in detail by a second review commissioned to address the same brief.[28] However, it is very important to tease apart the influences of variations in internal and external validity on effect size estimates to answer either set of questions.

Availability of estimates of effectiveness from RCTs and QEO studies

Both sets of questions can only be addressed by reviewing interventions for which both types of evidence (i.e. from RCTs and QEO studies, are available).[4] It is therefore important to consider the circumstances in which this is likely to be the case, since conclusions drawn from comparing evidence from RCTs and QEO studies may differ according to the circumstances and may not generalise to contexts when one or other is not available. It is normally the lack of RCT evidence that precludes the comparison, since there are almost no practical constraints on carrying out QEO studies. However, QEO study evidence may also be unavailable in some contexts where RCTs are the study design of first choice (e.g. evaluation of therapeutic benefit from pharmaceutical interventions).

Three sets of circumstances are considered in order to illustrate the potential problem:

- when RCTs are straightforward to do
- when RCTs are difficult to do or their findings may be of questionable relevance
- when RCTs are effectively impossible.

When RCTs are straightforward to do, QEO studies of the same research questions are likely to be rare. If QEO studies have been carried out in this situation, it is important to ask why they were carried out. The primary objective of a QEO study may be to assess a different outcome (e.g. a rare adverse event),[12] while collecting data on the primary outcome studied in RCTs as a secondary objective. (QEO studies carried out to test the generalisability of findings from RCTs are considered as examples of the second type of situation.) However, if both RCTs and QEO studies are found which truly address the same research question, the reason that both kinds of evidence are available may simply be that the researchers who carried out the QEO studies were less knowledgeable about the advantages of RCTs and the susceptibility of QEO study designs to bias, or less prepared to make the additional investment required. The very existence of QEO studies in this situation may therefore be a warning that the QEO studies are likely to be flawed, for example because they failed to take account of confounding[29] or took inadequate precautions to avoid bias.

RCTs and QEO studies are most likely to be available when RCTs are difficult to do, or when researchers perceive that RCTs alone are inadequate to answer a particular research question. Here, one justification for the QEO studies is likely to be that they answer a slightly different question from the RCT, for example about a more inclusive population, or about interventions when implemented by ordinary practitioners or in ordinary settings. This slight difference in the research question (i.e. the population or intervention) may mean that the evidence from RCTs and QEO studies is not directly comparable. A second justification may simply be the logistical difficulties of carrying out an RCT, for example a lack of equipoise when evaluating an established treatment or the possibility of contamination between groups. This latter situation is the one in which directly comparable evidence from RCTs and QEO studies is most likely to be available, although factors such as varying amounts of contamination in RCTs would also be expected to lead to discrepant estimates of effect size.

When RCTs are effectively impossible, by definition only QEO studies will be available, although there may be many QEO studies and variation between their results. One way to investigate the validity of QEO studies in this situation would be to explore factors associated with the variation in effect size between studies. However, the interpretation of such an investigation would depend on establishing which factors best characterise internal and external validity in circumstances where evidence from RCTs and QEO studies can be compared directly. It should be noted that even this approach makes the assumption that these factors exert similar influences in evaluations of interventions that cannot be evaluated by RCTs as they do in evaluations of interventions that can.

Thus, consideration of the motivations leading to different types of study emphasises the need to be able to separate the influences on effect size exerted by differences in internal and external validity.

Differences between estimates of effectiveness from RCTs and QEO studies

When both RCTs and QEO studies of the effectiveness of an intervention exist, differences between the findings of the two types of study design have often been reported.[30–34] The majority of authors

have reported that QEO studies tend to produce larger estimates of effectiveness than do RCTs,[30–33] although there are also reports of QEO studies producing smaller estimates.[34,35]

Attention has been focused on the tendency for QEO studies to produce larger effectiveness estimates, since this is the predicted direction of the influence of most biases and confounding. A few, frequently cited reviews that have compared the findings of RCTs and QEO studies of interventions[30–33] have therefore been very influential in reinforcing the perception that RCTs are the gold standard method for evaluating effectiveness. This perception has, in turn, led to the assumption that comparing effect size estimates from RCTs and QEO studies is a way of testing the internal validity of specific QEO studies.

Unfortunately these often-cited studies may themselves be subject to biases. Most are relatively crude comparisons with little or no attempt to quantify the magnitude of the differences between RCT and QEO study estimates of effectiveness, and without consideration of the precision of estimates. For example, in one case the findings of studies were simply classified as significant or not.[31] Data were inappropriately pooled in a review which compared RCT and QEO study estimates of the effectiveness of anticoagulants in following myocardial infarction.[30] None of the reviews considered the possibility that publication bias might affect RCT and QEO studies differentially, that RCTs and QEO studies may have been carried out over different time periods, or the possibility that discrepancies may result from differences in the precise research question addressed by individual primary studies (i.e. differences in population, outcome or intervention).

Other reviews which have examined sources of bias and confounding more carefully have not compared RCTs and QEO studies. Schulz and co-workers[16] reviewed RCTs identified by the Pregnancy and Childbirth Group of the Cochrane Collaboration[36] and demonstrated how a lack of blinding and a failure to conceal randomisation adequately led, on average, to important differences in prognostic factors between treatment groups which favoured the treatments being evaluated. Concato and co-workers[20] showed that apparent increases in mortality among men undergoing transurethral resection of the prostate (TURP) compared with open prostatectomy (the more invasive procedure) demonstrated by analysing observational databases were likely to arise from confounding by comorbidity despite researchers' attempts to control for confounding when analysing their data.

The above discussion points to the need to identify factors that influence effect size in RCTs and QEO studies. When these factors affect the different study designs differentially, they may give rise to differences between the respective effect size estimates. It is also important to consider the strength and likely direction of the influence for each factor.

The most important factors that influence effect size are well known, and are discussed below under the headings of internal and external validity. We also discuss some other factors which, it has been suggested, may influence effect size but which do not obviously fall under one of the headings. Finally, we consider sources of bias that can arise when reviewing discrepancies in effect size estimates between RCTs and QEO studies.

Factors which threaten internal validity

Factors affecting internal validity are considered under three main headings:

- information biases
- selection biases
- 'differential care' biases.

The likely effects of different kinds of bias on effect size estimates are summarised in *Table 1*.

Information biases

Information bias arises when there is misclassification of, or error in measuring, outcomes or confounding variables. In RCTs and cohort studies, misclassification and measurement error can affect the intervention and control groups equally (non-differential bias) or unequally (differential bias). In case–control studies, it is the misclassification of exposure and confounding variables among cases and controls that gives rise to information bias, which can similarly be non-differential or differential.

Independent non-differential misclassification of an **outcome** (RCT or cohort) or **intervention or exposure** (case–control) usually biases effect size estimates towards a null result, although there are circumstances when this is not the case.[37]

The effect of differential misclassification or measurement error of an outcome in an RCT or

TABLE I *Most likely effects of different kinds of bias on effect size estimates*

	More extreme	Less extreme	Either
Information bias			
Outcome: non-differential		✓	
Outcome: differential	✓		
Intervention: non-differential		✓	
ght Intervention: differential	✓		
Confounder: non-differential			✓
Confounder: differential	✓		
Selection bias or confounding			
RCTs and quasi-experimental cohort studies	✓		
Observational cohort studies			✓
Case–control studies			✓
Differential care bias	✓		

cohort study depends on the direction of the differential misclassification. However, the direction is usually towards a more extreme or beneficial effect[16] because of the probable underlying reason for the differential bias. For example, bias may arise when the person responsible for measuring the outcome (researcher or self-report by a study participant) is not blinded to the group allocation of the participant. The outcomes for participants in the intervention group can be differentially biased towards more benefit because of the researcher's desire to show a difference or because the participant experiences a placebo effect.

Differential misclassification or measurement error of an exposure in a case–control study gives rise to the same problems when the measurement (or self-reporting) of exposure is influenced by the researcher's (or participant's) knowledge of the case or control status of a participant.

Non-differential misclassification or measurement error of a confounding factor results in residual confounding;[20,37] that is, confounding that cannot be 'controlled for' by design features or when analysing the data. The direction of the effect is dependent on the direction of confounding. Differential misclassification or measurement error of a confounding factor is less likely, since the effect of the bias on the comparison between the groups being compared is less obvious. Nevertheless, when case-mix adjusted mortality rates were published for New York cardiac surgeons, the surgeons apparently reduced their thresholds for recording risk factors so that their case-mix adjusted outcomes would be improved.[38] Therefore, if differential misclassification or measurement error of a confounding factor exists, it is also likely to lead to a more extreme or beneficial effect.

Selection biases

Selection biases occur when important prognostic factors are not distributed equally among the groups being compared, and they can result in confounding. In prospective studies, selection biases are usually introduced by the biased allocation of patients to groups. In RCTs and quasi-experimental studies, where allocation is under the control of the researchers or the clinicians treating a patient, biased allocation is most likely to lead to a more extreme or beneficial effect. In observational studies, selection bias is likely to arise simply because clinicians offer treatments to patients most likely to benefit from them (i.e. confounding by clinical indication). The bias may arise in quasi-experimental studies because both researchers and clinicians have a vested interest in the outcome and they tend to select patients who are more likely to have better outcomes, or refuse to recruit those more likely to have poorer outcomes, for the group receiving the intervention under evaluation. However, selection bias can also lead to a less extreme effect, for example if researchers elect to try out a new treatment on more sickly patients who are judged to be unlikely to benefit from the standard treatment but who are also at higher risk of a poor outcome.

In RCTs, selection bias tends to occur when allocation is unconcealed or carried out using a pseudo-random method.[16] In quasi-experimental comparisons using contemporaneous controls, selection bias may be introduced when allocation to treatments is confounded by case mix. For example, a centre that has pioneered a new treatment may set out to compare its results with another centre

which is not offering the new treatment but which also has a different population and disease profile. Such comparisons may also be affected by a 'differential care' bias. In quasi-experimental comparisons using historical controls, the treatment comparison can be confounded both by case mix, as a result of researchers selecting less ill or otherwise less risky patients for the new treatment, and by calendar time, since there are likely to be general improvements in healthcare, and consequently in prognosis, as time passes.[29]

In truly observational studies patients are not, strictly speaking, allocated. However, selection biases can still occur in cohort studies because clinicians' choices of treatments for patients are typically influenced by clinical (often prognostic) and demographic information about the patient.[4] The direction of such biases is unpredictable, but tends to result in exaggeration of the benefit of an intervention since clinicians tailor treatments to patients in an attempt to optimise their outcome (confounding by clinical indication).

Selection biases arise in case–control studies from the unequal distribution of important prognostic factors among cases and controls. Their effects are also unpredictable since they depend on the nature of the association between the prognostic characteristics of patients and the probability of receiving the intervention of interest. When people with better prognoses are more likely to receive the intervention, confounding from selection biases will tend to exaggerate the benefit of the intervention. Conversely, when people with poorer prognoses are more likely to receive the intervention, confounding from selection biases will tend to underestimate the benefit of the intervention.

Researchers who use QEO study designs typically attempt to control for confounding by matching, at the design stage, or by stratification or regression methods at the stage of data analysis. It is important to point out that these methods are unlikely completely to eliminate confounding, because the confounding factor is measured either imperfectly or with insufficient precision. The residual confounding that remains can still bias the estimate of effect size to an extent that could be mistaken for a clinically important effect.

Selection biases can arise in RCTs by chance, typically when groups are small (< 50 patients per arm or stratum where stratified randomisation has been used). By definition, the direction of any influence of chance differences in prognostic factors between groups will be unpredictable. Chance may also give rise to unpredictable selection biases in QEO studies.

Differential care biases

The 'open' nature of prospective non-randomised studies may lead health practitioners to alter aspects of their care other than the intervention under investigation. Participants may also alter their health-related behaviour because of their knowledge of the treatment they have received. Because of clinicians' vested interests in a new treatment, and a tendency for patients to 'believe in' new treatments compared with standard ones, knowledge of patients' treatment allocations is likely to lead to changes in the behaviour of practitioners and patients that favour a new treatment.

Factors influencing external validity

Researchers rarely design new studies to be exact replications of previous ones, and usually change the definition of one or more of the key components of an evaluation (i.e. the population, the intervention (or control treatment) or outcome) in more or less subtle ways. Researchers may be motivated by wanting to test whether the change gives rise to a different answer or the same answer as previous studies, and either finding is possible.

Some changes introduced by researchers are likely to be associated with study design, because different study designs are more suited to some situations than others. QEO studies are likely to adopt more inclusive eligibility criteria and to recruit less specialist clinicians or settings than RCTs in order to test generalisability.

Broadening the eligibility criteria for patients usually means that sicker patients are included. However, the likely impact on effect size of studying an increased proportion of sicker patients is not clear. There are some medical conditions, such as diabetes, where pathology due to the advanced stage of disease is likely to be irreversible. Including diabetics with more advanced disease in an evaluation of a new diabetic treatment might therefore be expected to reduce the overall effect size compared with an evaluation in patients with less advanced disease, since the former have less capacity to benefit. On the other hand, the capacity of patients with cataract to benefit from cataract extraction, for example, is likely to be independent of the severity of the cataract (after taking account of confounding by age, if applicable).

Recruiting less specialist clinicians or settings to deliver the intervention might be expected to lessen the benefit if the success of a new treatment depends on the specialist expertise of the attending clinician (e.g. when evaluating a new surgical procedure) but perhaps not in other circumstances (e.g. when evaluating a new drug). Delivering a new intervention in less specialist settings may also require the intervention itself to be modified, although such a change may be difficult to distinguish from the dependency of an intervention on specialist expertise. Choosing a different outcome to study (e.g. side-effects rather than clinical benefit) represents an entirely different research question that may reverse the direction of effect (i.e. a new intervention may have both greater clinical benefits and more side-effects than a control treatment).

Studying the influences of variations in external validity was not an objective of this review. (Some of these factors have been reviewed by Britton and co-workers.[28]) However, it is important to note that the effect of differences in external validity between studies on effect size depends on what, precisely, differs between studies. The effect is often unpredictable and is likely to vary from one evaluation to another.

Preference, placebo and other psychosocial factors

RCTs are experiments in which patients are allocated to alternative treatments by chance. The highest quality RCTs are widely considered to be those that blind clinicians, researchers and patients to treatment allocations in order to prevent bias.[39] Patients are almost always required to give their consent to be recruited into RCTs and may have psychosocial responses as a result of their knowledge that they are participating in an experiment. All RCTs deprive participants of any choice in their treatment, and they may experience a sense of uncertainty and powerlessness when blinded. Unblinded RCTs are likely to induce different kinds of response after randomisation and these will be discussed separately.

Successful blinding means that patients' uncertainty about their treatment allocation continues, either until the end of the scheduled period of follow-up or until preset criteria (usually adverse events) are satisfied. The effect of uncertainty on health outcomes may depend on the condition and the intervention being studied, but is unlikely to be beneficial; the size of any psychosocial effect may also be increased when a patient has a strong preference and blinding is successful. Providing that blinding is maintained (allocation may become apparent to doctor or patient if a treatment is markedly effective or has noticeable side-effects), any bias towards less favourable outcomes will be non-differential and will tend to lead to underestimation of the true effect size.

Blinding also prevents any placebo effect, which is widely considered to be an advantage of RCTs. However, it is important to remember that the placebo effect contributes to the benefit that patients experience in usual clinical practice (and unblinded RCTs and QEO studies).[24] A RCT in which patients are blinded therefore measures the minimum benefit that might be expected from an intervention.[12] Unfortunately, in unblinded studies, it is almost impossible to distinguish between a true placebo effect and information or differential care biases.

In unblinded studies, uncertainty about treatment allocation ceases at the time of randomisation, but may be replaced by a sense of satisfaction if a patient is allocated to the treatment that was preferred at the outset. Alternatively, patients may experience a sense of disappointment if allocated to the treatment that was not preferred; this has been termed 'resentful demoralisation' by Cook and Campbell.[40] (Some patients may have no preference and therefore may not experience either satisfaction or disappointment.) Any influence on health outcome arising from these feelings will be non-differential if the probability of preferring the new treatment is 0.5, otherwise the bias will be differential.[24,41,42] For example, if the majority of patients prefer the new treatment and preference for the treatment has a beneficial effect on outcome, the effect size estimate will be greater than if the treatment allocation had been blinded. In many circumstances, a psychosocial effect of congruency between preferred and allocated treatments may be difficult to distinguish from differential care bias (e.g. effects mediated through better compliance or other changes in health behaviour). It is also conceptually difficult to distinguish between a placebo effect and other beneficial psychosocial effects of receiving a preferred treatment; both are likely to be dependent on a patient's belief in, or preference for, a particular treatment. One might expect preference effects to be enhanced by highlighting treatment comparisons (e.g. by obtaining informed consent).

Patients may experience a range of other emotions or changes in health behaviour that could influence health outcomes in a non-differential way:

- satisfaction from perceived altruism in participation
- satisfaction from a perception of 'better than usual' care or attentiveness arising from participation (including better rapport with carers or access to better information about treatments and prognosis)
- in blinded studies, satisfaction from the possibility of having been allocated to a new treatment which would not otherwise be available
- in blinded studies, overall improved compliance.

The extent to which psychosocial factors should be considered in truly observational studies is unclear. Such studies are not strictly blinded and patients are usually involved in their choice of treatment but, historically, they have often not been formally consented and have been unaware of the treatment comparison of interest to researchers. They are therefore less likely to change their health behaviour or experience particular satisfaction from receiving a preferred treatment; in so much as preference and placebo effects may be tapping the same underlying phenomenon, placebo effects may be less strong in observational studies.

Possible biases in reviews comparing RCTs and QEO studies

Reviews comparing estimates of effect size from RCTs and QEO studies can be biased by factors that are associated with the size of discrepancy and which are also associated with study design; this bias is analogous to confounding in primary studies[43] and we refer to it as **meta-confounding**. Such factors include:

- publication biases
- changes in the effectiveness of an intervention with calendar time
- variations in external validity
- differences in the types of intervention evaluated.

Publication bias may differentially affect RCT or QEO studies. For example, studies which conclude that there is no treatment difference may be more difficult to publish when based on a QEO study design rather than an RCT. A bias of this kind may change over time, as the reporting of well-designed trials that find interventions to be ineffective is encouraged and general suspicion of QEO studies of effectiveness increases.

Secular changes in the effectiveness of an intervention can introduce bias into reviews if there is a tendency for the majority of QEO studies to have been carried out either before or after RCTs of the same intervention. This bias may be more likely in areas in which RCTs are difficult to do (e.g. evaluations of the effectiveness of an established treatment), since RCTs may be perceived to be ethical only when a sufficient weight of observational evidence has accumulated.

As already discussed, there may be consistent variations in external validity between RCTs and QEO studies (e.g. with respect to eligibility criteria for both patients and clinicians or settings). Reviews of discrepancies in effect size between RCTs and QEO studies will be biased if these differences in external validity are associated with effect size. For example, if QEO studies tend to recruit sicker patients than RCTs and sicker patients benefit more from an intervention, the average discrepancy across studies will be overestimated. This bias can occur even when analyses of primary studies adjusted appropriately for confounding and when the primary studies reviewed were designed to evaluate the same[30,34] as well as different interventions.[32,33]

Where reviews include RCTs and QEO studies of many different interventions,[32,33] any tendency for certain kinds of intervention (e.g. drugs, surgical procedures, or health promotion or education interventions) to be evaluated more often by one or other study design can introduce bias if different kinds of intervention are more or less likely to be effective. Such biases are quite plausible, since it is easier to evaluate some interventions than others by RCTs.

The direction of biases that can occur when reviewing discrepancies between the effect sizes observed in RCTs and QEO studies are usually unpredictable. Information about the likely direction of bias from changes in effectiveness with calendar time may be gleaned by examining whether RCTs of an intervention tend to precede or follow QEO studies. The likely direction of bias arising from variations in external validity may be predictable if the effect of variations in external validity and their distribution in RCTs and QEO studies are known.

Summary

The above background to the review has attempted to show the varied ways in which discrepancies can arise between effect size estimates of the same

intervention derived from RCTs and QEO studies. Indeed, in these circumstances, perhaps one should expect discrepancies to exist. Understanding why a discrepancy exists is crucial if healthcare planners and providers are to act appropriately on both RCT and QEO evidence.

Factors that influence effect size can be broadly categorised as pertaining to internal or external validity, although some factors such as preference effects are difficult to classify in this framework. In general, factors that threaten internal validity lead to overestimates of effect size; this tendency may be less strong in the case of truly observational studies, especially case–control studies. The effect of variations in external validity are generally unpredictable, or depend on the specific population or intervention under evaluation.

The central problem is that the different influences on effect size are difficult to separate out. In reality, RCTs and QEO studies of exactly the same research question are extremely unlikely to be available. Where both types of study design have been used to evaluate the same intervention, it is likely that differences in external validity are central to the research questions. This review focuses on the internal validity of QEO study designs. The choice of specific objectives (see chapter 2) was primarily determined by the need to control as far as possible for differences in external validity between RCTs and QEO studies.

Chapter 2

Hypotheses tested

The starting point for the review was the assumption that the effectiveness of an intervention should not depend on the study design used to estimate it. If discrepancies exist, the challenge is to find explanations for them. From this perspective, observed discrepancies in effect size estimates from studies using different designs constitute data that can be used to investigate reasons for the discrepancies. Thus, in this review we aimed to investigate the size and direction of discrepancies between RCT and QEO study estimates, and the extent to which the discrepancies were associated with variations in methodological quality, external validity and other attributes of studies. We reasoned that being able to quantify such associations would be valuable to healthcare decision-makers in guiding the weight that they should attach to evidence about the effectiveness of interventions obtained from studies other than RCTs.

The primary strategy was to investigate in what circumstances, if any, estimates of effect size derived from QEO studies are internally valid. As part of this strategy, we also aimed to characterise attributes of QEO studies that could be shown consistently to threaten internal validity, with a view to recommending 'safety limits' for the interpretation of QEO studies with varying internal validity. As already described, characterising the effect of threats to internal validity requires the effect of variations in external validity to be minimised. We adopted two strategies to achieve this, as described below.

Strategy 1

The first strategy was to compare estimates of the effectiveness of an intervention as derived from RCT and QEO study design elements where both estimates were reported in a single paper. We reasoned that comparisons reported in such papers were more likely to compare like with like than were comparisons across papers where each paper reported an estimate for only one study design. Evaluations of any intervention were eligible for this strategy.

We intended to quantify

- the discrepancy between the effect size estimates for RCT and QEO study design elements
- threats to the internal validity of the observed estimates
- variations in external validity between the different study designs

and then to investigate the extent to which the latter measures could explain the size and direction of the observed discrepancies.

Strategy 2

The second strategy was to compare estimates of effectiveness derived from RCT and QEO study designs for interventions for which the intervention, population and outcome investigated were anticipated to be homogeneous across studies, and where RCT and QEO study estimates of effectiveness were reported in different papers. By careful selection of the interventions to be reviewed against criteria chosen to minimise variations in external validity, we hoped to be able to avoid possible biases that can affect reviews of primary evidence (see chapter 1).

As for strategy 1, we intended to quantify

- the discrepancy between the effect size estimates derived from RCTs and QEO study designs
- threats to the internal validity of the observed estimates
- any residual variations in external validity between the different study designs

and then to investigate the extent to which the latter measures could explain the size and direction of the observed discrepancies.

Hybrid designs and RCT variants

Because of the anticipated difficulty of 'controlling' for variations in external validity between primary studies that used different study designs, we were also interested in identifying 'hybrid' designs that combine both RCT and QEO study

design elements in the same or similar settings. This aim was extended to include variations of conventional RCTs, designed to overcome some of the problems typically experienced when carrying out RCTs. We sought to describe these designs, to describe the problems which the designs were intended to overcome and to review their advantages and disadvantages.

Chapter 3
Methods

Project administration

The project was based at the R&D Support Unit, Department of Social Medicine, University of Bristol. Rachel MacLehose was appointed as research associate in January 1996. A project steering group, which met four times during the review, was formed consisting of:

- Dr Barnaby Reeves (BCR), Department of Social Medicine, University of Bristol
- Dr Ian Harvey (IMH), Department of Social Medicine, University of Bristol
- Dr Andrew Black (AMB), Department of Anaesthetics, University of Bristol
- Professor Ian Russell (ITR), Department of Health Sciences, University of York
- Professor Trevor Sheldon (TAS), NHS Centre for Reviews & Dissemination, University of York
- Professor George Davey-Smith (GDS), Department of Social Medicine, University of Bristol
- Rachel MacLehose (RRM), Department of Social Medicine, University of Bristol.

Because the steering group members were drawn from Bristol and York, it was not possible for all members to attend meetings. GDS helped to choose possible interventions to review for strategy 2 (see chapter 4) at the start of the project, but was unable to continue as a steering group member due to other commitments.

The controversial nature of the review topic meant that some members of the steering group had quite strong opinions about the likely findings at the outset. In view of the difficulties that we experienced in identifying relevant literature for some strategies and our perception that many of the papers which we reviewed supported the authors' perspective at the outset, it is interesting to describe our own viewpoints before we started the review. Opinions of steering group members fell into three broad categories, with no single category dominating:

- QEO studies likely to give invalid and generally 'overoptimistic' results compared with RCTs (ITR, TAS and GDS)
- high-quality QEO studies likely to give valid results which are generally comparable to RCTs (BCR and AMB)
- no strong opinion (IMH and RRM).

Methods used for strategy 1

Literature searching

Electronic databases

Electronic searches of MEDLINE (1966 to June 1996) and EMBASE (1980 to June 1996) databases were carried out to identify relevant papers. The Cochrane RCT search strategy[44] was used in conjunction with search terms designed to identify observational studies (see appendix 2). This was not a fruitful strategy due to the inadequacy of the indexing of methodological aspects of studies.

MEDLINE has traditionally been indexed by medical subject terms (MeSH headings) rather than by terms relevant to study design. The Cochrane search strategy for RCTs[44] contains terms such as 'comparative studies' and 'prospective studies', which could apply equally to QEO studies as to RCTs. The term 'randomised controlled trial' is available as a publication type, but this classification was only introduced in 1992 and therefore covers only a small portion of our search period. RCTs published before 1992, which were identified by handsearching by review groups of the Cochrane Collaboration, are being re-indexed but this process is incomplete. Moreover, the use of the term 'randomised controlled trial' has not been consistently applied by reviewers at the National Library of Medicine, despite its availability.[44]

Because of these limitations, the use of an 'and' Boolean operator to identify papers that satisfied both the Cochrane RCT and the observational search criteria primarily detected studies which used a single study design. Search strategies with sufficient breadth to detect the majority of papers relevant to strategy 1 generated such a large number of references that the task of identifying

the very small proportion of relevant papers was like searching for a needle in a haystack.

Handsearching

Traditional handsearching methods were rejected for this strategy because:

- it was not obvious which journals should be searched
- it was considered that, as was found in the case of electronic searching, the ratio of relevant to irrelevant papers would be very low.

Handsearching would therefore have occupied a disproportionate amount of project time.

However, we did handsearch abstracts in four databases:

- a database containing 1535 references which had been constructed by another HTA Methodology Review project team[45] (because of the nature of the latter review, we suspected that this database might contain a higher proportion of relevant literature than would selected journals)
- Cochrane Review of Methodology[46]
- Cochrane Database of Systematic Reviews[47]
- Database of Abstracts of Reviews of Effectiveness (DARE).

In effect, these databases were searched for all review objectives simultaneously.

Expert knowledge

The project panel members identified relevant papers of which they were already aware, and searched their personal literature collections. Other experts were also contacted.

'Cascade' referencing

References that were cited in eligible papers for any of the review objectives and that were considered to be relevant were also obtained.

Systematicity

The project steering group considered that studies relevant to strategy 1 were central to addressing the brief for this review (see appendix 1). It was therefore our intention to identify all literature relevant to this strategy, irrespective of the intervention evaluated. However, for the reasons given above, we do not have high confidence that we have been successful in achieving this. A comparison of the papers that we identified for this strategy with those identified by Britton and co-workers,[28] who also sought papers relevant to this strategy, may make it possible to estimate the extent of relevant unidentified literature using capture–recapture methods.[48] The possible consequences of failing to identify a proportion of the relevant literature are discussed later (see chapter 8).

Assessment, data extraction and synthesis

Assessment

RRM read all the eligible papers that were identified for this strategy, and eligibility was confirmed by BCR. Papers reviewed in detail were assessed with respect to their quality by RRM, BCR and IMH. This assessment was not carried out using the instrument described under strategy 2 because the instrument was not designed for review studies or studies which contained multiple design elements. Instead, a simple assessment was carried out, focusing on three aspects of quality which were considered to be central to a valid comparison of RCT and QEO study effect sizes (see appendix 3):

- comparability of overall study populations, specifically with respect to the use of the same eligibility criteria (1 point), and the same time periods over which the study populations were recruited and followed up (1 point)
- adjustment of the QEO study estimate to take account of possible confounding by severity of disease (1 point), comorbidity (1 point) and other prognostic factors (1 point)
- blinding of the assessment of outcome or the use of an outcome, such as death from all causes, which is not susceptible to bias (1 point).

The scores for each reviewer were summed, giving a score out of 18.

Data extraction

RRM extracted information from the papers about the interventions and populations studied, sample sizes and outcome frequencies for RCT and QEO study elements, and RCT and QEO study estimates of effectiveness. BCR also checked these data. The two reviewers (RRM and BCR) who extracted data were not blinded to the findings obtained by the other. However, all extracted data were finally checked by RRM and BCR together, to resolve differences of opinion, because there were many uncertainties arising from poor reporting quality.

Data synthesis

No attempt was made to synthesise evidence across interventions. Instead, distributions of indices of

discrepancy between RCT and QEO study estimates were compared for high- and low-quality comparisons, as determined by the assessment made of each paper.

Methods used for strategy 2

Literature searching

Electronic databases

MEDLINE (1966 to June 1996) and EMBASE (1980 to June 1996) databases were searched for RCTs and QEO studies of the intervention that were chosen for review (see chapter 4). The Cochrane Database of Systematic Reviews,[47] Cochrane Controlled Trials Register[49] and DARE databases were also searched. Both primary studies and reviews were sought, since the latter were a source of primary evaluations. Abstracts of all papers identified by searches were read carefully and the full text of the original paper was obtained for any abstract that appeared relevant.

The MEDLINE searches used:

- the Cochrane RCT search strategy
- an observational search strategy devised by RRM
- MeSH terms and relevant text words for each of the interventions.

By using appropriate MeSH terms and text words for the respective interventions and associated pathologies, it was possible to design MEDLINE searches which yielded a relatively high ratio of relevant to irrelevant studies.

EMBASE searches were always carried out last. They followed the same principle as the MEDLINE searches but were designed to be less restrictive. They were intended to represent a final 'trawl' for papers that may have been missed by other searches. The searches were not limited to the English language. The MEDLINE and EMBASE electronic search strategies used are given in appendix 2.

Handsearching

Handsearching methods were rejected for this strategy because:

- relevant papers identified by electronic searches were published in a wide range of journals, giving no indication of which journals would be most appropriate for handsearching
- it was considered that the ratio of relevant to irrelevant papers would be very low and that handsearching would have occupied a disproportionate amount of project time.

Expert knowledge

Members of the panel identified relevant papers using their knowledge of the literature, and from their personal collections. Experts in each of interventions reviewed were also contacted. These experts included authors of papers that had already been identified, and researchers at relevant research institutes and charitable organisations.

'Cascade' referencing

This method of identifying relevant papers was also used.

Systematicity

We were primarily interested in the relationship between quality and effect size for each type of study design, rather than in a pooled estimate for each study type. Initially, we considered that it was sufficient to obtain a representative sample of studies using different designs and of varying quality to address strategy 2 in a valid manner. This view was taken by analogy with primary research – if an association were to exist between study quality and effect size among research studies of the two interventions, there should be no need to study the whole population of research studies, but only a representative sample.

On further consideration, it seems quite possible that publication bias could be more or less of a problem for studies of different quality or which used different designs (see chapter 1). We therefore searched intensively for all papers that evaluated each intervention chosen for review.

Assessment and data extraction

Assessment

Details of the journal, author, year of publication and all references were removed from papers chosen for review before they were distributed to three of the steering group panel for assessment of their methodological quality. If the methods of the study were reported in detail in a separate paper, relevant sections of the additional paper were also distributed with the paper under evaluation (also after removing any publication details).

All the papers were evaluated by RRM and by two other members of the project steering group. The papers were ordered randomly and then distributed in blocks to steering group members. One panel member (GDS) was unable to review the

papers allocated to him and these papers were redistributed to BCR and IMH.

The instrument

The instrument used for assessing methodological quality was based on a prototype of an instrument developed by Downs and Black[50] to provide a measure of internal and external validity for both randomised and observational study designs. The original instrument was subdivided into five sections:

- quality of the reporting of the study
- external validity
- internal validity – bias
- internal validity – confounding or selection bias
- power of the study to detect a clinically important difference.

The instrument was modified extensively for the purposes of this review in order to try to satisfy our aim of quantifying the relationship between study quality and effect size. Modifications were made, or additional questions developed, to quantify the extent to which factors likely to influence effect size (e.g. adjustment for confounders, resolution with which confounding variables were measured) were addressed by a study. The original questions about external validity were also merged into a single question. By assessing all the studies on the same question about external validity, we sought to obtain some indication of the external validity of different study designs and hence the extent to which differences in effect size could be attributed to this factor.

The instrument underwent piloting and several revisions before being used to assess the papers for strategy 2; the final version is described in full in appendix 4. Despite piloting, further ambiguities emerged during the assessment of papers. Additional revisions to take account of these problems are suggested in appendix 4.

Analysis of the performance of the instrument

We originally planned to use data reduction techniques to identify a smaller number of 'quality dimensions' from the items in the instrument. The techniques would have validated the original subdivisions of the instrument and possibly identified redundancy. Due to the small number of papers evaluated, this was not possible.

Instead, RRM, BCR and IMH identified items in the instrument that were considered to be the most important and unambiguous 'quality questions'. These questions were categorised according to the first four dimensions identified by Downs and Black:[50]

- quality of reporting of the study (REP)
- external validity analysis (EXV)
- susceptibility to bias (IVB)
- susceptibility to confounding (IVC).

We did not include the question about power from the prototype. This question simply regarded studies that had a sample size greater than an arbitrary cut-off limit as being of better quality than those which did not. The question in the prototype was modified in the version subsequently published.[50]

Scores from questions designated as pertaining to a dimension were summed. Summed scores for each dimension were also pooled to give an overall quality score. We made no attempt to weight the importance of different dimensions.

Inter-rater reliabilities for each question (pooled across all papers for the two health technologies) were determined by two methods:

- Unweighted κ values (calculated using the statistical package STATA Version 4.0; the use of three assessors precluded the calculation of weighted κ values for ordinal responses).
- Percent agreement: 3 points were awarded where all three assessors agreed, and 1 point where two of the three assessors agreed using the formula; points were summed across 36 papers (maximum 108) and expressed as a percentage.

Internal consistency of the items contributing to three of the four dimensions (REP, IVB, IVC) was described using Cronbach's α, stipulating that the direction of scores for each item should be maintained since zero had already been designated the lowest quality score for each item. Cronbach's α could not be calculated for EXV because this dimension was assessed by only one item.

An attempt was made to validate the four quality dimensions and the sum of the four dimensions by investigating predictors of quality scores. Study design (RCT, cohort study, case–control study), intervention and the interaction of study design and intervention were entered into the model. It was hypothesised that RCTs should have higher mean IVB and IVC scores than QEO studies but a poorer EXV score. Because there was only one question that assessed EXV and two dimensions assessed aspects of internal validity, we anticipated

that RCTs would also have a higher mean overall score.

Calculation of quality score

Items included in the different dimensions, and the scores assigned to individual items, are indicated in appendix 4. Possible answers to these questions were ranked as 'best', 'second best', etc., in order to calculate an overall score. The top rank was assigned a score of 1 (occasionally 2), the worst rank a score of 0, with intermediate ranks (if appropriate) assigned equally spaced scores. A score for each question for each paper was calculated by averaging the scores allocated by each reader. There were four reasons for excluding some questions from any of the quality dimensions:

- The question was not applicable to the interventions which were being reviewed (questions 4b, 8).
- The question was ambiguous. In some cases this was due to the wording of the question (e.g. questions 3a, 17a, 17b). In other cases, questions were applicable only for one or two study designs, but not others (e.g. questions 11, 16a, 16b, 17, 22, 23, 26).
- The question did not discriminate between studies because all studies scored the same on the question (e.g. questions 11).
- Some questions were considered not strictly to address a quality issue (e.g. questions 6c, 18b).

Some questions that were included were also not applicable to some study designs. It seemed inappropriate to penalise a study for not meeting a quality criterion which was not applicable. In such circumstances, it was necessary automatically to award the maximum score for the question to some study designs (see appendix 4).

Data extraction

RRM extracted additional information from the papers about the intervention, the population studied, the sample size, the outcome frequencies, the estimate of effectiveness and the confidence interval (CI), when available. BCR also checked these data. CIs were calculated when these were not reported, if sufficient data to do so were included in the paper.

Data synthesis

The effect size estimates for different studies were analysed by weighted or meta-regression, weighting each estimate by the inverse of the sampling variance.[51] This technique generated pooled effect size estimates for different study designs, and allowed investigation of associations between other attributes of studies (e.g. study quality and factors characterising the heterogeneity between study populations or interventions) and effect size. For these analyses, it was necessary to transform relative risk estimates into natural logarithms (ln(RR)) to give a linear outcome scale. The inverse of the precision of the relative risk estimate for a study $(1/(\text{standard error})^2)$ was used as its weight. Where necessary, standard errors were derived by calculating 'backwards' from the 95% CIs.

Methods used to identify hybrid study designs and RCT variants

Literature searching

Electronic databases

As hybrid studies contain both RCT and QEO study design elements, conventional search strategies for MEDLINE and EMBASE proved ineffective due to the methodological indexing problems of the electronic databases (see above). Therefore, although MeSH searches were attempted, they were unsuccessful for the same reasons as similar searches failed for strategy 1. Text word searches were carried out for the types of hybrid studies already known to the project steering group.

Handsearching

Handsearching was ruled out as it was considered that hybrid studies would appear in a wide selection of journals and be extremely rare. As for strategy 1 (see above), the Edinburgh HTA database of abstracts and the Cochrane Review of Methodology[46] database were handsearched, since we thought they might contain relevant references.

Expert knowledge

This strategy relied heavily on expert knowledge. ITR provided the project steering group with an initial list of hybrid studies. Other steering group members were requested to identify any studies thought to be relevant.

'Cascade' referencing

Papers thought to be relevant to the strategy were identified from the reference lists of papers found for this objective and for strategies 1 and 2. A high proportion of papers was found by this method.

Systematicity

We aimed to be systematic in documenting all hybrid designs and RCT variants, but had no intention to document all examples of each study design. Therefore, after finding a relevant example

we did not search for further examples of the same design. As in the case of strategy 1, we may not have been successful in achieving our aim of identifying all hybrids and RCT variants because of indexing problems with electronic databases. However, because several of the hybrids and variants which we identified represented variations on a small number of key 'themes', we suspect that there are unlikely to be other key themes which hybrid designs have been used to overcome.

Assessment

RRM read and identified all the relevant papers for this strategy. The descriptions and evaluation of the hybrid studies was then reviewed by two other members of the steering group (BCR and IMH).

Chapter 4

Studies included in the review

Eligibility criteria for strategy 1

There are three types of evidence that might be considered relevant to strategy 1:

- primary studies, which report both RCT and QEO study estimates of effect size for the same intervention
- secondary studies, which compare pooled RCT and QEO study estimates of effect size for the same intervention by synthesising evidence from primary studies which report either an RCT or QEO study estimate of effect size for the intervention (strategy 2 constituted this type of study for the two interventions reviewed)
- secondary studies, which compare pooled RCT and QEO study estimates of effect size for many interventions by aggregating evidence from primary studies that report either an RCT or QEO study estimate of effect size for more than one intervention.

Strategy 1 considered papers that fell into categories 1 and 2, but not 3. There were three reasons for this decision:

- When the review was planned, it was not clear how effect size estimates for different outcomes (e.g. events and continuous data) could be meaningfully combined. A method for expressing any measure of effect size using the Wilcoxon statistic[32,33] was identified while carrying out the review. However, we did not extend the eligibility criteria for strategy 1 because of other concerns about comparing RCT and QEO studies across interventions (see the two reasons below).
- A comparison of pooled RCT and QEO study estimates of effect size is itself an observational study and any finding may be confounded by intervention (see chapter 1). We considered that the problem of review bias was likely to be serious when contrasting pooled estimates of effect size for different study designs, since different interventions may be associated both with study design and study quality and the magnitude of the corresponding estimates of effectiveness.
- Pooled estimates of effect size for single interventions are usually only considered valid if they take account of all eligible evidence,[2] because of the possibility of publication biases.[2,52,53] Category 3 comparisons (i.e. secondary studies comparing pooled RCT and QEO study estimates of effect size across interventions) that we identified during the review[32,33] relied on identifying representative samples of RCT and QEO studies. This approach is only valid on the assumption that publication bias will have a uniform effect across all study designs. We believe that there are strong *a priori* reasons for doubting that this assumption is true. Particularly in recent years, when the importance of high quality evidence of ineffectiveness has been emphasised,[2,52,54] we suspect that there is a higher probability of negative RCT findings being published compared with negative QEO study findings.

Studies that fell into category 2 were considered not to be affected by the second point above, although they were still susceptible to confounding by publication bias.

We did not include studies which reviewed RCT and QEO studies for the same intervention in a comparative way, but which did not attempt to calculate a pooled estimate of effect size for different study designs.[31,55]

Eligibility criteria for strategy 2

Choice of interventions

A list of 31 potential areas of health technology was drawn up on the basis of the expertise of the project steering group, focusing on the availability of a substantial number of both RCTs and QEO studies of an intervention (see appendix 5).

Evidence of the amount of literature for each of the 31 interventions was sought by carrying out preliminary MEDLINE searches (1992 to January 1996). The search strategies were constructed in an identical manner for each intervention, and included:

- an abbreviated Cochrane RCT search strategy
- an observational strategy devised by RRM
- appropriate MeSH terms for each intervention.

Seven interventions (see appendix 6) were selected through this procedure for further consideration.

Our final choice of interventions for review under strategy 2 required three further criteria to be met by an intervention:

- the intervention under consideration should be uniform across studies
- the outcome by which the intervention under consideration was evaluated should be uniform across studies
- the populations in which the intervention was evaluated should be uniform across studies.

After further detailed MEDLINE searches the two interventions which best satisfied these criteria were selected from the seven:

- mammographic screening (intervention) for women aged 50–64 years (UK guidelines; population) to reduce mortality from breast cancer (outcome)
- periconceptional folic acid supplementation (intervention) for women trying to conceive (population) to prevent neural tube defects (outcome).

In addition to fulfilling all the criteria listed above, these areas were also currently under review within the NHS and were therefore considered to represent important public health measures. Reasons for rejecting the five other interventions on the shortlist are given in appendix 6.

Eligibility criteria for primary studies

Papers were eligible for strategy 2 if they reported primary evaluations of either of the interventions being reviewed, and if they matched the definitions for the intervention, population and outcome described above (see above). We considered this latter point to be very important since only by ensuring that all studies were evaluating the same intervention on the same population could we be confident of minimising review bias from differential variations in external validity between study designs (see chapter 1). However, some difficulties were experienced in applying these definitions (see chapter 6).

The instrument used to assess the quality of a study required the main confounding factors for an intervention to be specified. An additional information sheet (see appendix 7) was therefore circulated with the instrument, providing details of:

- the population, intervention and outcome which had been specified for the review (see appendix 4, questions 2, 3 and 4)
- the four most common confounding variables (see appendix 4, questions 5 and 25)
- up to four previously reported adverse effects of the intervention (see appendix 4, question 8).

Confounding factors for each intervention were selected by identifying all the confounding factors included in analyses or considered in the articles reviewed. The four most frequently cited confounding factors were then chosen. A similar approach was taken in identifying possible adverse effects of the interventions.

