Editors: **Detels, Roger; Beaglehole, Robert; Lansang, Mary Ann; Gulliford, Martin**

Title: ***Oxford Textbook of Public Health, 5th Edition***

6.14
Systematic Reviews and Meta-Analysis

Matthias Egger

George Davey Smith

Jonathan Sterne

Abstract

Systematic reviews are 'studies of studies' that are done using a systematic approach to minimize bias and random error. Similar to other research, the problem to be addressed and the collection and analysis of the data should be detailed in a study protocol. This should include eligibility criteria for studies to be included, a comprehensive search strategy for such studies, and an assessment of their methodological quality. Systematic reviews may, or may not, include meta-analysis, a statistical combination of results from several studies to produce a single estimate of the effect of an intervention. Systematic reviews allow for a more objective appraisal of the evidence than traditional, narrative reviews and may contribute to resolve uncertainty and identify areas where further studies are needed. Meta-analysis, if appropriate, will enhance the precision of estimates of intervention effects, leading to reduced probability of false negative results, and potentially to a timelier introduction of effective interventions. Meta-analyses are, however, liable to numerous biases both at the level of the individual trial ('garbage in, garbage out') and the dissemination of trial results (publication bias and other reporting biases). Meta-analysis should be performed only within the framework of carefully conducted systematic reviews. The thoughtful consideration of heterogeneity between study results is an important aspect of systematic reviews and meta-analyses, and particularly important in meta-analyses of observational studies.

The volume of data that need to be considered by practitioners and researchers is constantly expanding. In many areas it has become simply impossible for the individual to read, critically evaluate and synthesize the state of current knowledge, let alone keep updating this on a regular basis. Reviews have become essential tools for anybody who wants to keep up with the new evidence that is accumulating in his or her field of interest. However, since Mulrow (Mulrow 1987) drew attention to the poor quality of narrative review articles in the 1980s, it has become clear that conventional reviews are an unreliable source of information. Since then there has been increasing focus on formal methods of systematically reviewing studies, to produce explicitly formulated, reproducible, and up-to-date summaries of the effects of healthcare interventions.

This is illustrated by the sharp increase in the number of reviews that used formal methods to synthesize evidence (Fig. 6.14.1). This chapter discusses terminology and scope, provides some historical background, and examines the potentials and pitfalls of systematic reviews and meta-analysis.

Systematic review, overview, or meta-analysis?

A number of terms are used concurrently to describe the process of systematically reviewing and integrating research evidence, including 'systematic review', 'meta-analysis', 'research synthesis', 'overview', and 'pooling'. A systematic review is a review that has been prepared using a documented systematic approach to minimizing biases and random errors. A systematic review may, or may not, include a meta-analysis: A statistical analysis of the results from independent studies, which generally aims to produce a single, typical estimate of a treatment effect. The distinction between systematic review and meta-analysis is important because it is always

appropriate and desirable to systematically review a body of data, but it may sometimes be inappropriate, or even misleading, to statistically pool results from separate studies.

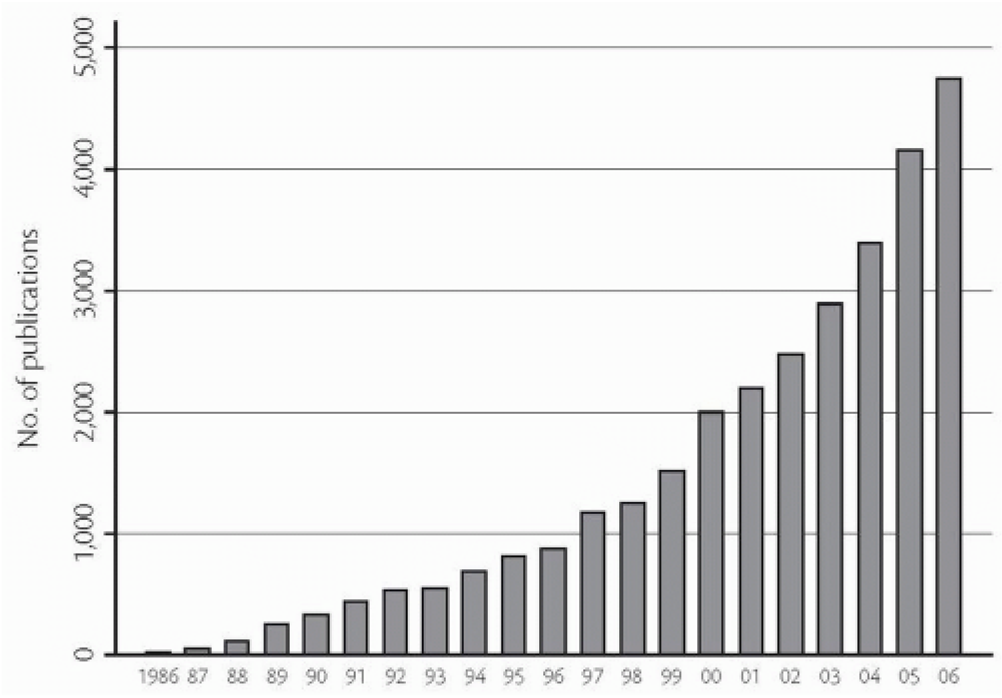

Fig. 6.14.1 Number of publications concerning systematic reviews and meta-analysis, 1986 to 2006. Results from MEDLINE search using text word and medical subject (MESH) heading 'meta-analysis' and text word 'systematic review'.

The scope of meta-analysis

A clear distinction should also be made between meta-analysis of randomized controlled trials and meta-analysis of epidemiological studies. Consider a set of trials of high methodological quality that examined the same intervention in comparable patient populations: Each trial will provide an unbiased estimate of the same underlying treatment effect. The variability that is observed between the trials can confidently be attributed to random variation and meta-analysis should provide an equally unbiased estimate of the treatment effect, with an increase in the precision of this estimate. A fundamentally different situation arises in the case of epidemiological studies, for example case-control studies, cross-sectional studies, or cohort studies. Due to the effects of confounding and bias, such observational studies may produce estimates of associations that deviate from the truth beyond what can be attributed to chance.

The fundamental difference that exists between observational studies and randomized controlled trials does not mean that the latter are immune to bias. As discussed below publication bias and other reporting biases may distort the evidence from both trials and observational studies. Bias may also be introduced if the methodological quality of clinical trials is inadequate. While systematic reviews have clear advantages over conventional reviews, it is crucial to understand the limitations of meta-analysis and the importance of exploring sources of heterogeneity and bias. Also, we believe that there continues to be a place for narrative reviews and editorials that express an informed but subjective opinion about how a particular body of evidence should be interpreted.

Historical notes

Efforts to compile summaries of research for medical practitioners who struggle with the amount of information that is relevant to medical practice are not new. Chalmers and Tröhler (Chalmers 2000) drew attention to two journals published in the eighteenth century in Leipzig and Edinburgh, *Comentarii de rebus in scientia naturali et medicina gestis* and *Medical and Philosophical Commentaries*, which published critical appraisals of important new books in medicine, including, for example, William Withering's now classic *Account of the Foxglove* (1785) on the use of digitalis for treating heart disease. The statistical basis of meta-analysis reaches back to the seventeenth century when in astronomy and geodesy intuition and experience suggested that combinations of data might be better than attempts to choose amongst them. In the twentieth century the distinguished statistician Karl Pearson was, in 1904, probably the first medical researcher reporting the use of formal techniques to combine data from different studies. The rationale for pooling studies put forward by Pearson in his account on the preventive effect of serum inoculations against enteric fever is still one of the main reasons for undertaking meta-analysis today: 'Many of the groups … are far too small to allow of any definite opinion being formed at all, having regard to the size of the probable error involved' (Pearson 1904). However, in contrast to psychology and educational research, such techniques were not widely used in medicine until the 1980s, when meta-analysis became increasingly popular in cardiology, oncology, and perinatal medicine. In the 1990s, the foundation of the Cochrane Collaboration (Box 6.14.1) facilitated numerous methodological developments and helped establish systematic reviews and meta-analysis as important tools in research and policy making.

Why do we need systematic reviews?

A patient with myocardial infarction in 1981

A likely scenario in the early 1980s, when discussing the discharge of a patient who had suffered an uncomplicated myocardial infarction, is as follows: A keen junior doctor asks whether the patient should receive a beta-blocker for secondary prevention of a future cardiac event. After a moment of silence the consultant states that this was a question which should be discussed in detail at the Journal Club on Thursday. The junior doctor is told to assemble and present the relevant literature. It is late in the evening when she makes her way to the library. The MEDLINE search identifies four clinical trials. When reviewing the conclusions from these trials, the doctor finds them to be rather confusing and contradictory (Table 6.14.1). Her consultant points out that the sheer amount of research published makes it impossible to keep track of and critically appraise individual studies. He recommends a good review article. Back in the library, the junior doctor finds an article which the *British Medical Journal* published in 1981. This narrative review concluded that 'Thus, despite claims that they reduce arrhythmias, cardiac work, and infarct size, we still have no clear evidence that beta-blockers improve long-term survival after infarction despite almost 20 years of clinical trials' (Mitchell 1981).

The junior doctor is relieved. She presents the findings of the review article, the Journal Club is a full success, and the patient is discharged without a beta-blocker.

Narrative reviews

Traditional narrative reviews have a number of disadvantages that systematic reviews may overcome. First, the classical review is subjective and therefore prone to bias and error. Mulrow (1987) showed that among 50 reviews published in the mid-1980s in leading general medicine journals, 49 reviews did not specify the source of the information and failed to perform a standardized assessment of the methodological quality of studies. Our junior doctor could have consulted another review of the same topic, published in the *European Heart Journal* in the same year. This review concluded that 'it seems perfectly reasonable to treat patients who have survived an infarction with timolol' (Hampton 1981). Without guidance by formal rules, reviewers will inevitably disagree about issues as basic as what types of studies it is appropriate to include and how to balance the quantitative evidence they provide. Selective inclusion of studies that support the author's view is common. It is thus hardly surprising that reviewers using traditional methods often reach opposite conclusions and miss small, but potentially important, differences (Mulrow 1987). In controversial areas the conclusions drawn from a given body of evidence may be associated more with the speciality of the reviewer than with the available data. By systematically identifying, scrutinising, tabulating, and perhaps integrating all relevant studies, systematic reviews allow a more objective appraisal, which can help to resolve uncertainties when the original research, classical reviews, and editorial comments disagree.

Box 6.14.1 The Cochrane Collaboration

Funded in 1993, the Cochrane Collaboration (www.cochrane.org) is a unique international organization whose aim is to help people make well-informed decisions by preparing, maintaining, and promoting systematic reviews in all areas of healthcare, including treatment, prevention, screening, and rehabilitation. At present, there are nearly 15 000 people participating in the collaboration, in nearly 100 countries. The main work is done in one of about 50 review groups that take on the task of preparing and maintaining reviews. These reviews generally focus on the findings from randomized trials, and most include one or several meta-analyses. There are also 12 Cochrane fields, including a Cochrane health promotion and public health field, which cut across the scope of review groups and help identify potential reviewers and topics. The coverage of Cochrane reviews is continually improving, with over 3000 reviews available at the beginning of 2008. The reviews are published in the Cochrane Database of Systematic Reviews, which is part of The Cochrane Library, an electronic publication available on Wiley Interscience (http://www.thecochranelibrary.com). Cochrane reviews are indexed in Medline. The Cochrane Library also includes a large register of controlled trials, the Cochrane Central Register of Controlled Trials, and a register of methodological studies.

The logo of the Cochrane Collaboration (Fig. 6.14.2) illustrates a systematic review of seven RCTs of a short, inexpensive course of a corticosteroid given to women about to give birth too early, comparing the intervention with placebo. A schematic representation of the forest plot is shown. The first of these RCTs was reported in 1972, the last in 1980. The diagram summarizes the evidence that would have been revealed, had the available RCTs been reviewed systematically a decade later: It indicates strongly that corticosteroids reduce the risk of babies dying from the complications of immaturity. Because no systematic review of these trials had been published until 1989, most obstetricians had not realized that the treatment was so effective, reducing the odds of the babies of these women dying from the complications of immaturity by 30-50 per cent. As a result, tens of thousands of premature babies have probably suffered and died unnecessarily, and needed more expensive treatment than was necessary. By 1991, seven more trials had been reported, and the picture had become still stronger.

A similar collaboration, the Campbell Collaboration, was set up to prepare systematic reviews of high-quality research conducted worldwide on effective methods and interventions in the fields of social welfare and social work, education and learning, and crime and delinquency (see www.campbellcollaboration.org).

Limitations of a single study

A single study often fails to detect, or exclude with certainty, a modest, albeit relevant, difference in the effects of two therapies. A trial may thus show no statistically significant treatment effect when in reality such an effect exists—it may produce a false negative result. In many trials there is a substantial probability of missing a clinically relevant difference in outcome (Freiman *et al*. 1992; Thornley & Adams 1998). The number of patients included in trials is thus often inadequate. In some cases, however, the required sample size may be difficult to achieve. A drug which reduces the risk of death from myocardial infarction by 10 per cent could delay many thousands of deaths each year in the United Kingdom alone. However, in order to detect such an effect with 90 per cent certainty over 10 000 patients in each treatment group would be needed.

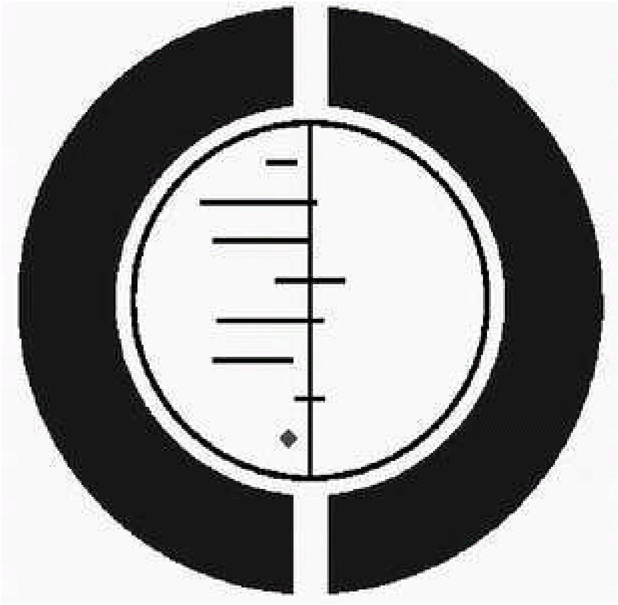

Fig. 6.14.2 The logo of the Cochrane Collaboration.

The meta-analytic approach appears to be an attractive alternative to such a large, expensive, and logistically problematic study. Data from patients in trials evaluating the same or a similar drug in a number of smaller, but comparable, studies are considered. In this way the necessary number of patients may be reached, and relatively small effects can be detected or excluded with confidence. Systematic reviews can also contribute to considerations regarding the applicability of study results. If many trials exist in different groups of patients, with similar results being seen in the various trials, then it can be concluded that the effect of the intervention under study has some generality. By putting together all available data, meta-analyses are also better placed than individual trials to answer questions regarding whether or not an overall study result varies

among subgroups—e.g. among men and women, older and younger patients, or participants with different degrees of severity of disease.

Table 6.14.1 Conclusions from four randomized controlled trials of beta-blockers in secondary prevention after myocardial infarction

- *'The mortality and hospital readmission rates were not significantly different in the two groups. This also applied to the incidence of cardiac failure, exertional dyspnoea, and frequency of ventricular ectopic beats'*. Reynolds *et al. British Heart Journal* (1972)

- *'Until the results of further trials are reported long-term beta-adrenoceptor blockade (possibly up to two years) is recommended after uncomplicated anterior myocardial infarction'*. Multicentre International Study, *BMJ* (1977)

- *'The trial was designed to detect a 50 per cent reduction in mortality and this was not shown. The non-fatal reinfarction rate was similar in both groups'*. Baber *et al. British Heart Journal* (1980)

- *'We conclude that long-term treatment with timolol in patients surviving acute myocardial infarction reduces mortality and the rate of reinfarction'*. The Norwegian Multicentre Study Group, *New England Journal of Medicine (1981)*

A more transparent appraisal

An important advantage of systematic reviews is that they render the review process more transparent. In traditional narrative reviews it is often not clear how the conclusions follow from the data examined. In an adequately presented systematic review it should be possible for readers to replicate the quantitative component of the argument. To facilitate this, it is valuable if the exclusion of potentially relevant studies is justified and the data included in meta-analyses are either presented in full or made available to interested readers. The increased openness required leads to the replacement of unhelpful descriptors such as 'no clear evidence, 'some evidence of a trend', 'a weak relationship', and 'a strong relationship'. Furthermore, performing a meta-analysis may lead to reviewers moving beyond the conclusions authors present in the abstract of papers, to a thorough examination of the actual data.

The epidemiology of results

The tabulation, exploration, and evaluation of results are important components of systematic reviews. As discussed in more detail below, this can be taken further to explore sources of heterogeneity and test new hypotheses that were not posed in individual studies. This has been termed the ‘epidemiology of results’ where the findings of an original study replace the individual as the unit of analysis (Jenicek 1989). Systematic reviews can thus lead to the identification of the most promising or the most urgent research question, and may permit a more accurate calculation of the sample sizes needed in future studies. This is illustrated by an early meta-analysis of four trials that compared different methods of monitoring the foetus during labour (Chalmers 1979). The meta-analysis led to the hypothesis that, compared with intermittent auscultation, continuous foetal heart monitoring reduced the risk of neonatal seizures. This hypothesis was subsequently confirmed in a randomized trial of almost seven times the size of the four previous studies combined (MacDonald *et al*. 1985).

What was the evidence in 1981?

What conclusions would our junior doctor have reached if she had had access to a systematic review and meta-analysis of the beta-blocker trials? A total of 13 such trials had in fact been published by the end of 1981. Using meta-analysis to combine the results of these 13 trials, the relative risk of mortality comparing patients treated with beta-blocker with those treated with placebo is estimated at 0.78 (95 per cent confidence intervals 0.69-0.88, P <0.001). Thus conclusive evidence of the life-saving potential of this treatment, though available, was ignored.

Steps in carrying out systematic reviews

Developing a review protocol

Systematic reviews should be viewed as observational studies of the evidence. The steps involved, summarized in Box 6.14.2, are similar to any other research undertaking: Formulation of the problem to be addressed, collection and analysis of the data, and interpretation of the results. Likewise, a detailed study protocol which clearly states the question to be addressed, the subgroups of interest, and the methods and criteria to be employed for identifying and selecting relevant studies and extracting and analysing information should be written in advance. This is important to avoid bias being introduced by decisions that are influenced by the data. For example, studies which produced unexpected or undesired results may be excluded by *post hoc* changes to the inclusion criteria. Similarly, unplanned data-driven subgroup analyses may produce spurious results. The review protocol should ideally be conceived by a group of reviewers with expertise both in the content area and the science of research synthesis.

Objectives and eligibility criteria

The formulation of detailed objectives is at the heart of any research project. This should include the definition of study participants, interventions, outcomes, and settings. As with patient inclusion and exclusion criteria in clinical studies, eligibility criteria can then be defined for the type of studies to be included. They relate to the quality of trials and to the combinability of patients, treatments, outcomes, and lengths of follow-up. Quality and design features of clinical trials can influence their results (see below). Ideally, only controlled trials with proper patient randomization which report on all initially included patients according to the intention-to-treat principle and with an objective, preferably blinded, outcome assessment would be considered for inclusion. However, the investigation of the influence of study quality may often be an important objective of a meta-analysis. Furthermore, assessing study quality can be a subjective process, especially since the information reported is often incomplete for this purpose (Schulz 1996). It is therefore generally preferable to define only basic inclusion criteria, to assess the methodological quality of component studies, and to perform a thorough sensitivity analysis, as illustrated below.

Literature search

The search strategy for the identification of the relevant studies should be clearly delineated. Identifying controlled trials has become more straightforward in recent years. Appropriate terms to index randomized trials and controlled trials were introduced in the widely used bibliographic databases MEDLINE and EMBASE by the mid-1990s. However, tens of thousands of trial reports had been included prior to the introduction of these terms. In a painstaking effort the Cochrane Collaboration checked the titles and abstracts of almost 300 000 MEDLINE and EMBASE records which were then re-tagged as clinical trials if appropriate. It was important to examine both MEDLINE and EMBASE: The majority of journals indexed in MEDLINE are published in the United States whereas EMBASE has better coverage of European journals. Also, the results of trials indexed only in EMBASE may differ from other trials (Sampson *et al*. 2003). Finally, thousands of reports of controlled trials have been identified by manual searches ('handsearching') of journals, conference proceedings, and other sources.

All trials identified in the re-tagging and handsearching projects have been included in the Cochrane Central Register of Controlled Trials (see Box 6.14.1). This register currently includes over 500 000 records and is the best single source of published trials for inclusion in systematic reviews. Searches of MEDLINE and EMBASE are, however, still required to identify trials that were published recently. Specialized databases, conference proceedings, and the bibliographies of review articles, monographs, and the located studies should be scrutinized as well. Finally, the searching by hand of key journals should be considered.

The search should be extended to include unpublished studies, as their results may systematically differ from published trials.

A systematic review which is restricted to published evidence may produce distorted results due to publication bias (see below). The registration of trials at the time they are established (and before their results become known) would eliminate the risk of publication bias. Trial registration has gained momentum in recent years. The 1997 US Food and Drug Administration (FDA) Modernization Act mandated registration of efficacy drug trials conducted under FDA regulations. Also, the International Committee of Medical Journal Editors (ICMJE) introduced a policy, effective from July 2005, that requires prospective trial registration as a condition of publication (De Angelis *et al*. 2004). Several web-based registers have since been established, including ClinicalTrials.gov (www.clinicaltrials.gov) and the International Standard Randomized Controlled Trial Number

Registry (http://isrctn.org). The World Health Organization has built an international search portal to facilitate access to the data from all major registers (www.who.int/ictrp). In addition to trial registries, colleagues, experts in the field, contacts in the pharmaceutical industry, and other informal channels can also be important sources of information on unpublished and ongoing trials. Finally, the proceedings of relevant FDA advisory panels and other FDA, which are also available at www.fda.gov, may be useful to identify unpublished studies or unpublished data, particularly on adverse effects of treatment (Jüni 2004).

Box 6.14.2 Steps in conducting a systematic review

- Formulate the review question
- Define inclusion and exclusion criteria, considering
    - Participants
    - Interventions and comparisons
    - Outcomes
    - Study designs and methodological quality
- Develop the search strategy to identify relevant studies, considering the following sources
    - MEDLINE, EMBASE, and other bibliographic databases
    - Cochrane Central Register of Controlled Trials
    - World Health Organization search portal of trial registers
    - Checking reference lists of relevant articles
    - Web site of Food and Drug Administration (FDA)
    - Search by hand of key journals
    - Personal communication with experts in the field
- Select studies
    - Have eligibility checked by >1 observer
    - Develop strategy to resolve disagreements
    - Keep log of excluded studies, with reasons for exclusions
- Assess study quality
    - Consider assessment by >1 observer
    - Use simple checklists rather than quality scales
    - Always assess concealment of treatment allocation, blinding, and handling of patient attrition
    - Consider blinding of observers to authors, institutions, and journals
- Extract data
    - Design and pilot data extraction form
    - Consider data extraction by >1 observer
    - Consider blinding of observers to authors, institutions, and journals
- Analyse and present results

    - Tabulate results from individual studies
    - Examine forest plot
    - Explore possible sources of heterogeneity and bias
    - Consider meta-analysis of all trials or subgroups of trials
    - Perform sensitivity analyses, examine funnel plots
    - Make list of excluded studies available to interested readers
- Interpret results
    - Consider limitations, including publication and related biases
    - Consider strength of evidence
    - Consider applicability
    - Consider numbers-needed-to-treat to benefit / harm
    - Consider economic implications
    - Consider implications for future research

Note: Points 1-7 should be addressed in the review protocol.

Selection of studies, assessment of methodological quality, and data extraction

Decisions regarding the inclusion or exclusion of individual studies often involve some degree of subjectivity. It is therefore useful to have two observers checking eligibility of candidate studies, with disagreements being resolved by discussion or a third reviewer.