Eligibility criteria for hybrid study designs and RCT variants

We originally defined hybrid studies as studies which included both an RCT and an observational design element. Because of our interest in the reasons why researchers advocated hybrids, this definition was broadened to include variations on RCTs without a specific observational element, but where the variant was designed to overcome problems when using a conventional RCT for health technology assessment. In addition to summarising the key points of each design, we aimed to describe the advantages and disadvantages and to report a relevant example of each hybrid design that we reviewed.

There is inevitably a blurred boundary between novel and traditional designs. We did not review RCT variants that we considered to be well established and in common use (e.g. cross-over trials).[56] We also did not review evaluations based on indirect comparisons.[57] We considered such comparisons to be essentially observational (i.e. no RCT element is involved), and therefore they did not satisfy the criteria for inclusion as hybrid designs.

Chapter 5

Comparisons of estimates of effectiveness from RCTs and QEO studies: strategy 1

Papers reviewed

Fourteen papers were identified as relevant to this strategy.[30,34,58–69] A further four papers were considered but rejected for various reasons[70–73] (see appendix 8). The papers were identified from a variety of sources (see *Table 2*), with only two papers (one included and one excluded) being identified from searches designed specifically for this strategy. The papers were also found in a wide range of journals (see *Table 3*), vindicating the decision not to handsearch particular journals; no more than two papers were found in any one journal.

Three of the 14 papers focused on a comparison between trial and non-trial patients,[66,68,69] but did not include all the data which were considered relevant to the review. In these cases, relevant data were extracted from earlier papers reporting the findings of the study by the same research groups.[74–76]

Study designs

QEO study design elements were of various types (see *Table 4*):

1. Reviews,[30,34,58] in which a pooled estimate of effect size from a number of primary RCTs was compared with a pooled estimate of effect size from a number of primary QEO studies.
2. Comprehensive cohort studies (CCS),[62–64] in which subjects who were eligible for an RCT but who were not recruited were compared with those who were randomised; the QEO study population could receive either the intervention or the control treatment.
3. Other studies in which members of the QEO study population could receive either the intervention or control treatment, but where the QEO study population was different (e.g. a different centre or the same centre during a different time period) from the RCT population (denoted by ICIC).[59–61,69]
4. Studies which compared one or both groups in an RCT with a single non-trial group, usually where the RCT found no difference between intervention and control groups.[65–68] In these studies, the QEO study population could either be the same (i.e. randomisable patients who received a single treatment) or different from the RCT population (denoted by ICI or ICC, depending on whether the QEO study population received the intervention or control treatment).

The inclusion of type (4) studies may be considered controversial, since they cannot provide a QEO study estimate of effect size directly. QEO study estimates for these studies were calculated by comparing the single QEO study outcome frequency with the outcome frequency from the RCT element of the study for the comparison group. We believe that this approach is justified because:

- the criteria for eligibility for this strategy meant that RCT and QEO study elements were likely to

TABLE 2 *Source of papers identified for strategy 1*

Source	Eligible	Ineligible
Aware of paper at outset	2	2
From searches specifically designed for strategy 1	1	1
From searches designed for other objectives	2	0
From handsearches of databases of abstracts	2	0
From reference lists of papers already obtained	6	1
From general reading	1	0
Total	14	4

be investigating similar populations and interventions (although this may not have been true for ICI and ICC comparisons reported in one review[34])
- if bias is introduced, the comparison will tend to overestimate the discrepancy between RCT and QEO study effect sizes (i.e. lead to a conservative conclusion about the internal validity of QEO studies;) comparing one QEO study group with the corresponding RCT group is extremely unlikely to generate systematic underestimates of discrepancies.

The majority of papers reported more than one comparison between RCT and QEO study effect size estimates, because authors often described estimates of effectiveness for more than one intervention, population or outcome. In total, 38 comparisons were reported in the 14 papers that were reviewed. We have chosen to report our findings for all 38 comparisons, although it is important to emphasise that multiple comparisons within papers are unlikely to be independent.

Quality of included studies

Studies were assessed for quality on six criteria by three members of the project steering group (BCR, RRM and IMH), as described in chapter 3 and appendix 3. Contemporaneity, eligibility and blinding of outcome assessment were considered to be matters of fact, and the few discrepancies between assessors that were found were resolved by discussion. The extent to which QEO study estimates of effect size were adjusted for confounding was considered to be a judgement, and disagreements between assessors were respected. Disagreements mainly concerned the adequacy of adjustment for co-morbidity and other prognostic factors, and almost certainly arose because assessors were asked to assign a dichotomous (yes/no) score. For example, in some papers authors may have adjusted for age but not for other prognostic factors (e.g. QEO study estimates of effect size were adjusted for one prognostic factor but not for all prognostic factors).

TABLE 3 *Journals in which papers identified for strategy 1 were published*

Journal	No. of papers
Acta Oncologica	1
American Journal of Epidemiology	1
American Journal of Medicine	1
American Heart Journal	1
Archives of Disease in Childhood	1
British Journal of Cancer	1
Journal of the American College of Cardiology	2
Journal of the National Cancer Institute	1
Lancet	1
New England Journal of Medicine	2
Statistics in Medicine	1
Evaluation Studies Review Manual	1

Quality scores for the six criteria for different types of comparison are shown in *Table 5*. There was little indication that quality was related to type of comparison except in the case of comparisons reported in reviews, which were awarded substantially lower scores. Since reviews, by definition, included a number of primary QEO studies, pooled data for these studies did not take account of possible confounding; the reviews also typically included studies on dissimilar populations at different periods of time. Two of the reviews[30,58] were carried out before appropriate methods were established for pooling data across studies, and reported unweighted means of outcome frequencies and risk differences for RCT and QEO study elements. One ICC comparison[65] was considered ineligible by one assessor, but eligible by the other two; this comparison has been retained in the analyses that follow.

The distribution of quality scores for all comparisons is shown in *Figure 1*. Scores out of 18 ranged from 0 to 18. For further analysis of the discrepancies between RCT and QEO study findings, comparisons were classified as high (scores > 9) or low (scores < 9) quality. In these analyses, the total quality score for comparison 19 (4/12) was averaged up in proportion to the scores assigned by the two assessors (to 6/18). The value of 9 was chosen as the cut-off point because:

- there was some indication of bimodality in the distribution of quality scores (see *Figure 1*)
- requiring a high-quality study to score > 9 meant that the QEO study estimate must have made some attempt to control for confounding.

Indices for assessing discrepancies between RCT and QEO study findings

There are no established indices for summarising the difference between the effect size estimates for two studies. An effect size can be described as a ratio or a difference, for example a relative risk

TABLE 4 *Study designs used by papers reviewed for strategy I*

Study design	Abbreviation*	No. of papers
1. Reviews of primary studies which used either an RCT or QEO study design	Review (ICIC and ICI)	3†
2. Comprehensive cohort studies	CCS (ICIC)	4
3. Other studies in which the QEO study population could receive both intervention and control treatments	ICIC	3
4. Studies in which the QEO study population received either the intervention or the control treatment	ICI	2
	ICC	2

* *The abbreviations denote the groups studied in the RCT and QEO study elements of the study: ICIC, both RCT and QEO study elements included both intervention and control groups; ICI, QEO study element included an intervention group only; ICC, QEO study element included a control group only*

† *One review*[34] *included both ICIC and ICI comparisons*

TABLE 5 *Quality scores (%) for different types of comparison (*n *= 38)*[*]

Quality criterion	Reviews (*n* = 13)	CCS (*n* = 8)	ICIC (*n* = 8)	ICI (*n* = 6)	ICC (*n* = 3)
Contemporaneity	0	100	38‡	0	75**
Eligibility	0	100	75	33	100**
Confounding by severity	0	50	85	100	75**
Confounding by comorbidity†	0	0	0	33§	75§ **
Confounding by other prognostic factors	0	29	100	100	75††
Blinding of outcome assessment†	23¶	13	100	33‖	100**
Overall quality score	4	49	67	50	83

* *The percentage for each cell is calculated from the number of quality items scored as present by three assessors for all comparisons of each type*

† *Only two papers mentioned blinding of outcome assessment;*[62,67] *this criterion was scored as 'yes' for other papers only where the outcome was unequivocal (i.e. survival/all-cause mortality)*

‡ *Three ICIC comparisons*[63] *were given credit for contemporaneity because the analyses included the 'date of entry' of each subject*

§ *Credit for adjustment of co-morbidity (and other confounders) was given to two ICI*[67] *and two ICC comparisons.*[66] *Because the RCT and QEO study samples were drawn from the same population and the QEO study population received only one treatment, the QEO study sample should have been balanced on all factors with both intervention and control RCT samples*

¶ *One review paper*[30] *considered a number of outcomes; credit for blinding of outcome assessment was only given for the outcome 'case fatality'*

‖ *Credit was given for blinding of outcome for only one of the outcomes considered in each of the two ICI studies (i.e. for endometrial cancer but not any disease event*[67]*), and for all-cause mortality, but not for intra-breast recurrence, distant metastasis or contralateral breast cancer*[68]

** *Scores are out of 8 rather than 9, because one comparison*[65] *(see* Table 8*, comparison ID No. 19) was considered ineligible by one assessor*

(RR) or a risk difference (RD). Similarly, discrepancies in effect size can be described as ratios or absolute differences, that is as the ratio of the relative risks calculated from RCT and QEO study populations (RR_{RCT}/RR_{QEO}) or the difference in risk differences calculated from RCT and QEO study populations ($RD_{RCT} - RD_{QEO} = \Delta RD$). When comparing the results of two studies, however, one can consider discrepancies in outcome frequencies (i.e. discrepancies in risks or rates) between the RCT and QEO study intervention (or control) groups as well as discrepancies in effect sizes. For difference measures, one can also consider absolute discrepancies or the size of the discrepancy as a proportion of the quantity being estimated. We attempted to construct the seven indices shown in *Table 6* for each comparison to reflect these different ways of quantifying discrepancies.

The rationale for comparing discrepancies in outcome frequencies arose from our primary objective of evaluating the internal validity of the QEO study estimates. The quality criteria involved

an assessment of whether the RCT and QEO study elements compared the same populations, as well as control for confounding. Under these circumstances, the outcome frequencies as well as the effect estimates should be the same for RCT and QEO study elements. If we had considered only effect size estimates, these might have disguised substantial differences in outcome frequencies. For example, if the outcome frequencies for the RCT and QEO study control groups were to differ, this would imply a different baseline risk in the RCT and QEO study populations.

It was not possible to calculate all indices for all comparisons considered, because some of the data required were not reported either in the papers which were reviewed or in previous papers by the same research teams describing the same investigations. At least one index was calculated for each comparison.

Where a QEO study population could only receive one treatment (i.e. ICI or ICC comparisons), indices 1 to 3 were calculated by comparing the available QEO study group with data for the relevant RCT group. For example, in a study to evaluate the effectiveness and safety of tamoxifen in reducing mortality from breast cancer,[67] all subjects in the QEO study population received tamoxifen; the relative risk and RD effect size measures were calculated by comparing the QEO study intervention group with the RCT control group. Two of the indices (4 and 6 for ICC comparisons, and 5 and 7 for ICI comparisons) could not be calculated for these studies.

Test statistics of the significance of discrepancies between RCT and QEO study elements were not calculated because they were not considered relevant to the aim of the review. Such statistics cannot quantify the discrepancy between RCT and QEO study findings. They are also dependent on sample size, which means that there is likely to be an inconsistent relationship between the size of a discrepancy and its significance across studies. (See below for a discussion of the extent to which sample size may have confounded the relationship between quality and the size and direction of discrepancies.)

Size of discrepancies between RCT and QEO study elements

Summary descriptions of each comparison are given in *Table 7*. Data on outcome frequencies (control outcome frequency (COF); intervention outcome frequency (IOF)) and relative risk and RD effect sizes for RCT and QEO study elements are reported in *Table 8*. For most comparisons these data were extracted directly from the paper being reviewed, but it was occasionally necessary to consult other papers written by the same research teams (see footnotes to *Table 8*).[74–76] The seven

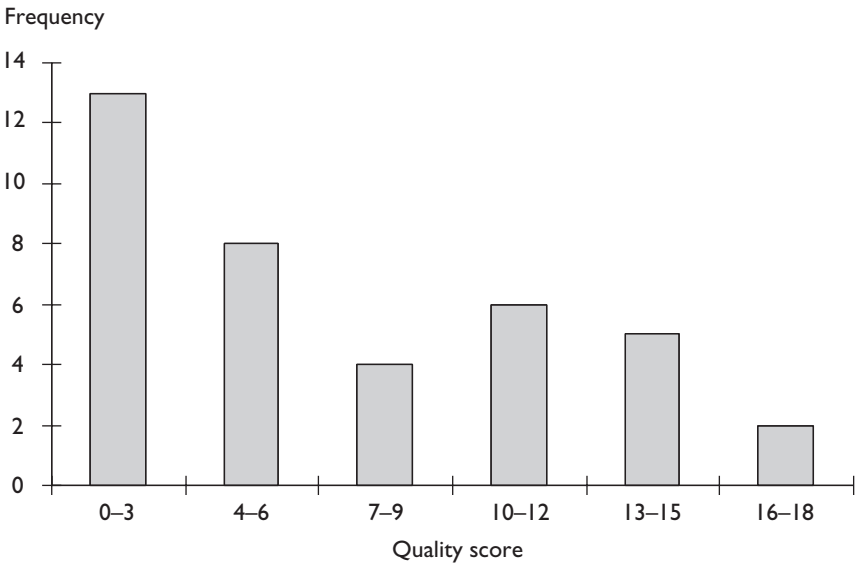


FIGURE 1 *Distribution of quality scores for comparisons*

TABLE 6 *Indices used to compare the findings of RCT and QEO study elements*

Index		Description
1	RR_{RCT}/RR_{QEO}	RR measure (risk, odds or rate ratio) for the RCT element divided by the RR measure for the QEO study element
2	ΔRD	Discrepancy in RD measures (risk or rate difference), i.e. RD for the QEO study element subtracted from the RD for the RCT element
3	ΔRD/mean RD	Discrepancy between the RDs for the RCT and QEO study elements (2), divided by the mean of the RDs for RCT and QEO elements
4	ΔIOF	Discrepancy in IOFs (risks, rates) for RCT and QEO study groups, i.e. the IOF for the QEO group subtracted from the IOF for the RCT group
5	ΔCOF	Discrepancy in COFs (risks, rates) for RCT and QEO study control groups, i.e. the COF for the QEO group subtracted from the COF for the RCT group
6	ΔIOF/mean IOF	ΔIOF (4) divided by the mean of the IOFs for the RCT and QEO study intervention groups
7	ΔCOF/mean COF	ΔCOF (5) divided by the mean of the COFs for the RCT and QEO control groups

indices used to compare RCT and QEO study elements are shown in *Table 9*. These indices were rarely reported by authors, but in most cases can be derived from the data given in *Table 8*.

In order to explore the extent to which the size of discrepancy was related to study quality, comparisons were classified as being of high (score > 9) or low (score ≤ 9) quality. Ordinal categories were also created for each index in order to visualise better the distributions of indices for high- and low-quality comparisons (see *Table 10*). The distributions in *Table 10* are transformed so that they are one-sided, i.e. by taking the reciprocal of values of index 1 which were less than 1, and by ignoring the sign of indices 2–7.

These distributions must be interpreted with caution because it is unlikely that comparisons within papers are independent. Discrepancies between RCT and QEO study findings may be more similar in size and direction within than between papers on account of similarity of interventions, populations or outcomes, or because of the perspective taken by authors. Quality scores assigned to comparisons within papers are also unlikely to be independent. Consequently, no analytical tests have been applied to investigate whether the distributions are significantly different. However, for all indices there is a clear tendency for high-quality studies to show much smaller discrepancies than low-quality studies.

An alternative approach would have been to consider only one comparison from each paper, in order to ensure independence. However, this would have reduced the number of comparisons to only 14, a sample size that would have resulted in very low power for analytical comparisons. Distributions of indices using this approach are shown in appendix 9. Insofar as a judgement can be made on the basis of such a small sample, the distributions appear to follow similar pattern, with high-quality studies showing smaller discrepancies.

Direction of discrepancies between RCT and QEO study elements

The direction of the discrepancy between RCT and QEO study elements was also of interest in view of the prevailing opinion that the biases which affect QEO study designs tend to produce more extreme effect sizes. RCT and QEO study effect sizes for each comparison were compared and classified as RCT more extreme, equal or QEO study more extreme. The number of low- and high-quality comparisons falling in each category are shown in *Table 11*.

The directions of the discrepancies between RCT and QEO study elements do not support the view that QEO studies give more extreme estimates of effect size. For high-quality comparisons, there was no evidence at all that QEO study estimates were more extreme than RCT ones (i.e. further from the expected value if there was no discrepancy, this being 1 for index 1 and 0 for indices 2–7). For low-quality comparisons, there was a tendency for QEO study estimates to be more extreme, but this was not marked.

The distribution of the directions of the discrepancies must again be interpreted with caution

TABLE 7 *Description of the interventions, populations and outcomes investigated in studies which were considered for strategy 1*[*]

Reference	Design[†]	RCT treatments	QEO study treatments	Population	Outcome	ID
Blichert-Toft *et al.*, 1988[61]	CCS	Breast conservation vs mastectomy	Breast conservation vs mastectomy	Women with breast cancer	Disease-free survival at 3 years	1
CASS, 1984[59]	CCS	CABG vs medical therapy	CABG vs medical therapy	Patients with angina or previous MI	Survival at 5 years	2
Chalmers *et al.*, 1977[30]	ICIC review	Anticoagulants vs placebo: 6 RCTs	Anticoagulants vs placebo: 8 studies with alternately assigned controls	Patients in hospital following an acute MI	(a) Risk of case fatality[†] (b) Risk of thrombosis[†] (c) Risk of haemorrhage[†]	3 4 5
	ICIC review	Anticoagulants vs placebo: 6 RCTs	Anticoagulant vs no anticoagulants: 18 studies with historical controls	Patients in hospital following an acute MI	(a) Risk of case fatality[†] (b) Risk of thrombosis[†] (c) Risk of haemorrhage[†]	6 7 8
Fisher *et al.*, 1994[67]	ICI	Tamoxifen vs placebo	Tamoxifen (no QEO control group)	Women who had had surgery for breast cancer	(a) 5-year rate/1000 of any disease event (b) Annual rate of endometrial cancer	9 10
Gray-Donald and Kramer, 1988[62]	ICIC	Traditional vs restricted formula supplementation	Formula supplementation vs no supplementation	Healthy neonates who had been initially breast-fed	% of mothers breast-feeding at 9 weeks	11
Hlatky *et al.*, 1988[63]	ICIC	CABG vs medical therapy: VA	CABG vs medical therapy: database[§]	Patients with coronary disease	Survival at 5 years	12
	ICIC	CABG vs medical therapy: ECSS	CABG vs medical therapy: database[§]	Patients with coronary disease	Survival at 5 years	13
	ICIC	CABG vs medical therapy: CASS	CABG vs medical therapy: database[§]	Patients with coronary disease	Survival at 5 years	14
Horwitz *et al.*, 1990[64]	ICIC	β blocker vs placebo	β blocker vs no β blocker	Patients who had had an acute MI; QEO study – all eligible for RCT	(a) Mortality at 2 years (b) Mortality at 3 years	15 16
	ICIC	β blocker vs placebo	β blocker vs no β blocker	Patients who had had an acute MI; QEO study – all eligible for RCT and others	(a) Mortality at 2 years (b) Mortality at 3 years	17 18
Kirke *et al.*, 1992[65]	ICC	Multivitamins with folic acid vs multivitamins without folic acid	No vitamins (no QEO study intervention group)	Women with a previous history of an NTD and planning a pregnancy	Risk of NTD affected pregnancy	19

Continued

TABLE 7 contd *Description of the interventions, populations and outcomes investigated in studies which were considered for strategy 1*[*]

Reference	Design[†]	RCT treatments	QEO study treatments	Population	Outcome	ID
Marubini *et al.*, 1996[68]	ICI	Breast conservation (QUART) vs mastectomy	Breast conservation (QUART) (no QEO study control group)	Women with breast cancer	(a) Survival at 10 years (b) Intra-breast recurrence (c) Distant metastasis (d) Contralateral breast cancer	20 21 22 23
Paradise *et al.*, 1984[60]	CCS	Surgery vs no surgery	Surgery vs no surgery	Children with severe recurrent throat infection	Episodes of throat infection: (a) during year 1 (b) during year 2 (c) during year 3	24 25 26
Reimold *et al.*, 1992[34]	ICIC review	Quinidine vs placebo to maintain sinus rhythm: 6 RCTs	Quinidine vs placebo to maintain sinus rhythm: 6 non-randomised studies	Patients with chronic atrial fibrillation following cardioversion	% of patients maintained in sinus rhythm: (a) at 3 months (b) at 6 months (c) at 12 months	27 28 29
	ICI review	Quinidine vs placebo to maintain sinus rhythm: 6 RCTs	Quinidine (no QEO study control group): 9 uncontrolled studies	Patients with chronic atrial fibrillation following cardioversion	% of patients maintained in sinus rhythm: (a) at 3 months (b) at 6 months (c) at 12 months	31 32 33
Schmoor *et al.*, 1996[69]	CCS	6 cycles vs 3 cycles chemotherapy	6 cycles vs 3 cycles chemotherapy	Women who had had surgery for breast cancer	Disease-free survival at 5 years	33
	CCS	Tamoxifen vs placebo	Tamoxifen vs placebo	Women who had had surgery for breast cancer	Disease-free survival at 5 years	34
	CCS	Chemotherapy and radiotherapy vs chemotherapy	Chemotherapy and radiotheraphy vs chemotherapy	Women who had had surgery for breast cancer	Disease-free survival at 5 years	35
Ward *et al.*, 1992[66]	ICC	5-Flourouracil and mitomycin C vs placebo	No treatment	Patients with operable stomach cancer	Survival at 1 year	36
		5-day induction, then 5-flourouracil and mitomycin C vs placebo	No treatment	Patients with operable stomach cancer	Survival at 1 year	37

Continued

TABLE 7 contd *Description of the interventions, populations and outcomes investigated in studies which were considered for strategy 1*[*]

Reference	Design[†]	RCT treatments	QEO study treatments	Population	Outcome	ID
Wortman and Yeaton, 1983[58]	ICIC review	CABG vs medical therapy: 9 RCTs	CABG vs medical therapy: 16 QEO studies	Patients with coronary artery disease	Survival/mortality (for longest follow-up reported by each primary study)	38

NTD, neural tube defect; CABG, coronary artery bypass graft; MI, myocardial infarction

[*] *Some studies included multiple comparisons between RCT and QEO study elements; ID identifies each comparison in subsequent tables*

[†] *CCS, comprehensive cohort study; ICIC, other studies in which subjects in the QEO study element could receive either the intervention or control treatments; ICI, studies in which subjects in the QEO study element only received the intervention treatment; ICC, studies in which subjects in the QEO study element only received the control*

[‡] *The duration for which risks were reported in this review[30] was not explicitly stated, but follow-up lasted until the end of hospitalisation for the index MI. This period was "at least 21 days in all studies"*

[§] *Hlatky* et al.[63] *compared the results of three RCTs with equivalent patients documented in a prospective database at another centre. VA, Veterans Administration Coronary Artery Bypass Surgery Cooperative Study Group;[77] ECSS, European Coronary Surgery Study Group;[78] CASS, Coronary Artery Surgery Study principal investigators and their associates[79]*

TABLE 8 *Summary of sample sizes, outcome frequencies and effect sizes for RCT and QEO study elements of studies that were considered for strategy 1*

Reference	ID No.	Sample size				Outcome frequency				Effect size			
		RCT		QEO study		RCT		QEO study		RR*		RD*	
		I	C	I	C	I	C	I	C	RCT	QEO	RCT	QEO
Blichert-Toft *et al.*, 1988[61]	1	313	306	60	76	0.800	0.760	0.830	0.950	0.833	3.400	–0.040	0.120
CASS, 1984[59]	2	390	390	570	745	0.950	0.920	0.940	0.920	0.625	0.750	–0.030	–0.020
Chalmers *et al.*, 1977†[30]	3	2106	1748	1517	1627	0.154	0.196	0.226	0.292	0.786	0.774	–0.042	–0.066
	4	2106	1748	1517	1627	0.111	0.213	0.125	0.232	0.521	0.539	–0.102	–0.107
	5	2106	1748	1517	1627	0.104	0.046	0.095	0.041	2.261	2.317	0.058	0.054
	6	2106	1748	4520	4570	0.154	0.196	0.223	0.383	0.786	0.582	–0.042	–0.160
	7	2106	1748	4520	4570	0.111	0.213	0.094	0.215	0.521	0.437	–0.102	–0.121
	8	2106	1748	4520	4570	0.104	0.046	0.065	0.020	2.261	3.520	0.058	0.045
Fisher *et al.*, 1994[67]	9	1419	1424	1220	–	194.8‡	315.6‡	178.5‡	–	0.617‡	0.566‡	–120.8‡	–137.1‡
	10	1419	1424	1220	–	1.6§	0.2§	1.4§	–	7.500§	7.000§	1.4§	1.2§
Gray–Donald and Kramer, 1988[62]	11	388	393	69	552	0.541	0.547	0.783	0.520	1.013	0.453	0.006	–0.236
Hlatky *et al.*, 1988[63]	12	332	354	← 719 →		0.830	0.780	0.855	0.809	0.773	0.759	–0.050	–0.046
	13	394	373	← 512 →		0.920	0.840	0.919	0.863	0.500	0.591	–0.080	–0.056
	14	390	390	← 250 →		0.950	0.920	0.930	0.872	0.625	0.547	–0.030	–0.058
Horwitz *et al.*, 1990[64]	15	1916	1921	417	205	0.073	0.092	0.076	0.097	0.793	0.784	–0.019	–0.021
	16	1916	1921	626	433	0.090	0.125	0.098	0.131	0.720	0.748	–0.035	–0.033
	17	1916	1921	417	205	0.073	0.092	0.102	0.144	0.793	0.708	–0.019	–0.042
	18	1916	1921	626	433	0.090	0.125	0.129	0.191	0.720	0.675	–0.035	–0.062
Kirke *et al.*, 1992[65]	19	172	89	–	103	0.000	0.011	–	0.029	0.000	–	–0.011	–0.029
Marubini *et al.*, 1996[68]	20	352	349	1408	–	0.790‖	0.760‖	0.768	–	0.875‖	0.805‖	–0.030‖	–0.050‖
	21	352	349	1408	–	NA	NA	NA	–	NA	NA	NA	NA
	22	352	349	1408	–	NA	NA	NA	–	NA	NA	NA	NA
	23	352	349	1408	–	NA	NA	NA	–	NA	NA	NA	NA
Paradise *et al.*, 1984[60]	24	38	35	44	34	1.24‖	3.09‖	1.77‖	3.09‖	NA	NA	–1.850‖	–1.320‖
	25	31	29	34	28	1.61‖	2.66‖	1.18‖	2.50‖	NA	NA	–1.050‖	–1.320‖
	26	22	20	15	13	1.77‖	2.20‖	1.47‖	3.15‖	NA	NA	–0.430‖	–1.680‖
Reimold *et al.*, 1992[34]	27	373	354	471	290	0.306	0.549	0.557	0.649	0.557	0.858	–0.236††	–0.139††
	28	373	354	471	290	0.423	0.667	0.728	0.812	0.634	0.897	–0.234††	–0.080††
	29	373	354	471	290	0.498	0.753	0.863	0.891	0.61	0.969	–0.244††	–0.035††
	30	373	354	751	–	0.306	0.549	0.414	–	0.557	–	–0.236††	–0.135††
	31	373	354	751	–	0.423	0.667	0.534	–	0.634	–	–0.234††	–0.133††
	32	373	354	751	–	0.498	0.753	0.641	–	0.61	–	–0.244††	–0.112††

Continued

TABLE 8 contd *Summary of sample sizes, outcome frequencies and effect sizes for RCT and QEO study elements of studies that were considered for strategy 1*

Reference	ID No.	Sample size				Outcome frequency				Effect size			
		RCT		QEO study		RCT		QEO study		RR*		RD*	
		I	C	I	C	I	C	I	C	RCT	QEO	RCT	QEO
Schmoor *et al.*, 1996[69]	33	235	238	133	114	0.520‡‡	0.490‡‡	NA	NA	0.900	0.900	NA	NA
	34	184	289	71	176	0.560‡‡	0.470‡‡	NA	NA	0.750	0.530	NA	NA
	35	98	101	41	88	NA	NA	NA	NA	0.790	0.760	NA	NA
Ward *et al.*, 1992[66]	36	74	69	–	493	0.590§§	0.530§§	–	0.560¶¶	NA	NA	–0.060	–0.030
	37	74	69	–	493	0.550§§	0.530§§	–	0.560¶¶	NA	NA	–0.020	0.010
Wortman and Yeaton, 1983†[58]	38	NA	NA	NA	NA	0.076	0.120	0.147	0.285	0.633	0.515	–0.044	–0.138

–, no data available because the QEO study element of the study studied only one group (intervention or control); I, intervention group; C, control group; NA, data not available because they were not reported or could not be calculated

* *For comparisons 1, 2, 12–14, 20, 27–32, 33–34, 36 and 37, RR and RD effect sizes were calculated as (1 – outcome frequency) rather than as outcome frequency (see text)*

† *Pooled estimates of outcome frequencies and risk differences were calculated as simple unweighted averages of the results for the primary studies which were reviewed*[30,57]

‡ *Cumulative 5-year rates, rate ratios and rate differences per 1000*

§ *Annual rates, rate ratios and rate differences per 1000*

¶ *Data obtained from Veronesi* et al.[74] *by reading survival probabilities off survival curves; these data do not take account of covariates. The QEO study ratio effect size was calculated from the RCT ratio effect size and the hazard ratio reported for being an 'out-trial' subject reported by Marubini et al.*[68] *in order to take account of covariates. The QEO study difference size was calculated directly from the outcome frequencies reported*[68,74] *(unlike the QEO study ratio effect size, see above) and therefore does not take account of covariates*

|| *Mean number of episodes of throat infection per year and differences in the mean number of episodes of throat infection per year*

** *Pooled rate differences were calculated using the method of DerSimonian and Laird;*[80] *therefore the rate differences do not correspond to the arithmetic differences between outcome frequencies for intervention and control groups*

‡‡ *Data obtained from Schumacher* et al.[75] *by reading survival probabilities off survival curves; these data do not take account of covariates*

§§ *Data obtained from Allum* et al.,[76] *by reading survival probabilities off survival curves; these data do not take account of covariates*

¶¶ *Data obtained from Ward* et al.,[66] *by reading survival probabilities off survival curves; these data do not take account of covariates*

TABLE 9 *Summary of comparisons of effect size and outcome frequency between RCT and QEO elements of studies which were considered for strategy 1*

Reference	ID No.	Significant difference[*]	Quality score	RR_{RCT}/RR_{QEO}	ΔRD	ΔRD/ mean RD	ΔIOF	ΔCOF	ΔIOF/ mean IOF	ΔCOF/ mean COF	More extreme?	
											Ratio	Difference[†]
Blichert-Toft *et al.*, 1988[61]	1	N	6	0.245	–0.160	–4.000	0.030	0.190	0.162	1.310	QEO	QEO
CASS, 1984[59]	2	N	13	0.833	–0.010	0.400	–0.010	0.000	0.182	0.000	RCT	RCT
Chalmers *et al.*, 1977[30]	3	Y	3	1.015	0.024	–0.444	–0.072	–0.096	–0.379	–0.393	QEO	QEO
	4	Y	0	0.967	0.005	–0.048	–0.014	–0.019	–0.119	–0.085	RCT	QEO
	5	Y	0	0.976	0.004	0.071	0.009	0.005	0.090	0.115	QEO	RCT
	6	Y	3	1.349	0.118	–1.168	–0.069	–0.187	–0.366	–0.646	QEO	QEO
	7	Y	0	1.192	0.019	–0.170	0.017	–0.002	0.166	–0.009	QEO	QEO
	8	Y	0	0.696	0.013	0.252	0.039	0.026	0.462	0.788	QEO	RCT
Fisher *et al.*, 1994[67]	9	Y	12[‡]	1.091	16.3	–0.126	16.3[§]	–	0.087	–	QEO	QEO
	10	Y	15[‡]	1.071	0.2	0.154	0.2[¶]	–	0.133	–	RCT	RCT
Gray-Donald and Kramer, 1988[62]	11	N	9	2.237	0.269	–2.091	0.241	–0.027	0.714	–0.058	QEO	QEO
Hlatky *et al.*, 1988[63]	12	N	15[**]	1.018	–0.004	0.083	0.025	0.029	0.159	0.141	QEO	RCT
	13	Y	15[**]	0.846	–0.022	0.353	–0.001	0.023	–0.012	0.155	RCT	RCT
	14	N	15[**]	1.143	0.028	–0.636	–0.020	–0.048	–0.333	–0.467	QEO	QEO
Horwitz *et al.*, 1990[64]	15	Y	12	1.013	0.002	–0.100	–0.003	–0.005	–0.040	–0.053	QEO	QEO
	16	Y	12	0.962	–0.002	0.059	–0.008	–0.006	–0.085	–0.047	RCT	RCT
	17	Y	9	1.120	0.023	–0.754	–0.029	–0.052	–0.331	–0.441	QEO	QEO
	18	Y	9	1.066	0.027	–0.557	–0.039	–0.066	–0.356	–0.418	QEO	QEO
Kirke *et al.*, 1992[65]	19	N	6[††]	2.636[‡‡]	0.018	–0.886	–	–0.018	NA	–0.886	QEO	QEO
Marubini *et al.*, 1996[68]	20	N	9	1.087	NA	NA	NA[§§]	–	NA	–	QEO	NA
	21	N	6	0.610	NA	NA	NA	–	NA	–	RCT	NA
	22	N	6	1.108	NA	NA	NA	–	NA	–	QEO	NA
	23	N	6	0.818	NA	NA	NA	–	NA	–	RCT	NA
Paradise *et al.*, 1984[60]	24	Y	6	NA	–0.530[¶¶]	0.334	–0.530[¶¶]	0.000[¶¶]	–0.352	0.000	NA	RCT
	25	Y	6	NA	0.270[¶¶]	–0.228	0.430[¶¶]	0.160[¶¶]	0.308	0.062	NA	QEO
	26	Y	6	NA	1.250[¶¶]	–1.185	0.300[¶¶]	–0.950[¶¶]	0.185	–0.355	NA	QEO
Reimold *et al.*, 1992[34]	27	Y	0	0.649	–0.097	0.901	–0.251	–0.100	–0.582	–0.167	RCT	RCT
	28	Y	0	0.707	–0.154	0.976	–0.305	–0.145	–0.530	–0.196	RCT	RCT
	29	Y	0	0.683	–0.209	1.604	–0.365	–0.138	–0.536	–0.168	RCT	RCT
	30	Y	0	0.739	–0.101	0.571	–0.108	–	–0.300	–	RCT	RCT
	31	Y	0	0.792	–0.101	0.589	–0.111	–	–0.232	–	RCT	RCT
	32	Y	0	0.777	–0.132	0.779	–0.143	–	–0.251	–	RCT	RCT

Continued

TABLE 9 contd *Summary of comparisons of effect size and outcome frequency between RCT and QEO elements of studies which were considered for strategy 1*

Reference	ID No.	Significant difference*	Quality score	RR_{RCT}/RR_{QEO}	ΔRD	ΔRD/ mean RD	ΔIOF	ΔCOF	ΔIOF/ mean IOF	ΔCOF/ mean COF	More extreme?	
											Ratio	Difference†
Schmoor *et al.*, 1996[69]	33	N	11	1.000	NA	NA	NA	NA	NA	NA	=	NA
	34	N	11	1.415	NA	NA	NA	NA	NA	NA	QEO	NA
	35	N	11	1.039	NA	NA	NA	NA	NA	NA	QEO	NA
Ward *et al.*, 1992[66]	36	N	18‡	0.936	–0.030	0.667	–	0.030	–	0.066	RCT	RCT
	37	N	18‡	0.936	–0.030	6.000	–	0.030	–	0.066	RCT	RCT
Wortman and Yeaton, 1983[58]	38	N	3	1.229	0.094	–1.033	–0.071	–0.165	–0.637	–0.815	QEO	QEO

–, no data available because the QEO study element of study studied only one group (intervention or control); NA, data not available because they were not reported or could not be calculated

* Y (yes) and N (no) in this column denote whether the effect size reported for the RCT was statistically significant or not

† These columns describe whether RR and RD measures of effect size were more extreme for the RCT or QEO study element of the study. RCT and QEO study RR effect sizes were equal (=) for one comparison

‡ These comparisons were considered to be fully adjusted for all confounders, since subjects recruited to the QEO study element met the eligibility criteria for the RCT and formed a single group (intervention or control)

§ Cumulative 5-year rate difference per 1000

¶ Annual rate difference per 1000

** Credit was given with respect to the quality criterion of contemporaneity for comparisons 12–14 because the date of entry of subjects into the relevant RCT or into the prospective database was included in the analyses

†† One assessor considered this study to be ineligible, because the QEO study control group received no treatment, whereas the RCT control group received multivitamins; the actual score of 4/12 has therefore been averaged up to 6/18

‡‡ The ratio of RR measures was calculated as the ratio of the outcome frequencies for the control groups of the RCT and QEO study elements, because there were no outcome events in the intervention arm of the RCT

§§ The difference between outcome frequencies for RCT and QEO study intervention groups was not calculated because the QEO study outcome frequency reported was not adjusted for covariates

¶¶ Differences between the mean number of episodes of throat infection per year

TABLE 10 *Distribution of indices used to compare the findings of RCT and QEO study elements across 38 comparisons, classified by study quality*[*]

Index 1: RR_{RCT}/RR_{QEO}[†]						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$1.00 \le x \le 1.10$	**$1.10 < x \le 1.25$**	**$1.25 < x \le 1.50$**	**$1.50 < x \le 2.00$**	**$x > 2.00$**	
Low (*n* = 25)	5	5	7	2	3	22
High (*n* = 13)	9	3	1	0	0	13
Index 2: ΔRD[‡]						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Low (*n* = 25)	5	3	2	6	2	18
High (*n* = 13)	4	4	0	0	0	8
Index 3: ΔRD/mean RD[§]						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Low (*n* = 25)	2	2	3	8	6	21
High (*n* = 13)	3	2	2	2	1	10
Index 4: DIOF[¶]						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Low (*n* = 25)	3	4	3	3	4	17
High (*n* = 13)	4	2	0	0	0	6

Continued

TABLE 10 contd *Distribution of indices used to compare the findings of RCT and QEO study elements across 38 comparisons, classified by study quality*[*]

Index 5: ΔCOF[‖]

Quality	Increasing disparity between RCT and QEO study elements →					Total
	$0.00 \leq x \leq 0.02$	$0.02 < x \leq 0.05$	$0.05 < x \leq 0.10$	$0.10 < x \leq 0.20$	$x > 0.20$	
Low (n = 25)	4	2	3	6	0	15
High (n = 13)	3	5	0	0	0	8

Index 6: $\Delta IOF/mean\ IOF$[**]

Quality	Increasing disparity between RCT and QEO study elements →					Total
	$0.00 \leq x \leq 0.10$	$0.10 < x \leq 0.25$	$0.25 < x \leq 0.50$	$0.50 < x \leq 1.00$	$x > 1.00$	
Low (n = 25)	1	5	9	5	0	20
High (n = 13)	4	3	1	0	0	8

Index 7: $\Delta COF/mean\ COF$[††]

Quality	Increasing disparity between RCT and QEO study elements →					Total
	$0.00 \leq x \leq 0.10$	$0.10 < x \leq 0.25$	$0.25 < x \leq 0.50$	$0.50 < x \leq 1.00$	$x > 1.00$	
Low (n = 25)	5	4	4	4	1	18
High (n = 13)	5	2	1	0	0	8

[*] *In order to present one-sided distributions, the reciprocal of values of index 1 which were less than 1, and the absolute values of indices 2–6, are shown*

[†] *This index was calculated for comparison 19 as the ratio of the outcome frequencies for the RCT and QEO study control groups; it could not be calculated for comparisons 24–26 because the outcome was expressed as the mean number of throat infections per year*

[‡] *This index could not be calculated for comparisons 9 and 10 because the outcome was expressed as a rate, for comparisons 23–26 and 33–35 because the data required were not available, and for comparisons 24–26 because the outcome was expressed as the mean number of throat infections per year*

[§] *This index could not be calculated for comparisons 23–26 and 33–35 because the data required were not available. The index was calculated for comparisons 9 and 10 and 24–26 because the outcome becomes irrelevant when ΔRD is expressed as a proportion of the mean RD*

[¶] *This index could not be calculated for comparisons 9, 10 and 24–26 because of the types of outcome reported, for comparisons 19, 36 and 37 because the study type was ICC, and for comparisons 23–26 and 33–35 because the data required were not available*

[‖] *This index could not be calculated for comparisons 9–10 and 24–26 because of the types of outcome reported, for comparisons 23–26 and 30–32 because the study type was ICI, and for comparisons 33–35 because the data required were not available*

[**] *This index could not be calculated for comparisons 19 and 36–37 because the study type was ICC, and for comparisons 20–23 and 33–35 because the data required were not available*

[††] *This index could not be calculated for comparisons 20–23 and 30–32 because the study type was ICI, and for comparisons 33–35 because the data required were not available*

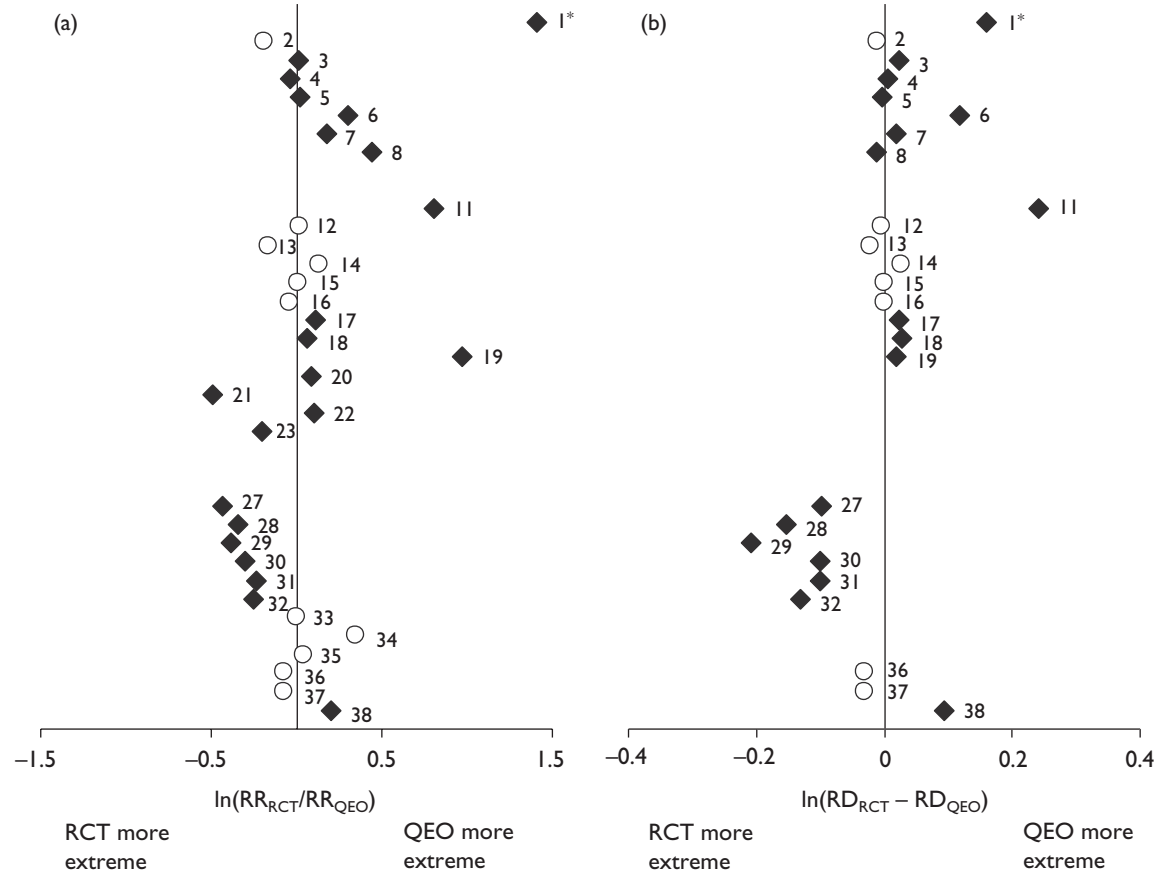
TABLE 11 *'Direction' of discrepancies between RCT and QEO study results*[*]

	RCT	=	QEO study	Total
Relative risk[†]				
Low quality	9	0	13	22
High quality	6	1	6	13
Risk difference				
Low quality	9	0	12	21
High quality	7	0	3	10

=, estimates of treatment effect that were identical for both RCT and QEO study elements

[*] *The number of comparisons in which the RCT and QEO study elements gave the more extreme estimate of effect size are tabulated separately for measures of relative risk (RR) and risk difference (RD). Findings are shown separately, firstly because one or other measure of effect size was unavailable for some comparisons, and secondly because, for four comparisons (4, 5, 8 and 12), the study design element which gave the most extreme element differed for the two measures of effect size (see* Table 9*). Measures of RD include any measures of the difference in outcome frequency between the intervention and control groups (i.e. risk, rate or difference between means in a continuous outcome). RD measures were not available for comparisons 20–23 and 33–35*

[†] *Measures of relative risk were not available for comparisons 24–26*



*Estimates of effect size from RCT and QEO study elements were in opposite directions for comparison 1, with the QEO study estimate being more extreme

FIGURE 2 *The size and direction of discrepancies*

because it is unlikely that comparisons within papers are independent. Similar tabulations of the directions of discrepancies including only one comparison from each paper are shown in appendix 9. Insofar as a judgement can be made on the basis of such a small sample, the distributions appear to follow a similar pattern.

The size and direction of discrepancies are shown together in *Figure 2*. The ratios of relative risk estimates have been transformed into natural logarithms to make the discrepancy scale linear. There are one or two large discrepancies where the QEO study estimate is more extreme (low-quality comparisons 1, 11 and 19), but there is no evidence that discrepancies for high-quality comparisons are larger when QEO study estimates are more extreme than when RCT estimates are more extreme.

Investigation of differences in RCT and QEO study populations

It is interesting to consider the relative extent to which poor internal validity and differences in the populations contributed to discrepancies. In two papers,[64,66] the authors reported how the effect size estimate for the QEO study element changed as different factors that might account for discrepancies (e.g. eligibility criteria and confounding) were taken into account. In both cases, restricting the QEO study population to subjects who would have been eligible for the RCT brought about the largest step in convergence between RCT and QEO study estimates, with adjustment for confounding producing relatively little subsequent change.

Investigation of possible meta-confounding

Investigation of the association between study quality and the size and direction of discrepancies between RCT and QEO study elements is 'observational' in nature (i.e. the populations studied for each comparison were not randomly allocated to the RCT and QEO study elements). It is therefore important to consider the extent to which any association between study quality and size and direction of discrepancies may be 'confounded' by other factors.

It was not possible to investigate this type of confounding formally, because of the likely non-independence of the discrepancies between RCT and QEO study design elements across comparisons within papers. Instead, we adopted one of the two following approaches:

- compare the relationship between quality and size and direction of discrepancies within strata of a possible confounding factor (see (1) and (3) below)
- describe the relationship between a possible confounding factor and the size and direction of discrepancies (see (2) below).

The former approach allows inspection of the relationship in the absence of the confounding factor; in the latter situation, if no relationship between the confounding factor and size or direction of discrepancy is apparent, confounding is extremely unlikely. We acknowledge that these simple descriptive analyses do not take account of interactions between the putative confounding factors and quality.

Three confounding factors were considered:

- confounding due to the inclusion of reviews
- confounding by whether or not the intervention is effective
- confounding by sample size.

Confounding due to the inclusion of reviews

By their nature, reviews were scored as being of low quality. If comparisons in review papers tend to give rise to larger discrepancies than comparisons made on the basis of primary data, the inclusion of reviews may be a confounding factor. It should also be pointed out that, because reviews included data from several primary studies, discrepancies for review comparisons were more likely to arise from differences in the characteristics of the populations studied. Since reviews were inevitably classified as low quality within the current scoring system, 'confounding' here refers to the possibility that the inclusion of reviews had a distorting effect on the size and direction of discrepancies for low-quality comparisons.

Appendix 9 shows the size and direction of discrepancies for all non-review comparisons. Since all the review comparisons were of low quality, the findings for the high-quality comparisons are unaltered. The distributions of the discrepancies for low-quality non-review comparisons continue to show a greater spread than for high-quality comparisons. Low-quality comparisons, but not high-quality ones, still showed a tendency to produce more extreme effect size estimates.

Confounding by whether or not the intervention is effective

One might expect the magnitude of discrepancies between RCT and QEO study effect size estimates to be smaller for ineffective interventions than for effective ones. For evaluations of ineffective interventions, indices 1–3 will be insensitive to differences in the composition of RCT and QEO study populations with respect to prognostic factors; discrepancies in these indices will only arise when prognostic factors are unevenly distributed between the intervention and control groups of the QEO study population. If the quality of comparisons of ineffective interventions tends, on average, to be higher than for comparisons of effective interventions, the association between quality and size and directions of discrepancy could be due in part to confounding by the effectiveness of the interventions being evaluated. Assuming that QEO studies yield exaggerated effect sizes (i.e. the prevailing view), one might also expect the discrepancies involving effective interventions to be larger, with those for the QEO studies being more extreme.

Confounding by the effectiveness of the intervention was explored by tabulating the size and direction of discrepancies for comparisons by a proxy marker of the effectiveness of an intervention, namely whether or not the RCT study element yielded a statistically significant result (see appendix 9). The distributions of the discrepancies for significant and non-significant comparisons show no marked difference, and the more extreme effect size estimate for a comparison appeared equally likely to be derived from the RCT or QEO study design element.

Confounding by sample size

One would expect a direct relationship between the size of discrepancies and sample size, because sample size (in conjunction with outcome frequency) determines the precision of an effect size estimate. If high-quality comparisons tend, on average, to be based on larger sample sizes, the relationship between quality and size of discrepancy might be substantially reduced by taking account of sample size. (It could be argued that sample size is itself a factor that should be considered in assessing quality, but it was not included in the quality assessment that was carried out for this strategy.) With respect to the direction of discrepancies, we would expect any general tendency for QEO study estimates to produce more extreme estimates of effect size to be more evident when the effect size estimates are more precise (i.e. for larger sample sizes).

Confounding by sample size was explored by tabulating size and direction of discrepancies for comparisons by low and high quality separately for 'small' and 'large' sample size strata (see appendix 9). Classification was based on the total RCT sample size only, because QEO study sample size was affected for some comparisons by the absence of either an intervention or a control group; a cut-off of 750 subjects was chosen since this created two groups with approximately equal numbers of comparisons in each.

Stratifying by sample size reduces the number of comparisons in each stratum and makes it more difficult to interpret the distributions of the discrepancies. Overall, however, the distributions of the size of discrepancies appear to be similar across strata, and show the same pattern as without stratification (i.e. larger discrepancies for low-quality comparisons). With respect to the direction of discrepancies, the tendency for QEO study estimates to be more extreme was most apparent for low-quality comparisons based on large sample sizes as predicted. Whether based on small or large sample sizes, there was no evidence from high-quality comparisons to support the view that QEO study estimates tend to give more extreme effect sizes than do RCT estimates.

Summary

Discrepancies between RCT and QEO study elements in relative risk and RD effect size estimates tend to be smaller for high- than for low-quality comparisons, as judged by the comparability of the RCT and QEO study populations and control of the QEO study effect size estimate for confounding. Although multiple comparisons within papers may not be independent, there was no evidence that this tendency disappeared when only one comparison per paper was considered.

Discrepancies for almost all high-quality comparisons were small, falling in one or other of the two least disparate ordinal categories for all indices except index 3. In contrast, discrepancies for low-quality studies were often substantial, with relative risks differing by as much as a factor of 2 and RDs and outcome frequencies by as much as 0.2. We made no attempt to explain discrepancies for particular comparisons as being due primarily to poor external or internal validity. However, we noted two studies that investigated this and found differences between the RCT and QEO study populations to be most important in accounting for discrepancies.[64,66]

For high-quality comparisons, there was no evidence that QEO study estimates of effect size were more extreme than RCT ones. For low-quality comparisons, there was a tendency for QEO study estimates of effect size to be more extreme than RCT ones. This tendency appeared more pronounced when considering comparisons based on larger sample sizes.

These findings lend some support to the view that QEO studies designed to evaluate interventions can yield valid effect size estimates. None of the high-quality comparisons reviewed could be considered to have adopted exceptional measures (relative to recommended epidemiological practices) to control for confounding. We consider whether or not this finding is generalisable in chapter 8.

Chapter 6

Comparisons of estimates of effectiveness from RCTs and QEO studies: strategy 2

Papers reviewed

A total of 34 papers were identified as relevant to this objective; 17 were studies of the effectiveness of mammographic screening to reduce mortality from breast cancer (MSBC)[81–97] and 17 were studies of the effectiveness of periconceptional folic acid supplementation to prevent neural tube defects (FAS).[65,98–113] Unlike the situation for strategy 1, all the papers were identified from literature searches (including contacts with experts; see appendix 10) related to this objective (*Table 12*).

As suggested by preliminary searches, the papers were found in a wide range of journals for both the interventions reviewed. The number of papers identified gave a misleading impression as to the number of individual studies that existed since there were often multiple publications reporting different aspects of the same projects and, for MSBC studies, different durations of follow-up. Only one publication per project was reviewed, although the multiplicity of reports from projects meant that earlier publications were sometimes referred to for details of the study design (see chapter 3).[114–119] There was one exception to this rule; the Diagnostich Onderzoek Mamma- carcinoma (DOM) project was analysed both as a case–control study[83,89] and as a cohort study,[89] and the two analyses were included as separate studies.

A range of study designs were used by the papers which were reviewed for each intervention (*Table 13*). Two studies used 'other' designs; one MSBC study used a 'case–cohort' design, and one FAS study used a cross-sectional design.

TABLE 12 *Source of papers identified for strategy 2*

Source	MSBC	FAS
From MEDLINE searches designed for strategy 2	16	9
From EMBASE searches designed for strategy 2	0	3
From reference lists of other references	1	3
From experts*	0	2
Total	17	17

* *See appendix 10 for lists of the experts who were contacted for MSBC and FAS*

Performance of instrument used to measure quality

The performance of the instrument used to measure study quality was investigated in some detail and is described here because it highlights some of the difficulties of assessing the quality of studies.

Interassessor agreement on instrument

Each paper was reviewed by RRM and two others (see appendix 11). There were no important differences between indices of interassessor agreement for the two interventions reviewed, and the indices are reported here for all papers together. *Table 14* shows the distribution of κ statistics and the percentage agreement for the items which were used to assess quality, based on the ratings of all MSBC and FAS papers (see also appendix 4). The majority of κ statistics indicated only slight (0–0.2) or fair (> 0.2–0.4) agreement.

It is clear from *Table 14* that there are also inconsistencies between the two indices of agreement; in contrast to the κ statistics, the percentage agreement was > 60% for the majority of items. The likely explanation for this finding is that the majority of papers were given the same rating on many items. κ is a 'chance corrected' measure of agreement, and is therefore usually preferred over simpler measures such as the percetage agreement. However, the interpretation of κ becomes difficult when the scores for the examples being rated are distributed across the available response categories in a grossly

TABLE 13 *Study designs used by papers reviewed for strategy 2*

Study design	MSBC	FAS
Individually assigned RCT	5	4
Cluster assigned RCT	3	0
Prospective cohort study	5	5
Retrospective cohort study	0	1
Unmatched case–control study	0	3
Matched case–control study	3	3
Other designs	1	1

unequal manner; the marginal totals become severely constrained, and a high degree of agreement is expected by chance. In these circumstances, a high value of κ is almost impossible to attain, and a low value may not necessarily imply poor agreement.

Internal consistency of subcomponents of quality score

The internal consistency of the items contributing to three of the four dimensions identified by the instrument (REP, EXV, IVB and IVC) were examined using Cronbach's α. The analyses were constrained to respect the direction of scales used to score questions since zero had already been designated to be the lowest quality score for each item on *a priori* grounds (*Table 15*). (Cronbach's α could not be calculated for EXV because it was based on only one item.) Cronbach's α was also calculated for the sum of EXV, IVB and IVC scores (giving a total 'analysis' score), and for the overall total ('total quality'). Analyses were first performed using the scores for the 17 papers for MSBC and FAS separately, and then using all 34 papers together.

The low α values indicate 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. Although some increase in the value of α is expected with an increasing number of items,[120] it was still surprising that the highest α value was obtained for all items, suggesting that different dimensions of quality were correlated (*Table 16*). There was a moderately strong correlation between REP and analysis, but only modest correlations between different aspects of analysis. We had expected a stronger inverse correlation between EXV and IVC (since RCTs, which are often

TABLE 14 *Distribution of κ statistics and percentage agreement for items used in quality assessment, and the relationship between the indices for the same items*

κ	Agreement (%)					Total (%)
	0–20	21–40	41–60	62–80	81–100	
0.0–0.2	0	1	6	6	2	15
> 0.2–0.4	0	0	4	9	1	14
> 0.4–0.6	0	1	4	4	0	9
> 0.6–0.8	0	0	0	0	1	1
> 0.8–1.0	0	0	0	0	0	0
Total	0	2	14	19	4	39

TABLE 15 *Cronbach's α scores for different quality scores and for MSBC and FAS studies separately and combined*

Cronbach's α	MSBC papers	FAS papers	MSBC and FAS papers
Reporting	0.325	0.468	0.356
Analysis	0.253	0.561	0.362
External validity	–	–	–
Susceptibility to bias	0.071	0.394	0.248
Susceptibility to confounding	0.441	0.601	0.473
Total quality	0.610	0.716	0.633

TABLE 16 *Correlations between different quality scores*

	Reporting	Analysis	External validity	Susceptibility to bias	Susceptibility to confounding
Analysis	0.66				
External validity	0.34	0.26			
Susceptibility to bias	0.28	0.53	0.38		
Susceptibility to confounding	0.56	0.82	–0.07	–0.03	
Total quality	0.90	0.92	0.33	0.45	0.76

TABLE 17 *Quality scores (standard deviation) for different types of study shown separately for MSBC studies, FAS studies and all studies combined*

MSBC studies				
	RCTs (*n* = 8)	**Cohort studies (*n* = 5)**	**Case–control studies (*n* = 3)**	**Other studies (*n* = 1)**
Reporting	10.5 (1.0)	7.7 (1.4)	10.4 (0.4)	8.0 (NA)
Analysis	10.3 (0.4)	7.4 (1.1)	8.9 (1.6)	7.0 (NA)
External validity	0.7 (0.2)	0.9 (0.2)	0.9 (0.2)	0.7 (NA)
Susceptibility to bias	4.7 (0.4)	4.7 (0.9)	5.4 (1.2)	4.3 (NA)
Susceptibility to confounding	4.8 (0.2)	2.0 (0.7)	2.8 (0.5)	2.0 (NA)
Total quality	20.7 (1.3)	15.2 (2.4)	19.2 (1.2)	15.0 (NA)
FAS studies				
	RCTs (*n* = 4)	**Cohort studies (*n* = 6)**	**Case–control studies (*n* = 6)**	**Other studies (*n* = 1)**
Reporting	10.7 (0.4)	8.7 (1.1)	10.1 (1.1)	5.7 (NA)
Analysis	10.9 (2.3)	8.1 (1.5)	9.0 (1.1)	6.2 (NA)
External validity	0.8 (0.3)	0.8 (0.2)	0.9 (0.1)	0.4 (NA)
Susceptibility to bias	5.0 (1.2)	5.2 (1.0)	5.7 (0.9)	4.3 (NA)
Susceptibility to confounding	5.0 (1.4)	2.1 (0.7)	2.4 (0.5)	1.5 (NA)
Total quality	21.6 (2.1)	16.7 (2.2)	19.1 (1.8)	11.9 (NA)
All papers				
	RCTs (*n* = 12)	**Cohort studies (*n* = 11)**	**Case–control studies (*n* = 9)**	**Other studies (*n* = 2)**
Reporting	10.6 (0.9)	8.2 (1.3)	10.2 (0.9)	6.9 (1.6)
Analysis	10.5 (1.3)	7.8 (1.3)	9.0 (1.2)	6.6 (0.6)
External validity	0.7 (0.2)	0.8 (0.2)	0.9 (0.1)	0.5 (0.2)
Susceptibility to bias	4.8 (0.7)	5.0 (0.9)	5.6 (1.0)	4.3 (0.0)
Susceptibility to confounding	4.9 (0.8)	2.0 (0.7)	2.5 (0.5)	1.7 (0.4)
Total quality	21.0 (1.6)	16.1 (2.3)	19.2 (1.6)	13.5 (2.2)

considered to have poor external validity, were automatically given credit for controlling for confounding), and were surprised that there was no correlation between IVB and IVC.

The α values would have been higher if the analyses had not been constrained to respect the polarity of the scores on each item, since the scores for some items would otherwise have been reversed. For example, the value of Cronbach's α for the IVC dimension would have been higher if the questions about the number of patients lost to follow-up (questions 28a and 28b) had scored studies with a large number of patients lost as being better than studies with few patients lost. This observation implies that, despite choosing the polarity of scores to be consistent, some items were inversely correlated with the majority of others. A more detailed description of the particular questions involved and possible reasons for the inverse correlations are given in appendix 12.

Quality of included studies

Average scores for different aspects of quality for each type of study design are shown in *Table 17*. Variation in the quality scores were investigated as a function of the intervention (i.e. MSBC or FAS), study design (i.e. RCT, cohort or case–control study; other designs were omitted from the analysis) and interaction of intervention and study design.

Regression analyses of total quality scores showed that both cohort and case–control studies had significantly poorer quality than RCTs. (Cohort studies: mean quality difference –4.9; 95% CI, –6.5 to –3.3; $p < 0.0001$. Case–control studies: mean quality

difference –1.8; 95% CI, –3.5 to –0.1; $p = 0.03$.) The difference in total quality between cohort and case–control studies was also significant (mean difference 3.1; 95% CI, 1.2 to 5.0; $p = 0.003$). There was no independent effect of intervention, and no interaction of intervention and study design.

For reporting quality, cohort studies, but not case–control studies, had worse scores than RCTs (mean difference –2.3; 95% CI, –3.2 to –1.4; $p < 0.0001$). Cohort studies also had poorer quality reporting than case–control studies (mean difference –2.0; 95% CI, –3.4 to –0.9; $p < 0.001$).

For IVC, both cohort and case–control studies had poorer scores than RCTs. (Cohort studies: mean quality difference –2.8; 95% CI, –3.4 to –2.3; $p < 0.001$. Case–control studies: mean quality difference –2.4; 95% CI, –3.0 to –1.8; $p = 0.001$.) Cohort and case–control studies did not differ significantly. This finding was not surprising, since all large RCTs were automatically given the maximum score for question 25 (see appendix 4).

There were no differences in external validity between study designs or interventions.

The finding that case–control studies scored higher than cohort studies is interesting in view of the widely accepted view that cohort studies provide stronger evidence than case–control studies. When reading papers, some assessors commented that truly observational studies, which were usually conducted by researchers with some epidemiological expertise, seemed to be of higher quality than quasi-experimental ones. Since quasi-experimental studies must use a cohort design, the finding is consistent with the perception of assessors. The observation may therefore having nothing to do with study design *per se*, but result from the fact that quasi-experimental studies tend to have been carried out by relatively unskilled researchers (e.g. researchers who might choose to carry out a trial with historic controls for convenience). A similar finding was recently reported in a meta-analysis of complications of endarterectomy.[121] The authors found significant heterogeneity between studies, one source of which was attributed to authorship; the risk of death or stroke was less if there was a neurologist was among the authors.

Heterogeneity of populations, interventions and outcomes

The number of papers was limited by the strict criteria that we laid down in order to achieve homogeneity of the intervention, population and outcome investigated. There were some difficulties in applying these criteria, and they were relaxed in some instances; we also found that studies differed in ways that we had not anticipated at the outset. Details of these sources of heterogeneity between studies are described below for each of the interventions in turn, and are summarised in *Tables 18* and *19* (see end of chapter) for MSBC and FAS, respectively.

Mammographic screening

We originally defined the population of interest to be women aged 50–64 years because this is the age range identified in the UK guidelines for mammographic screening. However, several of the studies identified did not limit their study populations to this age group, and did not report findings for this age separately. If the original population criterion had been applied strictly, several studies would have had to be excluded. The population criterion was therefore relaxed to include women aged 35–74 years (i.e. the range of ages investigated in the studies that we identified).

Variation in the intervention or exposure was a major source of heterogeneity between studies:

- Some studies used one-view mammography, some used two-view mammography, and others used both (on different screening visits, or when the screening protocol changed during the course of the study); in some studies, mammography was combined with breast examination by a physician or with the teaching of breast self-examination.
- The interval between screens varied between studies from once each year to once every 4 years; some studies used a variable screening interval.
- The duration of the intervention varied between studies from 4 to 17 years.
- Provision for the control group varied between studies; in some cases control participants received nothing, but in others they were taught breast self-examination.
- For case–control studies, the only exposure which could be compared across studies was 'ever screened' (compared with never screened); these studies were carried out in the context of an entire population being offered screening.
- Screening provision for participants at the end of the studies varied; in some studies, both intervention and control groups were offered screening, while in others both intervention and control groups received no screening.

A key factor that limited the number of MSBC studies reviewed for this objective was the choice of breast cancer mortality as the outcome measure. Many studies which were identified by the literature searches and which reported survival had to be excluded. Breast cancer mortality was chosen to minimise the problem of lead-time bias that affects survival studies of the effectiveness of screening interventions (Rothman and Greenland,[37] page 502); using mortality should tend to minimise discrepancies in effect size between QEO studies and RCTs. The decision was taken to use mortality because lead-time bias arises for relatively few healthcare interventions. MSBC was not chosen because we wanted to say something about QEO study evaluations of screening interventions, but rather because of the expected homogeneity of intervention and outcome.

Despite choosing breast cancer mortality as the outcome, there was some variation between studies in the precise outcome measured. The duration of follow-up varied considerably between studies, although we tried to minimise this source of heterogeneity by choosing the duration of follow-up that was closest to 10 years when findings were reported for different lengths of follow-up for a particular study. Alexander and co-workers[93] also highlighted the possibility that the definition of a breast cancer death might have varied between studies. We could not take account of this possibility because, with the exception of one study,[93] information about the precise definition used by researchers was not reported.

Folic acid supplementation

Eight studies investigated the use of FAS to prevent neural tube defects (NTDs) in women without a previous history of an NTD (i.e. occurrence),[102–104,108–111,113] and nine studied FAS in women with a previous history (i.e. recurrence).[65,98–101,105–107,112] Since our eligibility criteria did not specify whether FAS should be intended to prevent occurrence or recurrence, and because excluding either would have halved the number of studies available, both types of study were included. Researchers defined an NTD in different ways so that there may have been further heterogeneity within populations of women with a previous history of an NTD.

The definition of a control subject in case–control studies also varied. Some studies recruited a control group of healthy babies, some recruited malformed control babies without an NTD, and others recruited both types of control groups. Data are presented here for both control groups, if available.

As for studies of MSBC, several sources of heterogeneity in the interventions were observed:

- Some studies used folic acid alone and others used multivitamins which contained folic acid; one study used a factorial design to investigate the effect of both folic acid and multivitamins simultaneously.
- The dose of folic acid which was used varied from 0.3 to 5 mg/day.
- The time period during which women had to have taken FAS in order to be considered 'fully supplemented' varied, particularly with respect to supplementation prior to conception. If FAS was achieved prior to conception, researchers generally assumed that supplementation continued during the critical first 3 weeks after conception.
- In some studies the control group received multivitamins or a trace supplement and in others the control group received nothing. This variation was confounded by study design, since RCTs or quasi-experimental studies were more likely to give some intervention to the control group than were observational studies.

As in the case of MSBC, the search strategy for FAS papers identified some multiple publications. One paper reported analyses both of an RCT and a cohort study, where the intervention arm of the RCT was compared with a non-randomised group of unsupplemented women.[65] (This paper was also reviewed for strategy 1.) The results of the RCT are reported here because it was clear from assessors' ratings of this paper that they had focused on the RCT. Four papers are included from "the same continuous and continuing"[101] multicentre PCH study. The three papers by Smithells and co-workers[99–100,105] report data collected over different time periods. The paper by Seller and Nevin[101] includes both new data and some data reported previously by Smithells and co-workers.[99,100]

Investigation of factors associated with effect size

Sample sizes, outcome frequencies and relative risk estimates (rate, risk or odds ratios, depending on the study design) are shown for all MSBC and FAS studies in *Tables 20* and *21* respectively. *Table 21* includes data extracted for both normal and abnormal control groups for FAS case–control studies, if both were presented in the original papers.

Effect size was taken to be the relative risk estimate for a study. We did not distinguish between rate, risk and odds ratios because the outcome being considered was rare for both interventions. The highest outcome frequency observed in any study was approximately 0.09 (risk of recurrence of NTD in control subjects[98]) and most outcome frequencies were an order of magnitude lower than this.

Weighted regression[51] was used to investigate factors associated with the magnitude of the effect size estimates. Dummy variables were created to assess the effect of cohort and case–control study designs, using RCTs as the baseline comparison. Other independent variables characterising the quality of studies and their heterogeneity with respect to population and intervention were investigated.

Association between study attributes and effect size for MSBC studies

Independent variables that were investigated included:

- study design (cohort and case–control dummy variables)
- total quality (and components and total quality)
- programme duration
- duration of follow-up
- age of population studied (lowest age, mid-age and oldest age recruited)
- frequency of screening (two levels only, i.e. every year, or every 2 years).

A simple model containing only the study design variables was fitted first (*Table 22*). relative risk estimates for case–control studies were, on average, about 0.6 times smaller (i.e. indicating greater benefit) than for RCTs (mean difference ln(RR) = –0.50; 95% CI, –1.04 to 0.03; $p = 0.06$). In contrast, the mean relative risk estimate for cohort studies was the same as for RCTs (mean difference ln(RR) = –0.03; 95% CI, –0.34 to 0.28; $p = 0.82$). The overall model was a poor fit ($F = 2.19$; $p = 0.15$; r^2(adj) = 0.14).

Neither total quality, nor any component of the quality score, were significantly associated with effect size. Including any of these variables resulted in lower values of r^2(adj) without affecting the mean differences in effect size between different study designs. The lack of relationship between study quality and effect size for different study designs is summarised in *Figure 3*.

None of the other independent variables, either singly or in combination, were significantly associated with effect size. Nor did they affect the mean differences in effect size between different study designs. Coefficients for case–control and cohort studies varied little compared with their standard errors in different models (–0.47 to –0.61 and –0.12 to 0.13, respectively).

Association between study attributes and effect size for FAS studies

Independent variables which were investigated included:

- study design (cohort and case–control dummy variables)
- total quality (and components and total quality)
- previous history of NTD
- latest time when supplementation started in fully supplemented women (weeks prior to conception)
- earliest time when supplementation stopped in fully supplemented women (weeks after conception).

A simple model containing only the study design variables was fitted first (*Table 23*). Relative risk estimates for case–control studies were, on average, about 2.6 times higher (i.e. indicating less benefit) than for RCTs (mean difference 0.96; 95% CI, 0.02 to 1.89; $p = 0.05$). In contrast, the mean relative risk estimate for cohort studies was the same as for RCTs (mean difference –0.02; 95% CI, –1.13 to 1.08; $p = 0.96$). The overall model was a good fit ($F = 7.39$; $p = 0.007$; r^2(adj) = 0.46).

Neither total quality, nor any component of the quality score, were significantly associated with effect size. Including any of these variables resulted in lower values of r^2(adj), without affecting the mean differences in effect size between different study designs. The lack of relationship between study quality and effect size for different study designs is summarised in *Figure 4*.

Time of starting and stopping supplementation, but none of the other independent variables, were marginally associated with effect size, but including these variables had no impact on the coefficients for cohort and case–control studies. The relative risk was increased (i.e. less benefit) by 1.04 for each unsupplemented week prior to conception (coefficient = 0.04; 95% CI, –0.001 to 0.08; $p = 0.05$), and the relative risk was reduced by 0.95 (i.e. greater benefit) for each additional week supplemented after conception (coefficient = –0.05; 95% CI, –0.10 to 0.01; $p = 0.10$). These associations are not entirely consistent with the critical period for closure of the neural tube, but may simply reflect

that women who were supplemented for longer were more fully supplemented.

Summary

The above analyses for both interventions indicate discrepancies between relative risk estimates derived from case–control studies and other study designs. Interestingly, the direction of the discrepancy is not consistent across interventions; relative risk estimates from case–control studies tend to be more extreme than other study designs for MSBC, but less extreme for FAS. After taking account of different study designs, no association was found between study quality (or aspects of quality) and effect size. We acknowledge that the fit of the regression model for MSBC studies was poor; the model is reported because of the similarity of the models for MSBC and FAS apart from the direction of the discrepancy between case–control studies and other designs.

One can postulate reasons for the different direction of the discrepancy for the two interventions, although it should be recognised that any such explanations are *post hoc*. Case–control studies of MSBC provide estimates of screening with 100% coverage, albeit ever versus never screened; both RCTs and cohort studies included substantial proportions of unscreened women in their 'screened' group for analysis, since coverage for mammographic screening varied between 60% and 100% in RCTs and cohort studies that reported this information. Case–control studies of FAS also provide estimates of supplementation with 100% coverage. However, we suspect that 'coverage' among women assigned to supplementation in RCTs and cohort studies is likely to have been higher than for MSBC.

Almost all case–control studies also required women to recall their 'exposure'. Recalling supplementation, which often required remembering the particular supplement taken in order to confirm the folic acid content, is likely to have been much less reliable than asking women to recall whether they had ever had a mammogram. Unreliable recall could easily to lead to bias, with women whose pregnancy was affected by an NTD being more likely to report that they had taken supplementation when they had not than women who did not have an affected pregnancy.

We therefore conclude that our findings suggest that case–control studies designed to estimate effectiveness should be interpreted with caution, and that the direction of any discrepancy between relative risk estimates for case–control studies and other study designs is likely to depend on the intervention being evaluated. There was no evidence at all that discrepant estimates for case–control studies can be attributed to confounding by quality or sources of heterogeneity.

We were unable to demonstrate any independent effect of quality on effect size after taking account of study design. This finding is difficult to interpret because the failure to find an association could arise in a variety of ways (see chapter 8). We also did not observe any associations between characteristics of studies, considered likely to be associated with outcome for *a priori* reasons, and effect size. Despite being unable to demonstrate significant associations between quality, sources of heterogeneity and effect size, we are wary of pooling results for different study designs. Investigating reasons for discrepancies, rather than providing a pooled estimate, was the primary objective of review.

This objective highlighted the considerable problems that exist in measuring study quality, and other aspects of study design which may influence effect size. The instrument which we used was not entirely successful, largely because of the compromises and ambiguities which arose from using the same instrument for all study designs. Developing an instrument to assess and characterise different studies is an urgent priority (see also chapter 8).

TABLE 18 *Description of the interventions, populations and outcomes investigated in MSBC studies that were considered for strategy 2*

Reference	Design*	Population		Intervention†					Outcome: approx. follow-up (years)
		Age (years)	Inclusion and exclusion criteria	Intervention	Frequency (years)	Duration (years)	Control	End	
Alexander *et al.*, 1994[93]	RCTC	45–64	Resident in Edinburgh; no previous history of breast cancer	1/2, PE	2	10	N	M	10
Andersson *et al.*, 1988[85]	RCTI	45–69	Resident in Malmö	1/2	1.5–2	10	N	M	9
Collette *et al.*,1984,[83] 1992[89]	MCC	50–64	Resident in Utrecht	NS, PE	1–4	10	e/n	M	12
Collette *et al.*, 1992[89]	PCH	50–64	Resident in Utrecht	NS, PE	1–4	10	e/n	M	12
Dales *et al.*, 1979[81]	RCTI	35–54‡	Resident in San Francisco Bay; member of Kaiser Health Plan	NS	1	11	N	M	11
Frisell *et al.*, 1991[88]	RCTI	40–64	Resident in Stockholm	1	2	4.5	N	1	7
Hakama *et al.*, 1995[95]	PCH	40–47	Resident in Kotka and environs	1, PE	2	9	BSE	M	9
Miller *et al.*, 1992[91]	RCTI	40–49	Resident in Canada (15 urban centres); no mammography in previous 12 months; not pregnant; no previous history of breast cancer	NS, PE, BSE	1	4	BSE	N	9
Miller *et al.*[89]	RCTI	50–59	Resident in Canada (15 urban centres); no mammography in previous 12 months; not pregnant; no previous history of breast cancer	NS, PE, BSE	1	4	BSE	N	7
Morrison *et al.*, 1992[90]	PCH§	35–74	'White'; resident in the USA (29 centres)	2, PE	1	4	SD	N	9
Palli *et al.*, 1989[87]	MCC	40–70	Resident near Florence	2	~2.5	17	e/n	M	7–17
Peer *et al.*, 1995[96]	PCH	35–49	Resident in Nijmegen	1	2	15	N	M	10

Continued

TABLE 18 contd *Description of the interventions, populations and outcomes investigated in MSBC studies that were considered for strategy 2*

Reference	Design*	Population		Intervention†					Outcome: approx. follow-up (years)
		Age (years)	Inclusion and exclusion criteria	Intervention	Frequency (years)	Duration (years)	Control	End	
Shapiro *et al.*, 1982[82]	RCTI	40–64	Resident in Greater New York; no previous history of breast cancer; member of health insurance plan	2, PE	1	3	N	N	10
Tabar *et al.*, 1995[97]	RCTC	40–74	Resident in Sweden (two counties); no previous treatment for breast cancer	1	2	10	N	M	10
Thompson *et al.*, 1994[94]	CCH¶	≥ 50‖	Resident in Puget Sound; member of a group health cooperative	NS, PE	1–3	6	N	M	7
UK TEDBC Group, 1993[92]	PCH	45–64	Registered with a GP	1/2**	2	7	N	NS	10
Verbeek *et al.*, 1984[84]	MCC	35–65	Resident in Nijmegen	1	2	8	e/n	M	8

* *RCTI, individually randomised RCT; RCTC, cluster randomised RCT; PCH, prospective cohort study; MCC, matched case control study; CCH, case–cohort study*

† *Intervention: 1, one view; 1/2, one or two view, depending on screening round or patient characteristics; NS, number of views not specified; PE, physical or clinical breast examination; BSE, teaching of breast self-examination. Frequency: interval between screens. Duration: duration of intervention programme. Control intervention: N, nothing; BSE, teaching of breast self-examination; SD, secondary data used for comparison with exposed cohort; e/n, ever screened vs never screened comparison of exposure for case–control studies. End: M, groups maintained during entire follow-up (for case–control studies M simply denotes that the programme has continued); I, both groups received mammography; N, neither group received mammography; NS, not specified*

‡ *Mammography was only offered to women aged ≥ 48 years, but results given for breast cancer mortality for the whole group*[81]

§ *The cohort study reported by Morrison and co-workers*[86] *did not include a control group; estimates of effect size were calculated with respect to national data*[122]

¶ *Thompson and co-workers*[94] *analysed their case–cohort study using a 'case–cohort analysis', which resembled a Cox regression with the benefit expressed as a risk reduction of death from breast cancer if screened 1 year prior to diagnosis*

‖ *Mammography only offered routinely to women aged ≥ 50 years, although all women aged ≥ 40 years took part in the overall screening programme; no upper age limit was specified, and 3% of women were aged > 80 years. Results were presented for women aged ≥ 50 years separately, and these are presented in the table*

** *The UK Trial of Early Detection of Breast Cancer Group*[92] *compared three groups, mammography vs BSE vs nothing. Results were almost identical for mammography vs BSE and for mammography vs nothing. Only the latter are reported here since this comparison is most consistent with other studies*

TABLE 19 *Description of the populations, interventions, and outcomes investigated in FAS studies which were considered for strategy 2*

Reference	Design*	Population† PH?	Inclusion and exclusion criteria	Interventions‡				Outcome§
				Dose	Duration	Add.	Control	
Bower and Stanley, 1992[108]	MCC	N	Resident in Western Australia; index pregnancy affected by malformation (cases and controls 1) or not (controls 2); excluding pregnancy with NTD and other malformation	NS	–12 to 0	P	P	NTD(ns)
Chatkupt *et al.*, 1994[112] ¶	XS	Y	Resident in New Jersey; members of families with multiple cases of SBC	NS	–? to 0	P	N	SBC
Czeizel and Dudas, 192[109]	RCTI	N	Resident in Hungary; not currently pregnant; no previous history of delayed conception or infertility	0.8 mg	≤ –4 to +8	M	T	NTD(ns)
Kirke *et al.*, 1992[65]	RCTI	Y	Resident in Eire; previous history of A, I, E, SBA; previous history of impaired gastrointestinal absorption	0.4 mg	≤ –8 to +12	M	M	NTD; M A, I, E, SBA
Laurence *et al.*, 1981[98]	RCTI	Y	Resident in south Wales; age < 35 years and previous history of A, E or SBC	2.0 mg	–? to > +6	N	N	NTD; A, E, SBC
Martinez-Frias and Rodriguez-Pinilla, 1992[110]	UCC	N	Resident in Spain; index pregnancy affected by malformation (cases and controls)	≥ 0.3 mg	0 to +12	P	NS	NTD(ns)
Mills *et al.*, 1989[103]	MCC	N	Resident in California; index pregnancy affected by malformation (cases and controls 1) or not (controls 2); excluding pregnancy with non-NTD defects potentially related to vitamin use	RDA	–4 to +6	M	NS	NTD; A, E, I, L, M, MYN, R

Continued

TABLE 19 contd *Description of the populations, interventions, and outcomes investigated in FAS studies which were considered for strategy 2*

Reference	Design*	Population† PH?	Inclusion and exclusion criteria	Interventions‡				Outcome§
				Dose	Duration	Add.	Control	
Milunsky *et al.*, 1989[104]	RCH	N	Resident in USA; nulliparous, undergoing prenatal MSAFP in > 100 obstetric practices	0.1–1.0 mg	0 to 6	M	M	NTD; A, E, SB(ns)
MRC Vitamin Study Research Group, 1991[107]	RCTI (factorial)	Y	Resident in Australia, Canada, France, Hungary, Israel, UK, USSR; previous history of A, E, SBC	4.0 mg	–? to +12	P	P	NTD; A, E, SBC
Mulinare *et al.*, 1988[102]	MCC	N	Resident near Atlanta; index pregnancy affected by malformation (cases) or not (controls)	NS, ≥ 3 pw	–12 to +12	M	M, < 3 pw	NTD; AN, SB(ns)
Seller and Nevin, 1984[101]	PCH	Y	Resident in Northern Ireland or south-east England; previous history of A, E, I, M, MY, MYN	0.4 mg	–4 to +6	M	N	NTD; A, E, I, M, MY, MYN
Shaw *et al.*, 1995[113]	UCC	N	Resident in California; index pregnancy affected by NTD (cases) or not (controls)	Any, daily	–12 to > +4	P	P	NTD; A, I, R, SBC
Smithells *et al.*, 1981[99]	PCH	Y	Resident in six areas in the UK; previous history of NTD	0.4 mg	< –4 to +8	M	N	NTD; A, E, I, M, MY, MYN
Smithells *et al.*, 1983[100]	PCH	Y	Resident in six areas in the UK; previous history of NTD; excluding one ectopic pregnancy, three with unknown exposure	0.4 mg	≤ –4 to +8	M	N	NTD; A, E, I, M, MY, MYN
Smithells *et al.*, 1989[105] ‖	PCH	Y	Resident in Yorkshire; previous history of NTD; not pregnant but considering pregnant	0.4 mg	≤ –4 to +8	M	N	NTD(ns)

Continued

TABLE 19 contd *Description of the populations, interventions, and outcomes investigated in FAS studies which were considered for strategy 2*

Reference	Design*	Population† PH?	Inclusion and exclusion criteria	Interventions‡				Outcome§
				Dose	Duration	Add.	Control	
Vergel *et al.*, 1990[106]	PCH	Y	Resident in Cuba; previous history of NTD; not pregnant (exposed) or pregnant (< 5 weeks; unexposed)	5.0 mg	≤ –4 to +10	N	N	NTD(ns)
Werler *et al.*, 1993[111]	UCC	N	Resident in Boston, Philadelphia, Ontario; index pregnancy affected by malformation (cases and controls); excluding oral clefts	NS, daily	≤ –4 to +4	M	N	NTD; A, E, SB(ns)

* *RCTI, individually randomised RCT; PCH, prospective cohort study; RCH, retrospective cohort study; UCC, unmatched case–control study; MCC, matched case–control study; XS, cross-sectional study*

† *PH?: eligibility defined according to whether women had had a previous pregnancy affected by NTD (Y, yes; N, no)*
Inclusion and exclusion criteria: see 'Outcome' (below) for definitions of different types of NTD

‡ *Dose: dose of FAS in milligrams intervention/exposed group (NS, not specified; RDA, recommended daily allowance). Duration: duration of supplementation. Additional supplements taken by intervention group (Add.): N, nothing (i.e. FAS only); M, multivitamins; P, multivitamins for a proportion only. Control (supplementation given to control group): M, multivitamins; N, nothing; P, multivitamins for a proportion only; T, trace element supplementation; pw, number of times per week*

§ *Outcome: A, anencephaly; E, encephalocele; I, iniencephaly; L, lipomeningocele; M, meningocele; MY, myelocele; MYN, myelomeningocele; NTD, neural tube defect; NTD(ns), neural tube defect, not specified; R, rachischisis; SBA, spina bifida aperta; SBC, spina bifida cystica; SB(ns), spina bifida, not specified*

¶ *Chatkupt and co-workers[112] carried out a cross-sectional survey of affected families, comparing the affected pregnancies in which women took FAS with the proportion of pregnancies in the USA in which women took FAS*

‖ *The study by Smithells and co-workers[105] was carried out in one of the six areas included in the two previous studies by Smithells and co-workers.[99,100] The recruitment period was longer than, but overlapped with, the previous studies and some of the data in the most recent paper may have been included in the two previous reports*

TABLE 20 *Summary of sample sizes, outcome frequencies and effect sizes for MSBC studies which were considered for strategy 2*