Randomized controlled trials provide the best evidence of the efficacy of medical interventions but they are not immune to bias. The assessment of study quality is therefore an important component of systematic review. Trials with inadequate allocation concealment or lack of blinding (see Box 6.14.3 for a discussion of these concepts) tend to exaggerate estimates of intervention effects, compared with adequately concealed or adequately blinded trials (Jüni *et al*. 2001). Treatment effects may also be overestimated if some participants, for example, those not adhering to study medications, were excluded from the analysis. A large number of different scales and checklists are available to assess the quality of clinical trials. However, empirical evidence and theoretical considerations suggest that although summary quality scores may in some circumstances provide a useful overall assessment, scales should not generally be used to assess the quality of trials in systematic reviews (Jüni *et al*. 1999). Rather, the relevant methodological aspects should be identified in the study protocol, and assessed individually. Again, independent assessment by more than one observer is desirable. Blinding of observers to the names of the authors and their institutions, the names of the journals, sources of funding, and acknowledgments may also be considered but this is time consuming, and potential benefits may not always justify the additional costs.

It is important that two independent observers extract the data, so errors can be avoided. The extraction of data to calculate standardized mean differences (see below) has been shown to be particularly liable to errors. A standardized record form is needed for this purpose. Data extraction forms should be carefully designed, piloted, and revised if necessary. Electronic data collection forms and web-based forms have a number of advantages, including the combination of data abstraction and data entry in one step, and the automatic detection of inconsistencies between data recorded by different observers. However, the complexities involved in programming and revising electronic forms should not be underestimated.

Box 6.14.3 A crucial distinction: allocation concealment versus blinding in clinical trials

Allocation concealment refers to procedures that secure strict implementation of the schedule of random assignments by preventing foreknowledge of forthcoming allocations by study participants or by those recruiting them to the trial. It is always feasible to conceal allocation. Failure to conceal allocation may lead to biased selection of participants into intervention groups. Examples of procedures usually considered adequate include sequentially numbered drug containers of identical appearance; central allocation (including web-based or pharmacy-controlled randomization); and sequentially numbered, opaque, sealed envelopes. Examples of procedures usually considered inadequate include using an open random allocation schedule; assignment of envelopes without appropriate safeguards (for example, unsealed or non-opaque or not sequentially numbered); and alternation or rotation.

Blinding refers to procedures that prevent study participants, caregivers, or outcome assessors from knowing which intervention was received. Blinding of participants and caregivers may not be feasible: For example, in a trial of surgery versus radiotherapy for prostate cancer. In such circumstances it may still be possible to blind the assessment of outcomes. Blinding may reduce the risk that knowledge of the intervention received, rather than the intervention itself, affects outcomes and/or outcome measurements. Examples of procedures usually considered adequate include provision of indistinguishable placebo tablets, or use of a sham surgical procedure in the control group. An example of blinded outcome assessment is assessment of medical records to ascertain cause of death by an endpoints committee unaware of intervention status.

Meta-analysis: Presenting, combining, and interpreting results

Once studies have been selected, critically appraised, and data extracted, the characteristics of included studies should be presented in tabular form. For example, a meta-analysis of parallel group randomized trials that examined the effectiveness of beta blockers versus placebo or alternative treatment in patients who had had a myocardial infarction identified 31 trials of at least 6 months' duration, which contributed 33 comparisons of beta blocker with control groups (Freemantle *et al*. 1999). In the first table of the report, the authors presented the characteristics of each trial, including the trial acronym or first author and year of publication, average length of follow up, the name of the beta blocker tested, the level of blinding, concealment of allocation and the rate of loss to follow up.

Measures of treatment effect

The results from individual studies have to be expressed in a standardized format to allow for comparison between studies. If the endpoint is binary (for example, disease versus no disease, or dead versus alive) then relative risks or odds ratios are often calculated. The odds ratio has convenient mathematical properties, which allow for ease in the combination of data and the testing of the overall effect for statistical significance, but, the odds ratio will differ from the relative risk as the outcome becomes more common. Relative risks are more intuitively comprehensible to most people. However, as the outcome becomes more common the range of the relative risk is constrained while the odds ratio is not. The odds ratio has the further advantage that the odds ratio for non-occurrence of the outcome is exactly the inverse of the odds ratio for the outcome. Different measures such as the absolute risk reduction or the number of patients needed to be treated for one person to benefit are more helpful when applying results in clinical practice (see below).

If the outcome is continuous and measurements are made on the same scale (for example, blood pressure measured in mm Hg) the mean difference between the treatment and control groups is used. If trials measured outcomes in different ways, for example, pain on a 5-point ranking scale or on a 100-mm visual analogue scale, it is necessary to standardize the measurements on a uniform scale to allow their inclusion in meta-analysis. This is done by calculating the standardized mean difference for each study, i.e. the difference in means between the two groups divided by the pooled standard deviation of the measurements (Deeks *et al*. 2001).

Meta-analysis

Careful consideration of the combinability of the studies in question is an important step in systematic reviews (Box 6.14.2): It will not always be appropriate to combine the results from the different studies to produce a single estimate of the treatment effect. If, after careful consideration, a meta-analysis is deemed appropriate, the

next step consists in estimating a typical effect by combining the data. Two principles are important. First, simply pooling the data from different studies and treating them as one large study would fail to preserve the randomization, and introduce bias and confounding. For example, a ‘meta-analysis’ of the literature on the role of male circumcision in HIV transmission concluded that the risk of HIV infection was lower in uncircumcised men. However, the analysis was performed by simply pooling the data from 33 diverse studies. A re-analysis stratifying the data by study found that an intact foreskin was in fact associated with an increased risk of HIV infection (O'Farrell & Egger 2000). Confounding by study thus led to a change in the direction of the association (a case of ‘Simpson's paradox’ in epidemiological parlance). The study unit of analysis must therefore always be maintained when combining data. Of note, several randomized trials have since conclusively shown that circumcision is associated with a substantially reduced risk of HIV transmission (Busse *et al*. 2008).

Second, simply calculating an arithmetic mean would be inappropriate. The results from small studies are more subject to the play of chance and should, therefore, be given less weight. Let us assume that we have *k* studies, and have derived a treatment effect estimate $\hat{\theta}$(which might be a log odds ratio, log risk ratio, or mean difference) for each study (*i* = 1 to *k*). The 'fixed effects' model considers the variability between these treatment effect estimates as exclusively due to random variation, so that if all the studies were infinitely large they would give identical results. To derive a summary treatment effect estimate we calculate a weighted average of the treatment effect estimates in the individual studies:

$$\hat{\theta}_F = \frac{\sum w_i \hat{\theta}_i}{\sum w_i}$$

The subscript *F* denotes the fixed-effects assumption. Use of a weighted average accords with our first principle because individuals are only compared with other individuals in the same study. The usual choice of weight w_i for study *i*, which minimizes the variability of the summary treatment effect estimate, is inverse variance weight $w_i = 1/v_i$, where *vi* is the variance of the treatment effect estimate. This accords with our second principle because the larger the study, the smaller will be the variance of the treatment effect estimate from that study. The standard error of the summary effect estimate $\hat{\theta}_F$ is:

$$\mathrm{SE}(\hat{\theta}_F) = \frac{1}{\sqrt{\sum_{i=1}^{k} w_i}}$$

This can be used to derive confidence intervals, a *z* statistic, and hence a *P* value for the null hypothesis that the true treatment effect is zero. An alternative weighting scheme, which has been shown to be more robust when data are sparse, is to use Mantel-Haenszel weights to combine relative risks or odds ratios. More details on statistical methods for meta-analysis and meta-analysis software are given by Deeks *et al*. and Sterne *et al*. (2001).

Graphical display

Results from each trial are usefully displayed together with their confidence intervals in a 'forest plot', a form of presentation developed in the 1980s by Richard Peto's group in Oxford. Figure 6.14.3 represents the forest plot for the trials of beta-blockers in secondary prevention after myocardial infarction (Freemantle *et al*. 1999). Each study is represented by a black square whose centre corresponds to the treatment effect estimate, and a horizontal line representing the 95 per cent confidence intervals of the relative risk. The confidence interval of most studies cross this line. The area of the black squares is proportional to the weight of the study in the meta-analysis: Plots that use an equally sized symbol for each study unhelpfully draw attention to the widest confidence intervals and thus the smallest studies. The solid vertical line corresponds to no effect of treatment (relative risk 1.0). If the confidence interval includes 1, then the difference in the effect of experimental and control therapy is not statistically significant at conventional levels (*P* > 0.05). In Fig. 6.14.3, the confidence interval of most studies crosses this line.

The diamond at the bottom of the graph displays the result of the meta-analysis: The centre of the diamond corresponds to the summary treatment effect estimate, while its width corresponds to the 95 per cent confidence interval. The broken line also corresponds to the summary treatment effect estimate and is included to a visual assessment of the variability of the individual studies around the summary estimate.

A logarithmic scale was used for plotting the relative risk in Fig. 6.14.3. There are a number of reasons why ratio measures are best plotted on logarithmic scales. Most importantly, the value of a risk ratio and its reciprocal, for example 0.5 and 2, which represent risk ratios of the same magnitude but opposite directions, will be equidistant from 1.0. Studies with relative risks below and above 1.0 will take up equal space on the graph and thus visually appear to be equally important. Also, confidence intervals will be symmetrical around the point estimate.

Heterogeneity between study results

The thoughtful consideration of heterogeneity between study results is an important aspect of systematic reviews. As explained above, this should start when writing the review protocol, by defining potential sources of heterogeneity and planning appropriate subgroup analyses. Once the data have been assembled, simple inspection of the forest plot is informative. The results from the beta-blocker trials are fairly homogeneous, clustering between a relative risk of 0.5 and 1.0, with widely overlapping confidence intervals (Fig. 6.14.3). In contrast, trials of BCG vaccination for prevention of tuberculosis (Colditz *et al.* 1994) are clearly heterogeneous (Fig. 6.14.4). The findings of the British trial, which indicate substantial benefit of BCG vaccination are not compatible with those from the Madras or Puerto Rico trials which suggest no effect or only a modest benefit. There is no overlap in the confidence intervals of the three trials.

The fixed-effects summary estimate is based on the assumption that the true effect does not differ between studies, and statistical tests of homogeneity (also called tests of heterogeneity) assess the evidence against this. The null hypothesis is individual study results reflect a single underlying effect, so that the differences between treatment effect estimates in individual studies are a consequence of sampling variation and simply due to chance. The test statistic is:

$$Q = \sum_{i=1}^{k} w_i \sum w_i (\hat{\theta}_i - \hat{\theta}_F)^2,$$

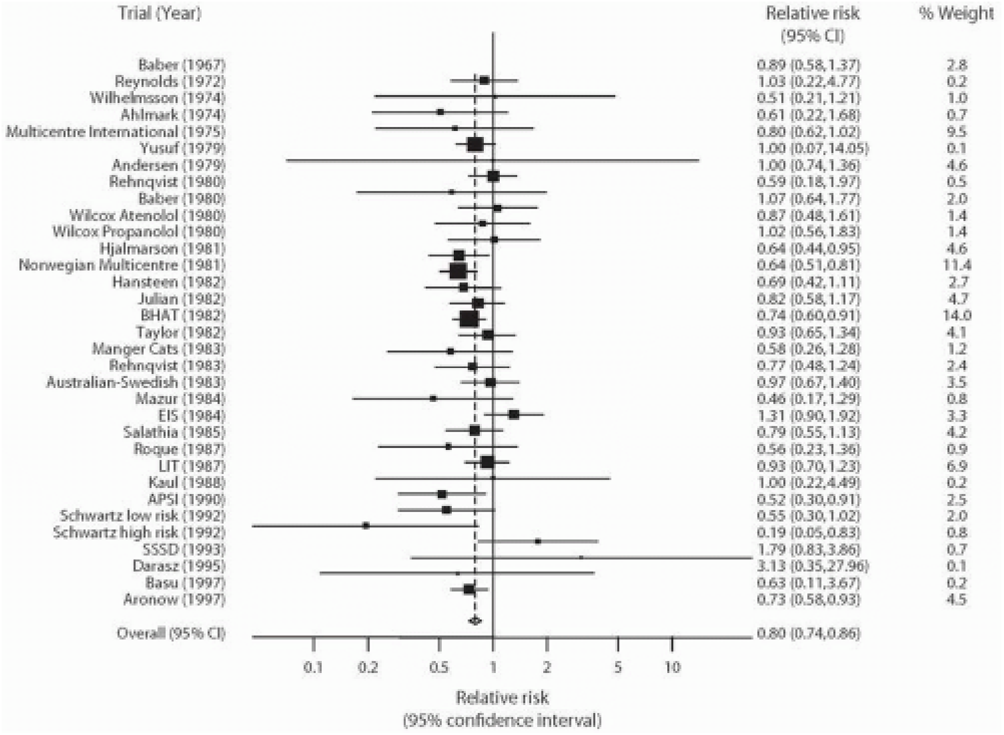

Fig. 6.14.3 ‘Forest plot’ showing mortality results from trials of beta-blockers in secondary prevention after myocardial infarction. Trials are ordered by year of publication. The black square and horizontal line correspond to the trials' risk ratio and 95 per cent confidence intervals. The area of the black squares reflects the weight each trial contributes in the meta-analysis. The diamond represents the combined relative risk with its 95 per cent confidence interval, from fixed effects meta-analysis, indicating a 20 per cent reduction in the risk of death.

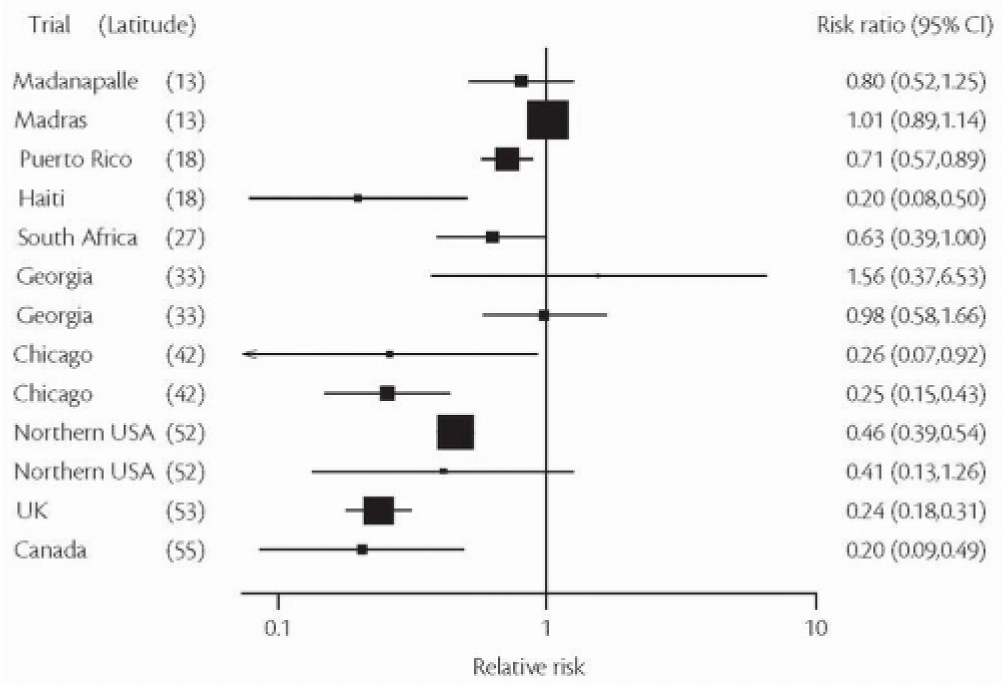

Fig. 6.14.4 Forest plot of trials of BCG vaccine to prevent tuberculosis. Trials are ordered according to the latitude of the study location, expressed as degrees from the equator. No meta-analysis is shown.

which is compared with the chi-squared distribution on (k−1) degrees of freedom. The greater the average difference between the individual study results and the summary estimate, the more evidence against the null hypothesis of a common fixed effect for all studies. The test of homogeneity gives P = 0.27 for the beta-blocker trials but P <0.001 for the BCG trials. The BCG trials are an extreme example, however, and a major limitation of statistical tests of homogeneity is their lack of power—they often fail to reject the null hypothesis of homogeneous results even if substantial between-study differences exist. Reviewers should therefore not assume that a non-significant test of heterogeneity excludes important heterogeneity.

An alternative to testing for heterogeneity is to quantify it. Higgins *et al*. (2003) developed a measure of the degree of inconsistency in the studies' results, called I^2, which describes the percentage of total variation across studies that is due to heterogeneity rather than chance. It is readily calculated as I^2 = 100 per cent×(Q−df)/Q where Q is the test statistic defined above and df the degrees of freedom. Negative values of I^2 are put equal to zero so that I^2 lies between 0 and 100 per cent. A value of 0 per cent indicates no observed heterogeneity, and larger values show increasing heterogeneity. The I^2 is 92 per cent for the BCG trials but only 12 per cent for the beta-blocker trials. Note that heterogeneity between study results should not be seen as purely a problem for systematic reviews, since it also provides an opportunity for examining why treatment effects differ in different circumstances, as discussed below.

Random-effects meta-analysis

There are a variety of statistical techniques available for meta-analysis, which can be broadly classified into 'fixed-effects' and 'random-effects' models (Deeks *et al*. 2001). Random-effects models (DerSimonian 1986) allow for between-study heterogeneity by assuming that the treatment effect varies between studies, and take this into consideration as an additional source of variation. The summary treatment effect from random-effect meta-analysis then estimates the mean about which the treatment effect in different studies is assumed to vary and thus should be interpreted differently from the results from a fixed-effects meta-analysis. In practice, random-effects estimates are derived simply by modifying the weights from the fixed-effects analysis. This leads to relatively more weight being given to smaller studies: This may be undesirable considering that small studies are more vulnerable to publication and other bias (see below). Because they assume an extra source of variability, random-effects estimates have wider confidence intervals than fixed-effects estimates.

While neither of the two models can be said to be 'correct', a substantial difference in the combined effect calculated by the fixed and random effects models will be seen only if studies are markedly heterogeneous, as in the case of the BCG trials (Table 6.14.2). Combining trials using a random-effects model indicates that BCG vaccination halves the risk of tuberculosis, whereas fixed-effects analysis indicates that the risk is only reduced by 35 per cent. This is essentially explained by the different weight given to the large Madras trial which showed no protective effect of vaccination (41 per cent of the total weight with fixed effects model, 10 per cent with random effects model, Table 6.14.2).

The use of random-effects models is often advocated if there is heterogeneity between study results. This is problematic: Rather than simply ignoring heterogeneity after allowing for it in a statistical model, a better approach is to scrutinize and attempt to explain it. As shown in Fig. 6.14.3, BCG vaccination appears to be effective at higher latitudes but not in warmer regions, possibly because exposure to certain environmental mycobacteria acts as a 'natural' BCG inoculation in warmer regions. In this situation it is more meaningful to quantify how the effect varies according to latitude than to calculate an overall estimate of effect which will be misleading, independent of the model used.

Table 6.14.2 Meta-analysis of trials of BCG vaccination to prevent tuberculosis using a fixed-effects and random effects model. Note the differences in the weight allocated to individual studies

Trial	Relative risk (95% confidence interval)	Fixed effects weight (%)	Random effects weight (%)
Madanapalle	0.80 (0.52-1.25)	3.20	8.88
Madras	1.01 (0.89-1.14))	41.40	10.22
Puerto Rico	0.71 (0.57-0.89)	13.21	9.93
Haiti	0.20 (0.08-0.50)	0.73	6.00
South Africa	0.63 (0.39-1.00)	2.91	8.75
Georgia	0.98 (0.58-1.66)	0.31	3.80
Georgia	1.56 (0.37-6.53)	2.30	8.40
Chicago	0.26 (0.07-0.92)	0.40	4.40
Chicago	0.25 (0.15-0.43)	2.25	8.37
Northern United States	0.41 (0.13-1.26)	23.75	10.12
Northern United States	0.46 (0.39-0.54)	0.50	5.05
United Kingdom	0.24 (0.18-0.31)	8.20	9.71

Canada	0.20 (0.09-0.49)	0.84	6.34
Combined relative risks (95% confidence interval)		0.65	0.49
		(0.60-0.70)	(0.35-0.70)

Cumulative meta-analysis

A useful way to show the accumulation of evidence over time is to perform a cumulative meta-analysis (Lau *et al*. 1992). Cumulative meta-analysis is defined as the repeated performance of meta-analysis whenever a new relevant trial becomes available for inclusion. This allows the retrospective identification of the point in time when a treatment effect first reached conventional levels of statistical significance.

Based on the systematic review by Freemantle *et al*. (1999), Fig. 6.14.5 shows mortality results from a cumulative meta-analysis of trials of beta-blockers in secondary prevention after myocardial infarction. A clear beneficial effect (*P* <0.001) was evident by the end of 1981. Subsequent trials in a further 15 000 patients simply confirmed this result. Similarly, Lau *et al*. (1992) showed that for the trials of intravenous streptokinase in acute myocardial infarction, a statistically significant (*P* = 0.01) combined difference in total mortality was achieved by 1973. The results of the subsequent 25 studies which included the large Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico-1 (GISSI-1) (Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI) 1986) and the Second International Study of Infarct Survival (ISIS-2) trials (ISIS-2 Collaborative Group 1988) and enrolled over 34 000 additional patients reduced the significance level to *P* = 0.001 in 1979, *P* = 0.0001 in 1986, and to *P* <0.00001 when the first mega-trial appeared, narrowing the confidence intervals around an essentially unchanged estimate of about 20 per cent reduction in the risk of death. This situation has been taken to suggest that further studies in large numbers of patients may be at best superfluous and costly if not unethical, once a statistically significant treatment effect is evident from meta-analysis of the existing smaller trials.

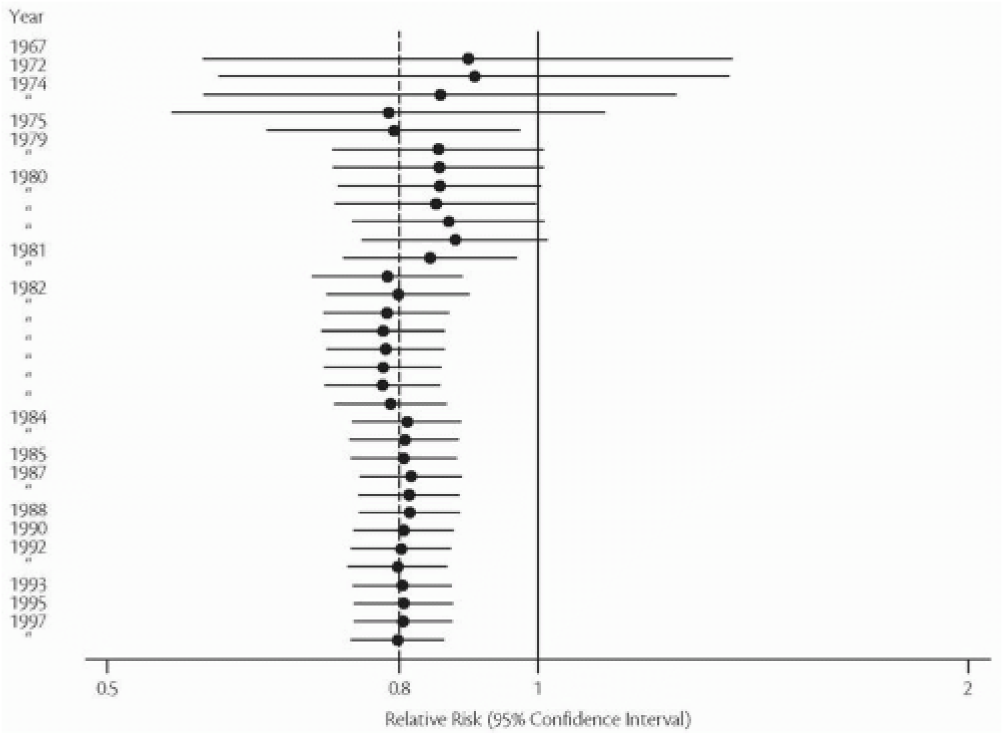

Fig. 6.14.5 Cumulative (fixed effects) meta-analysis of controlled trials of beta-blockers after myocardial infarction. A clear (*P* <0.001) reduction of mortality was evident by 1981.

Another application of cumulative meta-analysis has been to correlate the accruing evidence with the recommendations made by experts in review articles and textbooks. Antman *et al*. (1992) showed for thrombolytic drugs that recommendations for routine use first appeared in 1987, 14 years after a statistically significant (*P* = 0.01) beneficial effect became evident in cumulative meta-analysis. Conversely, the prophylactic use of lidocaine continued to be recommended for routine use in myocardial infarction despite the lack of evidence for any beneficial effect, and the possibility of a harmful effect being evident in the meta-analysis.