Reference	Age (years)	Duration of intervention (years)	Follow-up (years)	Sample size		Outcome frequencies			Relative risk		
				Intervention	Control	Measure	Intervention	Control	Measure	Point estimate	95% CI
Alexander *et al.*, 1994[93]	45–64	10	10	22,944	21,344	Rate × 10^{-4}	4.379	5.252	Rate ratio	0.83	0.63 to 1.11
Andersson *et al.*, 1988[85]	≥ 45	10	~9	21,088	21,195	Risk × 10^{-3}	2.987	3.114	Risk ratio	0.96	0.68 to 1.35
Collette *et al.*, 1984,[83] 1992[89]	50–64	10	12	116	348	*p*(exp)	0.405	0.339	Odds ratio	0.52	0.32 to 0.83
Collette *et al.*, 1992[89]	50–64	10	12	20,555	7,995	Rate × 10^{-4}	3.080	5.652	Rate ratio	0.55	0.36 to 0.83
Dales *et al.*, 1979[81]	35–54	11	11	2,791	2,914	Risk × 10^{-3}	5.016	4.804	Risk ratio	1.04	0.50 to 2.19
Frisell *et al.*, 1991[88]	40–64	4.5	~7	39,164	19,943	Rate × 10^{-4}	1.443	2.036	Rate ratio	0.71	0.4 to 1.2
Hakama *et al.*, 1995[95]	40–47	9	9	4,319	6,223	SMR	0.08	0.75	SMR ratio	0.11	0.00 to 0.71
Miller *et al.*, 1992[91]	40–49	4	7	25,214	25,216	Risk × 10^{-3}	1.507	1.110	Risk ratio	1.36	0.83 to 2.21
Miller *et al.*, 1992[90]	50–59	4	7	19,711	19,694	Risk × 10^{-3}	1.928	1.980	Risk ratio	0.97	0.62 to 1.52
Morrison *et al.*, 1988[86]	35–74	4	9	NA	NA	Rate × 10^{-4}	2.81	3.53	Rate ratio	0.80*	0.70 to 0.90
Palli *et al.*, 1989[87]	40–70	17	7–17	103	515	*p*(exp)	0.534	0.689	Odds ratio	0.53	0.33 to 0.85
Peer *et al.*, 1995[96]	35–49	15	10	14,200	13,200	Rate × 10^{-4}	4.510	4.802	Rate ratio	0.94	0.67 to 1.31
Shapiro *et al.*, 1982[82]	40–64	3	10	31,000	31,000	Risk × 10^{-3}	3.065	4.290	Risk ratio	0.71	0.55 to 0.93
Tabar *et al.*, 1995[97]	40–74	10	10	77,080	55,985	Risk × 10^{-3}	2.076	2.983	Risk ratio	0.70	0.55 to 0.87
Thompson *et al.*, 1994[94]	≥ 50	6	7	126	2,237		NA	NA	Relative risk†	0.61	0.23 to 1.62
UK TEDBC Group, 1993[92]	45–64	7	10	45,956	63,571	SMR	0.834	1.050	SMR ratio	0.79	0.66 to 0.95
Verbeek *et al.*, 1984[84]	35–65	8	8	46	230	*p*(exp)	0.565	0.704	Odds ratio	0.48	0.23 to 1.00

* *The rate ratio*[86] *was calculated by comparing the breast cancer mortality rate for a documented cohort of 'exposed' women with the rate for a much larger cohort derived from routine data, without reference to the number of person-years on which the latter outcome frequency was based. The CI quoted here was calculated conservatively, assuming an equal number of person-years in the unexposed cohort*

† *Data were analysed by conditional logistic regression, but the authors*[94] *argued that the analysis 'mimicked' a Cox proportional hazards analysis*

TABLE 21 *Summary of sample sizes, outcome frequencies and effect sizes for FAS studies which were considered for strategy 2*

Reference	Previous history of NTD	Dose (mg)	Duration (weeks)	Sample size		Outcome frequencies*			Relative risk reduction		
				Intervention/ case	Control	Measure	Intervention	Control	Measure	Point estimate†	95% CI
Bower and Stanley, 1992[108]	N	NS	–12 to 0	75	150 (N)	*p*(exp.)	0.013	0.056	Odds ratio	0.11	0.01 to 1.33
				75	77 (A)	*p*(exp.)	0.053	0.039	Odds ratio	0.69	0.06 to 8.53
Chatkupt *et al.*, 1994[112]	Y	NS	–? to 0	163	–	*p*(exp.)	0.288	–		NA	NA
Czeizel and Dudas, 1992[109]	N	0.8	≤ –4 to 8	2,108	2,052	Risk × 10^{-3}	0	2.924	Risk ratio	0.16	0.02 to 1.35
Kirke *et al.*, 1992[65]	Y	0.4	≤ –8 to 12	172	89	Risk × 10^{-3}	0	11.235	Risk ratio	0.52	0.03 to 8.18
Laurence *et al.*, 1981[98]	Y	2.0	–? to > 6	44	67	Risk × 10^{-3}	0	89.552	Risk ratio	0.25	0.03 to 2.04
Reanalysed by ITT analysis				60	51	Risk × 10^{-3}	33.333	78.431	Risk ratio	0.43	0.08 to 2.23
Martinez-Frias and Rodriguez-Pinilla, 1992[110]	N	≥ 0.3	0 to 12	285	8,276 (A)	*p*(exp.)	0.077	0.121	Odds ratio	0.61	0.38 to 0.96
Mills *et al.*, 1989[103]	N	RDA	–4 to 6	532	528 (N)	*p*(exp.)	0.126	0.121	Odds ratio	0.89	0.73 to 1.10
				532	520 (A)	*p*(exp.)	0.126	0.113	Odds ratio	0.93	0.76 to 1.12
Milunsky *et al.*, 1989[104]	N	0.1–1.0	0 to 6	10,713	3,157	Risk × 10^{-3}	0.933	3.484	Risk ratio	0.27	0.11 to 0.63
MRC Vitamin Study Research Group, 1991[107]	Y	4.0	–? to 12	593	602	Risk × 10^{-3}	10.118	34.883	Risk ratio	0.29	0.12 to 0.71
Mulinare *et al.*, 1988[102]	N	NS	–12 to 12	178	1,470 (N)	*p*(exp.)	0.135	0.275	Odds ratio	0.41	0.26 to 0.66
Seller and Nevin, 1984[101 §]	Y	0.4	–4 to 6	382	508	Risk × 10^{-3}	18.324	45.276	Risk ratio	0.40	0.18 to 0.93
Shaw *et al.*, 1995[113]	N	Any	–12 to ≥ 4	295	247 (N)	*p*(exp.)	0.298	0.397	Odds ratio	0.65	0.45 to 0.94

Continued

TABLE 21 contd *Summary of sample sizes, outcome frequencies and effect sizes for FAS studies which were considered for strategy 2*

Reference	Previous history of NTD	Dose (mg)	Duration (weeks)	Sample size		Outcome frequencies*			Relative risk reduction		
				Intervention/ case	Control	Measure	Intervention	Control	Measure	Point estimate†	95% CI
Smithells *et al.*, 1981[99] §	Y	0.4	≤ –4 to 8	195	295	Risk × 10^{-3}	5.128	44.068	Risk ratio	0.12	0.02 to 0.88
Smithells *et al.*, 1983[100] §	Y	0.4	≤ –4 to 8	234	215	Risk × 10^{-3}	8.547	51.162	Risk ratio	0.17	0.04 to 0.76
Smithells *et al.*, 1989[105] §	Y	0.4	≤ –4 to 8	150	320	Risk × 10^{-3}	6.667	56.250	Risk ratio	0.12	0.02 to 0.88
Vergel *et al.*, 1990[106]	Y	5.0	≤ –4 to 10	80	118	Risk × 10^{-3}	0	33.898	Risk ratio	0.37	0.04 to 3.24
Werler *et al.*, 1993[111]	N	NS	≤ –4 to 4	284	1,592 (A)	*p*(exp.)	0.120	0.213	Odds ratio	0.60	0.38 to 0.96

ITT, intention to treat; N, no; NA, not available; RDA, recommended daily allowance; Y, yes

* *Calculating outcome frequencies was not straightforward because some pregnancies resulted in twins. We attempted to report outcome frequencies in terms of pregnancies affected by an NTD, so the denominators used were total informative pregnancies. However, authors rarely stated whether twin pairs were concordant with respect to NTD outcome. N, normal control group for case–control study; A, abnormal control group for case–control study*

† *For studies where no outcomes were observed in the intervention group, point estimates were calculated by assuming a single outcome occurred in the intervention group*

§ *The three papers by Smithells and co-workers[99,100,105] and the one by Seller and Nevin[101] all report on "the same continuous and continuing"[101] multicentre prospective cohort study. The first two papers by Smithells and co-workers[99,100] cover two different recruitment periods in five or six centres. The third paper by Smithells and co-workers[105] includes data for only one of the centres, with recruitment starting after the end of recruitment to the second multicentre cohort. Seller and Nevin[101] reported data for two other centres; they state that some of their data were included in the first two papers by Smithells and co-workers, but some were obtained after the end of recruitment to the multicentre cohort. Data presented here for Seller and Nevin are combined for the two centres, rather than separately as published*

TABLE 22 *Meta-regression model of the effect of study design on effect size for MSBC studies*

Source	Sum of squares	df	Mean square	*F*	*p*	r^2(adj)
Model	0.312	2	0.156	2.19	0.15	0.137
Residual	0.927	13	0.071			
Total	1.239	15	0.083			

	Coefficient	95% CI	*t*	*p*
Cohort study	–0.033	–0.34 to 0.28	–0.23	0.82
Case–control study	–0.503	–1.04 to 0.03	–2.03	0.06
Constant	–0.163	–0.40 to 0.08	–1.47	0.17

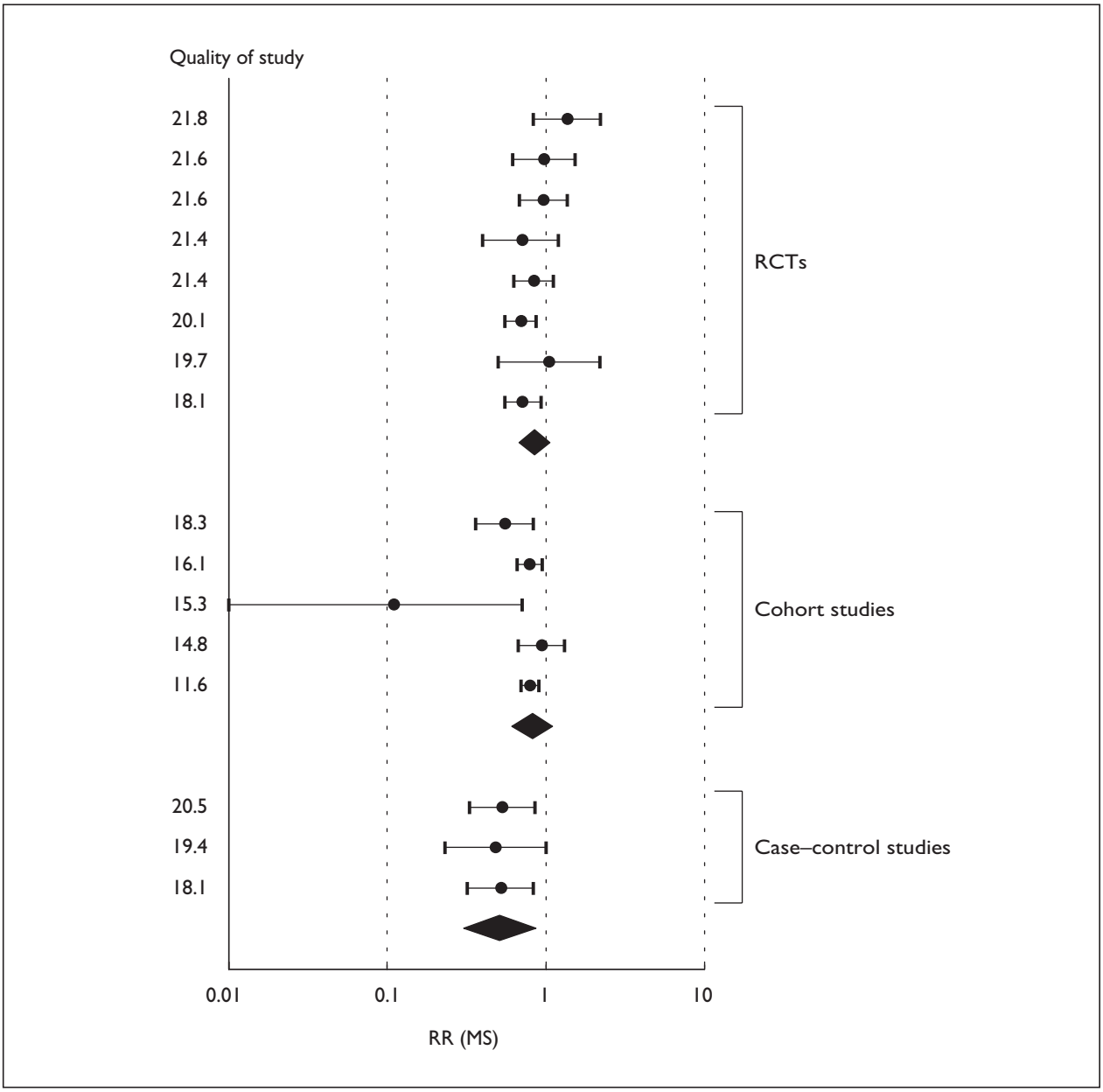


FIGURE 3 *Blobbogram showing the effect size point estimates and CIs for MS studies reviewed for strategy 2. Pooled estimates are shown separately for RCTs, cohort studies and case–control studies. Within each type of design, studies are ranked in order of quality*

TABLE 23 *Meta-regression model of the effect of study design on effect size for FAS studies*[*]

Source	Sum of squares	df	Mean square	*F*	*p*	r^2(adj)
Model	1.538	2	0.769	7.39	0.007	0.460
Residual	1.352	13	0.104			
Total	2.890	15	0.193			

	Coefficient	95% CI	*t*	*p*
Cohort study	–0.026	–1.13 to 1.08	–0.05	0.960
Case–control study	0.955	0.02 to 1.89	2.20	0.046
Constant	–1.288	–2.21 to –0.37	–3.03	0.010

[*] *The model included all studies except the one by Chatkupt and co-workers,*[112] *which did not provide an estimate of relative risk. Excluding two case–control studies*[110,111] *that had only abnormal control groups, did not affect the results*

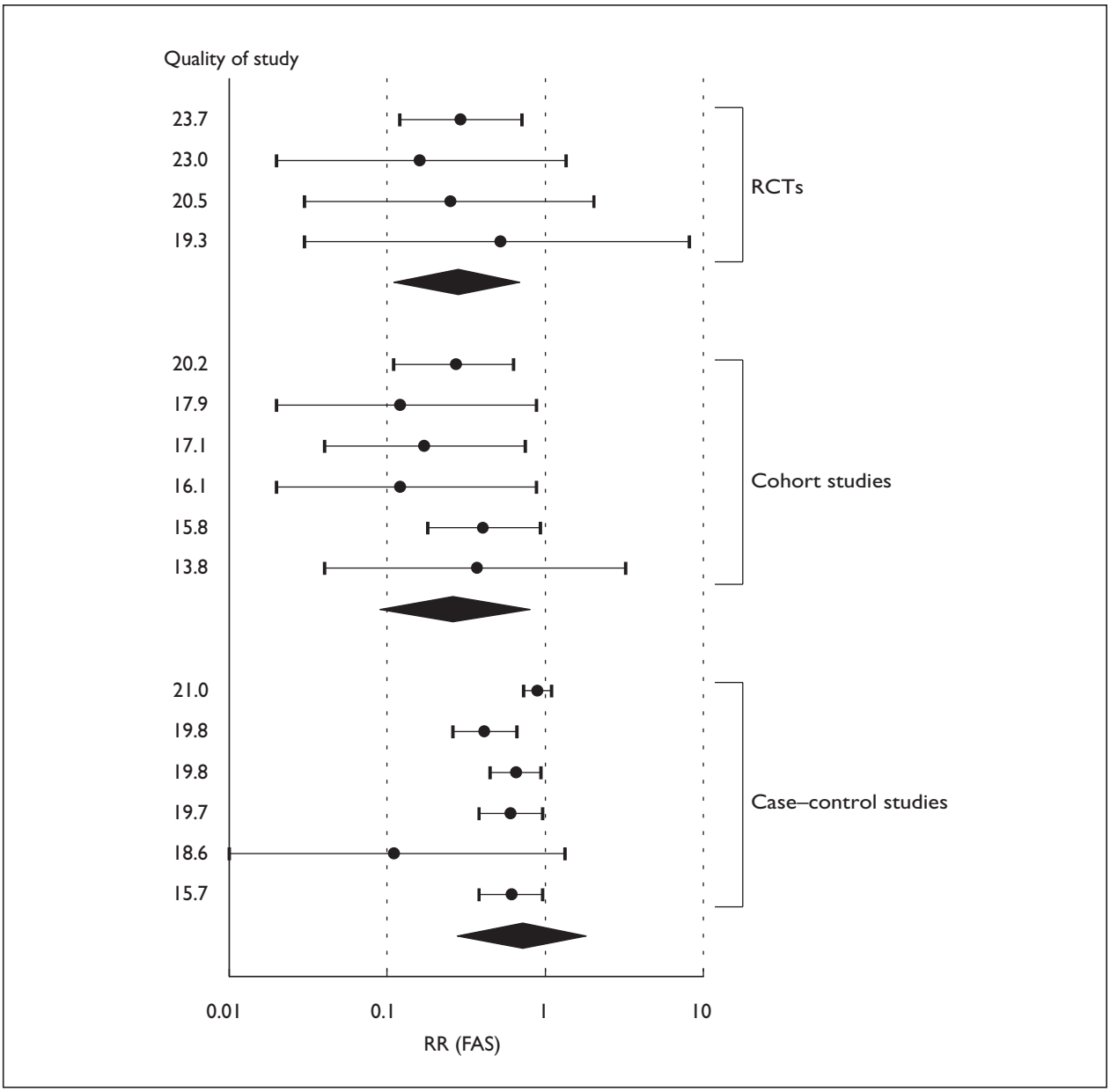


FIGURE 4 *Blobbogram showing the effect size point estimates and CIs for FAS studies reviewed for strategy 2. Pooled estimates are shown separately for RCTs, cohort studies and case–control studies. Within each type of design, studies are ranked in order of quality*

Chapter 7

Hybrid study designs and RCT variants

Study designs identified

Objective (3) sought to identify hybrid study designs and RCT variants that have been proposed to overcome difficulties experienced with conventional RCTs. Identifying studies that were eligible for this objective proved difficult. Ten study designs were located, which were classified as 'hybrids', if they were intended to provide both RCT and non-randomised estimates of effectiveness, or as 'RCT variants', if they adhered to the principle of randomisation but included some modification (*Box 2*).

Each of the study designs is described in detail below, including the authors' reasons for advocating the design, other views about the advantages and disadvantages of the design and, where available, an example of an evaluation carried out using the design.

Comprehensive cohort study

The design

The comprehensive cohort study (CCS) was first described formally by Olschewski and co-workers,[123] although previous examples of its use were cited;[59,124] the design was subsequently described in more detail.[125] The design proposes that all patients who are eligible for an RCT should be followed up, irrespective of whether they consent to be randomised or not (*Figure 5*); in effect, it is a prospective cohort study with an RCT nested in the cohort.

BOX 2 The hybrid study designs and RCT variants identified

Hybrid study designs

- Comprehensive cohort study
- Patient-preference trial
- Clinician-preferred-treatment trial
- Two-stage trial

RCT variants

- Single randomised consent trial
- Double randomised consent trial
- Randomised play-the-winner design
- Randomised discontinuation trial
- Placebo run-in trial
- Change to open-label trial

All eligible patients are asked to give informed consent to participate in the planned RCT. Patients who give consent are recruited to the RCT element of the CCS and are randomly allocated to one or other of the treatments being compared. Patients declining randomisation are given a choice between the treatments being compared and are followed up for the duration of the study to obtain the same outcomes as for the randomised patients. The prospective observation of both randomised and non-randomised patients is described as an essential prerequisite of the design.[125] Situations in which data from retrospective databases for non-randomised patients are combined with those from patients in a randomised trial should not be described as a CCS.

Olschewski and co-workers[123] contrasted the CCS with Zelen's double randomised consent design[126] (see also later in this chapter). They rejected the double randomised consent design on ethical grounds because the information given to patients at the time of obtaining consent may not be impartial if the treating clinician wants patients to agree to the treatment to which they have already been allocated by randomisation. Olschewski and co-workers[123] pointed out that it would be more ethical for clinicians to be blinded at the time of seeking consent, to obtain informed consent in the usual manner and simply to ask patients whether they accept randomisation. If they do, patients receive the treatment to which they were allocated by pre-randomisation and if they do not, patients are allowed to choose between alternative treatments. In effect, this modification creates a CCS.

In their description of the design, Olschewski and co-workers[123,125] envisage that the treatments being compared in the RCT are freely available outside the RCT on the basis of patients' preferences or other factors. However, studies were identified which resemble a CCS in all other respects but in which only one of the treatments compared in the RCT element, typically the control treatment, was available to patients who did not consent.[66,67] It is presumably also possible for patients who do

not consent to participate in the RCT to elect to receive a treatment that is not one of the treatments being compared in the RCT. This possibility is not considered explicitly by Olschewski and co-workers[123,125] and we assume that they would exclude such patients from a CCS.

It is suggested that the analysis of the entire cohort should proceed in stages.[123] First, prognostic factors should be considered. Second, the treatment main effect should be included. Third, interactions of prognostic factors and treatment should be investigated. Finally, a variable indicating whether a patient was randomised or not should be entered into the analysis. The analysis should only be considered 'stable' if the regression coefficients are not substantially altered at each step in the analysis. If the regression coefficient for the indicator variable is significantly different from zero, the results of the trial are heterogeneous and should not be generalised to the entire cohort.[123] Olschewski and co-workers[125] acknowledge that the interpretation of heterogeneous results may be difficult.

Advantages and disadvantages of the design

Olschewski and co-workers[123,125] described the following main advantages of the CCS:

- a CCS will recruit a larger total sample size than a simple RCT when it is anticipated that a large

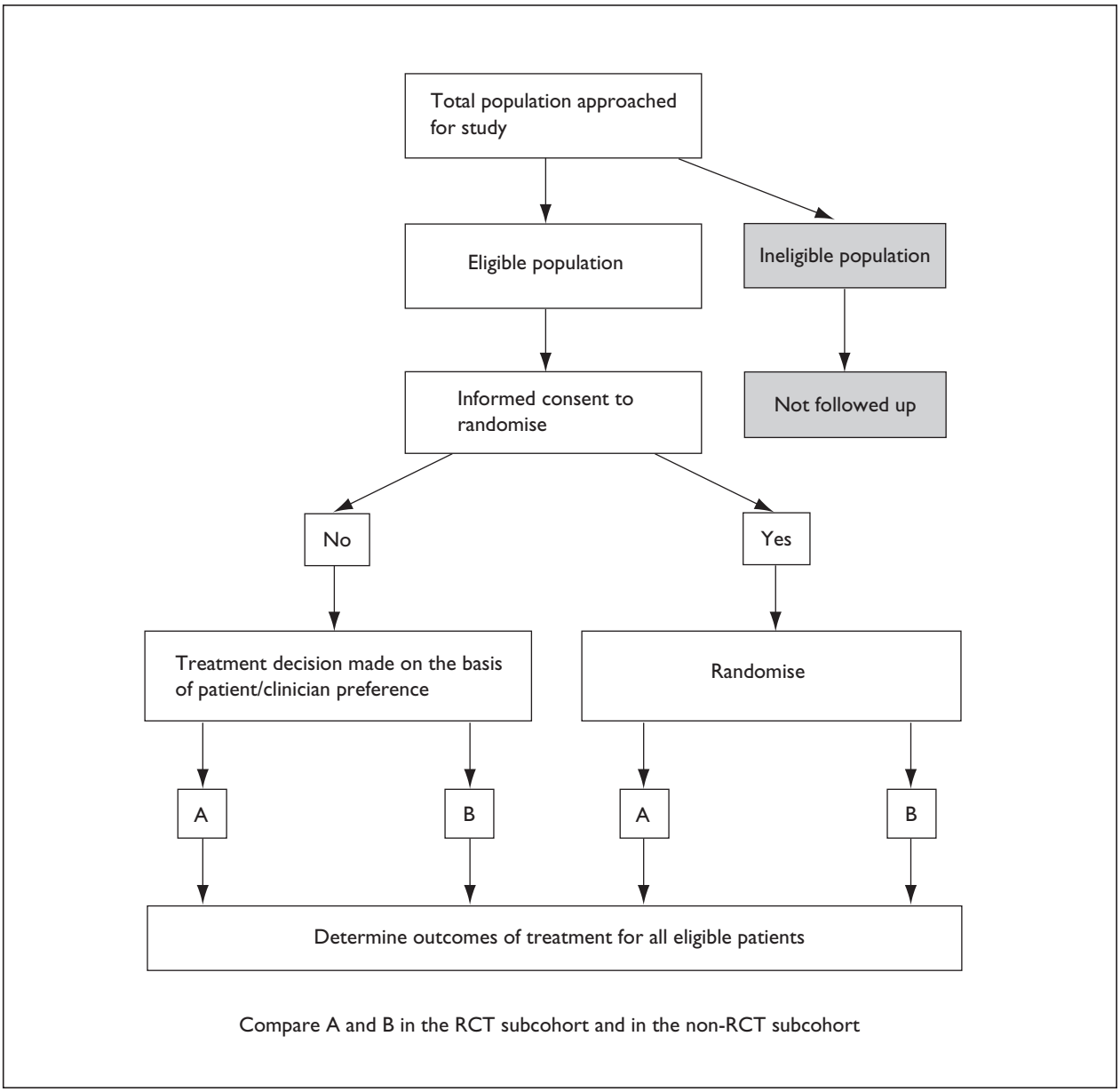


FIGURE 5 *Flow diagram of a comprehensive cohort study*

proportion of eligible patients will refuse to participate in an RCT[123]
- a CCS provides an alternative method of enhancing recruitment[123] without the ethical dilemmas posed by single or double randomised consent designs[126]
- a CCS provides an assessment of the external validity of the RCT.[125]

It should be noted that these advantages are not entirely compatible with each other. It is certainly true that a CCS is likely to recruit eligible patients to the overall cohort more quickly than to a simple RCT. However, if a CCS is to provide an assessment of the external validity of the RCT element, the randomised and non-randomised subcohorts need to be considered separately during analysis. For example, if the data for the entire cohort are analysed simultaneously, it is the interaction of the treatment with subcohort (i.e. randomised or non-randomised) that is of key interest. Therefore, the sample size for a CSS really needs to be chosen both (a) to be able to detect a clinically important difference in the randomised cohort alone and (b) to be able to detect a clinically important difference between the effect size estimates for the two subcohorts. An estimate of the proportion of eligible patients who are likely to accept recruitment to the RCT element is necessary in order to calculate the required sample size.

In neither of the descriptions of the analysis of a CCS[123,125] is it stated that the RCT element of a CCS should be analysed conventionally first (i.e. on an intention-to-treat basis). However, such an analysis seems to be an obvious first step, since the RCT element provides the most internally valid estimate of the difference in outcome between the treatments being compared.[127] Viewed in this way, a CCS requires a much larger sample size than a simple RCT and the faster rate of recruitment is a necessity not an advantage. The larger overall sample size required would be expected to increase significantly the costs of administration and follow-up in a CSS,[128] unless the necessary data collection can be carried out routinely (see chapter 8). This increase in costs must be weighed up against the potential benefits of including eligible patients who are not randomly allocated.

It is also important to note that the overall, observational analysis of an entire comprehensive cohort is difficult to interpret because the non-randomised subcohort may be subject to residual confounding.[127,129–131] Residual confounding can cause the estimates of effect derived from the two subcohorts to appear similar as well as discrepant.[127,131] One should therefore beware of the tendency to interpret similar estimates as evidence to support the external validity of the RCT element. The extent to which residual confounding may undermine the validity of the analysis of a CCS should be judged by conventional epidemiological standards, for example the care with which known prognostic factors have been measured and taken account of by stratification or regression modelling.[29]

The way in which patients who cross over from one treatment to another are handled during analysis is another important consideration in a CSS.[4,128] Olschewski and co-workers[125] state that the analysis of the whole cohort should be carried out on an intention-to-treat basis. The principle of intention-to-treat can be applied to the RCT element in a straightforward way. However, applying this principle to the non-randomised cohort can pose a problem if the date of recruitment, and a definitive treatment decision on this date, are not clearly documented. For example, when analysing data from registries or prospective database, it can be tempting to regard the date of initiation of a treatment as indicating both the time of recruitment and the treatment 'exposure'. An analysis conducted in this manner would confound the comparison of randomised and non-randomised cohorts with a comparison of intention-to-treat and explanatory analyses. The problem is well illustrated by the CASS study,[59] where some patients who were initially randomised to or who initially chose medical treatment subsequently underwent surgery.

There are other issues about a CCS that should be considered:

- The precise way in which recruitment is offered to patients may affect recruitment to the randomised subcohort. Ideally, consent for randomisation should be sought as in a simple RCT, with consent for follow-up only being sought from those who refuse. However, clinicians may enter more people into the observational element of the trial, rather than the RCT element, due to ethical concerns or fear of compromising the doctor–patient relationship.[127]
- If recruitment to a CCS is restricted to eligible patients,[119,121,123] it can only assess a limited number of aspects of external validity (i.e. those that arise from the refusal of patients to be randomised). Concern about the external validity of RCTs often arises from the restrictive eligibility criteria that are used rather than the

proportion of refusers; the CCS does not address this issue. A variation of the CCS could follow up patients who do not fulfil the study eligibility criteria but who might be considered for either treatment in normal clinical practice. One might also want to follow up all those who might be considered for the treatment of interest, including those who receive treatments other than those being compared, rather than only those who meet the eligibility criteria for an RCT (see chapter 8). Any conclusions about the effectiveness of treatments in such groups are likely to be severely limited by their small numbers.

- Systematic differences between those who accept randomisation and those who do not cannot be reliably assessed if patients who decline randomisation are not given a choice between the new and the standard treatment, but are given standard therapy in the observational element. This limitation arises because, if a new treatment is only available in the RCT, an incentive to join the RCT is created. When both treatments are available outside the trial, the subcohort comparison focuses on uncertainty versus preference. When only the standard treatment is available outside the trial, the RCT may include many control subjects who preferred the new treatment.

Example

The CASS study[59] was established in 1972 to compare coronary artery bypass surgery with conventional medical therapy for coronary artery disease. It was designed as a multicentre RCT, but all patients approached for the study were asked to consent to their medical records being included in a prospective registry. Eligible patients were asked to accept random assignment.

The 2099 eligible patients in the registry form a comprehensive cohort, of whom 780 accepted randomisation. In the absence of a documented, definitive treatment decision on recruitment, patients who refused to be randomised were classified as having had surgical treatment if they received surgery within 90 days or within the waiting time in which 95% of wait-listed patients in their hospital underwent their operations. In effect, this approach allowed both randomised and non-randomised cohorts to be analysed according to the principle of intention-to-treat. Surgery appeared to confer a slight, but non-significant, survival advantage among both randomised and non-randomised patients. The interaction of treatment and randomisation status did not approach significance.

Patient-preference trial

The design

The patient-preference trial (PPT) was first described formally by Bradley and Brewin[22,132] as a more appropriate method than a conventional RCT to compare the effectiveness of treatments when patients are likely to have strong treatment preferences. They argued that randomisation is not suitable when patients have strong preferences and need to be motivated to follow treatment regimens for a treatment to be successful:

> the greater the need for participation, the greater is the scope for motivation to influence outcome.
>
> (Brewin and Bradley,[132] page 313)

The problem can be expressed more formally as an interaction between physical and psychological effects of a treatment,[24] and can be described algebraically.[24,41,42] It should be noted that a preference effect of this kind (i.e. a strong belief in the efficacy of a treatment) is quite different[24] from a patient's choice for one treatment over another, because the patient has weighed up the respective balance of expected utilities from the known outcomes of the two treatments.[133] However, these two kinds of preference effect may, in practice, be difficult to distinguish. One consequence of preference effects is that an RCT might systematically underestimate the effectiveness of a treatment in practice.[12,42] In the event of a consistent preference among patients for a new treatment, the treatment may be found to be effective on psychological grounds of preference when it has no physiological benefit.[24]

As in the case of the CCS, all patients in a PPT who are eligible for the RCT element should be followed up, irrespective of whether they consent to be randomised or not (*Figure 6*). Given that the commonest reason for refusing randomisation is likely to be a preference for one or other treatment, the main difference between a PPT and a CCS appears to be the way in which participation is sought. In a PPT, patients are initially asked whether they have a preference for one or other of the treatments being compared in the RCT, as opposed to initially asking patients for consent to be randomised in a CCS. Patients who declare a preference receive their preferred treatment. Patients who have no preference are encouraged to accept random allocation to one or other treatment. Note, however, that it is not a simple matter to elicit from a patient whether the preference arises simply from weighing up the expected utilities or from a strong belief in the efficacy of

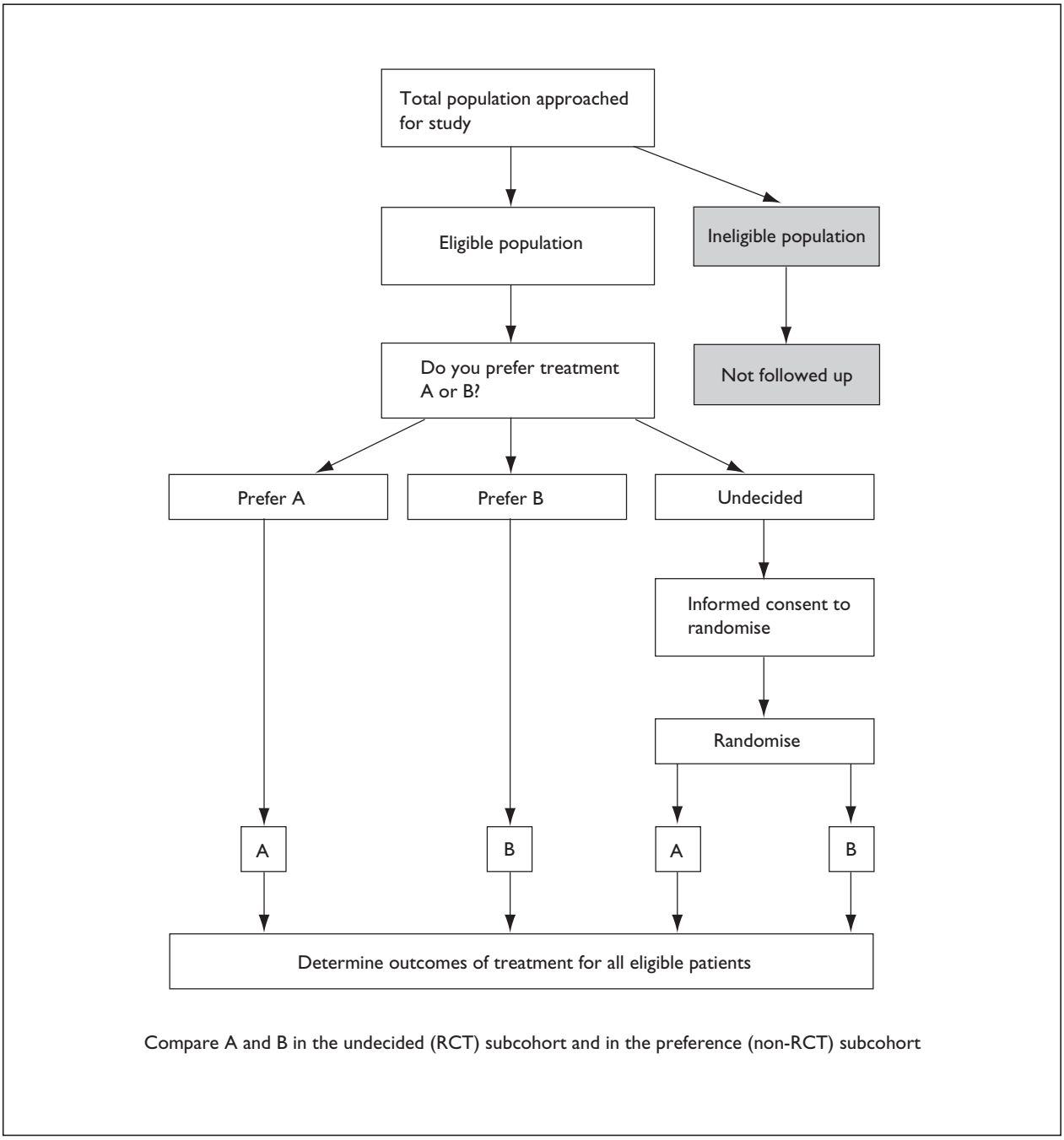


FIGURE 6 *Flow diagram of a patient-preference trial*

the preferred treatment. Bradley and Brewin[22,132] did not discuss how to analyse a PPT, but the issues to be considered appear essentially the same as for a CCS.

Advantages and disadvantages of the design

Bradley and Brewin[22,132] argued that the main advantages of a PPT are:

- a more valid estimate of the effectiveness of a treatment
- recruitment to a trial and patient compliance should improve if patients are receiving the treatment of choice.

A direct comparison of effect estimates from the preference and randomised subcohorts may be indicative of the difference between a physiological effect and a combined physiological and psychological effect of the treatment. However, as with a CCS, the non-randomised estimate may be confounded and residual confounding may cause estimates of effect for randomised and

preference cohorts to appear similar as well as discrepant.[24,130,134]

Two other concerns arise if the aim of the study is to derive separate estimates of effect for preference and randomised subcohorts. The first is the issue of the power of the study and the potential increase in costs from administering a larger study and following up more patients. The second is that, despite a likely overall improvement in recruitment rate with a PPT when patients have strong preferences, it may be even more difficult to recruit patients to the randomised subcohort when recruitment to the preference subcohort is offered at the outset (D Henry and colleagues (personal communication, 1999) have compared recruitment with a CCS and a PPT and observed a higher recruitment rate to the randomly allocated subcohort using a CCS.) Imbalance in the ratio of patients entering the two subcohorts increases the overall sample size required to detect a clinically important difference in effect size between the subcohorts.

Torgerson and co-workers[135] have suggested an alternative method for studying patient preferences. In the context of a conventional RCT to evaluate the effectiveness of a general exercise programme for patients with subacute back pain compared with standard treatment, they demonstrated that it was possible to recruit patients to the RCT while at the same time eliciting their preferences and the strength of their belief about the effectiveness of the new treatment. The authors suggested that preference effects can be estimated by entering indicator variables to represent concordance between allocated and preferred treatment as covariates in a conventional regression analysis. Although the interaction of treatment group and preference was not reported, patients who preferred the exercise programme were reported to have had a stronger belief in the exercise programme and more severe back pain. In view of the desirability for patients recruited to an RCT to be in equipoise with respect to the treatments being compared,[25,26] attempting to recruit patients who have expressed preferences to an RCT may raise ethical concerns.

Example

Henshaw and co-workers[136] used a PPT to assess women's preferences for and the acceptability of medical abortion (with mifepristone 600 mg followed 48 mg later by gemeprost 1 mg vaginal pessary) and vacuum aspiration in the early first trimester of pregnancy. Eligible women were offered a choice between surgical vacuum aspiration and medical abortion. Women with a preference (20% preferred medical abortion and 26% vacuum aspiration) received the method of their choice. Women not stating a preference (54%) were consented to random allocation to one or other method of abortion. Acceptability of the type of abortion was the primary outcome measure, which was assessed by the method a woman would choose if she ever had to have another termination in the future. The reasons for women's preferences, when preferences were expressed, were also studied.

The acceptability of the two methods of abortion differed in preference and randomised subcohorts. Only 4% of women with preferences for either method said that they would opt for a different method in the future. Of women randomised to treatment, only 2% of those who underwent vacuum aspiration, but 22% of those randomised to medical abortion, said that they would choose a different method. In women allocated to preferred treatments, only one of 12 bipolar adjectives about the acceptability of the method received was rated as significantly different between the groups of women receiving different methods (vacuum aspiration was less painful). In women allocated at random, medical abortion was rated significantly less acceptable on six of the 12 bipolar adjectives. The authors concluded that women with treatment preferences should be allowed to exercise their choice.

Clinician-preferred-treatment trial

The design

Korn and Baumrind[137] attempted to address the problem faced by clinicians who are asked to participate in RCTs when they already have existing preferences for, or opinions about, one or other treatment. Doctors with pre-existing treatment preferences may face ethical dilemmas when asked to enrol patients into a trial; such dilemmas may cause clinicians to subvert planned concealment of randomisation and hence introduce selection bias.[138]

In a clinician-preferred-treatment trial (CPTT), eligibility criteria are set at the outset of the study and eligible patients undergo an objective screening process designed to elicit whether patients have any clear indications for one or other treatment. Patients who do have clear clinical indications are allocated to the appropriate treatment. Patients who do not have clear indications for treatment are screened by a panel of at least two to four

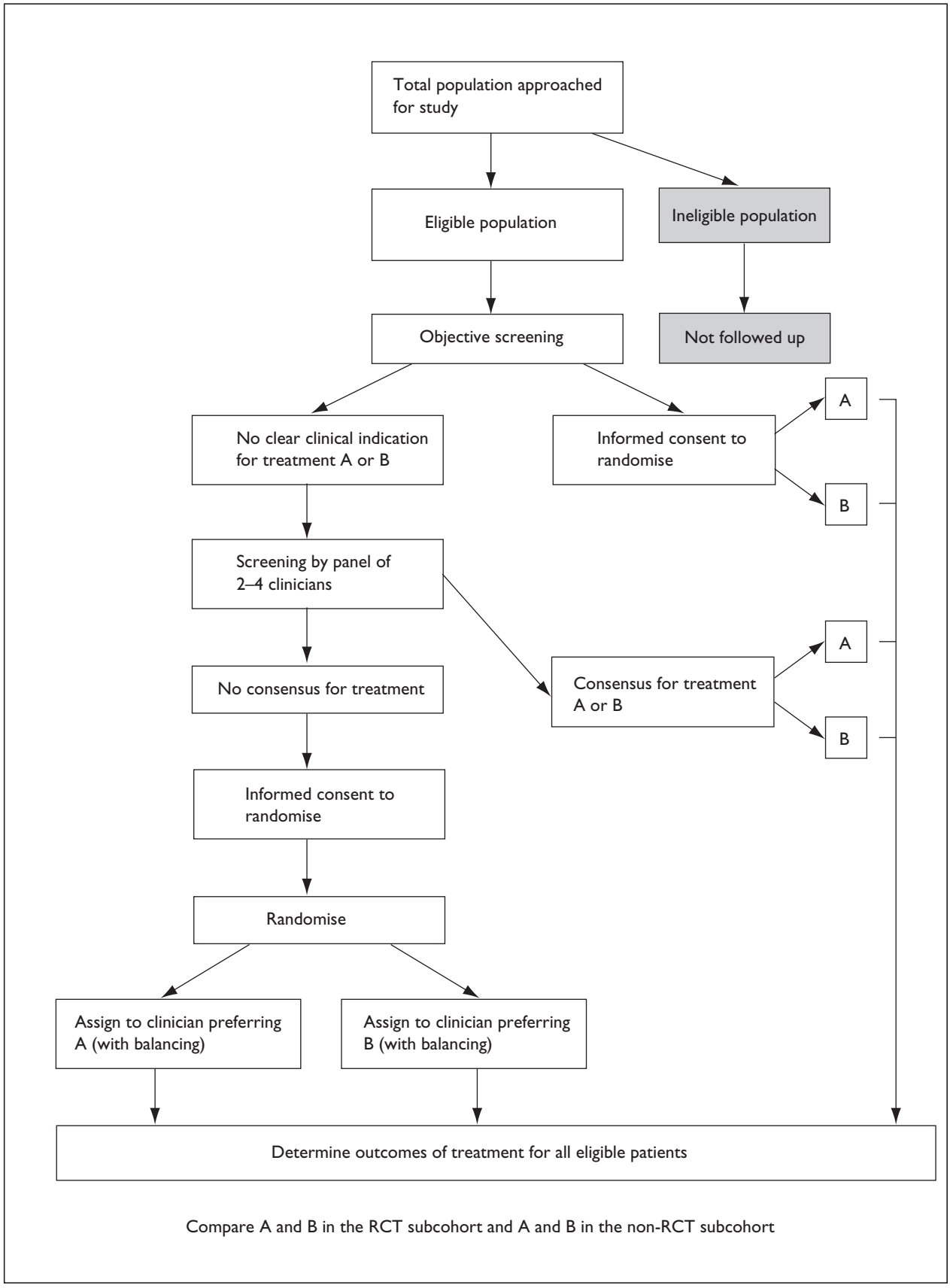


FIGURE 7 *Flow diagram of a clinician-preferred treatment trial*

doctors who are both willing and able to treat the patient. Doctors on the panel then state their preferred treatment. If the members of the panel agree about the treatment the patient should receive, the patient receives that treatment. If there is no agreement, the patient is asked for their consent to be randomised to either treatment. A randomly allocated patient is treated by a doctor who has preference for treating the patient with the allocated treatment. All patients are followed up. The CPTT is illustrated in *Figure 7*.