Bayesian meta-analysis

Some feel that a Bayesian approach to meta-analysis is more appropriate than the ‘classical’ approaches described above. Bayesian statisticians express their belief about the size of an effect by specifying some prior probability distribution before seeing the data— and then update that belief by deriving a posterior probability distribution, taking the data into account (Lilford & Braunholtz 1996). This is done by using Bayes theorem, named after the eighteenth century English clergyman Thomas Bayes. Bayesian models are available in both a fixed and random effects framework but published applications have usually been based on the random effects assumption. The confidence interval (or more correctly in Bayesian terminology: The 95 per cent credible interval which covers 95 per cent of the posterior probability distribution) will be slightly wider than that derived from using the conventional models. Bayesian approaches to meta-analysis can integrate other sources of evidence, for example, findings from observational studies or expert opinion, and are particularly useful for analysing the relationship between treatment benefit and underlying risk. The definition of prior probabilities may, however, involve subjective assessments and opinion, which runs against the principles of systematic review.

Deriving absolute measures of effect

The amount of between-study variability is usually lower for ratio than difference measures of treatment effects, so that meta-analyses are usually done using ratio measures. However, the absolute reduction in risk is a useful measure of the impact of treatment. For example, the relative risk of death associated with the use of beta-blockers after myocardial infarction is 0.80 (95 per cent confidence interval 0.74 to 0.86) (Fig. 6.14.3). The relative risk reduction, obtained by subtracting the relative risk from 1 and expressing the result as a percentage, is 20 per cent (95 per cent confidence interval 14 to 26 per cent). However, these relative measures ignore the underlying absolute risk. The risk of death among patients who have survived the acute phase of myocardial infarction varies widely.

The absolute risk reduction, or risk difference, reflects both the underlying risk without therapy and the risk reduction associated with therapy. Taking the reciprocal of the risk difference gives the number of patients who need to be treated to prevent one event, which is abbreviated to NNT or $\text{NNT}_{\text{benefit}}$ (Laupacis *et al*. 1988). The number of patients that need to be treated to harm one patient, denoted as NNH or, more appropriately, NNT_{harm} (Altman 1998) can also be calculated. It will usually be informative to calculate the risk difference, NNT or NNH for a range of baseline risks reflecting the range in the component studies of the meta-analysis.

Table 6.14.3 Beta-blockade in secondary prevention after myocardial infarction. Absolute risk reductions and numbers-needed-to-treat for 1 year to prevent one death, $\text{NNT}_{\text{benefit}}$, for different levels of control group mortality

One-year mortality risk among controls (%)	Absolute risk reduction	$NNT_{benefit}$
1	0.002	500
3	0.006	167
5	0.01	100
10	0.02	50
20	0.04	25
30	0.06	17
40	0.08	13
50	0.1	10

Calculations assume a constant relative risk reduction of 20 per cent.

For a baseline risk of 1 per cent per year, the absolute risk difference indicates that 2 deaths are prevented per 1000 treated patients (Table 6.14.3). This corresponds to 500 patients (1 divided by 0.002) treated for 1 year to prevent one death. Conversely, if the risk is above 10 per cent, less than 50 patients have to be treated to prevent one fatal event. Many clinicians would probably decide not to treat patients at very low risk, considering the large number of patients who would have to be exposed to the adverse effects of beta-blockade to prevent one death. Appraising the NNT from a patient's estimated risk without treatment, and the relative risk reduction with treatment, is a helpful aid when making a decision in an individual patient. A nomogram to determine NNTs at the bedside is available (Chatellier *et al*. 1996) and confidence intervals can be calculated (Altman 1998). The concept has been expanded to the number of health people needed to be screened to prevent one adverse outcome (Rembold 1998).

Combining absolute effect measures in meta-analysis is often inappropriate because the combined risk difference (and the NNT calculated from it) will be applicable only to patients at levels of risk corresponding to the typical control group risk of the trials analysed. It is generally more meaningful to use relative effect measures when summarizing the evidence while considering absolute measures when applying it to a specific clinical or public health situation.

Sources of bias in systematic reviews and meta-analysis

That there are limitations to the process of systematic review and meta-analysis is illustrated by meta-analyses that reviewed the same data but reached opposite conclusions. Examples include assessments of low molecular weight (LMW) heparins in the prevention of thrombosis following surgery (Leizorovicz *et al*. 1992; Nurmohamed *et al*. 1992) or of screening mammography (Kerlikowske *et al*. 1995; Gøtzsche & Olsen 2000). In the following sections, important sources of bias are discussed in detail.

Garbage in—garbage out?

The quality of component trials is of crucial importance: If the 'raw material' is flawed, then the findings of reviews of this material may also be compromised. The biases that threaten the validity of clinical trials relate to systematic differences in the patients' characteristics at baseline (*selection bias*), unequal provision of care apart from the treatment under evaluation (*performance bias*), biased assessment of outcomes (*detection bias*), and bias due to exclusion of patients after they have been allocated to treatment groups (*attrition bias*). Empirical evidence on specific trial characteristics associated with bias in intervention effect estimates has come from collections of meta-analyses assembled in so called 'meta-epidemiologic' studies. Several such studies have found that trials with inadequate allocation concealment or lack of blinding (see Box 6.14.3) tend to exaggerate estimates of intervention effects, compared with adequately concealed or adequately blinded trials.

Different ways of dealing (or not dealing) with the methodological quality of trials sometimes explain discrepancies in the results between different systematic reviews. For example, blinding of outcome assessments was important in trials comparing LMW weight heparin with standard heparin for the prevention of postoperative deep vein thrombosis: Trials that were not double-blind showed a benefit of LMW heparin that disappeared when restricting the analysis to trials with blinded outcome assessment (Jüni *et al*. 1999). This is not entirely surprising considering that the interpretation of fibrinogen leg scanning, which is used to detect thrombosis, can be subjective. One of the two reviews of LMW heparins mentioned came to discordant conclusions because the authors chose to ignore the quality of component trials, a practice that unfortunately is still fairly common (Gerber 2007).

A recent study of a large number of meta-analyses and trials found that the bias in intervention effect resulting from inadequate allocation concealment and lack of blinding, varied according to the type of outcome assessed (Wood 2008). There was little evidence of bias in trials with all-cause mortality outcomes, or other objectively assessed outcomes. In contrast, inadequate allocation concealment and lack of blinding were associated with over-optimistic estimates of intervention effects for subjectively assessed outcomes (Wood 2008). Efforts to minimize bias thus are particularly important when objective measurement of outcomes is not feasible.

Reporting biases

The dissemination of research findings is not a dichotomous event but a continuum ranging from the sharing of draft papers among colleagues, presentations at meetings, published abstracts, to papers in journals that are indexed in the major bibliographic databases. It has long been recognized that only a proportion of research projects reach full publication in an indexed journal and thus become easily identifiable for systematic review. For example, only about half of abstracts of studies presented at conferences are later published in full (von Elm *et al*. 2003). Reporting bias is introduced when the dissemination of research findings is influenced by the nature and direction of results. When discussing these reporting biases, which are summarized in Table 6.14.4, we will denote trials with statistically significant (*P* <0.05) and non-significant results trials as trials ‘positive’ and ‘negative’ results. However, the contribution made to the totality of the evidence by trials with non-significant results is of course potentially as important as that from trials with statistically significant results.

P.634

Table 6.14.4 Definitions of different reporting biases

Type of reporting bias	**Definition**
Publication bias	The *publication* or *non-publication* of research findings, depending on the nature and direction of the results
Time lag bias	The *rapid* or *delayed* publication of research findings, depending on the nature and direction of the results
Multiple (duplicate)	The *multiple* or *singular* publication of research findings, depending on

publication bias	the nature and direction of the results
Location bias	The publication of research findings in journals with different *ease of access* or *levels of indexing* in standard databases, depending on the nature and direction of results.
Citation bias	The *citation* or *non-citation* of research findings, depending on the nature and direction of the results
Language bias	The publication of research findings *in a particular language,* depending on the nature and direction of the results
Outcome reporting bias	The *selective reporting* of some outcomes but not others, depending on the nature and direction of the results

Publication bias

In a 1979 article 'The "file drawer problem" and tolerance for null results' Rosenthal (1979) described a gloomy scenario where 'the journals are filled with the 5 per cent of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95 per cent of the studies that show nonsignificant (e.g. $P > 0.05$) results'. The 'file drawer problem', more widely known as publication bias, has long been recognized in the social sciences: A review of psychology journals found that of 294 studies published in the 1950s, 97 per cent rejected the null hypothesis at the 5 per cent level ($P < 0.05$) (Sterling 1959). Similar results were later found for medical and public health journals. However, the proportion of all hypotheses tested for which the null hypothesis is truly false is unknown, and surveys of published results can only provide indirect evidence of publication bias. Direct evidence is available from studies of research proposals submitted to ethics committees or institutional review boards. Seven studies of proposals submitted to institutional committees in Oxford, Sydney, Baltimore, Bern, and to national ethics committees in France and Spain found rates of publication that ranged from 31 to 67 per cent (von Elm *et al*. 2008). Five of these studies compared the probability of publication of studies that produced positive results with those that did not. Meta-analysis of the results from these five studies indicates that the probability of publication is 2.6 times greater if results are statistically significant (odds ratio 2.6, 95 per cent confidence interval 2.0-3.4) (von Elm *et al*. 2008). These studies also showed that articles may appear in print many years after approval by the ethics committee, however, there is *time lag bias* (Table 6.14.4): The positive studies are published more rapidly than negative studies.

Other reporting biases

Among published studies, the probability of identifying relevant trials for a systematic review may also be influenced by their results (Table 6.14.4). *Multiple (duplicate) publication bias*, the production of multiple publications from single studies can lead to bias in a number of ways. Most importantly, studies with significant results are more likely to lead to multiple publications and presentations, which makes it more likely that they will be located and included in a meta-analysis. The inclusion of duplicated data may therefore lead to overestimation of treatment effects, as demonstrated for trials of the efficacy of ondansetron to prevent postoperative nausea (Tramèr 1997). It is not always obvious that multiple publications come from a single study, and one set of study participants may thus be included in an analysis twice. Indeed, it may be difficult if not impossible for reviewers to determine whether two papers represent duplicate publications of one trial or two separate trials: Two articles reporting the same trial may not share a single common author (von Elm *et al*. 2004).

The perusal of the reference lists of articles is widely used to identify other publications that may be relevant. However, retrieving literature by scanning reference lists may produce a biased sample of studies, introducing *citation bias*. Several studies have shown that trials with positive results tend to be cited more often than negative trials (Gøtzsche 1987). Of note, the association is not explained by superior methodological quality of cited articles. Sampson *et al*. (2003) found that trials published in journals that are indexed in EMBASE but not in MEDLINE tended to show smaller effects of treatments compared to trials indexed in MEDLINE or MEDLINE and EMBASE. *Location bias* may thus be introduced in systematic reviews exclusively based on MEDLINE searches, although another study found little difference in effect estimates between trials indexed and not indexed in MEDLINE (Egger *et al*. 2003).

Reviews may be exclusively based on trials published in English, although language restrictions have become less common in recent years (Gerber *et al*. 2007). Investigators working in a non-English speaking country publish some of their work in local journals. It is conceivable that authors are more likely to report in an international, English-language journal if results are positive whereas negative findings are published in a local journal (*language bias*). This has been demonstrated for the German language literature. When comparing pairs of articles published by the same first author, 63 per cent of trials published in English had produced significant (P <0.05) results as compared to 35 per cent of trials published in German (Egger *et al*. 1997b). However, when comparing the results of trials included in meta-analyses, trials published in languages other than English tended to show somewhat more beneficial effects of the intervention (Jüni *et al*. 2002).

It has been suspected for years that not only the reporting of entire studies with ‘positive’ results but also the inclusion or exclusion of outcomes within study reports are subject to selection mechanisms. Such *outcome reporting bias* has received much attention recently. Among studies approved by a research ethics committee in Denmark (Chan *et al*. 2004a) or funded by the Canadian Institutes of Health Research (Chan *et al*. 2004b) statistically significant outcomes were more likely to be reported than non-significant outcomes. Similarly, a review of published trials and survey of authors showed that incompletely or unreported outcomes were more likely to be non-significant than significant

(Chan & Altman 2005). Finally, several studies have shown that the reporting of adverse events and safety outcomes in clinical trials is often inadequate and selective (Ioannidis & Lau 2001; Melander *et al*. 2003).

How important are different sources of bias?