The result of the randomised subcohort, analysed according to the principle of intention-to-treat, represents a comparison between:

> treatment A as given by one who prefers it and treatment B as given by one who prefers it.
>
> (Korn and Baumrind,[137] page 510)

It applies directly to patients for whom there is collective equipoise among the panel. Korn and Baumrind[137] did not discuss in detail how to analyse data from a CPTT, although they indicated that patients for whom there were clear indications for treatment, or about whose treatment the expert panel agreed, should nevertheless be followed up.[137] Comparing the result from the randomised subcohort with the result(s) for the non-randomised cohorts (combined or separately) provides a test of the generalisability of the former result, albeit one that it is subject to the same limitations as for a CCS or a PPT with respect to residual confounding. These comparisons may be important, despite the clear indications for treatment or agreement among panel members, if clinicians' preferences are held in the absence of conclusive evidence about the effectiveness of alternative treatments in the non-randomised subcohorts.

Advantages and disadvantages of the design

Korn and Baumrind[137] describe three main advantages of the clinician-preferred treatment trial (CPTT):

- It overcomes ethical difficulties for clinicians who want to participate in an RCT but who are not in equipoise for all patients who satisfy the eligibility criteria.
- Compared with a conventional RCT, recruitment should be greater. Patients are more likely to accept a treatment recommended by a doctor[139] and obtaining informed consent may be easier, because the uncertainty of treatment allocation is not so great. Clinicians are more likely to recruit patients because the CPTT is less likely to be perceived as potentially compromising the doctor–patient relationship.
- The CPTT is suitable for evaluations of both existing and new treatments. Korn and Baumrind[137] cited evidence that clinicians may have strong preferences between alternative treatments in both situations.[140,141]

However, the CPTT creates logistical difficulties and is not applicable in all circumstances. It may be difficult to establish the 'objective criteria' required for the screening stage of the trial. If the members of the clinician panel change during the duration of the trial, disagreement over the criteria may emerge. Clinician preferences may change, throwing doubt over any subsequent enrolment.[137] The costs of a CPTT are likely to be high as more clinicians are involved in the study and the whole study will take longer than a conventional RCT; unless the vast majority of patients are randomised, a larger overall sample size will also be required to obtain sufficient patients in the randomised subcohort. Doctors cannot be blinded to treatment allocation. Acute interventions cannot be studied due to the lengthy processes involved in the design.

Patients for whom clear indications exist for one or other treatment may not be comparable, since the indications for treatment may be prognostic factors. If there is no overlap with respect to prognostic factors between patients allocated to treatment A and B, there is no opportunity to control for confounding. Interactions between clinician and treatment may exist, leading to difficulties in interpreting the trial conclusions. Clinician skill may also affect the results of a CPTT; this may be avoided by requiring doctors to treat an even number of patients from each group but doctors with strong preferences may by unwilling to administer both interventions.

The original description of the CPTT[137] does not consider patients who refuse to give informed consent or those who may have treatment preferences. However, Bradley[23] argued that this is unlikely to be a major limitation, since patients are likely to consent if the treatment is administered over a short period of time and has relatively few implications for the patient's life style. Bradley also suggested the possibility of combining clinician and PPTs,[23] but gave no detail of what such a design would involve. One possible 'combination' might look identical to a PPT, but with the treatment allocation in the preference cohort based on a negotiated preference of both patient and clinician.

Some of the features of the CPTT have similarities with other study designs that have tried to take account of clinicians' preferences. In an attempt to address the fact that clinicians may have individual zones of equipoise between alternative interventions, the Fetal Compromise Group has set up a conventional RCT in which clinicians set their own criteria for eligibility.[142] The study was designed to investigate the trade-off between continuing a pregnancy or delivering the fetus early, when the fetus showed certain signs of distress. This innovative design is suited to situations in which the threshold on a particular dimension (gestational age in the case of the above example) for switching from one treatment to the other varies between clinicians. The aim of the research is to identify the optimal threshold for switching. Varying thresholds between clinicians for switching treatments may help to explain the observation of collective equipoise in the absence individual equipoise;[26] the range of individual equipoise may be small and clinicians may be reluctant to acknowledge it.[142]

The feature of randomising eligible patients to clinicians who prefer the treatment to which a patient has been allocated may also have wider application. When an intervention may depend on the clinician delivering the treatment, for example when an operation is involved, this method of randomising may be more preferable than stratifying randomisation by participating clinicians, since clinicians may find it difficult to deliver two competing interventions with complete disinterest. This manoeuvre may also be useful for the evaluation of some organisational interventions. For example, patients could be randomly allocated to hospital wards that use different nursing practices. However, randomising patients to 'units' of healthcare delivery in this way will generate results that are unlikely to be generalisable, unless multiple 'units' providing each intervention are available.

Example

No examples of the study design described above could be found in the literature, perhaps because of the logistical difficulties of the design. However, Korn and Baumrind[137] illustrated the potential impact of doctor preferences by a practical demonstration of disagreement between orthodontists.

Ten patients were selected from a large pool of patients on the basis that orthodontists would probably disagree about the preferred treatment for the patients. The cases were presented to 14 orthodontists. Eleven orthodontists evaluated at least five of the cases. Disagreement between the orthodontists occurred in 90% of the cases.

Two-stage trial design

The design

The two-stage trial design (TSTD)[41] is designed to separate and quantify the physiological effects of a treatment, self-selection biases resulting from patients being free to choose their preferred treatment and the interaction between these two effects (the preference or psychological effect of a treatment[24]). Eligible patients are randomised into two 'arms' of the study (*Figure 8*). In one arm (the 'option group') patients are offered a free choice between two treatments undergoing evaluation; in effect, this arm of the TSTD is a PPT. Consent is sought from patients not expressing a preference to be randomised to either treatment. The second arm is the 'random group' and is equivalent to a conventional RCT.

The option and random arms include the same proportion of patients who have preferences, by virtue of randomisation. The TSTD therefore compares the outcome in a group of patients who have no choice (i.e. who are randomised) with a group of patients who are allowed to choose their treatments. Even if people with preferences can be recruited into a TSTD, as suggested by Torgerson and co-workers,[135] the estimation of independent physiological, selection and preference effects is extremely complicated.[24]

Advantages and disadvantages of the design

The advantage of a TSTD is that, in theory, it estimates the effects of physiological factors, selection and preference independently. However, it is not clear that the assumptions required for the results of a TSTD to be valid are likely to be satisfied in practice (see below). The analysis is also:

> highly laborious … [and] … large numbers are essential for sufficient precision, even for moderate effects.
>
> (McPherson and co-workers,[24] page 655)

The ability to separate and quantify the different effects depends on being able to recruit all patients in the random group. It is therefore necessary to describe the whole design to potential participants at the outset, that is, only patients who would accept randomisation in the 'random option' arm can be recruited. Given the difficulties in

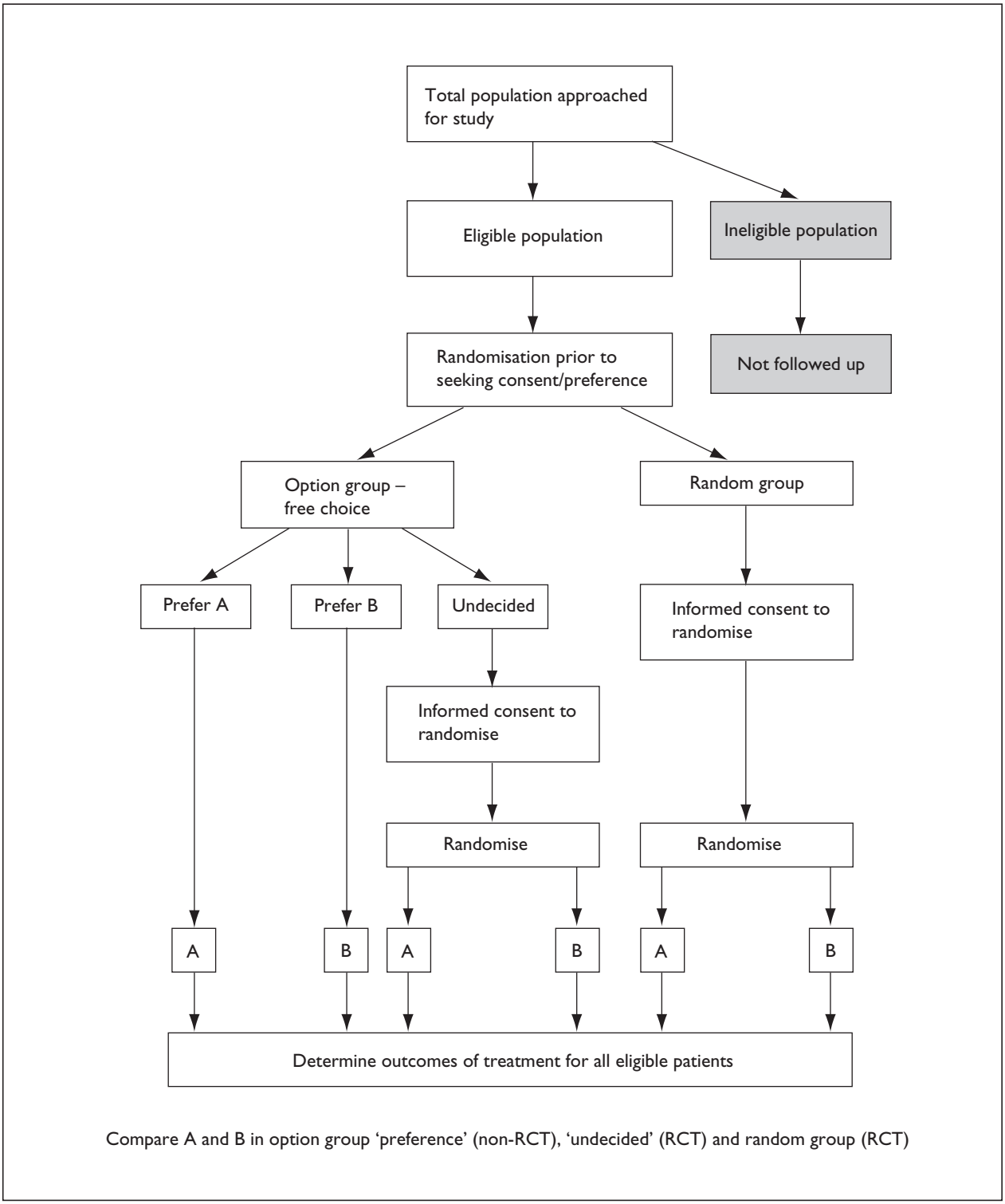


FIGURE 8 *Flow diagram of a two-stage trial design*

explaining a conventional RCT, it is likely to be even more of a problem to explain a TSTD, making informed consent difficult to obtain. Patients may be unwilling to be randomised to a non-choice group, although they may opt to participate in a TSTD even if they have a strong preference for one or other treatment, since they have 75% of getting their preferred treatment at the outset. (Patients would be even more likely to participate if their preferred treatment was not available outside the trial.) If it is not possible to randomise all patients in the random group, the advantages of the design are lost. The TSTD makes another assumption, namely that there is no effect

of random allocation to 'option' and 'random' groups on patients' responses to treatment.[41]

Example

We were unable to find any published report of an evaluation using the TSTD.

Single randomised consent design

The design

The single randomised consent design (SRCD) is intended to address the problem of obtaining informed consent in randomised trials and thereby to facilitate recruitment.[143] However, studies using Zelen-type methodologies were already in progress.[144]

The SRCD is only appropriate when a new treatment is being evaluated against 'current best practice' or 'no treatment' controls. The key feature of the design is that eligible patients are recruited and randomised into the trial prior to obtaining informed consent to participate. Patients allocated to 'current best practice' are not asked to give their consent to participate in a trial. Patients allocated to the new treatment are asked for their informed consent to receive the treatment. Patients who do not give their consent receive the control therapy (*Figure 9*).

The trial is analysed according to the principle of intention-to-treat basis (i.e. patients who do not consent are analysed in the 'new treatment' arm of the trial to which they were originally allocated). Secondary, explanatory analyses may be carried out according to the treatments that patients actually received, with caution, to investigate possible selection biases.

Advantages and disadvantages of the design

Zelen[143] has argued that the SRCD has several advantages:

- The SRCD avoids the need to describe the process of treatment allocation by chance, usually a fundamental part of seeking informed consent, which frequently deters many patients and doctors from participating in RCTs. Patients asked to consent to receiving a treatment, after a full discussion of the benefits and risks, face a much simpler choice than in a conventional RCT. Doctors often cite their dislike of discussions of uncertainty with patients and as it may compromise the doctor–patient relationship. In a study of physicians involved in the National Surgical Adjuvant Project for Breast and Bowel Cancers, researchers found that 73% of physicians who did not enrol patients in the trial did so because of fear that the doctor–patient relationship would be compromised. Thirty-eight per cent of the doctors also stated the difficulty of obtaining informed consent.[145]
- Compliance may be higher than a conventional RCT as patients have been able to exercise choice over treatment. Randomising patients prior to consent removes the dilemma for patients who have a preference when asked to participate in a conventional RCT and are, by chance, allocated to their non-preferred treatment.[126,146]
- A high refusal rate in the group allocated to the experimental therapy may indicate that it is premature to introduce a new experimental therapy into a clinical trial.[146]
- The SRCD allows the possible biases associated with patient selection for the experimental therapy to be investigated. The characteristics of patients allocated to the control treatment can be compared with those allocated to the new intervention who receive the control treatment after refusing the new treatment. If a comparison is made between the 'as-treated' groups, the role of self-selection can be examined.[143]

Zelen argued that the SRCD removes the need to discuss random allocation to treatment with a patient, implying that patients are not informed about the method used to choose their treatment.[126,143] Zelen's argument also implies that the comparison between the two alternative treatments is not discussed with patients allocated to the new treatment group:

> The proposed design has the desirable feature that physicians need only approach the patient to discuss a single therapy.
>
> (Zelen,[126] page 1429)

However, Zelen himself described two alternative methods of seeking informed consent from patients allocated to the new treatment, namely (a) asking patients to accept the new treatment after a discussion of the risks and benefits, and (b) offering patients a choice between the new and current treatments.[143]

Not obtaining consent from patients allocated to current treatment may be considered unethical. Zelen himself recognised that some ethicists believe patients should be informed when their assignment has been chosen randomly.[146] The

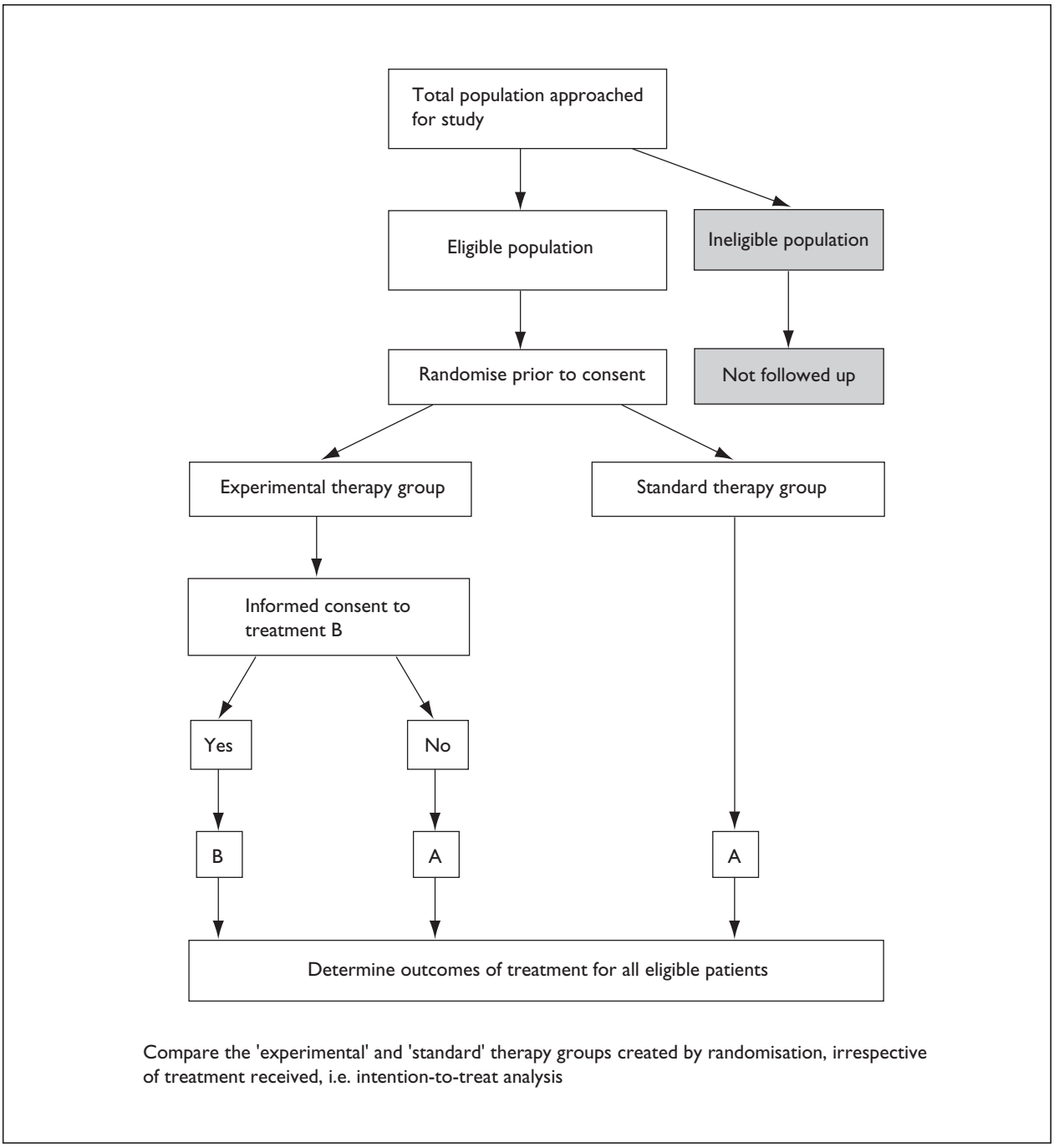


FIGURE 9 *Flow diagram of a single randomised consent design*

Code of Federal Regulations in the USA was revised in 1981 and 1983 to require informed consent to be obtained for 'best standard treatment' when a human subject is participating in research, which makes the SRCD inappropriate other than in exceptional circumstances.[146] The SRCD is also unacceptable to the UK Medical Research Council.[144]

The SRCD may have particular value when the majority of patients have a preference for a new treatment, knowledge of the treatment comparison being evaluated might bias the assessment of outcome and the intervention has an extremely low probability of causing harm. These circumstances may arise, for example, in evaluations of educational or rehabilitation interventions to improve the quality of life or functioning of patients with residual pathology that cannot be cured. Patients in this situation are often prepared to try any intervention that might make a difference and are therefore predisposed to 'prefer' a new treatment.

Assessment quality of life and functional outcomes, however, are usually assessed by the patients themselves and may be susceptible to preference and placebo effects, and it is often difficult to design an appropriate 'placebo' control. The SRCD allows patients to be randomised to the standard or new treatment without making the comparison explicit, thus minimising preference and placebo effects. A strong preference for the new treatment also means that the result from the SRCD should not be 'diluted' by refusals.[126,144]

Because of the close similarities and complementary nature of single and double randomised consent designs, issues that are relevant to both of these designs are discussed in more detail in a later section in this chapter.

Example

Korvick and co-workers[147] used a SRCD design in a multicentre prospective randomised trial of the effectiveness of adding rifampin to standard combination therapy (β-lactam and aminoglycoside) to treat *Pseudomonas aeruginosa* bacteraemia. Effectiveness of treatment was assessed by 'breakthrough bacteraemia' while a patient was receiving antibiotic treatment and by relapse after antibiotic treatment was stopped. A total of 121 consecutive patients with *P. aeruginosa* bacteraemia were recruited from four centres and were randomised to the new or standard therapy after stratification by two prognostic factors.

Sixty-three patients were randomised without consent to the standard therapy and 58 patients were randomised to the new therapy, of whom only six patients refused the new therapy. The analysis was carried out according to the principle of intention-to-treat. No difference in survival was observed, but the group of patients randomised to the new therapy (including refusers) demonstrated a significantly lower rate of breakthrough bacteraemia or relapse (2% versus 14%). The authors concluded that the SRCD design appeared to be well suited for comparative trials of antimicrobial agents.

Double randomised consent design

The design

The double randomised consent design (DRCD) was proposed by Zelen[126,146] for situations in which there is no control or best standard treatment, where the SRCD is not appropriate. Eligible patients are randomised into two groups (*Figure 10*). One group is allocated to the best standard therapy and the other to an experimental therapy. Patients are informed after randomisation has occurred that a study is underway and asked for their consent to take part in the study (see also the section below). Patients who refuse the treatment to which they have been allocated receive the alternative treatment. As in the case of the SRCD, the trial is analysed according to the principle of the intention-to-treat basis.

Advantages and disadvantages of the design

As well as being suitable when there is no control or best standard therapy, the design also addresses the problem of not attempting to obtain informed consent in the control arm of the SRCD. The DRCD therefore satisfies the legal requirements of the US Code of Federal Regulations and the UK Medical Research Council.[144]

Randomised consent designs have provoked a wide ranging debate,[144,148–153] the majority of which has focused on ethical concerns and on the relative efficiency of randomised consent designs compared with conventional RCTs.

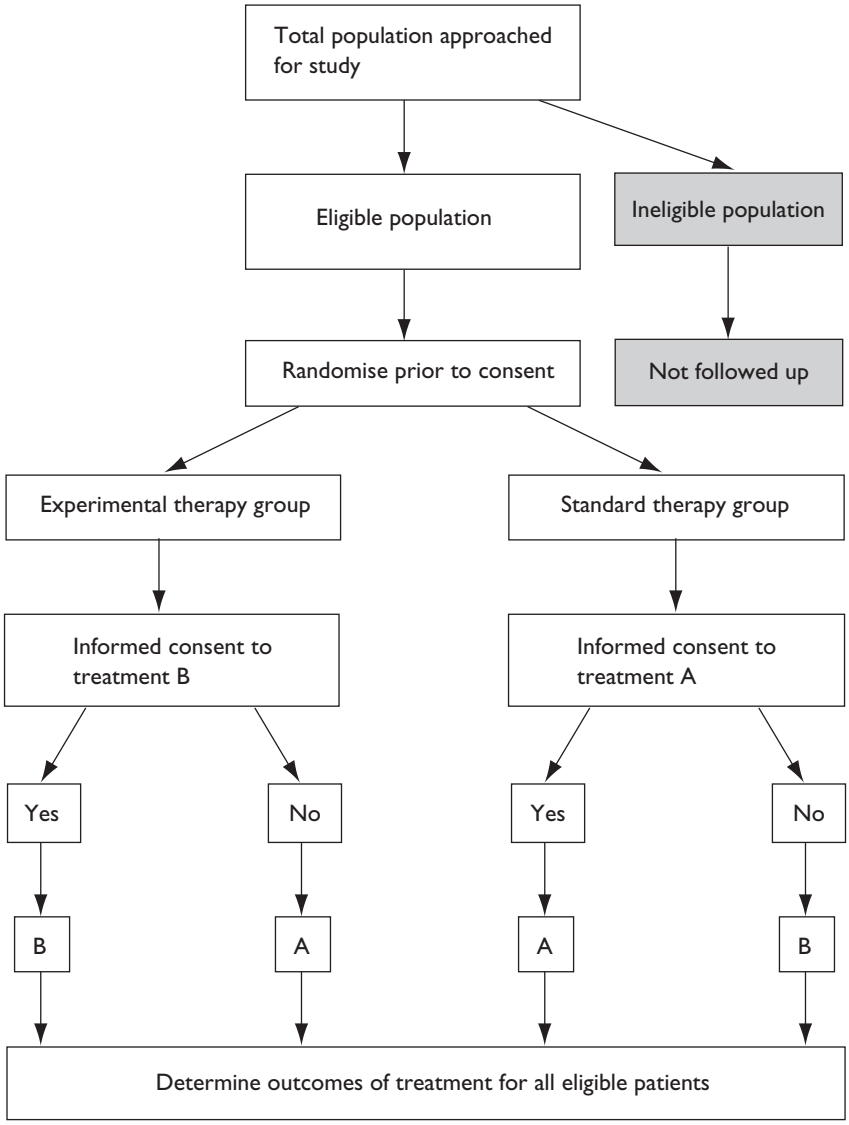
Ethical concerns have centred on the consenting process. Even with the DRCD, the consenting process used in randomised consent designs is not defined and is open to interpretation. A number of alternatives are possible (*Table 24*). The fact that several trials using randomised consent designs have been carried out and published demonstrates that different ethical perspectives exist about these designs. However, it is interesting that Altman and co-workers[144] found no published trials using a randomised consent design that had started after 1985, although this observation may be partially due to a normal lag in completing and publishing the results of trials that have been started after this date.

The relative efficiency of randomised consent designs was reviewed by Altman and co-workers.[144] Providing the acceptance rate to enter the study is higher than 50%, any differences observed may be attributed to the treatment under evaluation, but as increasing numbers of patients 'transfer' from their allocated treatment to the alternative treatment the efficiency of the trial decreases considerably. With only a 20% rate of transfer, a randomised consent design requires more than double the sample size of a conventional RCT.[144]

Although this dilution effect is the same for both the SRCD and the DRCD, the transfer rate for the

DRCD is likely to be higher because both arms of the trial have the opportunity to transfer.[154] The relatively poor efficiency of randomised consent designs neutralises one of their most important benefits, namely a faster rate of recruitment. Clearly, the increase in the rate of recruitment has to be sufficient at least to compensate for the rate of transfer; any benefit of a randomised consent design only accrues when the rate of recruitment more than compensates for the rate of transfer.

When making the decision to adopt a randomised consent design, it is therefore crucial to estimate these parameters from a pilot study. Perrone and co-workers[154] carried out a simulation to compare recruitment rates and transfer rates with the SRCD and DRCD in different circumstances. In the circumstances that were investigated, based on participants' hypothetical choices, the SRCD was found to be relatively efficient, and the DRCD inefficient, compared with a conventional RCT design. However, there may be circumstances



Compare the 'experimental' and 'standard' therapy groups created by randomisation, irrespective of treatment received, i.e. intention-to-treat analysis

FIGURE 10 *Flow diagram of a double randomised consent design*

TABLE 24 *Possible ways of obtaining informed consent in randomised consent designs*

Allocated to standard treatment	Allocated to new treatment
Single randomised consent design	
Patients not told that they are in a study; treatment comparison not stated; study outcomes must be 'routinely' available	Patients not told that they are in a study; consent sought for new treatment without stating explicit treatment comparison; study outcomes must be 'routinely' available
Patients told (in more or less detail) that they are in a study; treatment comparison may or may not be stated; permission sought for follow-up to determine relevant outcomes	Patients told (in more or less detail) that they are in a study; consent sought for new treatment; treatment comparison may or may not be stated; permission sought for follow-up to determine relevant outcomes
	Patients told (in more or less detail) that they are in a study; consent sought for new treatment; treatment comparison explicitly stated and treatment choice offered; permission sought for follow-up to determine relevant outcomes
Double randomised consent design	
Patients not told that they are in a study; consent sought for standard treatment without stating treatment comparison explicitly; study outcomes must be 'routinely' available	Patients not told that they are in a study; consent sought for new treatment without stating explicit treatment comparison; study outcomes must be 'routinely' available
Patients told (in more or less detail) that they are in a study; consent sought for standard treatment with or without stating treatment comparison explicitly; permission sought for follow-up to determine relevant outcomes	Patients told (in more or less detail) that they are in a study; consent sought for new treatment with or without stating treatment comparison explicitly; permission sought for follow-up to determine relevant outcomes
Patients told (in more or less detail) that they are in a study; consent sought for standard treatment; treatment comparison explicitly stated and treatment choice offered; permission sought for follow-up to determine relevant outcomes	Patients told (in more or less detail) that they are in a study; consent sought for new treatment; treatment comparison explicitly stated and treatment choice offered; permission sought for follow-up to determine relevant outcomes

when the DRCD is efficient. In a study by Chang and co-workers[155] investigating arthroscopy for osteoarthritis, patient recruitment for the trial increased six-fold when the design was changed from a conventional RCT design to a DRCD. The authors concluded that the design may be particularly appropriate for operative studies as participation rates in conventional RCTs of operative procedures are low.

Other considerations include:

- Double-blind trials are not possible with the SRCD or DRCD. The SRCD cannot be used for a placebo-controlled trial as consent is not obtained for those allocated to the placebo arm.[149] However, placebo treatment might be considered unnecessary if patients allocated to the control arm are unaware of the comparison being evaluated.
- Patients who are allocated to and accept the new treatment may be more conscientious about reporting outcomes or reporting for follow-up visits than those who receive standard treatment, introducing bias.
- Secondary analyses according to the treatment received may be difficult to interpret.[150,151] If the 'as-treated' analysis indicated a difference in treatment effect but the intention to treat analysis did not, it would be difficult to identify whether the difference was attributable to confounding arising from selection biases or to a treatment effect.

Example

Santen and co-workers[156] used a DRCD in a randomised trial of the effectiveness of surgical adrenalectomy compared with aminogluthemide plus hydrocortisone in women with advanced breast cancer. Effectiveness of treatment was assessed in terms of the survival of the women. A total of 96 women from two centres were randomised after stratification by several prognostic factors.

Forty-six patients were randomised to adrenalectomy and 50 to medical treatment prior to consent. After randomisation, the purpose of the study, the nature of the treatment in each arm, their possible risks and benefits and the exact nature of the randomisation process were

explained in full and written informed consent was obtained. At the end of the trial, 40 medical patients and 29 surgical patients were considered 'evaluable', although it is not clear why some of the 'inevaluable' patients were not included according to the principle of intention-to-treat. Amongst the 17 inevaluable surgical patients were seven patients randomised to adrenalectomy who refused to undergo surgery and a further eight patients in whom surgery was precluded because of rapid progression of the disease. Survival analysis showed no significant difference in survival between the two groups, although the small sample size and substantial attrition made the results difficult to interpret.

Randomised play-the-winner design

The design

The randomised play-the-winner design (RPWD) is an example of a response adaptive design. The use of response adaptive designs in clinical trials was reviewed by Rosenberger and Lachin,[157] although the concept was described as early as 1969[154,155] and subsequently by other commentators.[160–165] The principle of response adaptive designs is simple; new patients recruited to a trial are more likely to be placed on the treatment arm that currently appears to have better outcomes. As a trial progresses, treatment allocations are adapted accordingly.

There are various formulations of the RPWD. The simplest was described by Zelen[158] and is based on a simple gambling rule, namely that one should make the same (dichotomous) choice again, if successful, and only change the alternative choice in the event of a failure.[166] The primary outcome of interest in the trial must be dichotomous. The first patient entering a trial is allocated to the standard or new treatment with a probability of 0.5. A success with this patient leads to the next patient receiving the same treatment. The sequence of patients receiving the same treatment is terminated by a treatment failure, after which the next patient is allocated to the alternative treatment. Treatment allocation in the trial consists of series of varying length, each consisting of a run of successes and terminated by a failure.

This simple model is difficult to apply because it requires the outcome of the first patient to be known before the treatment allocation of the second patient can be decided. Zelen[158] described a modified rule in which the starting probabilities of receiving each treatment, typically 0.5, are modified as successes and failure become known; a treatment success increases the probability of the next patient receiving the successful treatment by a predetermined amount and a failure decrease the probability. Treatment allocation of new patients is based on the modified probabilities when a new patient is recruited, using a random-number tables or similar random-number generator.

Rosenberger and Lachin[157] described a more complex and flexible RPWD, illustrating the changing probabilities that underpin the RPWD design in terms of an urn containing a mixture of balls of two colours, one colour representing allocation to the standard treatment and the other allocation to the new treatment. The same number of balls of each colour are placed in the urn at the start. Patients are randomly allocated to a treatment by selecting a ball from the urn 'blindly', the probability of receiving each treatment being determined by the relative proportion of balls of each colour. These proportions are changed during the course of the trial to reflect the known outcomes of patients by adding balls of an appropriate colour to the urn. If the outcome of treatment is 'successful', n balls of the colour corresponding to the successful treatment colour are added to the urn. If the outcome of treatment is 'failure', n balls of the other colour are added. Assignment probabilities will consequently be skewed according to patient outcomes. The original number of balls, and number added to the urn can be adjusted, depending on the needs of the trial.

Advantages and disadvantages of the design

Proponents of the RPWD point to two main advantages:

- the design swiftly estimates the benefits (or lack) of a treatment
- the design minimises exposure of patients to the least effective treatment.

Despite these advantages, adaptive methods have seldom been used.[166]

Rosenberger and Lachin[157] argued that the RPWD is only a useful study design if adequate inferential techniques exist to provide a convincing test of the null hypothesis and to determine the magnitude of the treatment effect at the conclusion of the trial. In 1993, Rosenberger felt that the controversy over the appropriate analysis of a trial that used the RPWD (see the following section) was sufficient

evidence that these issues had not been resolved.[157] However, the increasing computer power and the development of simulation methods for estimating standard errors, for example bootstrap techniques,[167] are likely to mean that the above conditions can now be met, although the methods may not be straightforward to apply. These recent methods may also allow the RPWD to be used in trials where the primary outcome is polychotomous or continuous.[168]

The RPWD has a number of disadvantages:

- Allocation probabilities are determined by a single outcome (i.e. they cannot take account of secondary outcome measures). Analysis of secondary outcomes is also complex.[157]
- The design is unlikely to be feasible for trials of treatments for chronic diseases where outcomes may not be known for some time.
- The sample sizes required may be larger than for a conventional RCT. The number of patients receiving the apparently inferior treatment may be smaller, but this benefit can be outweighed by the number required for the alternative treatment. Consequently, the RPWD may, paradoxically, result in the general population not involved in the trial being exposed to the inferior treatment for a longer period.[160]
- RPWD designs assume that patients admitted to the trial are homogeneous with respect to characteristics that may affect treatment response,[160] although there seems to be no reason why separate randomisation sequences could not be established for separate strata. More of a problem is the possibility that the outcome of treatment for patients recruited early in the trial may differ from those recruited later due to differences in patient characteristics or the application of the treatment. The latter possibility implies that the RPWD should not be used by clinicians who are still on a learning curve with respect to a new treatment.
- Informed consent cannot truly be given in response adaptive designs because the probabilities of receiving treatments change as the trial progresses and ethical questions may be raised about allocating a patient to a treatment that is not doing as well as the alternative treatment. Disclosing current treatment probabilities when recruiting a new patient is equivalent to carrying out interim analyses of the results of the trial each time a patient is recruited.[157]
- Response adaptive designs can be susceptible to biased treatment allocation if clinicians responsible for recruiting patients know both the nature of the study design and the accumulating results of the trial. However, Berry and Eick[165] suggest that unbalanced randomisation can be used:

> unbalanced randomisation can be used in which the treatment opposite to the one assigned by the procedure is used with probability sufficiently great to ensure blindness but not so large that the advantage of the adaptive design is lost.
>
> (Berry and Eick,[165] page 232)

Example

Bartlett and co-workers[169] used the randomised play-the-winner design (RPWD) to evaluate extracorporeal membrane oxygenation (ECMO) compared with conventional treatment for moribund neonates weighing more than 2 kg within 7 days of birth. The RPWD "seemed ideal" to the authors because the outcome for each neonate was known quickly, a large difference in outcome between alternative treatments was expected (56% survival compared with 20%) and the design helped to assuage the ethical dilemma of withholding a potentially life-saving treatment from patients with an extremely high probability (> 90%) of dying.

Twelve neonates met the eligibility criteria during 18 months. The parents of all eligible patients gave consent to ECMO, after being informed about the potential risks and benefits.

The trial started from an allocation ratio of 1:1, with the outcome of each patient altering the ratio by one point. The first patient was randomly allocated to ECMO and survived, changing the allocation ratio to 2:1 in favour of ECMO. The second patient was randomly allocated to conventional treatment and died, changing the allocation ratio to 3:1. All the subsequent patients were randomly allocated to ECMO and survived. The trial was stopped at this point, in accordance with a predetermined stopping rule, because the researchers felt that the results gave evidence that efficacy had been proven.

The trial provoked criticism because the weighting was not large enough to cause sufficient patients to be allocated to the conventional treatment. Consequently, only a small number of patients were recruited. The inferential basis on which efficacy was claimed has also been questioned.[170]

Randomised discontinuation trial

The design

The randomised discontinuation trial (RDT), illustrated in *Figure 11*, was first described by Amery

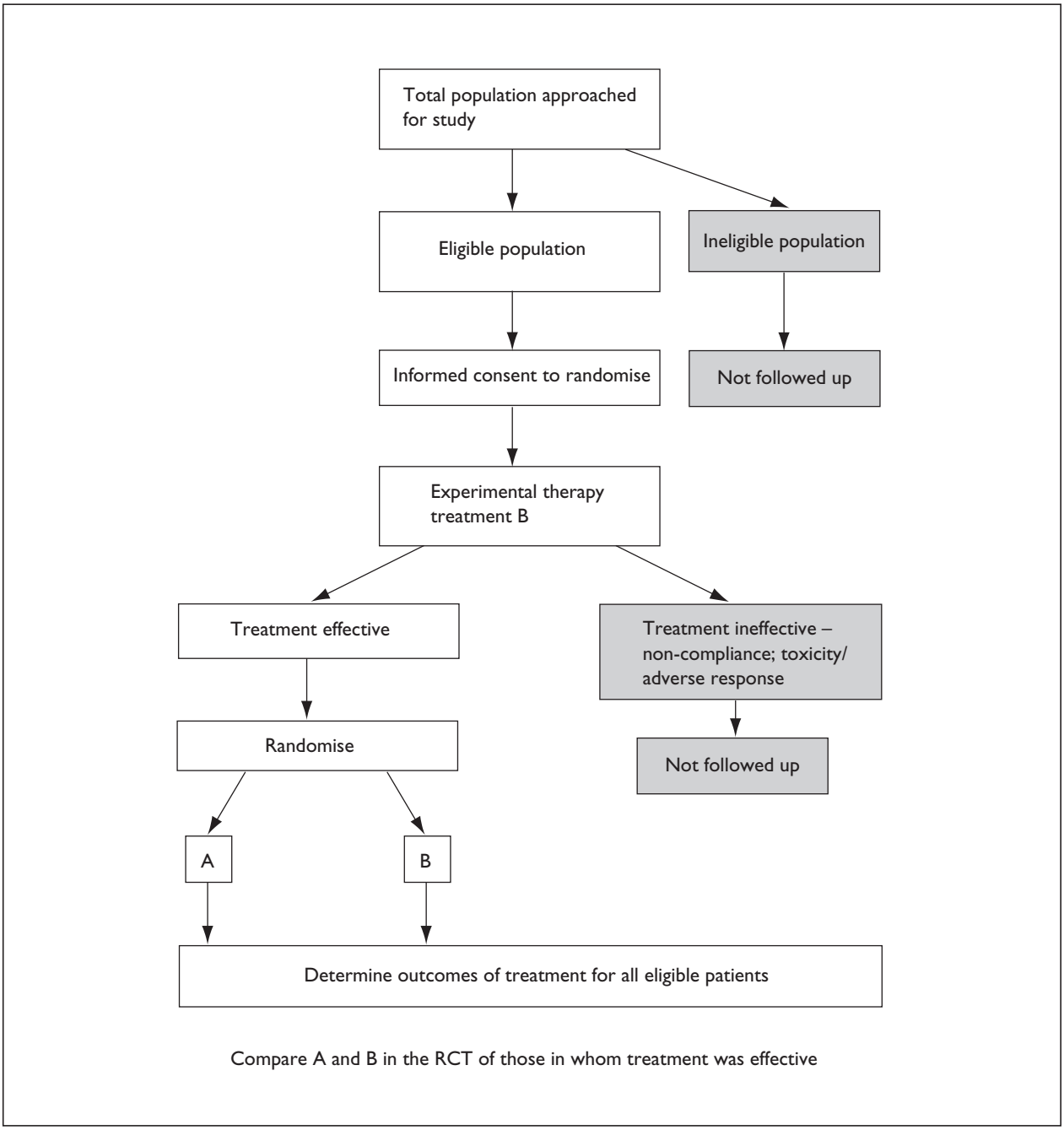


FIGURE 11 *Flow diagram of a randomised discontinuation trial*

and Dony.[171] Patients who have given informed consent are enrolled in the trial, which is divided into two phases. Phase 1 is an open phase, in which all patients are given the new treatment that the trial has been designed to evaluate. At the end of phase 1, the effects of the new active treatment are reviewed and recruited patients are divided into 'responders' and 'non-responders'. The latter group includes patients who suffer adverse health effects or fail to comply with the new treatment as well as those who fail to benefit from the new treatment.

Non-responders are excluded from the remainder of the trial. In phase 2, responders are randomised into two groups, as in a conventional RCT. Patients randomised to receive the intervention arm continue to receive the new treatment (as in the open period) and the patients randomised to the control arm receive a placebo, ideally in a double-blind manner. Formally, phase 2 is a test of the null hypothesis that all the improvements in outcome demonstrated in patients who respond to the new treatment during phase 1 are due to a placebo effect. It is suggested that the primary outcome for

an RDT should be a comparison between the number of patients relapsing or of 'survival' without relapse.

Amery and Dony[171] conceived the design in the context of trials of anti-angina treatments, where relapse in patients in whom active treatment is withdrawn is likely to be clearly identifiable within a relatively short period of time. However, they also argue that these circumstances are likely to apply to a range of other chronic conditions in which patients are maintained on active medications. Their original description was also confined to a comparison between an active treatment and an inactive placebo,[171] although there seems no reason why such a comparison should not be made against a background of best current treatment for a condition.

Advantages and disadvantages of the design

The main advantage of the RDT, emphasised by Amery and Dony,[171] is that it provides a method of minimising the exposure of patients to placebo treatment in an RCT, particularly for patients who are unlikely to benefit from the new treatment. Phase 1 also allows for the optimal dose to be established and any serious side-effects to be identified.

Kopec and co-workers[172] also argue that an RDT can be considerably more efficient than a conventional RCT. Non-responders and those who experience adverse effects of treatment, who are likely not to comply with treatment, will 'dilute' the magnitude of the estimate of the effect of a treatment in a conventional RCT. The sample size required in a conventional RCT will therefore need to be larger than for an RDT (for the same level of power), possibly by a factor of more than 2.[172] The need for a smaller sample size in phase 2 of an RDT is likely to have both logistical benefits, since poor recruitment is often a problem when carrying out an RCT, and cost benefits, particularly when a new treatment is expensive.

The relative efficiency of an RDT is achieved at the 'cost' of poor generalisability. Exclusion of non-responders in phase 2 means that the result of an RDT tends to an estimate of efficacy rather than effectiveness (i.e. the result only applies to responders, who are more likely to be fully compliant with a treatment regimen and who do not suffer serious side-effects). When healthcare practitioners have to decide whether to implement treatment for a patient, they do not usually know whether the patient is a responder or a non-responder. Hence, most organisations involved in health technology assessment, for example the Agency for Healthcare Research and Quality and the NHS R&D Directorate, now encourage evaluations of effectiveness.

An RDT cannot be used to evaluate surgical interventions or other 'curative' treatments where the effects of treatment are irreversible or persist for a long time. An RDT also only represents a fair test of the null hypothesis that benefits observed in phase 1 arise from a placebo effect if patients, at least, are 'blinded' during phase 2; patients might be expected to experience resentful demoralisation (an 'inverse' placebo effect)[40] if treatment is withdrawn. Some ethicists may also question whether it is ethical to withdraw a treatment from a patient who has experienced benefit from it.

The use of an outcome such as flare up or progression of disease may have wider application in the context of a conventional RCT (i.e. without a run-in period, which results in the exclusion of a selected subset of patients who are legitimate members of the target population for the intervention). Such a design has similarities with the change-to-open-label trial design discussed later in this chapter.

Example

The Canadian Hydroxychloroquine Study Group[173] carried out a trial using a design similar to an RDT to evaluate the effect of hydroxychloroquine sulphate compared with placebo for patients suffering from systemic lupus erythematosus. The drug was already in current use despite its efficacy not having been conclusively demonstrated. Patients with stable systemic lupus erythematosus, defined as clinical remission or minimal disease activity for at least 3 months, who had taken hydroxychloroquine at a dose of 100–400 mg/day for at least 6 months were eligible. The primary outcome measure was time to a clinical flare up of systemic lupus erythematosus or an increase in severity of disease, scored against defined criteria by a patient's clinician in a double-blind manner. This study differed from the description of an RDT published by Amery and Dony[171] in that phase 1 was implicit in the eligibility criteria rather than an explicit, prospective phase in the evaluation.

Forty-seven patients met the eligibility criteria and agreed to take part: 25 were randomised to continue with the hydroxychloroquine and 22 to receive a placebo. Clinical flare up was judged to have occurred in 16 patients on placebo and nine patients on hydroxychloroquine, which represented a 2.5-fold increase of flare up among patients who had discontinued the drug (Cox

proportional hazards analysis, $p = 0.02$). The authors concluded that the efficacy of hydroxychloroquine was proven, although the authors acknowledged that exclusion of those who had not been taking the drug previously may have resulted in the effectiveness of the drug being overestimated and its toxicity underestimated.

Placebo run-in trial

The design

The placebo run-in trial (PRIT) has often been used,[174,175] although we were unable to discover when it was first described. Because it tends to promote measures of efficacy rather than effectiveness (see below) and is relatively efficient compared with a conventional RCT, it may have originated in the pharmaceutical industry. Davis and co-workers[176,177] have commented more generally about the design.

Patients who have given informed consent are enrolled in the trial, which is divided into two phases. In phase 1 all patients are given a placebo treatment in a single-blind manner. At the end of phase 1, compliance with the placebo treatment and 'side-effects' are reviewed, and recruited patients are divided into 'good compliers' and 'poor compliers'. Poor compliers are excluded from phase 2, in which good compliers are randomised to the new treatment or to continuing with the placebo treatment, as for a conventional RCT (*Figure 12*). The analysis is carried out as for a conventional RCT (i.e. according to the principle of intention-to-treat).

Advantages and disadvantages of the design

The main advantage of the PRIT is considered to be its relative efficiency, on the assumption that patients who comply poorly during the placebo-run-in period will also comply poorly in phase 2 of the trial.[177] It will be most efficient when a high rate of non-compliance to treatment is expected,[178] or when poor adherence is associated with a substantial reduction of therapy.[179] The PRIT is also advantageous when an assessment of efficacy is needed, for example when carrying out an equivalence trial where non-compliance will bias the result towards equivalence.

As in the case of an RDT, efficiency is achieved at the 'cost' of poor generalisability. Exclusion of poor compliers in phase 2 means that the result of a PRIT tends towards an estimate of efficacy rather than effectiveness, which can only be applied to good compliers. Such a result has to be interpreted with caution, since decisions about providing treatment for an individual (or population) are usually taken without knowledge of whether an individual is likely to comply or not (or the degree of compliance in a population).

Other disadvantages of the PRIT include:[177]

- a long recruitment period if compliance is low (although the recruitment may be shorter under certain conditions[179])
- the need for effective methods of identifying non-compliers (e.g. by trace elements detectable in the urine or blood or by counting 'left over' pills), to avoid misclassification of patients at the end of phase 1.

Davis and co-workers[177] carried out an empirical evaluation of the efficiency and other consequences of the PRIT, in the context of an evaluation of a cholesterol-lowering drug among people over 65 years of age.[180] A placebo-run-in period was included, but poor compliers were not excluded from phase 2. It was observed that:

- 15% of patients (classified as poor compliers) took < 80% of the pills that they were supposed to have taken
- good and poor compliers differed with respect to the proportion that had been educated beyond high school and their mean level of triglycerides, but were otherwise strikingly similar
- compliance during the placebo-run-in period, as measured by left over pills, predicted compliance during phase 2 of the trial
- compliance was a good predictor of the effect on cholesterol lowering among patients randomised to the active treatment.

The authors concluded that the presumed increase in statistical power from the use of a placebo-run-in period would have been small in their study, whereas the effect on recruitment would have been substantial, especially among poorly educated people.

The importance of compliance was demonstrated in the Coronary Drug Project.[181] This conventional RCT evaluated the safety and effectiveness of lipid-lowering drugs in delaying death in patients with coronary heart disease. No difference in the risk of death was found between intervention and control groups, but a significant difference in outcome between compliers and non-compliers, which favoured compliers, was

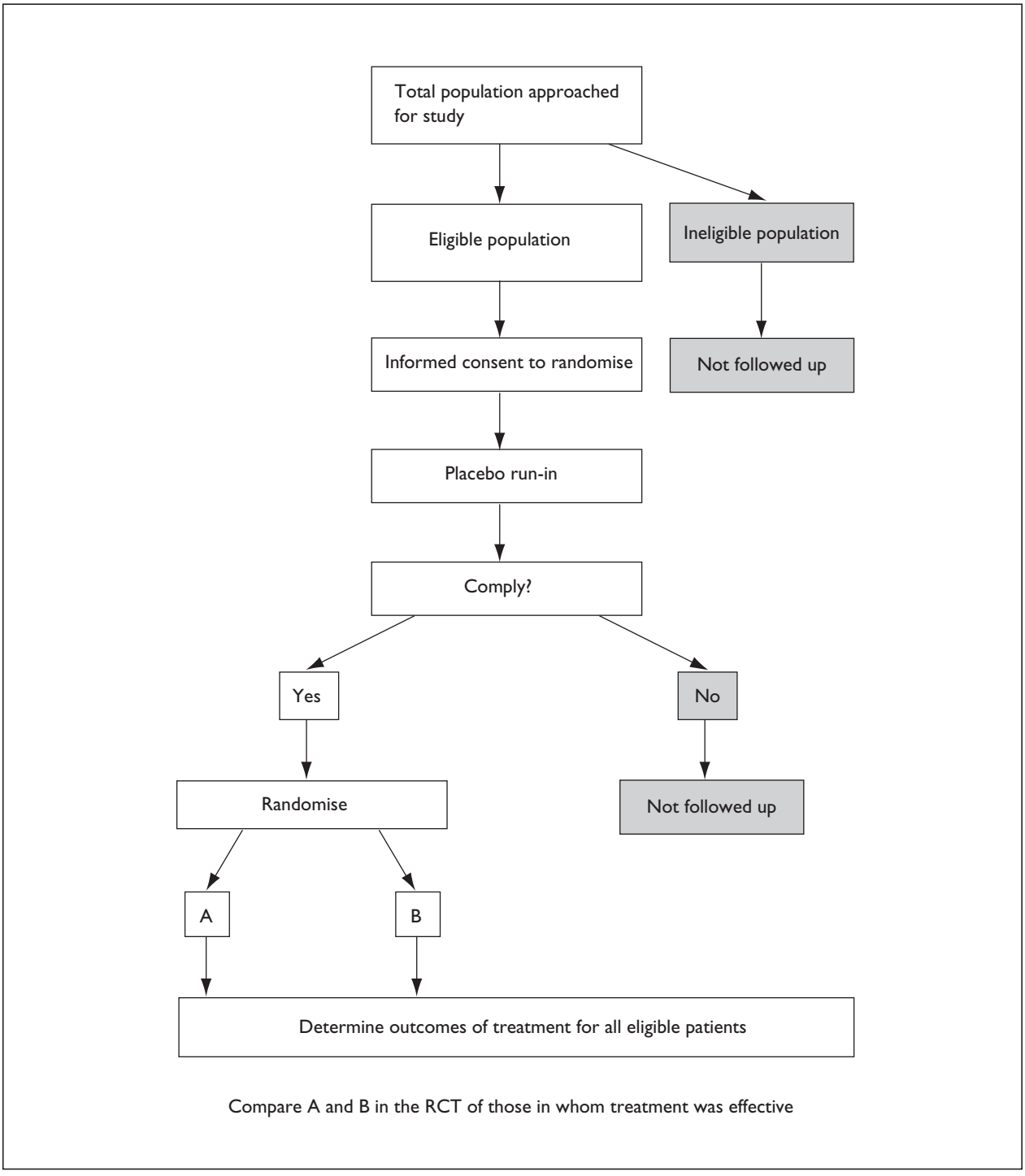


FIGURE 12 *Flow diagram of a placebo run-in trial*

reported. The size of the difference was unaffected when all measured confounding factors were taken into account. In so much as compliance may be associated with 'belief' in a treatment, this effect may be related to the possible psychological effect of a treatment discussed by McPherson and co-workers.[24] By excluding non-compliers, a PRIT would have been unable to observe this phenomenon.

Example

A PRIT was used by the SOLVD investigators[174] to evaluate the effect of enalapril, an angiotensin-converting enzyme inhibitor, on mortality in patients with reduced left ventricular ejection fractions and congestive heart failure. The RCT consisted of a short, single-blind run-in period on enalapril, a single-blind placebo run-in period and followed by a double-blind placebo-controlled

phase. Patients were excluded at both stages, so the trial combined aspects of a randomised discontinuation trial with a PRIT.[177]

A total of 7402 eligible patients were entered in phase 1 of the trial (i.e. treatment with 2.5 mg enapril for 2–7 days); 310 (4.2%) patients were excluded at the end of this phase because of non-compliance, because their condition worsened or because of symptomatic hypotension. Patients were then treated with placebo for 14–17 days; a further 295 (4.2%) patients were excluded after the placebo stage because of non-compliance or because their congestive heart failure worsened. Thirty-eight patients died during the run-in periods. In phase 3, the remaining patients were entered in either a treatment or a preventive double-blind RCT. The results of the treatment RCT showed that enapril significantly reduced mortality and hospitalisation in patients suffering from heart failure.[174]

Change-to-open-label design

The design

The change-to-open-label (COLA) design[182,183] was proposed to address various concerns about conventional RCTs:

- ethical concerns about the use of an inactive placebo in a conventional RCT, especially when there is a long duration of follow-up
- the difficulty posed for analysis by 'drop-outs'
- the lack of generalisability of the results of RCTs when participants represent an atypical subgroup of the target population for the treatment under evaluation.

The COLA design closely resembles a conventional RCT. Informed consent is requested from patients in the usual way and those who agree to participate are then randomised to alternative treatments, ideally in a double-blind manner. Patients are told that they, or their doctor, can request a change to 'open' treatment at any point in the study. The outcome measure is the time until a patient requests open treatment, analysed using survival methods.

Advantages and disadvantages of the design

Hogel and co-workers[182] pointed out several anticipated advantages:

- Recruitment to a COLA trial may be easier than to a conventional RCT, if patients know they can easily change to open treatment; this advantage should also promote the generalisability of the results.
- The outcome is quantitative and clearly defined.
- The drop-out rate should be almost zero since, after a minimum time limit, drop-outs can be treated as censored observations in survival analysis.
- Separate COLA trials of the same treatment would be more comparable, since the outcome is uniform, improving the validity of meta-analyses.

The advantage of improved recruitment may be difficult to sustain, since patients should always be allowed to change to open treatment or to drop out in conventional trials. Analysing such events by survival methods appears to be attractive, as long as bias is not introduced. Careful consideration needs to be given to censoring drop-outs, since these may be associated with the treatment allocation of patients. Differential loss to follow-up can lead to "informative drop out", which has been discussed recently in the context of combining quality of life and true survival data to evaluate trade-offs between length and quality of life.[184] Depending on whether selective drop-out arose from the ineffectiveness of a control treatment or side-effects of a new treatment, treating drop-outs as censored observations could underestimate or overestimate the effectiveness of a new treatment. It would be better to investigate reasons for drop-out, if at all possible, and to attribute these to a patient's treatment allocation (analysed as a end-point) or not (analysed as a censored observation), in a reliable and predefined manner.

Hogel and co-workers[182] also discussed likely limitations of the design. They suggested that a COLA trial would almost certainly be inappropriate:

1. to evaluate a treatment that requires some time for benefits to become apparent
2. to evaluate a treatment where side-effects predominate in the short term
3. to evaluate a treatment where beneficial medical effects are not obvious to the patient
4. to evaluate or compare preventive treatments
5. to compare treatments with varying times of onset of their effects (either beneficial or adverse).

Relevant examples in each of the above categories might be: (1) antidepressant medications; (2) radiotherapeutic or chemotherapeutic treatments for patients with cancer; (3) antiglaucoma medications; (4) treatments to improve patients' lipid

TABLE 25 *Advantages of hybrid designs and RCT variants*

	Hybrid designs				RCT variants					
	CCS	**PPT**	**CPTT**	**TSTD**	**SRCD**	**DRCD**	**RPWD**	**RDT**	**PRIT**	**COLT**
Increases generalisability of results	Y	Y		Y						Y
Takes account of patient preferences	Y	Y		Y						
Takes account of clinicians' preferences			Y							
Enhances recruitment	Y	Y	Y	Y	Y	Y				Y
Estimates physiological and preference effects				Y						
Estimates efficacy rather than effectiveness								Y	Y	
Minimises patients' exposure to placebo treatment							Y	Y		Y
Facilitates obtaining informed consent					Y	Y				
Promotes patients' compliance with treatment	Y	Y	Y						Y	Y

TABLE 26 *Disadvantages of hybrid designs and RCT variants*

	Hybrid designs				RCT variants					
	CCS	**PPT**	**CPTT**	**TSTD**	**SRCD**	**DRCD**	**RPWD**	**RDT**	**PRIT**	**COLA**
Increased sample size/cost resources compared with a conventional RCT	Y	Y	Y	Y	Y	Y				Y
A QEO element of a hybrid design susceptible to confounding	Y	Y	Y	Y						
Decreased recruitment to a RCT element of a hybrid design		Y	Y							
Design likely to encounter logistical difficulties			Y	Y						
Ethical/informed consent difficulties				Y	Y		Y			
Effect size likely to be underestimated due to cross-over					Y	Y				
Blinding of patients or researcher not possible*	(Y)	(Y)	Y	(Y)	Y	Y				
Not suitable if long-term outcomes are of interest							Y			Y
Not suitable if interventions are irreversible								Y		Y
Not suitable for preventive interventions or when benefits are not apparent to patients										Y
Generalisability compromised because design tends to estimate efficacy								Y	Y	

* *It is possible to blind the RCT element of hybrid designs but not the QEO element. In a CPTT it is also not possible to blind the clinician treating the patient*

profiles to prevent cardiovascular disease; and (5) a comparison between an allopathic and an 'alternative' treatment (e.g. homeopathy) (H Walach, personal communication, 1999). A COLA trial would also be inappropriate for any irreversible treatment such as a surgical procedure. These limitations affect the diseases that are appropriate for study in a COLA trial. The design is unlikely to be appropriate for rapidly progressive conditions with fatal or serious, irreversible consequences.

Hogel and co-workers[182] discussed in more detail the possible use of a COLA trial for treatments that have both desirable and undesirable effects. There are clearly difficulties when these effects appear 'unequal' in some sense to the patient, for example because of their different time courses ((2) above) or their varying impact on a patient's consciousness ((3) above). However, where such difficulties do not exist, a COLA trial appears attractive as an indirect, but an entirely valid and practical, assessment of the joint utility attached to the adverse and beneficial effects of a treatment by a patient. Alternative health economic methods that attempt to access and quantify such utilities can appear artificial and threatening to patients.

The COLA principle may be applicable more widely by interpreting the end-point of open-label in a broader manner. For example, the design might be useful to evaluate an established treatment against no treatment, for a chronic or slowly progressive condition, by using a highly sensitive criterion for 'progression' as the end-point. The condition could even be asymptomatic and the established treatment potentially irreversible, as in the case of surgery for glaucoma, providing the measure and definition of progression was acceptable to patients and the control treatment was no treatment or placebo.

Example

We were unable to find any published report of an evaluation using the COLA design.

Summary

Researchers have proposed designs to overcome a range of problems (*Table 25*), although the advantages have not always been substantiated. These designs also have disadvantages (*Table 26*). Apart from the TSTD, which has not been used, hybrid designs have tended to assume that estimates of effectiveness derived from non-randomised subcohorts are unbiased and that discrepancies between estimates from randomised and non-randomised subcohorts therefore reflect the factors of interest (e.g. willingness to be randomised or treatment preference). This assumption seems dangerous in view of the many examples demonstrating the near impossibility of ruling out residual confounding. It is important to note that residual confounding can explain both concordance and discrepancy between estimates for randomised and non-randomised subcohorts. Nevertheless, there may be advantages in carrying out comprehensive cohort studies or PPTs providing that the costs of doing so are proportionate to the benefits that are likely to accrue (see chapter 8).

RCT variants can be broadly classified into three categories:

- randomised consent designs (SRCD and DRCD)
- trial designs that use a run-in period (RDT or PRIT)
- designs which use the failure of treatment or a request to change treatment as the main outcome (RDT or COLA).

Randomised consent designs are limited by ethical concerns and their relative inefficiency. The SRCD may be a valuable design in special circumstances, if ethical approval can be obtained. The DRCD is unlikely to be an efficient design if patients are consented in the normal manner (i.e. by explaining both treatments in full to a patient and allowing them to make a choice between or to be randomly allocated to one or other treatment). Trial designs that use a run-in period tend to estimate efficacy rather than effectiveness and therefore go against current policy. They may be useful in phase 2/phase 3a clinical trials, providing that their limitations are appreciated by healthcare decision-makers. The open-label design is to all extents and purposes a conventional RCT. Using a request to change treatment or the failure of treatment as the main outcome may well be a useful strategy in certain circumstances, providing bias can be avoided.

Chapter 8

Summary and conclusions

Summary of findings

The results of the first strategy suggest that QEO study estimates of effectiveness can be valid (as measured by agreement with the results of RCTs), providing that account is taken of important confounding factors using standard epidemiological methods of analysis, that is stratification or regression modelling. The small size of the discrepancies for high-quality comparisons between RCT and QEO study estimates also suggests that psychological factors, for example treatment preferences or willingness to be randomised, had a negligible effect on outcome.

We are extremely cautious about generalising this finding for several reasons (see later in this chapter) and therefore cannot conclude that it is 'safe' to act on evidence from high-quality QEO studies. However, based on a simple measure of quality, low-quality QEO study evidence that makes little or no attempt to take account of confounding factors does appear to be misleading. Such studies were found to have a tendency to overestimate the effectiveness of interventions compared with RCTs, confirming the findings of other researchers,[30–33,55] although we made no formal attempt to explain discrepancies between RCT and QEO study estimates as arising from poor internal validity or differences in external validity.

Previous comparisons reported in the literature may have overemphasised the differences between RCT and QEO study estimates of effectiveness because of the poor quality of much of the non-randomised, and especially quasi-experimental, evidence. 'Pooled' comparisons between RCT and QEO study estimates of effectiveness based on reviews of primary RCTs and QEO studies are likely to be misleading, because in the past reviewers have rarely considered possible 'confounding' factors such as publication biases and other systematic differences between RCTs and QEO studies.

The second strategy found no difference in effect size between RCTs and cohort studies for both interventions that were reviewed, suggesting that evidence from cohort studies can be valid when this evidence is considered collectively. However, for both interventions, significantly different relative risk effect size estimates were obtained for case–control studies compared with either RCTs or cohort studies. Interestingly, the direction of the discrepancy was not consistent across interventions, with relative risk estimates from case–control studies indicating on average more benefit than those from RCTs and cohort studies for MSBC, but less benefit for FAS. No association was found between study quality and effect size for either MSBC or FAS interventions, after taking account of different study designs.

There are several possible reasons for the different direction of the discrepancy for the two interventions, although it should be recognised that any such explanations are *post hoc*. We conclude, in common with standard methodological texts,[185] that case–control studies designed to estimate effectiveness should be interpreted with caution. The direction of any discrepancy between relative risk estimates for case–control studies and other study designs is likely to depend on the intervention being evaluated. There was no evidence at all that discrepant estimates for case–control studies can be attributed to confounding by quality or sources of heterogeneity.

We were unable to demonstrate any independent effect of quality on effect size after taking account of study design. This finding is difficult to interpret because the failure to find an association could arise in a variety of ways:

- the analyses had inadequate power to detect an important association
- study quality is highly correlated with study design
- study quality is not associated with relative risk in a predictable way
- the instrument did not adequately characterise aspects of study quality.

We believe that the first two possibilities are unlikely, but we are unable to distinguish between the third and fourth possibilities. With respect to the power of the analyses, there was no evidence of any trend in effect with total study quality, nor of associations between components of study quality and effect size. With respect to the correlation

between quality and study design, differences in the mean quality scores for RCTs, cohort and case–control designs were not consistent with the pooled effect sizes. Studies using the same design also showed variation in quality, especially for FAS. REP and IVC were associated with effect size for FAS in the absence of dummy variables representing cohort and case–control study designs, but the coefficients were not consistent; better reporting was associated with less benefit, whereas less IVC was associated with more benefit. None of the four quality dimensions was associated with effect size in the analysis of MSBC.

There are similar difficulties in interpreting the failure to show associations between aspects of heterogeneity considered likely to be associated with outcome for *a priori* reasons and effect size. The analyses may have had inadequate power, or the independent variables included in the analysis may have been unreliable measures of the relevant attributes. Information about other important sources of heterogeneity (e.g. dose of folic acid) was not always available or was inadequately specified.

Therefore, despite being unable to demonstrate significant associations between quality, sources of heterogeneity and effect size, we are wary of reporting the pooled results for different study designs. Investigating reasons for discrepancies, rather than providing a pooled estimate, was the primary objective of the review.

Limitations of the review

Our aim was to establish 'safety limits' for QEO studies of different levels of quality, that is the size of effect which would be unlikely to be explained by bias, in order to guide interpretation of QEO study evidence. This aim was thwarted for both strategies, for different reasons. For strategy 1, we felt unable to draw strong conclusions because of the paucity of evidence, and the potentially unrepresentative nature of the evidence we reviewed. For strategy 2, the aim was thwarted by our inability to distinguish, and measure reliably, the influences of variations in internal and external validity between studies.

One aspect of the unrepresentative nature of the evidence reviewed for strategy 1 is the possibility that we failed to find all of the relevant evidence. Searching for eligible papers was extremely difficult. Comparing the literature reviewed for strategy 1 by ourselves, by Britton and co-workers[28] and by Kunz and Oxman[35] shows substantial discordance. Some of the discordance may be attributable to varying eligibility criteria adopted by different reviewers, but the small number of papers in common to the three reviews highlights the difficulty of searching and our failure to identify several potentially eligible studies.

The problems we experienced in measuring quality were, to some extent, due to our attempt to design (or modify) an instrument at the same time as we were reviewing literature. Modification of the original instrument[50] was necessary because we wanted to quantify in more detail factors that are widely regarded as compromising internal validity. Despite piloting the instrument, several ambiguities arose when we used the instrument to assess studies for strategy 2. Many of the problems related to using the same instrument to try to assess both cohort and case–control studies (see later in this chapter).

The variance of effect size estimates for different study designs was also potentially of interest to our aim of establishing safety limits for QEO studies. For example, even if QEO studies, on average, give the same answer as RCTs, it might be the case that distributions of estimates from QEO studies are intrinsically more 'noisy'. If so, one might want to inflate the conventional statistical CI for a QEO study. It may be theoretically possible to investigate the variance of effect size estimates independently of the sample size of studies, but we did not pursue this question because of the small number of RCTs and QEO studies that were available for direct comparison.

Implications of the findings

We do not recommend generalising the results found using the first strategy to other contexts, for three main reasons. First, few papers were reviewed, and our findings may depend on the specific interventions evaluated in these papers. For example, evaluations of interventions for cardiovascular disease predominated among the comparisons that were reviewed.

Second, most high-quality comparisons studied RCT and QEO study populations with the same eligibility criteria; this may have had the effect of creating relatively uniform risk strata, reducing the possibility of confounding. In contrast, QEO studies typically use more relaxed selection criteria than RCTs, recruiting populations which are more heterogeneous with respect to prognostic factors.

This tendency highlights the importance of considering the availability of different study designs, discussed in chapter 1. The situations in which healthcare decision-makers are most in need of guidance about whether or not to act on QEO study evidence are usually exactly those situations in which RCT evidence is unlikely ever to be obtained. The situations in which both RCT and high-quality QEO study estimates of effectiveness exist may be extremely constrained.

Third, the papers that were reviewed may have been subject to some form of publication bias. We have already mentioned the difficulty of identifying all published studies that were eligible for strategy 1, but have not discussed the possibility of the existence of unpublished comparisons between RCTs and QEO studies. Authors of the papers that we reviewed did not appear to be disinterested about the comparison between RCT and QEO study estimates and findings appeared to support authors' viewpoints. Therefore the papers may not be a representative sample of all instances in which researchers have set out to compare RCT and QEO study effect size estimates (i.e. researchers may have chosen to report examples that supported their points of view).

We have similar reservations about generalising the results found using the second strategy. With respect to the validity of findings from cohort studies, only two interventions were reviewed and our findings may depend on these specific interventions. For example, both interventions were preventive and were only applicable to women. Although important risk factors exist for both outcomes that the interventions are designed to prevent, neither topic required consideration of issues of comorbidity and disease severity, which may be much more serious potential confounding factors.

Three recent examples highlight the need for caution. Shadish and Ragsdale[186] reviewed published and unpublished RCTs ($n = 60$) and quasi-experimental cohort studies ($n = 36$) of marital or family psychotherapy and found that the pooled effect size for the RCTs was significantly larger than for the cohort studies (standardised effect size (RCT) = 0.60 versus standardised effect size (cohort) = 0.08). The analysis was carried out by weighted regression. The authors also investigated possible confounding factors (but not heterogeneity in the psychotherapeutic intervention) and, although the discrepancy was reduced when some of these were accounted for, the pooled RCT estimate still remained significantly larger than the pooled cohort estimate.

Egger and co-workers[51] reported meta-analyses of the association between β-carotene and cardiovascular mortality separately for RCTs and cohort studies. RCT participants were supplemented, whereas cohort participants had varying β-carotene intake or serum β-carotene concentration. The pooled relative risks for both meta-analyses were significantly different from unity but in opposite directions. The evidence from cohort studies was consistent with a strong protective effect (RR = 0.69; 95% CI, 0.59 to 0.80) whereas the evidence from the RCT was consistent with a modest increase in the risk of cardiovascular death (RR = 1.12; 95% CI, 1.04 to 1.22). Although RCTs and cohort studies recruited rather different populations, it seems unlikely that this factor alone could account for such a large difference.

The recent publication of a large RCT of the effectiveness of hormone replacement therapy in preventing coronary heart disease in postmenopausal women[187] allows a third comparison to be made with the results of QEO studies.[188] Stampfer and Colditz[188] meta-analysed the evidence from QEO studies, producing separate pooled estimates for different designs. Case–control studies that used hospital controls gave a relative risk estimate that did not differ from unity, but these studies were considered to be affected by selection biases. Case–control studies that used population controls and which were presumed to be less affected by selection biases demonstrated a modest but significant protective effect. QEO study designs traditionally regarded as demonstrating stronger evidence of causality (i.e. prospective cohort studies with a concurrent control) showed a highly significant protective effect with a narrow CI (RR = 0.58; 95% CI, 0.48 to 0.69).

In complete contrast, the large RCT showed no protective effect at all, with a reasonably narrow CI around the estimate (RR = 0.99; 95% CI, 0.80 to 1.22). As in the case of the β-carotene example, the RCT and QEO studies investigated different populations, with the RCT recruiting women with existing heart disease and the QEO studies primarily women who had elected to take hormone replacement therapy. However, again it seems unlikely that this factor alone could account for such a large discrepancy between the result of the RCT and the pooled estimate for cohort studies.

These examples, together with the results of the strategy 2, show that discrepancies of all kinds are

possible between the estimates of RCTs and QEO studies. This conclusion is consistent with the observations of Kunz and Oxman.[35] Given that all types of discrepancy are possible, it is worth considering how to interpret situations, like our own, in which RCTs and cohort studies give similar pooled estimates. As in the case of the QEO study element of a CCS or PPT, the pooled estimate for QEO studies is always susceptible to residual confounding and confounding can explain both a similar and a discrepant result compared to RCTs. O'Rourke[131] has argued that, paradoxically, combining RCT and QEO study evidence when there is no evidence of heterogeneity actually increases rather than decreases (because of the increased sample size) the imprecision of the overall pooled estimate. Thus, while one might take some reassurance from similarity between the findings of RCTs and QEO studies, it is logical to disregard the QEO study evidence and to act on the pooled RCT estimate alone.

Comparison with other reviews

There have been four recent reviews comparing the effect sizes from RCTs and QEO studies of the similar interventions[28,35,189,190] and in this section we discuss their findings and conclusions in relation to our own. One review[28] used a method similar to our strategy 1 to compare the estimates from pairs of RCTs and QEO studies. Three reviews[35,189,190] use a method similar to strategy 2, either to carry out reviews of their own[189,190] or to summarise reviews carried out by others.[35]

Since comparisons of the effects observed in RCTs with those observed in QEO studies require many issues to be considered (e.g. meta-confounding; see chapter 1), it is also important to describe the extent to which these other reviews have considered these issues. The extent to which these issues were considered by the reviews is summarised in *Table 27*.

The findings of the reviews are summarised in *Table 28*, separately for 'strategy 1' reviews and for 'strategy 2' reviews. There appears to be no striking conflict between their results. All reviews found some instances when RCTs gave a more extreme estimate of effect size and some instances when QEO studies gave a more extreme estimate of effect size. This is, perhaps, not a surprising result in view of the diverse influences on estimates of effect size (described in chapter 1), the issues specifically affecting comparisons between RCTs

TABLE 27 *Issues considered by different reviews of the effect sizes derived from RCTs and QEO studies*

Issue	Strategy 1		Strategy 2			
	This review	Britton *et al.*[28]	This review	Kunz and Oxman[35]	Benson and Hartz[189]	Concato *et al.*[190]
Differential publication bias	✓		✓	✓		
Differential dates of publication	✓		✓		✓	
Differences in the interventions studied	✓	✓	✓		✓	
Differences in the populations studied	✓		✓		✓	
Differences in the outcomes studied	✓		✓		✓	
Differences in the settings studied		✓				
Assessment of quality of primary studies	✓		✓	?*		?†
Distinguished different QEO study designs			✓			✓‡
Conclusion based on size of discrepancy	✓		✓	✓§		✓
Conclusion based on the statistical significance of discrepancy		✓	✓	✓§	✓¶	✓

✓, The issue was referred to by the authors; their results and conclusions could often still have been affected by the issue

* *Kunz and Oxman[35] reported assessing the quality of studies that they reviewed, but appeared not to consider quality in their review of randomised versus non-randomised studies*

† *Concato* et al.[190] *stated that they assessed the quality of primary studies, but they did not report their findings; it appears that the quality assessment focused on RCTs*

‡ *Concato* et al.[190] *reported two comparisons of RCTs versus case–control studies and three comparisons of RCTs versus cohort studies. However, they did not consider the QEO study type as a possible factor affecting the size or direction of any discrepancy*

§ *Kunz and Oxman[35] summarised the findings of the reviews that were identified, which were not reported consistently*

¶ *Benson and Hartz[189] described two comparisons as being discrepant, where the QEO point estimate fell outside the 95% CI for the RCT point estimate*

TABLE 28 *Summary of the findings of reviews comparing the effect sizes derived from RCTs and QEO studies*

Review	Effect size greater for RCTs	Effect size the same	Effect size greater for QEO studies
Strategy 1			
This review	19[*]	1[‡]	15[¶]
Britton *et al.*[28]	2[†]	7[§]	8[‖]
Strategy 2			
This review (cohort studies)		2[§]	
This review (case–control studies)	1[†]		1[‖ **]
Kunz and Oxman[35]	2[*]	1[‡]	5[¶]
Benson and Hartz[189]	1[† ††]	17[§]	1[‖ ††]
Concato *et al.*[190] (cohort studies)	1[*]		1[¶]
Concato *et al.*[190] (case control studies)	2[*]		1[¶]

[*] *Effect size greater for RCTs, but not necessarily significantly greater*
[†] *Effect size significantly greater for RCTs*
[‡] *Effect size identical for RCTs and QEO studies*
[§] *Effect size not signficantly different for RCTs and QEO studies*
[¶] *Effect size greater for QEO studies, but not necessarily significantly greater*
[‖] *Effect size significantly greater for QEOs*
[**] *The significance of the difference between the pooled estimates for case–control studies and RCTs was of borderline significance (p = 0.06)*
[††] *Benson and Hartz[189] describe two comparisons as being discrepant, where the QEO study point estimate fell outside the 95% CI for the RCT point estimate*

and QEO studies, and the variable extent to which different reviewers considered or took account of these issues.

With respect to the authors' conclusions, both reviewers who used strategy 1 (this review and Britton and co-workers[28]) concluded that there was no evidence that QEO studies systematically overestimate effect sizes. This was also the conclusion of the reviewers of three of the four reviews that used strategy 2 (Benson and Hartz,[189] Concato and co-workers[190] and this review). The exception was the review of Kunz and Oxman,[35] who concluded that:

> On average, non-randomised trials and randomised trials with inadequately concealed allocation result in overestimates of effect.
>
> (Kunz and Oxman,[35] page 1189)

However, the latter review included four types of comparison, two between non-randomised and randomised studies and two between high- and low-quality randomised trials. The 'headline' conclusion appears to have been based primarily on the comparisons of high- and low-quality randomised trials. The authors also qualified their conclusion:

> This bias, however, can go in either direction, can reverse the direction of effect, or can mask an effect.
>
> (Kunz and Oxman,[35] page 1189)

There was less agreement between the more detailed conclusions of the reviewers, arising from differences in focus of the specific reviews and differences in interpretation. Britton and co-workers[28] focused on differences in external validity between RCTs and QEO studies, and concluded that it should not be assumed that the results of RCTs apply to all potential patients. Kunz and Oxman[35] concluded that their results supported the current policy of valuing the results of RCTs and meta-analyses of RCTs most strongly, and of being extremely cautious about evidence from QEO studies. Benson and Hartz[189] concluded that observational studies usually do provide valid information. Concato and co-workers[190] simply concluded that cohort and case–control studies do not systematically overestimate effect sizes.

The limited evidence available does not support the view that effect size estimates from QEO studies are systematically biased. However, this does not imply that estimates of effect size from QEO studies are 'usually valid', although adherence to well-recognised study design principles will promote the validity of QEO studies. RCTs should remain the preferred study design for evaluating health technologies, but high-quality QEO study designs should be considered when RCTs are impracticable.

Recommendations for future research

Our recommendations for future research relate directly to the problems we experienced in carrying out the review.

Our first recommendation is a general one. For both strategies, many of the quasi-experiments that were reviewed were of poor quality, both with respect to the conduct and reporting of the research. Given that it is possible to conduct and report QEO studies to a high standard, we strongly recommend that the use of quasi-experimental designs to evaluate healthcare interventions should not be rejected on the basis of past QEO study evidence.

Many more quality factors need to be considered when reviewing QEO studies compared with RCTs and, arguably, this makes it even more important to establish and keep under review standards for the reporting of such studies, similar to the Consolidation of Standards for Reporting Trials (CONSORT) statement.[191] We therefore recommend the development of standards for reporting QEO studies; these would almost certainly have an impact both on the conduct of future studies as well as on the reporting of current ones. Creating the consensus required for such a statement would also be likely to provide the basis for an instrument to assess the quality of QEO studies (see below).

The poor quality of much of the evidence that we reviewed leads to our third recommendation, namely that there is a need for more direct evidence about the comparability of findings from RCTs and QEO studies. Given the difficulties we identified of separating the influences of internal and external validity, and the limitations of all hybrid designs, this recommendation in turn raises the question of how best to obtain such evidence.

Despite the limitation of residual confounding of the QEO study element, we believe that the CCS is the best study design to use to obtain such evidence. Studies need to be carried out in areas where RCTs are the preferred design, as well as areas where RCTs are problematic, in order to assess the generalisability of evidence about the validity of QEO study evidence. It should be possible both to inform the debate about QEO study evidence and to provide evidence of importance to healthcare decision-makers, by carrying out such studies on topics that have been prioritised. Analyses of comprehensive cohort studies should focus on using the RCT estimate of effectiveness, and estimates of the effect of other prognostic factors from the entire cohort, to predict outcome frequencies for different groups of patients who were not randomised. Close agreement between predicted and observed results would provide reassurance about the validity of QEO studies. However, this approach cannot take account of interactions between an intervention and prognostic factors.

Comprehensive cohort studies are expensive, since they need at least double the sample size of a conventional RCT. It would therefore be attractive to nest RCTs in established high-quality prospective databases, where relevant prognostic factors and outcomes are routinely recorded and where large numbers of patients can be studied at reasonable cost.[192] Where high-quality prospective databases are established, and the marginal cost of collecting data for an additional patient is negligible, the aim should be to collect data for all patients.

Consideration also needs to be given to the optimal way of obtaining consent and offering randomisation or treatment choices to patients in comprehensive cohort studies. Other issues, such as careful definition of treatment decisions and their timing, need to be considered in the design of prospective databases to maximise the validity of subsequent comparisons between randomised and non-randomised groups of patients.

A fourth recommendation concerns the difficulty in identifying relevant literature for strategy 1, particularly our inability to design effective searches of electronic databases to detect studies according to the design or designs used. At least three solutions are possible:

- index QEO studies by the design used with the same care as for RCTs
- compile a register of relevant studies
- design innovative search strategies.

Because of ambiguities in the ways in which study designs can be described and the difficulty of classifying some designs, a reliable method of indexing study designs in electronic databases may be a remote prospect. It is also likely that highly skilled staff would be needed to carry out the classification, given the history of indexing of RCTs on MEDLINE.

A more optimistic possibility is some form of collaboration between interested groups, such as was acknowledged by Kunz and Oxman,[35] in order

to compile a register of relevant studies similar to the Cochrane Controlled Trials Register, detailing important attributes of included studies in a consistent manner. Such a register could provide an important research resource, similar to the Cochrane Controlled Trials Register, and would differ from the existing database of methodological research held in the Cochrane Library. A register would go some way to preventing the selective citation of articles to support researchers' points of view, but would not overcome the problem of methodological researchers potentially not publishing results that are contradictory to their views.

Researchers working on other methodological topics prioritised by the NHS R&D Executive HTA Programme provide an example of the third possibility (HTA Programme Methods Group, personal communication, 1999). Papers, already known as being important to the topic under review, have been used as 'seeds' for searching forward in time using the Science Citation Index. Relevant papers that are found (directly or in the reference lists of other papers) can be used to re-seed other searches in 'snowball' fashion. Investigating systematically the value of novel methods of searching could make an important contribution in improving the 'effectiveness' of future reviewers tackling methodological topics for which relevant literature may be difficult to identify.

Our fifth recommendation arises from the considerable problems we experienced in measuring study quality and other aspects of study design which may influence effect size. The instrument that we used was not entirely successful, largely because of the compromises and ambiguities that arose from using the same instrument for all study designs. It is not clear that it is worthwhile comparing the results of evaluations using different designs for other interventions (i.e. strategy 2) until a more suitable instrument has been developed to assess different attributes of studies.

The instrument must be able to assess all aspects of study design that may influence effect size and should make explicit the aspect of quality that each item is intended to assess and the anticipated direction of its influence on effect size. If possible, the inclusion of items should be based on empirical evidence of its effect. It may be preferable simply to define relevant items rather than to pool items to give an overall score. Separate instruments are likely to be required for cohort and case–control studies, if reviewers intend to include the latter, necessitating some method of standardising quality scores when comparisons are made across study designs.

Our ambivalence about recommending the development of separate quality assessment instruments for different study designs arises because of the potential difficulty of then judging the quality of different study designs side by side. There are also difficulties concerning the scoring of questions that are not obviously applicable to some designs. We believe it may be possible to develop an instrument to assess the quality of both RCTs and prospective cohort studies (where the quality issues are broadly similar), which would be advantageous when reviewing evidence about a particular intervention obtained using both types of design. However, even in this situation, there are difficult issues about the extent to which RCTs should be 'credited' with good quality for some items (e.g. confounding), and about the assessment of QEO studies for items such as blinding and analysis by intention-to-treat. Retrospective cohort studies cause problems with other items, such as loss to follow-up and refusals. The problems are much more extensive with case–control studies.

Although this review has not achieved its aim of establishing 'safety limits' to guide interpretation of QEO study evidence, it has provided evidence that it is possible to design valid QEO studies to evaluate interventions which cannot be evaluated by RCTs. We did not identify any special features, over and above those already well known to clinical epidemiologists, that should be taken into consideration in the design of such studies, although meticulous attention to quality control in the conduct of such studies is essential to minimise their susceptibility to bias.

Acknowledgements

This study was commissioned by the NHS R&D Executive's HTA Programme. We are extremely grateful for the comments of the referees, which were most helpful. We would also like to thank the following: staff at the postgraduate library, Bristol Royal Infirmary, for help in obtaining the literature required; Karin Richards, for setting up and entering references into the bibliography; Sandra Kiauka, for sharing the database of references for their review (93/43/02); Sara Downs and Nick Black, for advice about their instrument; Clare Swinburn and Kate Hutton, for help in maintaining the bibliography; Mary Sargeant, for preparing the figures for chapter 7; and Ole Olsen and colleagues of the Cochrane Non-Random Studies Methods Group, for important references, advice, and stimulating discussions.
The views expressed in this report are those of the authors, who are responsible for any errors.

References

1. Cochrane AL. Effectiveness and efficiency. London: Nuffield Provincial Hospitals Trust; 1972.
2. Chalmers I. Evaluating the effects of care during pregnancy and childbirth. In: Chalmers I, editor. Effective care in pregnancy and childbirth. Oxford: Oxford University Press; 1989. p. 3–38.
3. Jaeschke R, Sackett DL. Research methods for obtaining primary evidence. *Int J Technol Assess Health Care* 1989;**5**:503–19.
4. Byar DP. Problems with using observational databases to compare treatments. *Stat Med* 1991; **10**:663–6.
5. Davey Smith G. Cross design synthesis: a new strategy for studying medical outcomes? *Lancet* 1992;**340**:944–6.
6. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *Ann NY Acad Sci* 1993;**703**:314–40.
7. Sheldon T. Please bypass the PORT. *BMJ* 1994;**309**: 142–3.
8. Hlatky MA. Using databases to evaluate therapy. *Stat Med* 1991;**10**:647–52.
9. Wennberg JE. What is outcomes research? In: Gelijns AC, editor. Medical innovation at the crossroads. Vol. 1. Washington: National Academy Press; 1990. p. 33–46.
10. Hennekens C, Buring JE. Observational evidence. *Ann NY Acad Sci* 1993;**703**:18–24.
11. Black N. Experimental and observational methods of evaluation. *BMJ* 1994;**309**:540.
12. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996; **312**:1215–18.
13. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. 2nd ed,. Toronto: Little Brown and Company; 1991.
14. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* 1993;**270**:2598–601.
15. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. 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? *JAMA* 1994;**271**:59–63.
16. 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.
17. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *J Clin Epidemiol* 1995;**48**:23–40.
18. Doll R. Doing more good than harm: the evaluation of health care interventions. *Ann NY Acad Sci* 1993: **703**:310–13.
19. Sackett DL, Wennberg JE. Choosing the best research design for each question. *BMJ* 1997; **315**:1636.
20. Concato J, Horwitz RI, Feinstein AR, Elmore JG, Schiff SF. Problems of comorbidity in mortality after prostatectomy. *JAMA* 1992;**267**:1077–82.
21. Ellenberg JH. Cohort studies. Selection bias in observational and experimental studies. *Stat Med* 1994;**13**:557–67.
22. Bradley C. Clinical trials – time for a paradigm shift? *Diabet Med* 1988;5:107–9.
23. Bradley C. Designing medical and educational intervention studies. A review of some alternatives to conventional randomised controlled trials. *Diabetes Care* 1993;**16**:509–17.
24. McPherson K, Britton AR, Wennberg JE. Are randomized controlled trials controlled? Patient preferences and unblind trials. *J R Soc Med* 1997; **90**:652–6.
25. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;**317**:141–5.
26. Lilford RJ, Jackson J. Equipoise and the ethics of randomization. *J R Soc Med* 1995;**88**:552–9.
27. Ukuomonne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess* 1999:**3**(Pt 5):1–92.
28. 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**(Pt 13):1–70.