Empirical research has shown that the importance of reporting bias and bias due to inadequate quality of trials in a particular meta-analysis is often unpredictable, and the exploration of potential sources of bias is therefore an important step in any systematic review and meta-analysis. It is nevertheless worthwhile to study and compare the overall, average size and direction of different biases by analysing many meta-analyses. A 'meta-meta-analysis' of these studies found that, on average, published trials and trials published in languages other than English will overestimate treatment effects by about 10 per cent (Egger *et al*. 2002). Larger effects are seen for concealment of allocation and blinding: Trials with inadequate or unclear concealment and trials that are not double-blind overestimate treatment effects by about 30 and 15 per cent, respectively. In general, the quality of trials thus appears to be a more important source of bias than publication bias and other reporting biases (Egger *et al*. 2002). Of note, although this has improved in recent years, many meta-analyses do not assess the quality of component studies and explore the influence of study quality (Gerber *et al*. 2007; Moher *et al*. 2007).

Investigating and dealing with bias and heterogeneity

There will often be diverging opinions on the correct method for performing a particular meta-analysis. The robustness of the findings to different assumptions, the presence of bias, and possible sources of heterogeneity should therefore always be examined.

Sensitivity analysis

A thorough sensitivity analysis of the beta-blocker after myocardial infarction meta-analysis (Freemantle *et al*. 1999) is illustrated in Fig. 6.14.6. First, the overall effect was calculated by different statistical methods, using both a fixed and a random effects model. It is evident from the figure that the overall estimate is virtually identical and that confidence intervals are only slightly wider when using the random effects model. This is explained by the relatively small amount of between trial heterogeneity present in this meta-analysis.

Methodological quality was assessed in terms of concealment of allocation of study participants to beta-blocker or control groups and blinding of patients and investigators. Figure 6.14.6 shows that the estimated treatment effect was similar in studies with concealment of treatment allocation, or studies that were described as double-blind. Publication bias is more likely to affect small studies and may therefore be examined by stratifying the analysis by study size. If publication bias is present, it is expected that of published studies, the larger ones will report the smaller effects: Smaller effects can be statistically significant in larger studies. The figure shows that this is indeed the case, with the 11 smallest trials (25 deaths or less) showing the largest effect. However, exclusion of the smaller studies has little effect on the overall estimate. Studies varied in terms of length of follow-up but this again had little effect on estimates. Finally, two trials were terminated earlier than anticipated on the grounds of the results from interim analyses. Estimates of treatment effects from trials which were stopped early because of a significant treatment difference are liable to overestimate treatment effects (Montori 2005). Bias may thus be introduced in a meta-analysis which includes such trials. Exclusion of these trials again affected the overall estimate only marginally.

The sensitivity analysis thus shows that the results from the beta-blocker meta-analysis are robust to the choice of the statistical method and to the exclusion of trials of lesser quality or of studies terminated early. It also suggests that publication bias is unlikely to have distorted its findings.

Funnel plots

Funnel plots are scatter plots in which the treatment effects estimated from individual studies on the horizontal axis are plotted against a measure of study size on the vertical axis. Such plots have long been proposed as a means of detecting publication bias (Light & Pillemer 1984). In the absence of bias, the plot should resemble a symmetrical inverted funnel, with the results of smaller studies being more widely scattered than those of the larger studies. If the plot shows an asymmetrical shape, publication bias may be present. This usually takes the form of a gap in the wide part of the funnel which indicates the absence of negative small studies. The funnel plot for the meta-analysis of the trials of beta-blockade in secondary prevention after myocardial infarction is shown in the upper panel of Fig. 6.14.7. The plot is fairly symmetrical. In contrast, the funnel plot of controlled trials of magnesium infusion in acute myocardial infarction (lower panel of Fig. 6.14.7) is clearly asymmetrical. This is an example where publication bias may explain the discrepancy between meta-analyses of smaller trials, which showed a clear treatment effect, and a very large trial (the ISIS-4 trial, ISIS-4 (Collaborative Group 1995) that showed no effect (Egger & Davey Smith 1995).

Funnel plot asymmetry does not prove the presence of bias in a meta-analysis: In interpreting funnel plots, reviewers should consider the different reasons for asymmetry listed in Table 6.14.5 (Egger *et al*. 1997a). Other types of bias can lead to asymmetry. Smaller studies are, on average, conducted and analysed with less methodological rigour than larger studies, and trials of lower quality tend to show larger effects. Heterogeneity between the treatment effects in different trials may lead to funnel plot asymmetry if the true treatment effect is larger in the smaller trials. Interventions may have been implemented less thoroughly in larger trials, thus explaining the more positive results in smaller trials. This is likely, for example, in trials of complex interventions in chronic diseases, such as rehabilitation after stroke or multifaceted interventions in diabetes mellitus. Thus the funnel plot should be seen as a generic means of examining ‘small study effects’ (the tendency for the smaller studies in a meta-analysis to show larger treatment effects) (Sterne *et al*. 2000). Other graphical representations, discussed in detail elsewhere, are useful to investigate bias and heterogeneity. These include Galbraith plots (Galbraith 1988) and L'Abbé plots (Song 1999).

Tests for funnel plot asymmetry

Visual inspection of funnel plots is inherently subjective. Statistical tests for funnel plot asymmetry, which examine whether the association between estimated intervention effects and a measure of study size is greater than expected by chance, can be useful to assess and quantify the evidence for asymmetry. For continuous outcomes this is straightforward. We can perform a weighted linear regression of the intervention effect estimates, for example, differences in blood pressure, against their standard errors (Egger *et al*. 1997a). When outcomes are dichotomous, this approach is more problematic because the standard error of the log odds ratio is mathematically linked to the size of the odds ratio, even in the absence of small study effects (Sterne *et al*. 2000). Simulation studies have shown that this may lead to false-positive test results, particularly when interventions have very large effects (Sterne *et al*. 2000; Peters *et al*. 2006). Tests that avoid the association between the log odds ratio and its standard error are therefore preferred for binary outcomes, including the modified test proposed by Harbord *et al*. (2006) or the approach proposed by Peters *et al*. (2006).

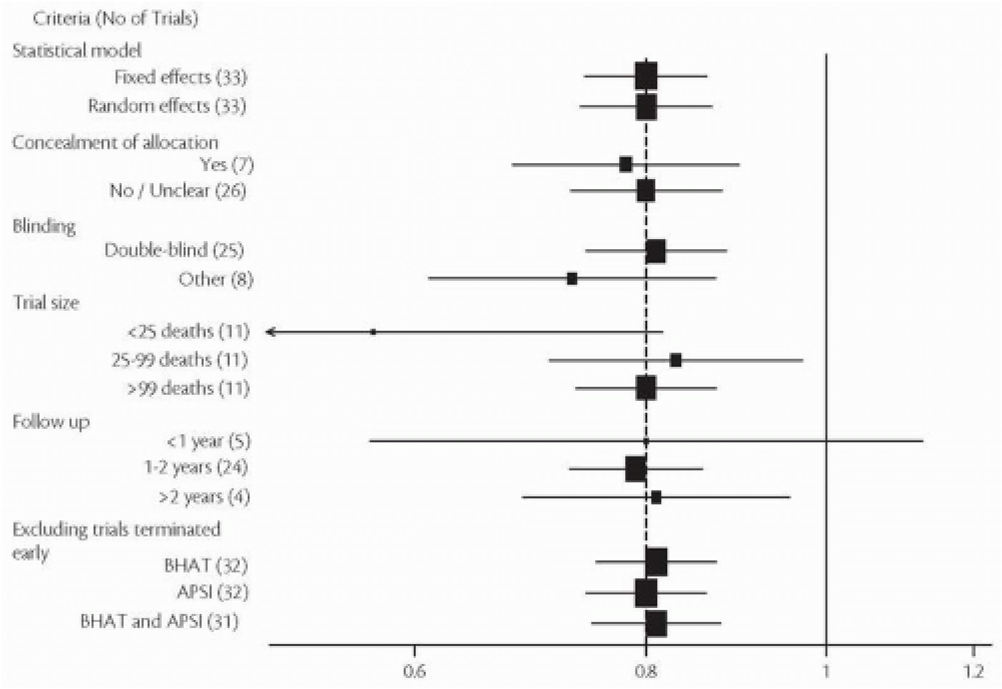

Fig. 6.14.6 Sensitivity analyses examining the robustness of the effect on total mortality of beta-blockers in secondary prevention after myocardial infarction. The dotted vertical line corresponds to the combined relative risk from the fixed effects model (0.8).

Table 6.14.5 Possible sources of asymmetry in funnel plots

1. Selection biases

    Publication bias

        Delayed publication (also known as ‘time-lag’ or ‘pipeline’) bias

        Location biases

            Language bias

Citation bias

Multiple publication bias

Selective outcome reporting

2. Poor methodological quality leading to spuriously inflated effects in smaller studies

Poor methodological design

Inadequate analysis

Fraud

3. True heterogeneity

Size of effect differs according to study size (for example, due to differences in the intensity of interventions or differences in underlying risk between studies of different sizes)

4. Artefactual

In some circumstances, sampling variation can lead to an association between the intervention effect and its standard error.

5. Chance

The power of tests will often be limited because many meta-analyses include few trials only (Gerber *et al*. 2007; Ioannidis & Trekalinos 2007) Therefore, even when a test does not provide evidence of funnel plot asymmetry, bias cannot be excluded with confidence. As a rule of thumb, tests for funnel plot asymmetry should be used only when there are at least 10 studies, and they should not be used if substantial heterogeneity is present between trials. Results should be interpreted in the light of visual inspection of the funnel plot. Do small studies tend to lead to more or less beneficial intervention effect estimates? Are there outliers with markedly different intervention effect estimates, or studies that are highly influential in the meta-analysis? When there is evidence of small-study effects, publication bias should be considered as one of a number of possible explanations (see Table 6.14.5). Although funnel plots, and tests for funnel plot asymmetry, may alert review authors to a problem, they do not provide a solution.

Correcting for publication bias?

The process that determines which results are published and which are not can be modelled (Iyengar & Greenhouse 1988) and such ‘selection models’ have been extended to estimate treatment effects corrected for the estimated publication bias. For example, Copas and Shi (2000) used such an approach to show that in epidemiological studies of passive smoking and lung cancer, allowing for the possibility of publication bias reduces the estimate of relative risk associated with passive smoking. The ‘trim and fill’ method is a graphical approach to identify and correct for publication bias

(Duval & Tweedie 2000). The method ‘trims’ (removes) the smaller studies causing funnel plot asymmetry, estimates the ‘true’ centre of the funnel, and then ‘fills’ (replaces) the omitted studies and their missing counterparts around the centre.