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

30. Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence favouring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1977;**297**:1091–6.

31. Sacks HS, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. *Am J Med* 1982;**72**:233–9.

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

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

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

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

36. Chalmers I, editor. Effective care in pregnancy and childbirth. Oxford: Oxford University Press; 1989.

37. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia: Lippincott-Raven; 1998. p. 125–34.

38. Green J, Wintfeld N. Report cards on cardiac surgeons. *N Engl J Med* 1995;**332**:1229–32.

39. 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**(Pt 10): 1–62.

40. Cook TD, Campbell DT. Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin; 1979.

41. Rucker G. A two-stage trial design for testing treatment, self-selection and treatment preference effects. *Stat Med* 1989;**8**:477–85.

42. McPherson K. The best and the enemy of the good: randomised controlled trials, uncertainty, and assessing the role of patient choice in medical decision making. *J Epidemiol Community Health* 1994; **48**:6–15.

43. Davey Smith G, Egger M, Phillips AN. Meta-analysis: beyond the grand mean? *BMJ* 1997;**315**:1610–14.

44. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994; **309**:1286–91.

45. Prescott RJ, Counsell CE, Gillespie WJ, Grant AM, Russell IT, Kiauka S, *et al.* Factors that limit the quality, number and progress of randomised controlled trials. *Health Technol Assess* 1999;**3**(Pt 20).

46. Cochrane review of methodology. Cochrane Library; 1996. Issue 2.

47. Cochrane database of systematic reviews. Cochrane Library; 1996. Issue 2.

48. Spoor P, Airey M, Bennett C, Greensill J, Williams R. Use of the capture–recapture technique to evaluate the completeness of systematic literature searches. *BMJ* 1996;**313**:342–3.

49. Cochrane controlled trials register. Cochrane Library; 1996. Issue 2.

50. 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 Community Health* 1998;**52**:377–84.

51. Egger M, Schneider M, Davey Smith G. Meta-analysis: spurious precision? Meta-analysis of observational studies. *BMJ* 1998;**316**:140–4.

52. Egger M, Davey Smith G, Scheider M, Minder CE. Bias in meta-analysis detected by a simple graphical test. *BMJ* 1997;**315**:629–34.

53. Stern MS, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997;**315**:640–5.

54. Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. VI. How to use an overview. *JAMA* 1994;**272**:1367–71.

55. Chalmers TC, Celano P, Sacks HS, Smith H. Evidence favouring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1983;**309**:1358–61.

56. Pocock SJ. Clinical trials. Chichester: Wiley; 1983.

57. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol* 1997;**50**:683–91.

58. Wortman PM, Yeaton WH. Synthesis of results in controlled trials of coronary artery bypass graft surgery. In: Light R, editor. Evaluation studies review annual. London: Sage Publications; 1983. p. 536–57.

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

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

61. Blichert-Toft M, Brincker H, Andersen JA, Andersen KW, Axelsson CK, Mouridsen HT, *et al.* A Danish randomized trial comparing breast preserving therapy with mastectomy in mammary carcinoma. Preliminary results. *Acta Oncol* 1988; **27**:671–7.

62. Gray-Donald K, Kramer MS. Causality inference in observational vs experimental studies. An empirical comparison. *Am J Epidemiol* 1988;**127**:885–92.

63. Hlatky MA, Califf RM, Harrell FE, Lee KL, Mark DB, Pryor DB. Comparison of predictions based on observational data with the results of randomized controlled clinical trials of coronary artery bypass surgery. *J Am Coll Cardiol* 1988;**11**:237–45.

64. Horwitz RI, Viscoli CM, Clemens JD, Sadock RT. Developing improved observational methods for evaluating therapeutic effectiveness. *Am J Med* 1990;**89**:630–8.

65. Kirke PN, Daly LE, Elwood JH, for the Irish Vitamin Study Group. A randomised trial of low dose folic acid to prevent neural tube defects. *Arch Dis Child* 1992;**67**:1442–6.

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

67. Fisher B, Costantino JP, Redmond CK, Fisher ER, Wickerman DL, Cronin WM, *et al.* Endometrial cancer in tamoxifen-treated breast cancer patients: findings from the national surgical adjuvant breast and bowel project (NSABP) B-14. *J Natl Cancer Inst* 1994;**86**:527–37

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

69. Schmoor C, Olschewski M, Schumacher M. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Stat Med* 1996;**15**:263–71.

70. Shaikh W, Vayda E, Feldman W. A systematic review of the literature on evaluative studies of tonsillectomy and adenoidectomy. *Pediatrics* 1976;**57**:401–7.

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

72. Law MR, Wald NJ, Wu T, Hackshaw A, Bailey A. Systematic underestimation of association between serum cholesterol concentration and ischaemic heart disease in observational studies: data from the BUPA study. *BMJ* 1994;**308**:363–6.

73. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ* 1994;**308**:367–73.

74. Veronesi U, Banfi A, Salvadori B, Luini A, Saccozzi R, Zucali R, *et al.* Breast conservation is the treatment of choice in small breast cancer: long-term results of a randomized trial. *Eur J Cancer* 1990;**26**:668–70.

75. Schumacher M, Bastert G, Bojar H, Hubner K, Olschewski M, Sauerbrei W, *et al.* A randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *J Clin Oncol* 1994;**12**:2086–93.

76. Allum WH, Hallissey MT, Kelly KA, for the British Stomach Cancer Group. Adjuvant chemotherapy in operable gastric cancer. 5 year follow-up if first British Stomach Cancer Group trial. *Lancet* 1989; **i**:571–4.

77. Veterans Administration Coronary Artery Bypass Surgery Cooperative Study Group. Eleven-year survival in the veterans administration randomized trial of coronary bypass surgery for stable angina. *N Engl J Med* 1984;**311**:1333–9.

78. European Coronary Surgery Study Group. Long-term results of prospective randomised study of coronary artery bypass surgery in stable angina pectoris. *Lancet* 1982;**ii**:1173–80.

79. CASS Principal Investigators and their Associates. Myocardial infarction and mortality in the Coronary Artery Surgery Study (CASS) randomized trial. *N Engl J Med* 1984;**310**:750–8.

80. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177–88.

81. Dales LG, Friedman GD, Collen MF. Evaluating periodic multiphasic health check ups: A controlled trial. *J Chron Dis* 1979;**32**:385–404.

82. Shapiro S, Vent W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982;**69**:349–55.

83. Collette HJA, Rombach JJ, Day NE, de Waard F. Evaluation of screening for breast cancer in a non-randomized study (the DOM project) by means of a case–control study. *Lancet* 1984;**i**:1224–6.

84. Verbeek ALM, Holland R, Sturmans F, Hendriks JHCL, Mravunac M, Day NE. Reduction of breast cancer mortality through mass screening with modern mammography. First results of the Nijmegen project, 1975–1981. *Lancet* 1984;**i**:1222–4.

85. Andersson I, Aspergren K, Janzon L, Landberg T, Lindholm K, Linell F, *et al.* Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. *BMJ* 1988;**297**:
943–8.

86. Morrison AS, Brisson J, Khalid N. Breast cancer incidence and mortality in the Breast Cancer Detection Demonstration Project. *J Natl Cancer Inst* 1988;**80**:1540–7.

87. Palli D, Rosselli Del Turco M, Buiatti E, Ciatto S, Crocetti E, *et al.* Time interval since last test in a breast cancer screening programme: a case–control study in Italy. *J Epidemiol Community Health* 1989; **43**:241–8.

88. Frisell J, Eklund G, Hellstrom L, Lidbrink E, Rutqvist LE, Somell A. Randomized study of mammography screening – preliminary report on mortality in the Stockholm trial. *Breast Cancer Res Treat* 1991;**18**:49–56.

89. Collette HJA, de Waard F, Rombach JJ, Collette C, Day NE. Further evidence of benefits of a (non-randomized) breast cancer screening programme: the DOM project. *J Epidemiol Community Health* 1992;**46**:382–6.

90. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Can Med Assoc J* 1992;**147**:1477–88.

91. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Can Med Assoc J* 1992;**147**:1459–76.

92. UK Trial of Early Detection of Breast Cancer Group. Breast cancer mortality after 10 years in the UK trial of early detection of breast cancer. *Breast* 1993;**2**:13–20.

93. Alexander FE, Anderson TJ, Brown HK, Forrest APM, Hepbrun W, Kirkpatrick AE, *et al.* The Edinburgh random- ized trial of breast cancer screening: results after 10 years of follow-up. *Br J Cancer* 1994;**70**:542–8.

94. Thompson RS, Barlow WE, Taplin SH, Grothaus L, Immanual V, Salazar A, *et al.* A population-based case–cohort evaluation of the efficacy of mammographic screening for breast cancer. *Am J Epidemiol* 1994;**140**:889–901.

95. Hakama M, Pukkala E, Kallio M, Godenhjelm K, Svinhufvud U. Effectiveness of screening for breast cancer in women under 50 years at entry: the Kotka pilot project in Finland. *Int J Cancer* 1995;**63**:55–7.

96. Peer PGM, Werre JM, Mravunac M, Hendriks JHCL, Holland R, Verbeek ALM. Effect on breast cancer mortality of biennial mammographic screening of women under age 50. *Int J Cancer* 1995;**60**:808–11.

97. Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, *et al.* Efficacy of breast cancer screening by age. New results from the Swedish two-county trial. *Cancer* 1995;**75**:2507–17.

98. Laurence KM, James N, Miller MH, Tennant GB, Campbell H. Double-blind randomized controlled trial of folate treatment before conception to prevent recurrence of neural-tube defects. *BMJ* 1981;**282**:1509–11.

99. Smithells R, Sheppard S, Schorah CJ, Seller MJ, Nevin NC, Harris R, *et al.* Apparent prevention of neural tube defects by periconceptional vitamin supplement- ation. *Arch Dis Child* 1981;**56**:911–18.

100. Smithells R, Seller MJ, Harris R, Fielding DW, Schorah CJ, Nevin NC, *et al.* Further experience of vitamin supplementation for prevention of neural tube defect recurrences. *Lancet* 1983;**i**:1027–31.

101. Seller MJ, Nevin NC. Periconceptional vitamin supplementation and the prevention of neural tube defects in south-east England and Northern Ireland. *J Med Genet* 1984;**21**:325–30.

102. Mulinare J, Cordero JF, Erickson JD, Berry RJ. Periconceptional use of multivitamins and the occurrence of neural tube defects. *JAMA* 1988; **260**:3141–5.

103. Mills JL, Rhoads GG, Simpson JL, Cunningham GC, Conley MR, Lassman MR, *et al.* The absence of a relation between the periconceptional use of vitamins and neural-tube defects. *N Engl J Med* 1989;**321**:430–5.

104. Milunsky A, Jick H, Jick SS, Bruell CL, MacLaughlin DS, Rothman KJ, *et al.* Multivitamin/folic acid supplementation in early pregnancy reduces the prevalence of neural tube defects. *JAMA* 1989;**262**:2847–52.

105. Smithells R, Sheppard S, Wild J, Schorah CJ. Prevention of neural tube defect recurrences in Yorkshire: final report. *Lancet* 1989;**ii**:498–9.

106. Vergel RG, Sanchez LR, Heredero BL, Rodriguez PL, Martinez AJ. Primary prevention of neural tube defects with folic acid supplementation: Cuban experience. *Prenat Diagn* 1990;**10**:149–52.

107. MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council vitamin study. *Lancet* 1991;**338**:131–7.

108. Bower C, Stanley FJ. Periconceptional vitamin supplementation and neural tube defects; evidence from a case–control study in Western Australia and a review of recent publications. *J Epidemiol Community Health* 1992;**46**:157–61.

109. Cziezel AE, Dudas I. Prevention of the first occurrence of neural-tube defects by peri-conceptional vitamin supplementation. *N Engl J Med* 1992;**327**:1832–5.

110. Martinez-Frias ML, Rodriguez-Pinilla E. Folic acid supplementation and neural tube defects. *Lancet* 1992;**340**:620.

111. Werler MM, Shapiro S, Mitchell AA. Peri-conceptional folic acid exposure and risk of occurent neural tube defects. *JAMA* 1993; **269**:1257–61.

112. Chatkupt S, Skurnick JH, Jaggi M, Mitruka K, Koenigsberger MR, Johnson WG. Study of genetics, epidemiology, and vitamin usage in familial spina bifida in the United States in the 1990s. *Neurology* 1994;**44**:65–70.

113. Shaw G, Schaffer D, Velie EM, Morland K, Harris JA. Periconceptional vitamin use, dietary folate, and the occurrence of neural tube defects. *Epidemiology* 1995;**6**:219–26.

114. Shapiro S, Strax P, Venet L. Evaluation of periodic breast cancer screening with mammography. Methodology and early observations. *JAMA* 1966; **195**:731–8.

115. UK Trial of Early Detection of Breast Cancer Group. Trial of early detection of breast cancer: description of method. *Br J Cancer* 1981;**44**:618–27.

116. Roberts MM, Alexander FE, Anderson TJ, Forrest APM, Hepburn W, Huggins A, *et al.* The Edinburgh randomised trial of screening for breast cancer: Description of method. *Br J Cancer* 1984;**50**:1–6.

117. Tabar L, Gad A, Holmberg A, Ljunguist U, Fagerberg CJG, Baldetorp L, *et al.* Reduction in mortality from breast cancer after mass screening with mammography. *Lancet* 1985;**i**:829–33.

118. Palli D, Rosselli Del Turco M, Buiatti E, Carli S, Ciatto S, Toscani L, *et al.* A case–control study of the efficacy of a non-randomized breast cancer screening programme in Florence (Italy). *Int J Cancer* 1986;**38**:501–4.

119. Peeters PHM, Verbeek ALM, Hendriks JHCL, van Bon MJH. Screening for breast cancer in Nijmegen. Report of 6 screening rounds, 1975–1986. *Int J Cancer* 1989;**43**:226–30.

120. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd ed. Oxford: Oxford Medical Publications; 1995. p. 64–5.

121. Rothwell PM, Slattery J, Warlow CP. A systematic review of the risks of stroke and death due to endarterectomy for symptomatic carotid stenosis. *Stroke* 1996;**27**:260–5.

122. Young J, Percy CL, Asire AJ, Berg JW, Cusano MM, Gloeckler LA, *et al.* Cancer incidence and mortality in the United States, 1973–77. *Natl Cancer Inst Monogr* 1981;**57**:110, 124, 984–5.

123. Olschewski M, Scheurlen H. Comprehensive cohort study: an alternative to randomized consent design in a breast preservation trial. *Methods Inf Med* 1985; **24**:131–4.

124. Francis T, Korns RF, Voight RB, Boisen M, Hemphill FM, Napier JA, *et al.* An evaluation of the 1954 poliomyelitis vaccine trials. *Am J Public Health* 1955;**45**:1–63.

125. Olschewski M, Schumacher M, Davis KB. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Controlled Clin Trials* 1992; **13**:226–39.

126. Zelen M. Alternatives to classic randomized trials. *Surg Clin North Am* 1981;**61**:1425–32.

127. Schmoor C, Olschewski M, Schumacher M. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Stat Med* 1996;**15**:263–71.

128. Davis K. The comprehensive cohort study: the use of registry data to confirm and extend a randomized trial. *Recent Results Cancer Res* 1988; **11**:138–48.

129. Byar DP. Why data bases should not replace randomized clinical trials. *Biometrics* 1980;**36**: 337–42.

130. Harvey I, West R, Newcombe R. Patient preferences and randomised clinical trials. *BMJ* 1989;**299**:684–5.

131. O'Rourke K. Meta-analysis and the combination of unbiased and biased estimates: is it better to widen or narrow confidence intervals from a RCT when a non-randomised study provides a similar estimate. Paper presented at the annual meeting of the International Society for Technology Assessment in Health Care, Edinburgh; 1999.

132. Brewin CR, Bradley C. Patient preferences and randomised clinical trials. *BMJ* 1989;**299**:313–15.

133. Kassirer JP. Incorporating patients' preferences into medical decisions. *N Engl J Med* 1994;**330**: 1895–6.

134. Silverman WA, Altman DG. Patients' preferences and randomised trials. *Lancet* 1996;**347**:171–4.

135. Torgerson DJ, Klaber-Moffett J, Russell IT. Patient preferences in randomised trials: threat or opportunity. *J Health Services Res Policy* 1996;**1**:194–7.

136. Henshaw RC, Naiji SA, Russell IT, Templeton AA. Comparison of medical abortion with surgical vacuum aspiration: women's preferences and acceptability of treatment. *BMJ* 1993;**307**:714–17.

137. Korn EL, Baumrind S. Randomised clinical trials with clinician-preferred treatment. *Lancet* 1991;**337**: 149–52.

138. Kennedy A, Grant A. Local sealed envelope randomisation in a multicentre trial: a cautionary tale. Paper presented at 5th Annual Cochrane Colloquium, Amsterdam; 1997.

139. Schafer A. The ethics of the randomized clinical trial. *N Engl J Med* 1982;**307**:719–24.

140. Mackilop WJ, Ward GK, O'Sullivan B. The use of expert surrogates to evaluate clinical trials in non-small cell lung cancer. *Br J Cancer* 1986;**54**:661–7.

141. O'Rourke PP, Crone RK, Vacanti, Ware JH, Lillehei CW, Parad RB, *et al.* Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn: a prospective randomised study. *Pediatrics* 1989;**84**:957–63.

142. Lilford R for the Fetal Compromise Group. Formal measurement of clinical uncertainty: prelude to a trial in perinatal medicine. *BMJ* 1994;**308**:111–12.

143. Zelen M. A new design for randomized clinical trials. *N Engl J Med* 1979;**300**:1242–5.

144. Altman DG, Whitehead J, Parmar MKB, Stenning SP, Fayers PM, Machin D. Randomised consent designs in cancer clinical trials. *Eur J Cancer* 1995; **31**:1934–44.

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

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

147. Korvick JA, Peacock JE, Muder RR, Wheller RR, Yu VL. Addition of rifampin to combination antibiotic therapy for *Pseudomonas aeruginosa* bacteremia: prospective trial using the Zelen protocol. *Antimicrob Agents Chemother* 1992;**36**:620–5.

148. Curran WJ. Reasonableness and randomisation in clinical trials: fundamental law and government regulation. *N Engl J Med* 1979;**300**:1273–4.

149. Horwitz RI, Feinstein AR. Advantages and drawbacks of the Zelen design for randomized clinical trials. *J Clin Pharmacol* 1980;20:425–7.

150. Anbar D. The relative efficiency of Zelen's prerandomization design for clinical trials. *Biometrics* 1983;**39**:711–18.

151. Ellenberg SS. Randomisation in comparative clinical trials. *N Engl J Med* 1984;**310**:1404–8.

152. Rosner F. The ethics of randomised clinical trials. *Am J Med* 1987;**82**:283–90.

153. Matts J, McHugh R. Randomization and efficiency in Zelen's single-consent design. *Biometrics* 1987; **43**:885–94.

154. Perrone F, De Placido S, Giusti C, Gallo C. Looking for consent in randomised clinical trials: a randomised trial with surrogate patients. *Epidemiol Prev* 1995;**19**:282–90.

155. Chang RW, Falconer J, Stulberg SD, Arnold WJ, Dyer AR. Prerandomization: an alternative to classic randomization. The effects on recruitment in a controlled trial of arthroscopy for osteo- arthrosis of the knee. *J Bone Joint Surg* 1990;**72**:1451–5.

156. Santen RJ, Worgul TJ, Somojlik E, Interrante A, Boucher AE, Lipton A, *et al.* A randomized controlled trial comparing surgical adrenalectomy with aminoglutethimide plus hydrocortisone in women with advanced breast cancer. *N Engl J Med* 1981;**305**:545–51.

157. Rosenberger WF, Lachin JM. The use of response-adaptive designs in clinical trials. *Controlled Clin Trials* 1993;**14**:471–84.

158. Zelen M. Play the winner rule and the controlled clinical trial. *J Am Stat Assoc* 1969;**64**:131–46.

159. Cornfield J, Halperin M, Greenhouse SW. An adaptive procedure for sequential trials. *J Am Stat Assoc* 1969;**64**:759–70.

160. Byar D, Simon RM, Friedwald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, *et al.* Randomized clinical trials. Perspectives on some recent ideas. *N Engl J Med* 1976;**295**:74–80.

161. Simon R. Adaptive treatment assignment methods and clinical trials. *Biometrika* 1977;**33**:743–9.

162. Wei LJ, Durham. The randomized play-the-winner rule in medical trials. *J Am Stat Assoc* 1978;**73**:840–3.

163. Weinstein MC. Allocation of subjects in medical experiments. *N Engl J Med* 1981;**291**:1278–85.

164. Simon R. A decade of progress in statistical methodology. *Stat Med* 1991;**10**:1789–1817.

165. Berry DA, Eick SG. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Stat Med* 1995;**14**:231–46.

166. Robbins H. Some aspects of the sequential design of experiments. *Bull Am Math Soc* 1952;**58**:527–35.

167. Efron B. Estimating the error rate of a prediction rule: improvements on cross-validation. *J Am Stat Assoc* 1983;**78**:316–31.

168. Rosenberger WF. Asymptotic inference with response adaptive treatment allocation designs. *Ann Stat* 1993;**21**:2098–107.

169. Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal circulation in neonatal respiratory failure: a prospective randomised study. *Pediatrics* 1985; **76**:479–87.

170. Begg CB. On inferences from Wei's biased coin design for clinical trials. *Biometrika* 1990;**77**:467–84.

171. Amery W, Dony J. A clinical trial design avoiding undue placebo treatment. *J Clin Pharmacol* 1975; **15**:674–9.

172. Kopec J, Abrahamowicz M, Esdaile JM. Randomized discontinuation trials: utility and efficiency. *J Clin Epidemiol* 1993;**46**:959–71.

173. The Canadian Hydroxychloroquine Study Group. A randomized study of the effect of withdrawing hydroxychloroquine sulfate in systemic lupus erythematosus. *N Engl J Med* 1991;**324**:150–4.

174. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N Engl J Med* 1991;**325**:293–302.

175. Buring JE, Hennekens CE. Cost and efficiency in clinical trials: the US Physicians' Health Study. *Stat Med* 1990;**9**:29–33.

176. Davis CE. Prerandomization compliance screening: a statistician's view. In: Shumaker SA, Schram EB, Okun JK, editors. The handbook of health behaviour change. New York: Springer; 1991. p. 342–7.

177. Davis CE, Applegate WB, Gordon DJ, Curtis RC, McCormick M. An empirical evaluation of the placebo run-in. *Controlled Clin Trials* 1995;**16**:41–50.

178. Brittain E, Wittes J. The run-in period in clinical trials: the effect of misclassification on efficacy. *Controlled Clin Trials* 1990;**11**:327–38.

179. Schechtman KB, Gordon ME. A comprehensive algorithm for determining whether a run-in strategy will be a cost-effective design modification in a randomized clinical trial. *Stat Med* 1993;**12**: 111–28.

180. LaRosa JC, Applegate WB, Crouse JR, Hunninghake DB, Grimm RH, Knopp RH, *et al.* Cholesterol lowering in the elderly: results of the Cholesterol Reduction in Seniors Program (CRISP) pilot study. *Arch Intern Med* 1994;**154**:529–39.

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

182. Hogel J, Walach H, Gaus W. Change-to-open label design. Proposal and discussion of a new design for clinical parallel-group double-masked trials. *Arzneimittelforschung* 1994;**14**:97–9.

183. Walach H. Das 'change-to-open-label' design: Anpassung und Veränderung des parallelgruppen-Blinddesigns für die klinische Forschung. *Z Klin Psychol* 1994;**23**:213–18.

184. Billingham LJ, Abrams K, Jones DR. Quality of life assessment and survival data. *Health Technol Assess* 1999;**3**(Pt 10).

185. Elwood M. Critical appraisal of epidemiological studies and clinical trials. 2nd ed. Oxford: Oxford University Press; 1998.

186. Shadish WR, Ragsdale K. Random versus nonrandom assignment in controlled experiments: do you get the same answer? *J Consult Clin Psychol* 1996;**64**:1290–305.

187. Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, *et al.* for the Heart and Estrogen/ Progestin Replacement Study (HERS) Research Group. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA* 1998;**280**:605–13.

188. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med* 1991;**20**:47–63.

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

190. Concato J, Shah N, Horwitz RI. Randomized controlled trials, observational studies and hierarchy of research designs. *N Engl J Med* 2000;**342**:1878–86.

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

192. Black N. High-quality clinical databases: breaking down barriers. *Lancet* 1999;**353**:1205–6.

Appendix 1

Commissioning brief

Area 93/45: Comparing the use of randomised controlled trial (RCT) designs with quasi-experimental and/or observational studies for assessing the effectiveness of interventions and comparing the quality of care (including methods for improving and assessing the adequacy of adjustment for case mix).

The RCT is seen as the gold standard design for assessing the efficacy of health technologies because it has increased internal validity (i.e. it reduces confounding and bias) and results can more reliably be used to attribute causality. However, sometimes it is not so easy to have randomised control groups or to find appropriate controls. In these circumstances a range of quasi-experimental and observational designs have been developed. There are advocates for the increased use of observational studies (often making use of more easily available, routine or administrative databases) in health technology assessment and in monitoring variations in quality of care between providers. However, others have argued that, because of the problems of bias and confounding, observational data are of little use in health technology assessment. Controversy has focused on the validity of constructing hospital mortality 'league tables'.

The debate over the relative merits of RCTs, quasi-experimental and observational studies has become very polarised and would benefit from objective investigation.

Research is needed to:

- explore the validity of using quasi-experimental and/or observational techniques where RCTs are difficult or impossible
- identify barriers to using observational data
- assess the validity of using observational studies for effectiveness research

assess the degree to which better measures of case mix (patient severity, diagnosis, age and co-morbidity) can be developed generally and for particular conditions and explore the potential and limitations of using severity adjusted observational data in health technology assessment and in comparing the quality of providers.

Appendix 2

Electronic search strategies

MEDLINE

Cochrane RCT search strategy[41]

1. randomized controlled trial.pt
2. randomized controlled trials/
3. random allocation/
4. double-blind method/
5. single-blind method/
6. 1 or 2 or 3 or 4 or 5
7. limit 6 to human
8. limit 6 to animal
9. 7 and 8
10. 8 not 9
11. 6 not 10
12. clinical trial.pt.
13. exp clinical trials/
14. clin$ trial$.ti.
15. clin$ trial$.ab.
16. (singl$ or doubl$ or trebl$ or tripl$) adj (blind$ or mask$)
17. placebos/
18. placebo$.ti.
19. placebo$.ab.
20. random$.ti.
21. random$.ab.
22. exp research design/
23. or/12–22
24. limit 23 to human
25. limit 23 to animal
26. 24 and 25
27. 25 not 26
28. 23 not 27
29. 28 not 11
30. comparative study/
31. exp evaluation studies/
32. follow-up studies/
33. prospective studies/
34. (control$ or prospectiv$ or volunteer$).ti,ab,sh.
35. 30 or 31 or 32 or 33 or 34
36. limit 35 to human
37. limit 35 to animal
38. 36 and 37
39. 37 not 38
40. 35 not 39
41. 40 not (11 or 29)

Observational search strategy (devised by RRM)

42. "cohort studies"/
43. "longitudinal studies"/
44. "prospective studies"/
45. "follow-up studies"/
46. "cross-sectional studies"/
47. "retrospective studies"/
48. "case-control studies"/
49. cohort studies.tw.
50. longitudinal studies.tw.
51. prospective studies.tw.
52. follow up studies.tw.
53. cross sectional studies.tw.
54. retrospective studies.tw.
55. case control studies.tw.
56. or/42–55

MeSH and text words used for MSBC search

57. exp "mammography"/
58. exp "breast neoplasms"/
59. exp "mass screening"/
60. mammography. tw.
61. breast neoplasms. tw.
62. 58 or 61
63. 62 and 59
64. 57 or 60
65. 62 and 64
66. 63 or 64
67. limit 66 to human

MeSH and text words used for folic acid search

57. exp "folic acid"/
58. folic acid. tw.
59. 57 or 58
60. exp "neural tube defects. tw.
61. neural tube defects. tw.
62. 60 or 61
63. 59 and 62

Search strategy limits

- Human.
- Review articles.
- Abstracts.
- Publication types:
  - abstract
  - classical article
  - clinical conference
  - clinical trial
  - clinical trial, Phase I

- clinical trial, Phase II
- clinical trial, Phase III
- clinical trial, Phase IV
- comment
- controlled clinical trial
- editorial
- guideline
- historical article
- journal article
- letter
- meta-analysis
- monograph
- multicentre study
- randomised controlled trial
- retracted publication
- review
- review of literature
- review of reported cases
- review, academic
- review, multicase
- review, tutorial.

EMBASE

Folic acid

1. neural tube defect*
2. folic acid
3. 1 + 2
4. supplementation
5. 3 + 4

Breast cancer

1. mammography
2. breast neoplasm*
3. breast cancer
4. mass screening
5. screening
6. mortality
7. 2, 3
8. 1, 7
9. 4, 5
10. 6 + 8 + 9

Search strategy limits

- 1980–1997.
- All languages selected.
- Document types selected:
  - article
  - conference
  - review
  - letter
  - editorial
  - book/monogram
  - abstract
  - preliminary communication
  - book
  - conference paper
  - report.
- Unselected:
  - short survey
  - erratum
  - note.

Appendix 3

Instructions for assessing the quality of studies for strategy 1

Eligibility criteria

Were the same eligibility criteria applied for the patients included in RCT and QEO study elements of the study? For some studies, the QEO study element may constitute only one arm (either intervention or control), which is compared with the alternative treatment group in the RCT; if patients in the single QEO study arm are recruited using the same eligibility criteria as the RCT, answer 'yes'.

(The answer will almost certainly be 'no' for 'review' studies, even if this is difficult to judge from the paper under consideration.)

Contemporaneity

Were the recruitment periods for RCT and QEO study elements of the study contemporaneous? This question should be answered 'no', even if the period of recruitment for one design element falls entirely within a longer recruitment period for the other design element. If the recruitment periods for RCT and QEO study elements are different, but this is taken account of by including date of recruitment as a potential confounding factor in the analysis, answer 'yes'.

Blinding of outcome assessment

Was the assessment of outcome blinded? Answer 'yes' for outcome measures which could not be biased (e.g. all-cause mortality). Answer 'no' if only one of the study elements used blinded outcome assessment.

(This question does **not** refer to blinding of patients and researchers with respect to treatment received.)

Confounding – severity

Was the analysis of the QEO study element of the study (and the RCT element, if differences in prognostic factors were documented as the result of randomisation) adjusted for severity of disease?

Confounding – co-morbidity

Was the analysis of the QEO study element of the study (and the RCT element, if differences in co-morbidity were documented as the result of randomisation) adjusted for co-morbidity?

Confounding – other prognostic factors

Was the analysis of the QEO study element of the study (and the RCT element, if differences in prognostic factors were documented as the result of randomisation) adjusted for other prognostic factors (e.g. age, other risk factors for the outcome of interest)?

Appendix 4

The instrument for assessing the quality of a study in strategy 2

Questions on the reporting, the external validity, the internal validity (bias and confounding) of the study are listed in *Tables 29* to *32*, and some additional comments on the instrument are given in *Table 33*.

TABLE 29 *Reporting of the study*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
1	**Is the hypothesis/aim/objective of the study clearly described?** This question refers to a clear statement of the objective, i.e. to measure the effectiveness of *x* in population y with respect to z, even if *x*, *y* and *z* are not clearly described (see questions 2, 3 and 4) (a) Yes (1) (b) No (0)	78	–0.007	Reporting
2	**Are the main outcomes to be measured clearly described in the Introduction or Methods section?** If the main outcomes are first mentioned in the Results section, the question should be answered 'no'. In case–control studies the case definition should be considered the outcome (a) Yes (1) (b) No (0)	81	–0.627	Reporting
3a	**Are the characteristics of the patients included in the study clearly described in the Introduction or Methods section?** Is a statement describing the population given? A 'yes' answer does not require a detailed table of characteristics, but only a simple text description of the study population (a) Yes (b) No	72	–0.007	–
3b	**Are the inclusion and exclusion criteria clearly stated in the Introduction or Methods section?** If the inclusion and exclusion criteria are implicitly given in the description of the characteristics of the population (question 3a), answer 'yes' (a) Yes (1) (b) No (0)	79	0.118	Reporting

Continued

TABLE 29 contd *Reporting of the study*[*]

No.	Question					Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
4a	**Are the interventions of interest described in detail in the Introduction or Methods section?** Treatments (e.g. intervention and control) that are to be compared should be clearly described					81	–0.091	Reporting
	(a) Yes (1) (b) No (0)							
4b	**Are the different interventions appropriate?** For example, the answer should be 'no' if a placebo was used as the control when it is current practice to use some standard treatment					83	–0.049	–
	(a) Yes (b) No							
5	**Are the distributions of the principal confounders clearly described for each group of patients to be compared?** Are measures of central tendency (mean or median) and dispersion (standard deviation or interquartile range) reported for the confounders in both the intervention and control groups?							
5a	**Intervention group**							
		Yes	No	Not reported				
	Confounder 1	a	b	c	(a = 1)	44	0.392[‡]	Reporting
	Confounder 2	a	b	c	(b = 0)	62	0.536[‡]	Reporting
	Confounder 3	a	b	c	(c = 0)	43	0.475[‡]	Reporting
	Confounder 4	a	b	c		54	0.533[‡]	Reporting
5b	**Control group**							
		Yes	No	Not reported				
	Confounder 1	a	b	c	(a = 1)	50	0.429[‡]	Reporting
	Confounder 2	a	b	c	(b = 0)	40	0.536[‡]	Reporting
	Confounder 3	a	b	c	(c = 0)	44	0.475[‡]	Reporting
	Confounder 4	a	b	c		54	0.283[‡]	Reporting
6a	**Was a primary outcome identified?**					93	0.389[‡]	Reporting
	(a) Yes (1) (b) No (0)							

Continued

TABLE 29 contd *Reporting of the study*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
6b	**Was a power calculation reported for the primary outcome?** (a) Yes (1) (b) No (0) (c) Can't tell (0)	85	0.667[‡]	Reporting
6c	**Were other outcome measures (secondary outcomes) described?** If the paper refers to other papers (associated with the same study) but not reporting different outcome measures, answer 'yes' (a) Yes (b) No	75	0.258[‡]	
6d	**Are the main findings of the study clearly described?** Simple outcome data (including denominators and numerators) should be reported for all major findings so that the reader can check the major analyses and conclusions. (This question does not cover statistical tests, which are considered below) (a) Yes (2) (b) No (0)	78	0.071	Reporting
7	**Are estimates of the random variability of the main outcomes clearly described for each group of patients to be compared?** In non-parametric data the interquartile range should be reported. In parametric data the standard error and/or standard deviation, or confidence intervals should be reported. If the distribution of the data (parametric or not) is not described and can			
	Intervention group (a) Yes (0.5) (b) No (0) (c) Can't tell (0)	47	–0.085	Reporting
	Control group (a) Yes (0.5) (b) No (0) (c) Can't tell (0)	47	–0.044	Reporting

Continued

TABLE 29 contd *Reporting of the study*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
8	**Have all reported adverse events that may be a consequence of the intervention been reported?** This should be answered 'yes' if the study demonstrates that there was a comprehensive attempt to measure adverse events within the context of the study duration. A list of possible adverse events is provided (a) Yes (b) No	81	0.280[‡]	
9	**Have the numbers of patients lost to follow-up been described?** (a) Yes (1) (b) No (0)	74	0.253[‡]	Reporting
10	**Have 95% CIs and/or actual probability values (i.e. 0.035 rather than < 0.05) been reported for the main outcomes, except where the probability value is less than 0.001?** (a) Both CI and *p* value (0.67) (b) Either CI or *p* value (0.33) (c) Neither (0)	80	0.467[‡]	Reporting
11	**Have the authors considered whether the effects of patients preferences and expectations of treatment may affect the outcome** (a) Yes (b) No	78	–0.125	–

[*] *The points allocated to each question are given in parentheses*
[†] *Rater agreement by question for all papers (see text for calculation)*
[‡] *Indicates that the κ value is significant at* $p < 0.05$
[§] *EV, external validity; IVB, internal validity/bias; IVC, internal validity/confounding. If an item is not indicated for analysis or reporting, it was not included in the analyses reported in the text*

TABLE 30 *External validity*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
12a	**What proportion of subjects who were approached were *ineligible* to participate?**[¶] 'Ineligible' are those who do not meet the inclusion criteria. Answer 0–5% for case–control studies where the cases and controls would not be approached if they were ineligible What is the actual % of ineligibles reported in the paper? (1) **or** What numbers (*x*/*n*) were reported as being ineligible? (1) **or** If this is not reported in the paper, answer 'N/A' (0)	63	0.209[‡]	Reporting
12b	**What proportion of subjects who were eligible *refused* to participate?**[¶] For case–control studies answer < 5%. The refusal to participate in case–control studies gives rise to problems of selection bias, not external validity/generalisability What is the actual % of refusal reported in the paper? (1) **or** What numbers (*x*/*n*) were reported for refusal? (1) **or** If this is not reported in the paper, answer 'N/A' (0)	63	0.250[‡]	Reporting
13	**Were the staff, places and facilities where the patients were treated representative of the treatment the majority of patients receive? Please rate on the 5-point scale given** For the question to be answered 'representative' the reader should be confident that the findings of the study would apply in a range of different settings (e.g. teaching hospital and DGH). The reader should exercise his/her judgement, taking into account (a) Representative (1) (b) ↓ (0.75) (c) ↓ (0.5) (d) ↓ (0.25) (e) Not representative (0)	42	0.085	EV

[*] *The points allocated to each question are given in parentheses*
[†] *Rater agreement by question for all papers (see text for calculation)*
[‡] *Indicates that the κ value is significant at* $p < 0.05$
[§] *EV, external validity; IVB, internal validity/bias; IVC, internal validity/confounding. If an item is not indicated for analysis or reporting, it was not included in the analyses reported in the text*
[¶] *Questions 12a and 12b were used for reporting only, therefore the percentage agreement and κ statistics refer only to whether or not the percentage of patients refusing was reported*

TABLE 31 *Internal validity – bias*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
14	**Was an attempt made to blind patients to the intervention they received?** For studies where the patients would have no way of knowing which intervention they received this should be answered 'yes' (a) Yes (1) (b) No (0) (c) Can't tell (1) (d) Not applicable (1)	63	0.085	IVB
15	**Was an attempt made to blind those measuring the main outcomes of the intervention?** If the outcome was determined from a routine data source (e.g. all-cause mortality), answer 'yes', as there is no opportunity for bias to occur. (Note: Cause-specific mortality might be biased because the assignment of cause could be subject to prejudice) (a) Yes (1) (b) No (0) (c) Can't tell (0)	35	0.026[‡]	IVB
16a	**Was an attempt made to blind those *performing* the intervention?** (a) Yes (b) No (c) Can't tell (d) Not possible	60	0.000[‡]	–
16b	**If the answer to 16a was 'yes', did the study attempt to assess the level of success in blinding?** Answer 'not applicable' if the answer to 16a is 'no', 'can't tell' or 'not possible' (a) Yes (b) No (c) Not applicable	92	0.824[‡]	–

Continued

TABLE 31 contd *Internal validity – bias*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
17a	**If any of the results of the study were based on 'data dredging', was this made clear?** Any analyses that had not been planned at the outset of the study should be clearly indicated. If no retrospective or unplanned subgroup analyses were reported, then answer 'yes' (a) Yes (b) No	54	–0.072	–
17b	**Were any analyses that were planned at the outset subject to bias?** An example would be planned comparisons across subgroups which were not stratified during randomisation. Bias may occur if the number of subjects in the smallest subgroup involved in a comparison is < 50 (a) Yes (b) No	61	0.214[‡]	–
18a	**For all large studies, was the 'average' duration of follow-up between the groups the same?** To be considered a 'large study', the smallest arm must have no fewer than 100 subjects (a) Yes (1) (b) No (0) (c) Can't tell (0)	79	0.448[‡]	IVB
18b	**If 'yes', what was the analysis based on?** (a) Person time (person-years/survival analysis) (b) Risk/odds ratio (c) Not applicable (question 18a = yes)	59	0.483[‡]	–

Continued

TABLE 31 contd *Internal validity – bias*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
18c	**If 'no', or small size, did the analysis take account of the differences in the duration of follow-up?** In case–control studies this may have been done using 'calendar time' (a) Yes (1) (b) No (0) (c) Can't tell (0) (d) Not applicable (question 18a = yes) (1)	76	0.084	IVB
19	**Were the statistical tests used to assess the main outcomes legitimate?** The statistical tests used must be appropriate to the data. When sample sizes are small, 'exact methods' should be used. Where little statistical analysis has been undertaken but where there is no evidence of bias, the question should be answered 'yes' (a) Yes (1) (b) No (0) (c) Can't tell (0)	66	0.021[‡]	IVB
20	**What proportion of patients in each group completed the allocated treatment regimen?** This question is about the proportion of subjects who 'crossed over' with respect to the treatment they received or who did not comply with their allocated treatment, rather than loss to follow-up			
	Intervention group	67	0.330[‡]	IVB
	What % completed the allocated treatment regimen? (2: > 95%) **or** What number (*x*/*n*) completed the allocated treatment regimen? (1: 85–95%) **or** Answer 'can't tell' if it is impossible to tell (0: < 85%)			
	Control group	57	0.146[‡]	IVB
	What % completed the allocated treatment regimen? (2: > 95%) **or** What number (*x*/*n*) completed the allocated treatment regimen? (1: 85–95%) **or** Answer 'can't tell' if it is impossible to tell (0: < 85%)			

Continued

TABLE 31 contd *Internal validity – bias*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
21	**Were the main outcome measures used valid and reliable?** For studies where the outcome measures are clearly described and likely to be both valid and reliable (e.g. 'dead or alive'), the question should be answered 'yes'. For studies which refer to other work or that demonstrates the outcome measures are accurate, the question should be answered as 'yes'. For case–control studies, treat the case definition as the outcome measure (a) Yes (1) (b) No (0) (c) Can't tell (0)	56	–0.086	IVB
22	**Were the patients in the different intervention groups (trials and cohort studies) or the cases and controls (case–control studies) recruited from the same source population?** For example, patients for all comparison groups should be selected from the same hospitals. The question should be answered 'can't tell' for cohort studies where there is no information concerning the source of the patients included in the study. It should be considered that for some case–control studies that the use of the same local population for cases and controls may be inappropriate. A subjective judgement has to be made in the instance of case–control studies, as on occasion the use of the same local population for controls as well as cases may be inappropriate (a) Yes (b) No (c) Can't tell	91	0.533[‡]	–
23	**Were the patients in the different treatment group (trials and cohort studies) or the cases and controls (case–control studies) recruited over the same period of time?** Answer 'yes' for all RCTs. For a study which does not specify the time period over which patients were recruited, the question should be answered 'can't tell' (a) Yes (b) No (c) Can't tell	79	0.141[‡]	–

Continued

TABLE 31 contd *Internal validity – bias*[*]

[*] *The points allocated to each question are given in parentheses*
[†] *Rater agreement by question for all papers (see text for calculation)*
[‡] *Indicates that the* κ *value is significant at* $p < 0.05$
[§] *EV, external validity; IVB, internal validity/bias; IVC, internal validity/confounding. If an item is not indicated for analysis or reporting, it was not included in the analyses reported in the text*