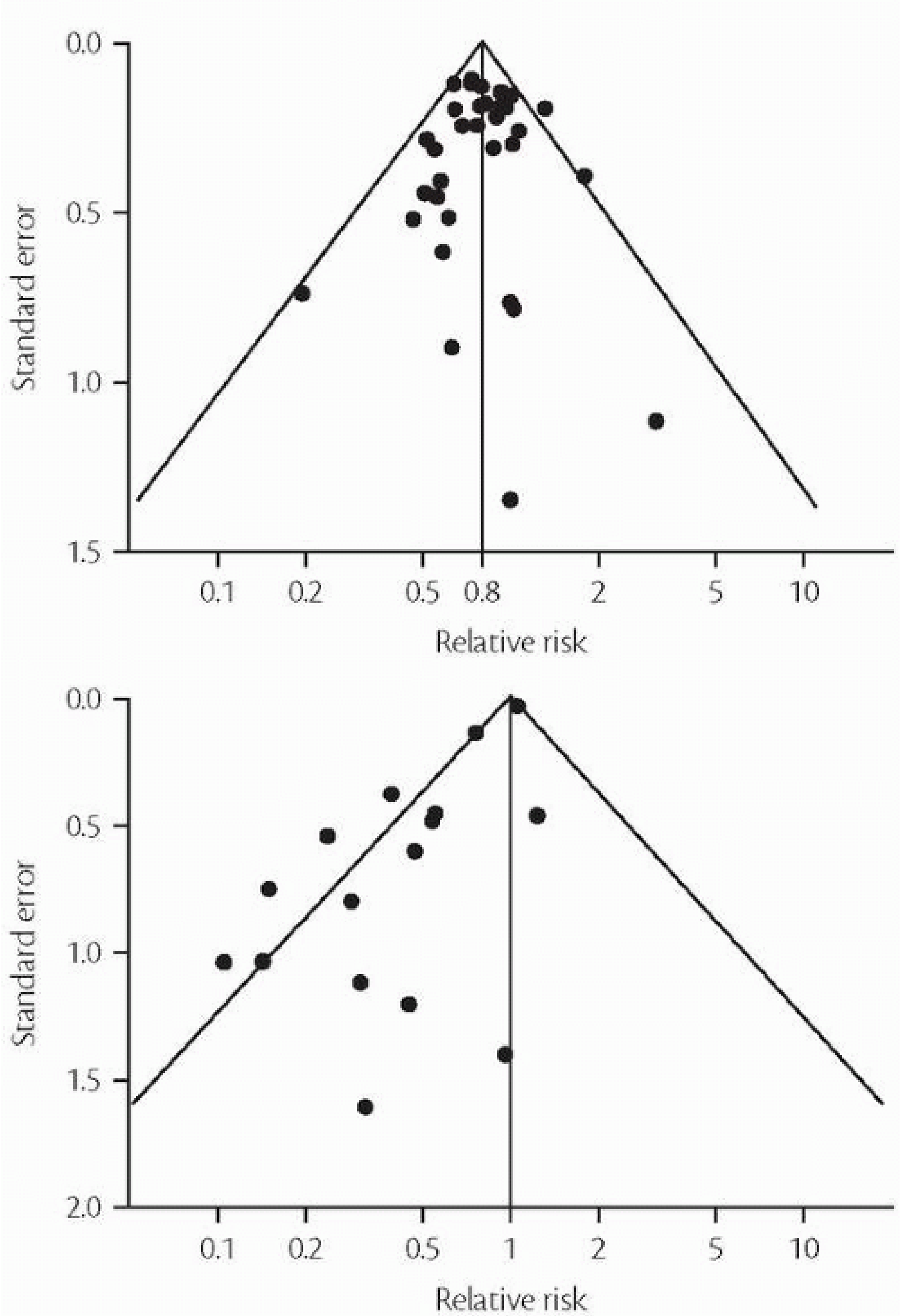

Fig. 6.14.7. Funnel plots of trials of beta-blockers in secondary prevention after myocardial infarction (upper panel) and of trials of magnesium infusion in acute myocardial infarction (lower panel). The relative risk is plotted on a logarithmic scale, to ensure that effects of the same magnitude but opposite directions will be equidistant from 1.0. Plotting against the standard error of the treatment effect emphasizes differences

between the smaller studies among which publication and other biases are most likely to occur. The vertical line shows the summary estimate from the fixed effects model, diagonal lines show the expected 95 per cent confidence intervals around the summary estimate.

Whatever the method used, correcting for publication bias is problematic because the true mechanism for publication bias is unknown. Equally importantly, other reasons for small study effects other than publication bias are ignored. Therefore, ‘corrected’ estimates from these methods should be interpreted with great caution. Sensitivity analyses excluding smaller studies, and studies of lower methodological quality (see above and Fig. 6.14.6), will often be more helpful than either a *P* value from a test of funnel plot asymmetry, or an estimate from a statistical model that is supposedly ‘corrected’ for publication bias.

Spurious precision? Meta-analysis of observational studies

The randomized controlled trial is the principal research design in the evaluation of medical interventions. However aetiological hypotheses, for example, those relating common exposures to the occurrence of disease, cannot generally be tested in randomized experiments. Does breathing other people's tobacco smoke propagate the development of lung cancer, drinking coffee cause coronary heart disease, and eating a diet rich in unsaturated fat induce breast cancer? Studies of such ‘menaces of daily life’ (Feinstein 1988) employ observational designs, or examine the presumed biological mechanisms in the laboratory. In these situations the risks involved are generally small, but once a large proportion of the population is exposed, the potential public health implications of these associations—if they are causal—can be striking.

Analyses of observational data also have a role in medical effectiveness research. The evidence that is available from clinical trials will rarely answer all the important questions. Most trials are conducted to establish efficacy and safety of a single agent in a specific clinical situation. Due to the limited size of such trials, less common adverse effects of drugs may only be detected in case-control studies, or in analyses of databases from post-marketing surveillance schemes. Also, because follow-up is generally limited, adverse effects occurring many years later will not be identified. If years later established interventions are incriminated with adverse effects, there will be ethical, political, and legal obstacles to the conduct of a new trial. Examples for such situations include the controversy surrounding intramuscular administration of vitamin K to newborns and the risk of childhood cancer, or oral contraceptive use and breast cancer.

Meta-analysis, by promising a precise and definite answer when the magnitude of the underlying risks are small, or when the results from individual studies disagree, appears an attractive proposition both in aetiological studies and in observational effectiveness research.

Confounding, residual confounding and bias

The overall effect calculated from a group of sensibly combined and representative randomized trials will provide an essentially unbiased estimate of the treatment effect, with an increase in the precision of this estimate. A fundamentally different situation arises in the case of observational studies. Such studies yield estimates of association which may deviate from true underlying relationships beyond the play of chance. This may be due to the effects of confounding factors, the influence of biases, or both.

Those exposed to the factor under investigation may differ in a number of other aspects that are relevant to the risk of developing the disease in question. Consider, for example, smoking as a risk factor for suicide. Virtually all cohort studies have shown a positive association, with a dose response relationship being evident between the amount smoked and the probability of committing suicide. A meta-analysis of these cohorts produces very precise and statistically significant estimates of the increase in suicide risk that is associated with smoking different daily amounts of cigarettes: Relative rate for 1-14 cigarettes 1.43 (95 per cent confidence interval 1.06-1.93); for 15-24 cigarettes 1.88 (95 per cent confidence interval 1.53-2.32); 25 or more cigarettes 2.18 (95 per cent confidence interval 1.82-2.61) (Egger *et al*. 1998).

Based on established criteria, many would consider the association to be causal—if only it were more plausible. Indeed, it is improbable that smoking is causally related to suicide. Rather, it is the social and mental states predisposing to suicide that are also associated with the habit of smoking (Davey Smith *et al*. 1992). Factors that are related to both the exposure and the disease under study, confounding factors, may thus distort results. If the factor is known and has been measured, the usual approach is to control for

its influence in the analysis. For example, any study assessing the influence of coffee consumption on the risk of myocardial infarction should control for smoking, since smoking is generally associated with drinking larger amounts of coffee and smoking is a cause of coronary heart disease. However, even if adjustments for confounding factors have been made in the analysis, residual confounding remains a potentially serious problem in observational research. Residual confounding arises whenever a confounding factor cannot be measured with sufficient precision—a situation that often occurs in epidemiological studies (Phillips & Davey Smith 1991). Confounding is the most important threat to the validity of results from cohort studies whereas more difficulties, in particular selection biases, arise in case-control studies.

Plausible but equally spurious findings?

Implausibility of results, like in the case of smoking and suicide, rarely protects us from reaching misleading claims. It is generally easy to produce plausible explanations for the findings from observational research. For example, observational studies have consistently shown that people eating more fruits and vegetables, which are rich in beta-carotene, and people having higher serum beta-carotene concentrations have lower rates of cardiovascular disease and cancer (Jha *et al*. 1995). Beta-carotene has antioxidant properties and could thus plausibly be expected to prevent carcinogenesis and atherogenesis by reducing oxidative damage to DNA and lipoproteins

(Jha *et al*. 1995). Contrary to many other associations found in observational studies, this hypothesis could be, and was, tested in experimental studies.

A meta-analysis of the findings for cardiovascular mortality, comparing the results from six observational studies with those from four randomized trials is shown in Fig 6.14.8. For observational studies results relate to a comparison between groups with high and low beta-carotene intake or serum beta-carotene level, whereas in trials participants randomized to beta-carotene supplements were compared with participants randomized to placebo. The meta-analysis of the cohort studies shows a significantly lower risk of cardiovascular death (relative risk 0.69, 95 per cent confidence interval 0.59-0.80, *P* <0.0001). The results from the randomized trials, however, indicate a moderate adverse effect of beta-carotene supplementation (relative risk 1.12, 95 per cent confidence interval 1.04-1.22, *P* = 0.005). Similarly discrepant results between epidemiological studies and trials were observed for other antioxidant vitamins (Lawlor *et al*. 2004). This example illustrates that in some meta-analyses of observational studies, the analyst may well be simply producing narrow confidence intervals around spurious results. This is particularly a problem in the study of dietary factors because dietary habits are associated with a large number of lifestyle factors as well as socioeconomic position.

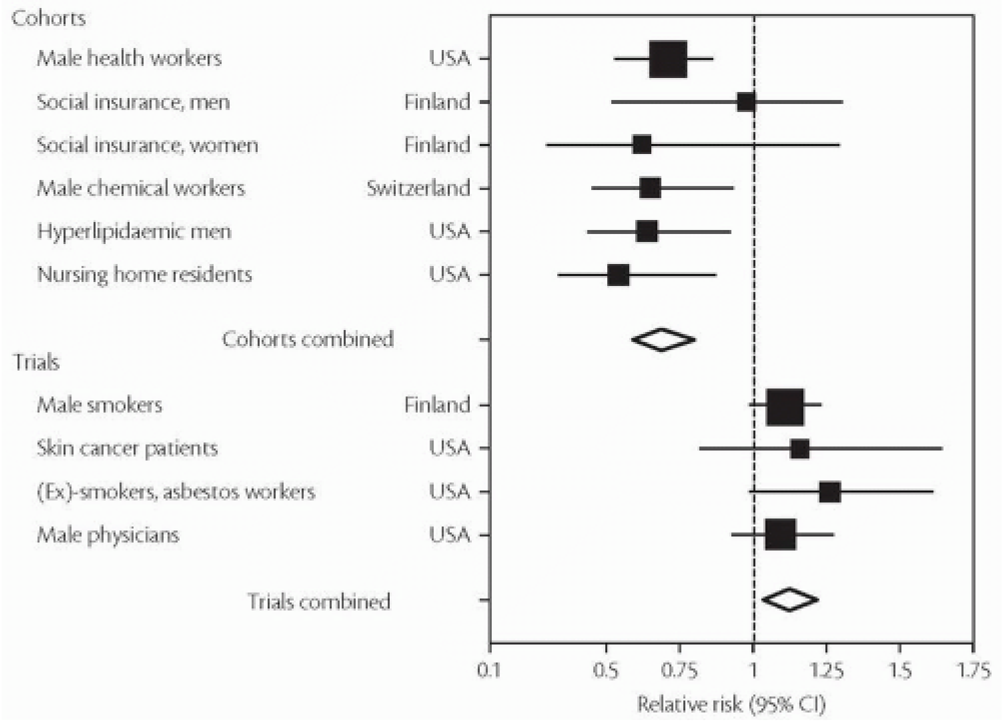

Fig. 6.14.8 Meta-analysis of the association between beta-carotene intake and cardiovascular mortality: The results from observational studies are compared to the findings of four large trials. The observational studies indicate considerable benefit whereas the findings from randomized controlled trials show an increase in the risk of death. Meta-analysis by fixed-effects model.

Exploring sources of heterogeneity

Some observers suggest that meta-analysis of observational studies should be abandoned altogether (Shapiro 1994). We disagree, but think that thorough consideration of possible sources of heterogeneity between observational study results will often provide more insights than the mechanistic calculation of an overall measure of effect, which will often be biased. An example relating to diet and breast cancer is depicted in Fig. 6.14.9. The hypothesis from ecological analyses (Armstrong & Doll 1975) that higher intake of saturated fat could increase the risk of breast cancer generated much observational research, often with contradictory results. A meta-analysis (Boyd *et al*. 1993) showed an association for case-control but not for cohort studies (odds ratio 1.36 for case-control studies versus relative rate 0.95 for cohort studies, comparing highest with lowest category of saturated fat intake, $P = 0.0002$ for difference, Fig. 6.14.9). The most likely explanation for this situation is that biases in the recall of dietary items, and in the selection of study participants, have produced a spurious association in the case-control comparisons.