TABLE 32 *Internal validity – confounding*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
24	**Were patients randomised to intervention groups?** Studies should be classified by method of randomisation. Group randomised study should be ticked if, for example, units of healthcare delivery such as general practices have been randomised rather than individual patients (a) Truly random **and** concealed (1) (b) Truly random but **not** concealed (0.5) (c) Group randomised study (1) (d) Not random/cohort/case–control (0)	76	0.507[‡]	IVC
25	**Was there adequate adjustment for the effects of confounding in the analysis from which the main findings were drawn?** Refer to the list of confounders provided. Please indicate the resolution with which adjustment was carried out for each confounder using the 4-point scale (a–d); answer 'none' (e) if the analysis did not adjust for the confounder (For example, is age adjusted for using only 2 'strata' (< 50 years or > 50 years) in 10 or 5 year age strata of as a continuous variable? Is alcohol consumption categorised as low or high, or broken down into units per day?) If case–control studies were matched, the variable on which it was matched should be treated as a confounder. (See also question 18 regarding the length of follow-up) Resolution of adjustment: High → → Low None (a) (b) (c) (d) (e)			
	Confounder 1 (max. 0.5)	53	0.296[‡]	IVC
	Confounder 2 (max. 0.5)	75	0.396[‡]	IVC
	Confounder 3 (max. 0.5)	76	0.319[‡]	IVC
	Confounder 4 (max. 0.5)	77	0.273[‡]	IVC
	(RCTs = 2)			

Continued

TABLE 32 contd *Internal validity – confounding*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
26	**Were the main confounding variables used valid and reliable?** For studies where the confounding variables are clearly described and likely to be both valid and reliable, the question should be answered 'yes'. For studies which refer to other work or that demonstrates the confounding variables are accurate, the question should be answered as 'yes' (a) Yes (b) No	56	0.183[‡]	–
27	**Are the main conclusions of the study based on an intention-to-treat analysis rather than on an analysis of treatments actually received?** For studies in which all patients received the treatment to which they were allocated, answer 'yes' rather than 'not applicable' (a) Yes (1) (b) No (0) (c) Not applicable (case–control studies and retrospective cohorts: = 1)	63	0.393[‡]	IVC
28	**How many subjects were lost to follow-up?** If the number of patients lost to follow-up is not reported, the question should be answered as 'can't tell'			
28a	**Intervention group** What is the percentage of patients lost to follow-up? (2: < 5%) **or** What is the number (*x*/*n*) of patients lost to follow-up? (1: 5–15%) **or** Can't tell (0: > 15%)	55	0.162[‡]	IVC
28b	**Control group** What is the percentage of patients lost to follow-up? (2: < 5%) **or** What is the number (*x*/*n*) of patients lost to follow-up? (1: 5–15%) **or** Can't tell (0: > 15%)	58	0.216[‡]	IVC

Continued

TABLE 32 contd *Internal validity – confounding*[*]

No.	Question	Agreement (%)[†]	κ (unweighted)	Reporting, EV, IVB, IVC[§]
29	**If substantial losses to follow-up occurred (i.e. > 5%) was a comparison made of the characteristics of those lost to follow-up and those followed up?** (a) Yes (1) (b) No (0) (c) Not applicable (1)	77	0.218[‡]	Reporting

[*] *The points allocated to each question are given in parentheses*
[†] *Rater agreement by question for all papers (see text for calculation)*
[‡] *Indicates that the κ value is significant at* $p < 0.05$
[§] *EV, external validity; IVB, internal validity/bias; IVC, internal validity/confounding. If an item is not indicated for analysis or reporting, it was not included in the analyses reported in the text*

TABLE 33 *Additional comments about the instrument*

Question	Included in quality score?	Additional comments
1	✓	
2	✓	
3a	✗	This question was not included because it was considered too ambiguous in its present format
3b	✓	
4a	✓	We suggest that 'interventions' be replaced by 'interventions/exposures' in order to clarify how the question should be answered for case–control studies
4b	✗	This question was not considered to be relevant to the two health technologies being considered for objective 2. The question may be relevant for other health technologies
5a, 5b	✓	The question needs to be modified to clarify when 'no' and 'not reported' should be used. The question should be answered 'no' when confounders are reported, but no measures of central tendency, etc., are given. 'Not reported' should be used when the confounders are not reported The question assumes that the confounders are continuous variables, and therefore needs to be modified to indicate what information should be reported for categorical confounding variables; we suggest that reporting the frequency/proportion of observations in each category should be sufficient for a confounder to scored as 'yes' Additionally, the question needs to take account of possible confounding factors on which cases are matched in a matched case–control study. We suggest that variables used for matching should be awarded the same score as for 'yes'
6a	✓	The question requires an additional statement to the effect that implicit primary outcomes should be scored as 'yes'
6b	✓	On the assumption that a power calculation will only be reported for a primary outcome, we suggest that the order of questions 6a and 6b should be reversed. Question 6a would become 'Was a power calculation reported?' (with possible answers of 'yes', 'no' and 'can't tell'). Question 6b would become "If no, was a primary outcome identified? Where a primary outcome is implicit, answer 'yes' " (with possible answers of 'yes' or 'no')
6c	✗	This was not considered to be a relevant 'quality question' for the review. There are no grounds for regarding a study as poor merely because it does not include secondary outcomes
6d	✓	This question was considered very important from the point of view of reporting, hence the score of 2
7a, 7b	✓	The question does not explain how binary outcomes should be regarded. We suggest adding the following statement: 'For binary outcomes confidence intervals should be reported for each group' This is the first of several questions which refer to 'intervention' and 'control' groups. It is important that all such questions make explicit how these terms should be interpreted for case–control studies. In this case it only makes sense to interpret intervention and control groups as 'case' and 'control' groups (since the relative numbers of cases and controls, i.e. the 'main outcome', among exposed and unexposed groups is dependent on the ratio of cases to controls). We therefore suggest adding the following statement: 'For case–control studies, confidence intervals should be reported for the proportions of cases and controls who are exposed'

Continued

TABLE 33 contd *Additional comments about the instrument*

Question	Included in quality score?	Additional comments
8	✗	This question was not considered relevant to the review. However, it might be very important for HTs for which there was a much more salient trade-off between effectiveness and harm from the HT This question is also difficult to answer for case–control studies; because they are retrospective, information is unlikely to be collected about adverse events, and serious adverse events might even preclude a patient being included (unless the adverse event was the outcome of interest to the study). We suggest that the question should be modified to indicate that case–control studies should be scored as 'no'
9	✓	This question should be modified to indicate that 'loss to follow-up' should be interpreted as 'refusal to take part by eligible subjects' for case–control and retrospective cohort studies (see questions 28 and 29)
10	✓	
11	✗	This question needs to be modified to acknowledge that it is only applicable to randomised or non-randomised trials, where a patient has been allocated to treatment, and where the patient is aware of the different treatments being compared. There does not appear to be an equivalent issue with observational studies
12a	✓	
12b	✓	
13	✓	
14	✓	The question needs to be clarified for observational studies. We suggest adding the statement: "Where study participants were unaware of the treatment comparison/exposure of interest, answer 'yes'". For the review, all observational studies were given a score of 1, because the question did not indicate clearly whether such studies should be scored as 'can't tell' or 'not applicable'
15	✓	Clarification is required for case–control studies: 'For case–control studies, this question should be taken as referring to the measurement of exposure'
16a, 16b	✗	Consideration needs to be given to the scoring of question 16b. Scoring 'not applicable' as 'no' doubly penalises studies in which blinding may not have been possible. However, scoring 'not applicable' as 'yes' means that a blinded intervention where the success of blinding was not evaluated scores no better than an intervention which was not blinded at all. Clarification is also required for case–control and retrospective cohort studies, for which this question is not applicable These questions were excluded from our analysis as they were not considered relevant to the two HTs chosen for objective 2
	✗	TAS suggested that an additional question should be inserted here regarding the validity of the measurement of exposure, especially when measurements were self-reported or retrospective
17a	✗	This question proved difficult to answer, since papers frequently failed to state explicitly the analyses which were planned at the outset. The question should be modified: "Answer 'yes' only if there is an explicit statement about the planned analyses and if no other analyses were carried out"
17b	✗	The question should be modified to indicate that all observational and non-RCT designs should be scored 'yes'
	✓	'Yes' and 'not applicable' should be combined for scoring purposes. As the question currently stands it is unclear how to proceed if the study is not large

Continued

TABLE 33 contd *Additional comments about the instrument*

Question	Included in quality score?	Additional comments
18a 18b 18c	✓ ✗ ✓	The different parts of this question are poorly structured; the worst aspect is that it is unclear how to proceed if the study is not large. We suggest combining the different parts into a single question: 'Was the duration of follow-up the same for all groups being compared?'. Three responses should be used: (a) 'yes', (b) 'no but accounted for' and (c) 'no and not accounted for' For the review, scores were assigned as follows: large studies with the same average duration of follow-up for all groups, 2; large or small studies where differences in the average duration of follow-up were accounted for, 1; large or small studies where differences in the duration of follow-up were not accounted for, 0
19	✓	
20a 20b	✓ ✓	Clarification is needed for case–control studies. Since their retrospective nature means that, by definition, both the exposure and outcome status of participants cannot change, we suggest adding the statement that: 'All case–control studies should be scored as 100%' This question caused problems when scoring some of the breast cancer studies, since there was no way to take account of screening attendance rates in multiple screening rounds
21	✓	
22	✗	This question was extremely difficult to answer for case–control studies because the question of what constitutes an appropriate control group is always controversial for case–control studies
23	✗	Clarification is needed for case–control studies, where the important issue is that both cases and controls were potentially at risk of exposure over the same period of time
24	✓	Options (a) and (c) were combined for scoring purposes. However, one might not wish to score group randomised studies as highly as individually randomised studies, particularly if the number of groups randomised was small
25	✓	This question gave rise to problems for matched case–control studies. Such studies should now be analysed by conditional logistic regression, and matching variables should be included in the regression model. However, many older studies often carried out simple 'paired' analyses (e.g. McNemar tests). The question needs to be modified to give some credit for adjusting for confounding in these cases: 'When case–control studies match for one or more confounding variables, score the precision of the matching as a proxy for the resolution of adjustment'. All large RCTs (> 50 per treatment group) were credited with maximum points for this question
26	✗	The question needs a third response category for papers where no confounding variables were adjusted for: '(c) No adjustment for confounding variables'. This question was not included in the quality score for the review. If it were to be used, credit should automatically be given to all large RCTs
27	✓	This question poses difficulties with respect to the scoring for case–control and cohort studies. By virtue of the way in which 'exposed' and 'unexposed' groups are defined, the issue of intention-to-treat does not arise for these observational studies. We therefore scored these studies as 'yes', i.e. as if they had been analysed on an intention-to-treat basis. However, the homogeneity of 'exposed' and 'unexposed' groups might be considered to be artificial, if RCTs and non-RCTs of the same HT were to experience substantial non-compliance or crossing over between treatment groups. In such cases, the observational studies would be expected to give different (more extreme) effect size estimates than RCTs and non-RCTs

Continued

TABLE 33 contd *Additional comments about the instrument*

Question	Included in quality score?	Additional comments
28a, 28b	✓	Clarification is needed for case–control and retrospective cohort studies. Eligible subjects who refuse to participate should be regarded as having been lost to follow-up (see question 12b). We suggest adding the statement: "For retrospective studies, interpret 'lost to follow-up' as meaning refusal by eligible cases/intervention subjects or controls to take part"
29	✓	Clarification is needed for case–control and retrospective cohort studies. We suggest adding the statement: "For retrospective studies, interpret 'lost to follow-up' in the same way as for question 28"

Appendix 5

Health technologies initially considered for strategy 2

Some of these health technologies were rejected at an early stage. Therefore population, intervention and outcome were not necessarily well defined in all cases. When this list was drawn up, the project steering group were also considering reviewing one or more areas in which RCTs were extremely unlikely to be carried out.

1. Treatments for chronic low back pain to reduce pain.
2. Postoperative tamoxifen for women with breast cancer to improve survival.
3. Treatments for benign prostatic hyperplasia.
4. Laparoscopic versus conventional cholocystectomy.*
5. Faecal occult blood screening for colorectal cancer.
6. Digital rectal examination for colorectal cancer.
7. Mammographic population screening to reduce mortality from breast cancer.
8. Comparisons of angioplasty, coronary artery bypass grafting and medical treatment for coronary artery disease.*
9. Cervical smear population screening to reduce mortality from cervical cancer.
10. Laparoscopic versus conventional inguinal hernia repair.*
11. Rehabilitation to improve physical functioning and quality of life in stroke patients.
12. Comparisons of alternative treatment regimens for perinatal care.
13. Comparisons of alternative drug treatment regimens for hypertension.
14. Intrauterine interventions to treat foetal conditions.
15. Comparisons of alternative operative strategies for total hip replacement.
16. Comparison of phakoemulsification versus extracapsular cataract extraction
17. Thrombolysis in patients who have had a myocardial infarction to reduce the risk of a second myocardial infarction or cardiovascular death.
18. Comparisons of alternative treatment regimens for peripheral vascular disease.
19. Comparisons of alternative treatment regimens for schizophrenia.
20. HLA versus non-HLA corneal tissue-typing for corneal transplantation.
21. Kidney transplantation versus renal dialysis for chronic renal failure.
22. Grommets for 'glue' ear in children.
23. Hormone replacement therapy in post-menopausal women to reduce the incidence of osteoporosis.
24. Hormone replacement therapy in post-menopausal women to reduce the incidence of heart disease.
25. MMR vaccine to prevent mumps, measles and rubella.
26. Health promotion strategies (e.g. to prevent people taking up smoking, increase uptake of condom use).
27. Folic acid/multivitamin supplementation for women trying to conceive, to prevent neural tube defects.
28. Comparison of home versus hospital delivery for elective pregnancies.
29. Clofibrate for patients with hypercholesteraemia to prevent myocardial infarction or other cardiovascular events.
30. Lithotripsy to treat renal calculi.
31. Dietary salt restriction to reduce or prevent hypertension.

* We liaised with the reviewers at the London School of Hygiene and Tropical Medicine (Britton and co-workers[28]) when choosing health technologies to review for strategy 2. To prevent duplication, asterisked interventions were not considered further because Britton and co-workers[28] indicated a strong interest in reviewing these technologies.

Appendix 6

The seven health technologies shortlisted for strategy 2

TABLE 34 *Details of the seven health technologies shortlisted for strategy 2*

Health technology	Main outcome	Comments
Mammographic population screening to reduce mortality from breast cancer	Breast cancer mortality	Uniform intervention Uniform outcome Uniform population Therefore included in objective 2
Folic acid supplementation for women trying to conceive to prevent neural tube defects	Neural tube defects	Uniform intervention Uniform outcome Uniform population Therefore included in objective 2
Clofibrate to reduce morbidity and mortality from cardiovascular disease	Myocardial infarction, mortality	Few available trials that looked at clofibrate independently of other lipid-reducing drugs
Nicotine replacement therapy (patches, gum, community programmes) to promote smoking cessation	Smoking cessation	Heterogeneous interventions Outcome difficult to define Heterogeneous populations
Low back pain Physiotherapy, manual, 'back school', drugs, education	Pain index Mortality	Heterogeneous interventions Heterogeneous outcomes (e.g. different pain and mobility indexes)
Dietary salt restriction to reduce or prevent high blood pressure	Blood pressure	Difficulty in defining a uniform intervention Heterogeneous populations
Cervical smear population screening to reduce cervical cancer mortality	Cervical cancer mortality	Uniform intervention Uniform outcome Uniform population No RCTs available

Appendix 7

Additional information supplied to assessors for strategy 2

Mammographic screening to reduce mortality from breast cancer

Factors of interest to the review (see questions 2, 3 and 4):

- Outcome: age-specific breast cancer mortality.
- Population: women aged 50–64 years (UK Guidelines).*
- Intervention: mammography.

The four most common confounders (see questions 5 and 25):

- age
- family history of breast cancer
- parity
- height/weight ratio.

Possible adverse effects of mammographic screening for breast cancer (see question 8):

- Anxiety from a false-positive screening test result.
- Unnecessary treatment arising from diagnosis of clinically irrelevant non-invasive cancers.
- Harmful effect of exposure to X-ray radiation during mammography.

Periconceptional folic acid supplementation to prevent neural tube defects

Factors of interest to the review (see questions 2, 3 and 4):

- Outcome: recurrent or occurrent neural tube defects.
- Population: women attempting to conceive or becoming pregnant unintentionally.
- Intervention: periconceptional folic acid or multivitamin (with a folic acid content) supplementation.

The four most common confounders (see questions 5 and 25):

- age
- race/ethnicity
- maternal education
- pregnancy history.

Possible adverse effects of periconceptional folic acid supplementation (see question 8):

- toxicity
- masking of vitamin B_{12} deficiency.

* The age limit of 50–64 years was relaxed when reviewing papers relevant to this topic because of the paucity of available papers.

Appendix 8

Reasons for excluding four papers identified as possibly relevant to strategy 1

A paper that reported the results of an RCT of standard Schwarz vaccine also monitored vaccination and national vaccination campaigns.[71] Although RCT and QEO study elements were compared in this paper, we considered that the elements differed in many more respects than simply study design. The authors themselves attributed differences between results from the different study designs primarily to factors related to the efficiency with which the intervention was delivered (e.g. the quality of the 'cold chain' for maintaining the integrity of the vaccine) rather than to factors relating to internal and external validity.

Two papers reporting the effects of reducing serum cholesterol gave effect sizes for both RCT and observational studies, but the observational studies were not evaluating any intervention.[72,73] The cohort studies reviewed in these papers classified subjects according to their naturally occurring serum cholesterol levels at the outset, and then followed them to determine their risk of ischaemic heart disease.

Shaikh and co-workers[70] reviewed primary studies of tonsillectomy and adenoidectomy that used different designs. 'Quality' scores were assigned depending on the study design and other factors. Analyses considered the extent to which the findings of studies were associated with quality, but made no attempt to synthesise separate estimates of effect size for RCTs and QEO studies.

Appendix 9

Size and direction of discrepancies by quality

The size and direction of discrepancies by quality are shown for the following comparisons in the tables indicated:

- using only one comparison per paper (*Tables 35* and *36*)
- excluding comparisons from review papers (*Tables 37* and *38*)
- for 'effective' and 'ineffective' interventions (*Tables 39* and *40*)
- stratified by sample size (*Tables 41* and *42*).

TABLE 35 *Distribution of indices used to quantify discrepancies between RCT and QEO study effect sizes classified by quality, for comparisons 1–3, 10–12, 15, 19, 20, 24, 27, 33, 36 and 38*[*]

Index 1: RR_{RCT}/RR_{QEO}						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$1.00 \le x \le 1.10$	**$1.10 < x \le 1.25$**	**$1.25 < x \le 1.50$**	**$1.50 < x \le 2.00$**	**$x > 2.00$**	
Low (*n* = 8)	2	1	0	1	3	7
High (*n* = 6)	5	1	0	0	0	6
Index 2: ΔRD						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Low (*n* = 8)	1	1	2	1	1	6
High (*n* = 6)	3	1	0	0	0	4
Index 3: ΔRD/ mean RD						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Low (*n* =8)	0	0	2	2	3	7
High (*n* = 6)	2	1	1	1	0	5
Index 4: ΔIOF						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Low (*n* = 8)	0	1	2	0	1	5
High (*n* = 6)	2	1	0	0	0	3

[*] *The number of comparisons for each index may be less than 14, for reasons given in* Table 10

Continued

TABLE 35 contd *Distribution of indices used to quantify discrepancies between RCT and QEO study effect sizes classified by quality, for comparisons 1–3, 10–12, 15, 19, 20, 24, 27, 33, 36 and 38*[*]

***Index 5:* ΔCOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \leq x \leq 0.02$	**$0.02 < x \leq 0.05$**	**$0.05 < x \leq 0.10$**	**$0.10 < x \leq 0.20$**	**$x > 0.20$**	
Low (*n* = 8)	1	1	1	3	0	6
High (*n* = 6)	2	2	0	0	0	4
***Index 6:* ΔIOF/mean IOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \leq x \leq 0.10$	**$0.10 < x \leq 0.25$**	**$0.25 < x \leq 0.50$**	**$0.50 < x \leq 1.00$**	**$x > 1.00$**	
Low (*n* = 8)	0	1	2	3	0	6
High (*n* = 6)	1	3	0	0	0	4
***Index 7:* ΔCOF/mean IOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \leq x \leq 0.10$	**$0.10 < x \leq 0.25$**	**$0.25 < x \leq 0.50$**	**$0.50 < x \leq 1.00$**	**$x > 1.00$**	
Low (*n* = 8)	2	1	1	2	1	7
High (*n* = 6)	3	1	0	0	0	4

[*] *The number of comparisons for each index may be less than 14, for reasons given in* Table 10

TABLE 36 *'Direction' of discrepancies between RCT and QEO study results for comparisons 1–3, 10–12, 15, 19, 20, 24, 27, 33, 36 and 38*

Quality	**RCT**	**=**	**QEO**	**Total**
Relative risk				
Low (*n* = 8)	1	0	6	7
High (*n* = 6)	3	1	2	6
Risk difference				
Low (*n* = 8)	2	0	5	7
High (*n* = 6)	4	0	1	5

TABLE 37 *Distribution of indices used to quantify discrepancies between RCT and QEO effect sizes classified by quality, for comparisons 1, 2, 9–26, and 33–37*[*]

Index 1: RR_{RCT}/RR_{QEO}						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$1.00 \le x \le 1.10$	**$1.10 < x \le 1.25$**	**$1.25 < x \le 1.50$**	**$1.50 < x \le 2.00$**	**$x > 2.00$**	
Low (n = 12)	2	3	0	1	3	9
High (n = 13)	9	3	1	0	0	13
Index 2: ΔRD						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Low (n = 12)	1	2	0	1	1	5
High (n = 13)	4	4	0	0	0	8
Index 3: ΔRD/mean RD						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Low (n = 12)	0	1	1	3	3	8
High (n = 13)	3	2	2	2	1	10
Index 4: ΔIOF						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Low (n = 12)	0	3	0	0	1	4
High (n = 13)	4	2	0	0	0	6

[*] *The number of comparisons for each index may be less than 25, for reasons given in* Table 10

Continued

TABLE 37 contd *Distribution of indices used to quantify discrepancies between RCT and QEO effect sizes classified by quality, for comparisons 1, 2, 9–26, and 33–37*[*]

***Index 5:* ΔCOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Low (*n* = 12)	1	1	2	1	0	5
High (*n* = 13)	3	5	0	0	0	8
***Index 6:* ΔIOF/mean IOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Low (*n* = 12)	0	2	4	0	1	7
High (*n* = 13)	4	3	1	0	0	8
***Index 7:* ΔCOF/mean IOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Low (*n* = 12)	3	0	3	1	1	8
High (*n* = 13)	5	2	1	0	0	8

[*] *The number of comparisons for each index may be less than 25, for reasons given in* Table 10

TABLE 38 *'Direction' of discrepancies between RCT and QEO study results for comparisons 1, 2, 9–26, and 33–37*[*]

Quality	**RCT**	**=**	**QEO**	**Total**
Relative risk				
Low (*n* = 12)	2	0	7	9
High (*n* = 13)	6	1	6	13
Risk difference				
Low (*n* = 8)	1	0	7	8
High (*n* = 6)	7	0	3	10

[*] *The number of comparisons for each index may be less than 25, for reasons given in* Table 10

TABLE 39 *Distribution of indices used to quantify discrepancies between RCT and QEO study effect sizes, classified by whether or not the RCT effect size was statistically significant*[*]

Index 1: RR_{RCT}/RR_{QEO}						
Significance	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$1.00 \le x \le 1.10$	**$1.10 < x \le 1.25$**	**$1.25 < x \le 1.50$**	**$1.50 < x \le 2.00$**	**$x > 2.00$**	
Not significant (n = 16)	6	5	1	1	3	16
Significant (n = 22)	8	3	7	1	0	19
Index 2: ΔRD						
Significance	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Not significant (n = 16)	3	3	1	1	1	9
Significant (n = 22)	6	4	1	5	1	17
Index 3: ΔRD/mean RD						
Significance	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Not significant (n = 16)	1	0	1	3	4	9
Significant (n = 22)	4	4	4	7	3	22
Index 4: ΔIOF						
Significance	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Not significant (n = 16)	1	3	1	0	1	6
Significant (n = 22)	6	3	2	3	3	17

[*] *The number of comparisons for each index may be less than 38, for reasons given in* Table 10

Continued

TABLE 39 contd *Distribution of indices used to quantify discrepancies between RCT and QEO study effect sizes, classified by whether or not the RCT effect size was statistically significant*[*]

***Index 5:* ΔCOF**

Significance	Increasing disparity between RCT and QEO study elements →					Total
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Not significant (*n* = 16)	2	5	0	2	0	9
Significant (*n* = 22)	5	2	3	4	0	14

***Index 6:* ΔIOF/mean IOF**

Significance	Increasing disparity between RCT and QEO study elements →					Total
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Not significant (*n* = 16)	0	3	1	2	0	6
Significant (*n* = 22)	5	5	9	3	0	22

***Index 7:* ΔCOF/mean IOF**

Significance	Increasing disparity between RCT and QEO study elements →					Total
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Not significant (*n* = 16)	4	1	1	2	1	9
Significant (*n* = 22)	6	5	4	2	0	17

[*] *The number of comparisons for each index may be less than 38, for reasons given in* Table 10

TABLE 40 *'Direction' of discrepancies between RCT and QEO study results classified by whether or not the RCT effect size was statistically significant*[*]

	RCT	=	QEO	Total
Relative risk				
Not significant (*n* = 16)	5	0	10	16
Significant (*n* = 22)	10	1	9	19
Risk difference				
Not significant. (*n* = 16)	4	0	5	9
Significant (*n* = 22)	12	0	10	22

[*] *The number of comparisons in each table may be less than 38, for the reasons given in* Table 11

TABLE 41 *Distribution of indices used to quantify discrepancies between RCT and QEO study effect sizes classified by quality*[*]

Index 1: RR_{RCT}/RR_{QEO}						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$1.0 \le x \le 1.10$	**$1.1 < x \le 1.25$**	**$1.2 < x \le 1.50$**	**$1.5 < x \le 2.00$**	**$x > 2.00$**	
Sample size ≤ 750						
Low (*n* = 15)	1	2	5	2	2	12
High (*n* = 6)	5	0	1	0	0	6
Sample size > 750						
Low (*n* = 9)	4	2	2	0	1	9
High (*n* = 7)	4	3	0	0	0	7
Index 2: ΔRD						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.02$	**$0.02 < x \le 0.05$**	**$0.05 < x \le 0.10$**	**$0.10 < x \le 0.20$**	**$x > 0.20$**	
Sample size ≤ 750						
Low (*n* = 15)	1	0	1	5	1	8
High (*n* = 6)	1	2	0	0	0	3
Sample size > 750						
Low (*n* = 9)	4	3	0	1	1	9
High (*n* = 7)	3	2	0	0	0	5
Index 3: ΔRD/mean RD						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	$0.00 \le x \le 0.10$	**$0.10 < x \le 0.25$**	**$0.25 < x \le 0.50$**	**$0.50 < x \le 1.00$**	**$x > 1.00$**	
Sample size ≤ 750						
Low (*n* = 15)	0	1	1	6	3	11
High (*n* = 6)	1	0	0	1	1	3
Sample size > 750						
Low (*n* = 9)	2	1	2	2	2	9
High (*n* = 7)	2	2	2	1	0	7

[*] *The sample size stated is for the RCT element. The number of comparisons in each table may be less than 21 and 16, respectively, for reasons given in* Table 10

Continued

TABLE 41 contd *Distribution of indices used to quantify discrepancies between RCT and QEO study effect sizes classified by quality*[*]

***Index 4:* ΔIOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	0.00 ≤ *x* ≤ 0.02	**0.02 < *x* ≤ 0.05**	**0.05 < *x* ≤ 0.10**	**0.10 < *x* ≤ 0.20**	***x* > 0.20**	
Sample size* ≤ *750						
Low (*n* = 15)	0	1	0	3	3	7
High (*n* = 6)	0	1	0	0	0	1
Sample size* > *750						
Low (*n* = 9)	3	3	2	0	1	9
High (*n* = 7)	4	1	0	0	0	5

***Index 5:* ΔCOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	0.00 ≤ *x* ≤ 0.02	**0.02 < *x* ≤ 0.05**	**0.05 < *x* ≤ 0.10**	**0.10 < *x* ≤ 0.20**	***x* > 0.20**	
Sample size* ≤ *750						
Low (*n* = 15)	1	0	0	4	0	5
High (*n* = 6)	0	3	0	0	0	3
Sample size* > *750						
Low (*n* = 9)	3	2	3	1	0	9
High (*n* = 7)	3	2	0	0	0	5

***Index 6:* ΔIOF/mean IOF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	0.00 ≤ *x* ≤ 0.10	**0.10 < *x* ≤ 0.25**	**0.25 < *x* ≤ 0.50**	**0.50 < *x* ≤ 1.00**	***x* > 1.00**	
Sample size* ≤ *750						
Low (*n* = 15)	0	3	4	3	0	10
High (*n* = 6)	0	1	0	0	0	1
Sample size* > *750						
Low (*n* = 9)	1	2	5	1	0	9
High (*n* = 7)	4	2	1	0	0	7

[*] *The sample size stated is for the RCT element. The number of comparisons in each table may be less than 21 and 16, respectively, for reasons given in* Table 10

Continued

TABLE 41 contd *Distribution of indices used to quantify discrepancies between RCT and QEO study effect sizes classified by quality*[*]

***Index 7:* ΔCOF/mean COF**						
Quality	**Increasing disparity between RCT and QEO study elements →**					**Total**
	0.00 ≤ *x* ≤ 0.10	**0.10 < *x* ≤ 0.25**	**0.25 < *x* ≤ 0.50**	**0.50 < *x* ≤ 1.00**	***x* > 1.00**	
Sample size ≤ 750						
Low (*n* = 15)	2	3	1	1	1	8
High (*n* = 6)	2	1	0	0	0	3
Sample size > 750						
Low (*n* = 9)	3	4	3	2	0	9
High (*n* = 7)	3	1	1	0	0	5

[*] *The sample size stated is for the RCT element. The number of comparisons in each table may be less than 21 and 16, respectively, for reasons given in* Table 10

TABLE 42 *'Direction' of discrepancies between RCT and QEO study results classified by quality*[*]

Quality	**RCT**	**=**	**QEO**	**Total**
Relative risk				
Small sample size				
Low (*n* = 15)	8	0	4	12
High (*n* = 6)	2	1	3	6
Large sample size				
Low (*n* = 9)	1	0	8	9
High (*n* = 7)	4	0	3	7
Risk difference				
Small sample size				
Low (*n* = 15)	7	0	4	11
High (*n* = 6)	3	0	0	3
Large sample size				
Low (*n* = 9)	2	0	7	9
High (*n* = 7)	4	0	3	7

[*] *The sample size stated is for the RCT element. The number of comparisons in each table may be less than 21 and 16, respectively, for the reasons given in* Table 11

Appendix 10

Experts approached about information for strategy 2

The following experts on breast cancer were approached for information about literature.

- The Secretary, Action Against Breast Cancer, UK
- The Secretary, Association for International Cancer Research, UK
- The Secretary, Breakthrough Breast Cancer, UK
- The Secretary, Breast Cancer Campaign, UK
- The Secretary, Breast Cancer Care, UK
- The Secretary, Breast Cancer Research Trust, UK
- Dr S Feig, Breast Imaging Centre
- Dr DD Dershaw, Breast Imaging Section
- The Secretary, Cancer Research Campaign
- Dr S Moss, Cancer Screening Evaluation Unit
- Dr A Vandenbroucke, Centre des Tumeurs et de Radiotherapie – UCL
- Dr H Collette, Department of Epidemiology, University of Utrecht
- Mr J van Dijck, Department of Epidemiology, University of Nijmegen
- Dr P Peer, Department of Medical Informatics, University of Nijmegen
- Dr FE Alexander, Department of Public Health Sciences, University of Edinburgh
- Dr P Peeters, Department of Epidemiology, University Hospital, Nijmegen
- Dr Frisell, Department of Surgery, Sodersjukhuset, Stockholm
- Dr I Andersson, Department of Diagnostic Radiology, Malmö University Hospital
- Dr DB Kopans, Department of Radiology, Ambulatory Care Center, Massachusetts General Hospital, Boston, Massachusetts
- Dr S Fletcher, Department of Ambulatory Care and Prevention, Harvard Medical School, Boston, Massachusetts
- Dr L Nystrom, Department of Epidemiology, Public Health and Clinical Medicine, Umea University
- Dr S Shapiro, Department of Health Policy, Johns Hopkins Hospital, Baltimore, Maryland
- Dr RS Thompson, Department of Preventive Care, Group Health Cooperative of Puget Sound, Seattle, Washington
- Dr C Baines, Department of Preventive Medicine and Biostatistics, University of Toronto
- Dr AB Miller, Department of Preventive Medicine and Biostatistics, University of Toronto
- Dr P Glasziou, Department of Social and Preventive Medicine, University of Queensland Medical School, Brisbane
- Dr KC Chu, Early Detection Branch, National Cancer Institute, Bethesda, Maryland
- Dr C Byrne, Environmental Epidemiology Branch, National Cancer Institute, Bethesda, Maryland
- Dr D Palli, Epidemiology Unit, Centro per lo studio e la Prevénzione Oncologia, Florence
- The Secretary, Imperial Cancer Research Fund
- Dr DG Sienko, Ingham County Health Department, Lansing, Michigan
- The Secretary, Ludwig Institute for Cancer Research
- Dr L Tabar, Mammography Department, Central Hospital, Falun
- Mr S Duffy, MRC Biostatistics Unit
- Dr M Quinn, National Cancer Registration Bureau, OPCS, London
- The Secretary, Tenovus
- The Secretary, The Association for International Cancer Research
- Dr M Hakama, The Finnish Cancer Registry, Helsinki
- The Secretary, The Wellcome Trust
- Professor John Double, War on Cancer
- Dr CW Blackwell, Women's Cancer Control Program
- Dr CR Smart, Department of Radiology, University of Colorado Health Sciences Center, Denver, Colorado

The following experts on folic acid were approached for information about literature:

- Secretary to the Director, Association for Spina Bifida
- Dr M Khoury, Birth Defects and Genetic Diseases Branch, Centers for Disease Control, Atlanta, Georgia
- Dr Shurtleff, Birth Defects Clinic, Division of Congenital Defects, Children's Hospital and Medical Center, University of Washington, Seattle, Washington
- March of Dimes, Birth Defects Foundation, USA

- Dr JG Hall, Department of Paediatrics, University of British Columbia, B.C. Children's Hospital, Vancouver
- Professor N Wald, Department of Environmental and Preventive Medicine, Medical College of St Bartholomew's Hospital, London
- Dr A Gillies, Department of General Practice, Medical School, University of Birmingham
- Professor RW Smithells, Department of Paediatrics and Child Health, University of Leeds
- Dr A Copp, Neural Development Unit, University College London
- Dr J Mulinare, Division of Birth Defects and Developmental Disabilities, Centers for Disease Control, Atlanta, Georgia
- Dr Lynn Bailey, Department of Food Science and Human Nutrition, University of Florida, Gainsville, Florida
- Dr A Cziezel
- Dr CK Langley, Head of Information Services
- Dr PN Kirke, Health Research Board, Dublin
- Dr RG Vergel, National Center of Medical Genetics, Havana
- Dr JL Mills, Department of Epidemiology, Statistics and Preventative Research, National Institute of Child Health and Human Development, Bethesda, Maryland
- Dr MJ Seller, Division of Medical and Molecular Genetics, United Medical School, Guy's Hospital, London
- Dr K Laurence, Department of Child Health, University Hospital of Wales and Cardiff
- Dr M Super, Royal Manchester Children's Hospital, Manchester
- Dr C Schorah, University of Leeds
- Dr M Werler, Slone Epidemiology Unit, Boston University School of Medicine, Brookline, Massachusetts
- L Buxton, Tommy's Campaign
- Dr C Bower, Western Australian Research Institute for Child Health, Princess Margaret Hospital for Children

Appendix 11

Allocation to reviewers of the papers included for strategy 2

TABLE 43 *The papers assessed by each steering group member*

	Member*			
	BCR	**IMH**	**TAS**	**ITR**
MSBC studies				
Alexander *et al.*, 1994[93]	✓		✓	
Andersson *et al.*, 1988[85]		✓		✓
Collette *et al.*, 1984, 1992[85,89]		✓	✓	
Collette *et al.*, 1992[89]	✓			✓
Dales *et al.*, 1976[81]	✓			✓
Frisell *et al.*, 1991[88]		✓		✓
Hakama *et al.*, 1995[95]		✓	✓	
Miller *et al.*, 1992[90]	✓		✓	
Miller *et al.*, 1992[91]	✓			✓
Morrison *et al.*, 1988[86]	✓	✓		
Palli *et al.*, 1989[87]	✓			✓
Peer *et al.*, 1995[96]	✓		✓	
Shapiro *et al.*, 1982[82]		✓	✓	
Tabar *et al.*, 199597		✓		✓
Thompson *et al.*, 1994[94]	✓		✓	
UK TEDBC Group, 1993[92]		✓		✓
Verbeek *et al.*, 1984[84]		✓		✓
Total	9	9	7	9
FAS studies				
Bower and Stanley, 1992[108]		✓		✓
Chatkupt *et al.*, 1994[112]	✓	✓		
Czeizel and Dudas, 1992[109]	✓		✓	
Kirke *et al.*, 1992[65]	✓	✓		
Laurence *et al.*, 1981[98]			✓	✓
Martinez-Frias and Rodriguez-Pinilla, 1992[110]	✓	✓		
Mills *et al.*, 1989[103]	✓			✓
Milunsky *et al.*, 1989[104]	✓			✓
MRC Vitamin Study Research Group, 1991[107]	✓	✓		
Mulinare *et al.*, 1988[102]		✓	✓	
Seller and Nevin, 1984[101]	✓	✓		
Shaw *et al.*, 1995[113]		✓	✓	
Smithells *et al.*, 1981[99]			✓	✓
Smithells *et al.*, 1983[100]	✓	✓		
Smithells *et al.*, 1989[105]	✓		✓	
Vergel *et al.*, 1990[106]	✓			✓
Werler *et al.*, 1990[111]			✓	✓
Total	11	9	7	7

* *BCR, Dr Barnaby Reeves, Department of Social Medicine, University of Bristol; IMH, Dr Ian Harvey, Department of Social Medicine, University of Bristol; TAS, Professor Trevor Sheldon, NHS Centre for Reviews & Dissemination, University of York; ITR, Professor Ian Russell, Department of Health Sciences, University of York*

Appendix 12

Possible reasons for inverse correlations between items in the instrument for assessing quality

For reporting quality, questions about the detail with which authors reported the intervention, outcome and main confounding factors (questions 4a, 5 and 6b; see appendix 4) would have been reversed. However, we suspect that the need to reverse these items may have arisen from limitations in the instrument, and rules which were developed for scoring. The items are ones on which RCTs would be expected to score more highly than QEO study designs; indeed, all RCTs were automatically given credit for question 5. It may have been the case that RCTs scored less highly than QEO studies on other aspects of reporting quality (e.g. questions 7, 9, 12a, 12b and 29); these items are likely to have been well reported by observational studies, but have often been poorly reported for RCTs.[14,15]

Similar problems are likely to have arisen for the IVB and IVC quality dimensions. Questions about blinding, the appropriateness of the choice of statistical tests and 'cross-over' between intervention and control groups (questions 14, 19, 20a and 20b; see appendix 4) should have been reversed for IVB. These aspects of quality are not always relevant to observational studies, especially case–control studies which were automatically given credit for items 14, 20a and 20b. Items 28a and 28b (see appendix 4), which referred to the number of patients lost to follow-up, should have been reversed for IVC; these items were again of primary relevance to RCTs and cohort studies; case–control studies were given credit for 100% follow-up.

Methodology Group

Members

Methodology Progamme Director
Professor Richard Lilford
Director of Research and Development
NHS Executive – West Midlands, Birmingham

Chair
Professor Martin Buxton
Director, Health Economics Research Group
Brunel University, Uxbridge

Professor Douglas Altman
Professor of Statistics in Medicine
University of Oxford

Dr David Armstrong
Reader in Sociology as Applied to Medicine
King's College, London

Professor Nicholas Black
Professor of Health Services Research
London School of Hygiene & Tropical Medicine

Professor Ann Bowling
Professor of Health Services Research
University College London Medical School

Professor David Chadwick
Professor of Neurology
The Walton Centre for Neurology & Neurosurgery
Liverpool

Dr Mike Clarke
Associate Director (Research)
UK Cochrane Centre, Oxford

Professor Paul Dieppe
Director, MRC Health Services Research Centre
University of Bristol

Professor Michael Drummond
Director, Centre for Health Economics
University of York

Dr Vikki Entwistle
Senior Research Fellow,
Health Services Research Unit
University of Aberdeen

Professor Ewan B Ferlie
Professor of Public Services Management
Imperial College, London

Professor Ray Fitzpatrick
Professor of Public Health & Primary Care
University of Oxford

Dr Naomi Fulop
Deputy Director,
Service Delivery & Organisation Programme
London School of Hygiene & Tropical Medicine

Mrs Jenny Griffin
Head, Policy Research Programme
Department of Health
London

Professor Jeremy Grimshaw
Programme Director
Health Services Research Unit
University of Aberdeen

Professor Stephen Harrison
Professor of Social Policy
University of Manchester

Mr John Henderson
Economic Advisor
Department of Health, London

Professor Theresa Marteau
Director, Psychology & Genetics Research Group
Guy's, King's & St Thomas's School of Medicine, London

Dr Henry McQuay
Clinical Reader in Pain Relief
University of Oxford

Dr Nick Payne
Consultant Senior Lecturer in Public Health Medicine
ScHARR
University of Sheffield

Professor Joy Townsend
Director, Centre for Research in Primary & Community Care
University of Hertfordshire

Professor Kent Woods
Director, NHS HTA Programme, & Professor of Therapeutics
University of Leicester

HTA Commissioning Board

Members

Programme Director
Professor Kent Woods
Director, NHS HTA Programme, &
Professor of Therapeutics
University of Leicester

Chair
Professor Shah Ebrahim
Professor of Epidemiology of Ageing
University of Bristol

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

Professor Douglas Altman
Director, ICRF Medical Statistics Group
University of Oxford

Professor John Bond
Director, Centre for Health Services Research
University of Newcastle-upon-Tyne

Ms Christine Clark
Freelance Medical Writer
Bury, Lancs

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

Dr Andrew Farmer
General Practitioner &
NHS R&D Clinical Scientist
Institute of Health Sciences
University of Oxford

Professor Adrian Grant
Director, Health Services Research Unit
University of Aberdeen

Dr Alastair Gray
Director, Health Economics Research Centre
Institute of Health Sciences
University of Oxford

Professor Mark Haggard
Director, MRC Institute of Hearing Research
University of Nottingham

Professor Jenny Hewison
Senior Lecturer
School of Psychology
University of Leeds

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

Dr Donna Lamping
Head, Health Services Research Unit
London School of Hygiene & Tropical Medicine

Professor David Neal
Professor of Surgery
University of Newcastle-upon-Tyne

Professor Gillian Parker
Nuffield Professor of Community Care
University of Leicester

Dr Tim Peters
Reader in Medical Statistics
University of Bristol

Professor Martin Severs
Professor in Elderly Health Care
University of Portsmouth

Dr Sarah Stewart-Brown
Director, Health Services Research Unit
University of Oxford

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

Dr Gillian Vivian
Consultant in Nuclear Medicine & Radiology
Royal Cornwall Hospitals Trust
Truro

Professor Graham Watt
Department of General Practice
University of Glasgow

Dr Jeremy Wyatt
Senior Fellow
Health Knowledge Management Centre
University College London

Current and past membership details of all HTA 'committees' are available from the HTA website (see inside front cover for details)

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.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278