Conclusions

Systematic review including, if appropriate, a formal meta-analysis is clearly superior to the narrative approach to reviewing research. Systematic reviews involve structuring the processes through which a thorough review of previous research is carried out. The issues of the completeness of the evidence identified, the quality of component studies, and the combinability of evidence are made explicit. The unprecedented effort to inject scientific principles into the process of research synthesis, which has taken place over the past decade, has improved the quality of reviews published in recent years (Gerber *et al*. 2007) although considerable room for further improvement remains (Moher *et al*. 2007). Some shortcomings of systematic review and meta-analysis are, however, a consequence of more general failings. Although various initiatives, including reporting guidelines for randomized trials (Altman 2001) and observational studies (Vandenbroucke 2007) mean that the identification and assessment of studies have become an easier task, the conduct of research and dissemination of its results continue to be an imperfect process. Finally, the suggestion that formal meta-analysis of observational studies can be misleading and that insufficient attention is often given to heterogeneity does not mean that a return to the previous practice of highly subjective narrative reviews is called for. Many of the principles of systematic reviews remain: A study protocol should be written in advance, complete literature searches should be carried out, and studies selected and data extracted in a reproducible and objective fashion. This allows for differences and similarities of the results found in different settings to be inspected, hypotheses to be formulated, and the need for future studies, including randomized controlled trials, to be defined. In summary:

- Systematic reviews allow for a more objective and reproducible appraisal of the evidence than traditional narrative reviews. The definition of eligibility criteria for trials to be included, a comprehensive search for such trials, and an assessment of their methodological quality are central components.
- Meta-analysis, if appropriate, will enhance the precision of estimates of treatment effects, leading to reduced probability of false negative results, and potentially to a more timely introduction of effective interventions. Meta-analysis is a two-stage process involving the calculation of an appropriate summary statistic for each of a set of studies followed by the combination of these statistics into a weighted average.
- Inadequate quality of studies and publication bias and other reporting biases may distort results. Concealment of treatment allocation, blinding of outcome assessment, and handling of patient attrition should be assessed and funnel plots should be examined.
- The thoughtful consideration of heterogeneity between study results is an important aspect of systematic reviews, and particularly important in systematic reviews and meta-analyses of observational studies.

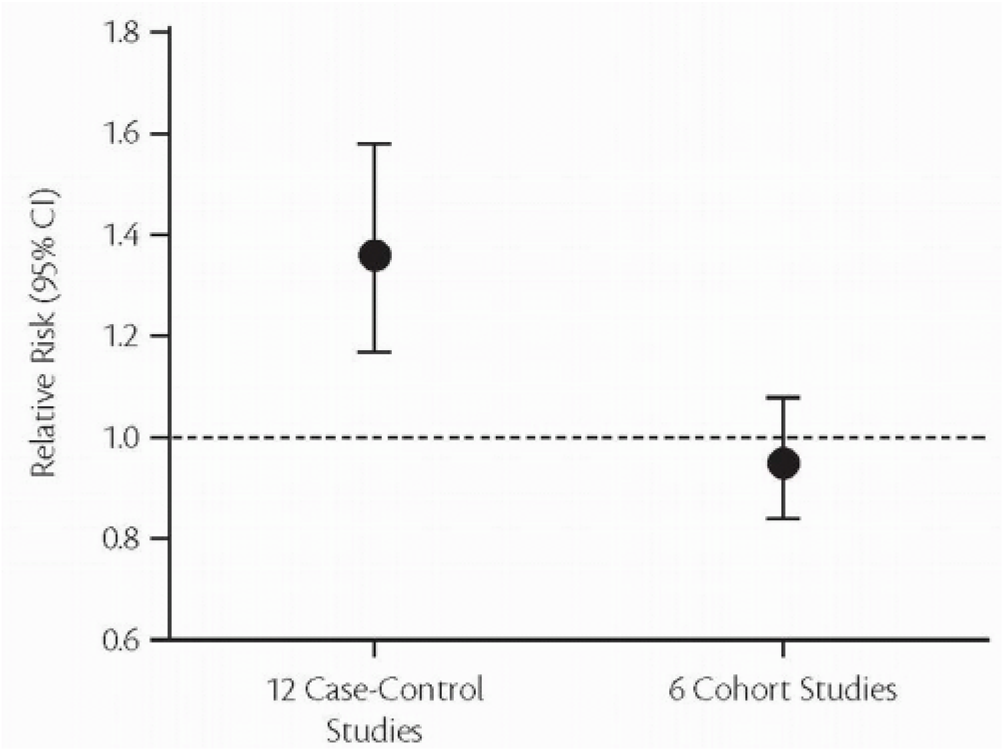

Fig. 6.14.9. An example of heterogeneity in a meta-analysis of observational studies: Saturated fat intake and cancer.

References

Altman, D.G. (1998). Confidence intervals for the number needed to treat. *BMJ*, **317**, 1309-12.

Altman, D.G., Schulz, K.F., Moher, D. *et al*. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine*, **134**, 663-94.

Antman, E.M., Lau, J., Kupelnick, B. *et al*. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *JAMA*, **268**, 240-8.

Armstrong, B. and Doll, R. (1975). Environmental factors and cancer incidence and mortality in different countries with special reference to dietary practices. *International Journal of Cancer*, **15**, 617-31.

Baber, N.S., Wainwright Evans, D., Howitt, G. *et al*. (1980). Multicentre post-infarction trial of propranolol in 49 hospital in the United Kingdom, Italy and Yugoslavia. *British Heart Journal*, **44**, 96-100.

Boyd, N.F., Martin, L.J., Noffel, M. *et al*. (1993). A meta-analysis of studies of dietary fat and breast cancer. *British Journal of Cancer*, **68**, 627-36.

Busse, J.W., Mills, E., Dennis, R. *et al*. (2009). Clinical epidemiology. In *Oxford Textbook of Public Health* (eds. R. Detels, R. Beaglehole, M.A. Lansang and M. Gulliford), Oxford University Press, Oxford.

Chalmers, I. (1979). Randomised controlled trials of fetal monitoring 1973-1977. In *Perinatal medicine* (eds. O. Thalhammer, K. Baumgarten, and A. Pollak), p. 260. Thieme, Stuttgart.

Chalmers, I. and Tröhler, U. (2000). Medical and philosophical commentaries, 1773-1795: A 200-year old response to the challenge of keeping abreast of the medical literature. *Annals of Internal Medicine*, **133**, 238-43.

Chan, A.W. and Altman, D.G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *BMJ*, **330**, 753.

Chan, A.W., Hrobjartsson, A., Haahr, M.T. *et al*. (2004a). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, **291**, 2457-65.

Chan, A.W., Krleza-Jeric, K., Schmid, I. *et al*. (2004b). Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal*, **171**, 735-40.

Chatellier, G., Zapletal, E., Lemaitre, D. *et al*. (1996). The number needed to treat: A clinically useful nomogram in its proper context. *BMJ*, **312**, 426-9.

Colditz, G.A., Brewer, T.F., Berkley, C.S. *et al*. (1994). Efficacy of BCG vaccine in the prevention of Tuberculosis. *JAMA*, **271**, 698-702.

Copas, J.B. and Shi, J.Q. (2000). Reanalysis of epidemiological evidence on lung cancer and passive smoking. *BMJ*, **320**, 417-18.

Davey Smith, G., Phillips, A.N., and Neaton, J.D. (1992). Smoking as 'independent' risk factor for suicide: illustration of an artifact from observational epidemiology. *Lancet*, **340**, 709-11.

De Angelis, C.D., Drazen, J.M., Frizelle, F.A. *et al*. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *JAMA*, **292**, 1363-4.

Deeks, J.J., Altman, D.G., and Bradburn, M.J. (2001). Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In *Systematic reviews in health care: Meta-analysis in context* (eds. M. Egger, D.G. Smith, and D.G. Altman),. p. 285. BMJ Books, London.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177-88.

Duval, S., and Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**, 455-63.

Egger, M. and Davey Smith, G. (1995). Misleading meta-analysis. Lessons from "an effective, safe, simple" intervention that wasn't. *BMJ*, **310**, 752-4.

Egger, M., Davey Smith, G., Schneider, M. *et al*. (1997a). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, **315**, 629-34.

Egger, M., Zellweger-Zähner, T., Schneider, M. *et al*. (1997b). Language bias in randomised controlled trials published in English and German. *Lancet*, **350**, 326-9.

Egger, M., Schneider, M., and Davey Smith, G. (1998). Spurious precision? Meta-analysis of observational studies. *BMJ*, **316**, 140-5.

Egger, M., Ebrahim, S., and Smith, G.D. (2002). Where now for meta-analysis? *International Journal of Epidemiology*, **31**, 1-5.

Egger, M., Jüni, P., Bartlett, C. *et al*. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment*, **7**, 1-76.

Feinstein, A.R. (1988). Scientific standards in epidemiological studies of the menace of daily life. *Science*, **242**, 1257-63.

Freemantle, N., Cleland, J., Young, P. *et al*. (1999). Beta blockade after myocardial infarction: systematic review and meta regression analysis. *BMJ*, **318**, 1730-7.

Freiman, J.A., Chalmers, T.C., Smith, H. *et al*. (1992). The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial. In *Medical uses of statistics (2)* (eds. J.C. Bailar and F. Mosteller), p. 357. NEJM Books, Boston, MA.

Galbraith, R. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, **7**, 889-94.

Gerber, S., Tallon, D., Trelle, S. *et al*. (2007). Bibliographic study showed improving methodology of meta-analyses published in leading journals 1993-2002. *Journal of Clinical Epidemiology*, **60**, 773-80.

Gøtzsche, P.C. (1987). Reference bias in reports of drug trials. *BMJ*, **295**, 654-6.

Gøtzsche, P.C. and Olsen, O. (2000). Is screening for breast cancer with mammography justifiable? [see comments]. *Lancet*, **355**, 129-34.

Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI) (1986). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet*, **1**, 397-402.

Hampton, J.R. (1981). The use of beta blockers for the reduction of mortality after myocardial infarction. *European Heart Journal*, **2**, 259-68.

Harbord, R.M., Egger, M., and Sterne, J.A. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, **25**, 3443-57.

Higgins, J.P., Thompson, S.G., Deeks, J.J. *et al*. (2003). Measuring inconsistency in meta-analyses. *BMJ*, **327**, 557-60.

Ioannidis, J.P. and Lau, J. (2001). Completeness of safety reporting in randomized trials: An evaluation of 7 medical areas. *JAMA*, **285**, 437-43.

Ioannidis, J.P. and Trikalinos, T.A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, **176**, 1091-6.

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet*, **2**, 349-60.

ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group (1995). ISIS-4: A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction. *Lancet*, **345**, 669-87.

Iyengar, S. and Greenhouse, J.B. (1988). Selection models and the file drawer problem. *Statistical Science*, **3**, 109-35.

Jenicek, M. (1989). Meta-analysis in medicine. Where we are and where we want to go. *Journal of Clinical Epidemiology*, **42**, 35-44.

Jha, P., Flather, M., Lonn, E. *et al*. (1995). The antioxidant vitamins and cardiovascular disease. *Annals of Internal Medicine*, **123**, 860-72.

Jüni, P., Witschi, A., Bloch, R. *et al*. (1999). The hazards of scoring the quality of clinical trial for meta-analysis. *JAMA*, **282**, 1054-60.

Jüni, P., Altman, D.G., and Egger, M. (2001). Assessing the quality of controlled clinical trials. *BMJ*, **323**, 42-6.

Jüni, P., Holenstein, F., Sterne, J. *et al*. (2002). Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *International Journal of Epidemiology*, **31**, 115-23.

Jüni, P., Nartey, L., Reichenbach, S. *et al*. (2004). Risk of cardiovascular events and rofecoxib: Cumulative meta-analysis. *Lancet*, **364**, 2021-9.

Kerlikowske, K., Grady, D., Rubin, S.M. *et al*. (1995). Efficacy of screening mammography. A meta-analysis [see comments]. *JAMA*, **273**, 149-54.

Lau, J., Antman, E.M., Jimenez-Silva, J. *et al*. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. The *New England Journal of Medicine*, **327**, 248-54.

Laupacis, A., Sackett, D.L., and Roberts, R.S. (1988). An assessment of clinically useful measures of the consequences of treatment. *The New England Journal of Medicine*, **318**, 1728-33.

Lawlor, D.A., Davey, S.G., Kundu, D. *et al.* (2004). Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? *Lancet*, **363**, 1724-7.

Leizorovicz, A., Haugh, M.C., Chapuis, F.R. *et al.* (1992). Low molecular weight heparin in prevention of perioperative thrombosis. *BMJ*, **305**, 913-20.

Light, R.J. and Pillemer, D.B. (1984). *Summing up. The science of reviewing research*. Harvard University Press, Cambridge, Massachusetts, and London, England.

Lilford, R.J. and Braunholtz, D. (1996). The statistical basis of public policy: A paradigm shift is overdue. *BMJ*, **313**, 603-7.

MacDonald, D., Grant, A., Sheridan-Pereira, M. *et al.* (1985). The Dublin randomised controlled trial of intrapartum fetal heart rate monitoring. *American Journal of Obstetrics and Gynecology*, **152**, 524-39.

Melander, H., Ahlqvist-Rastad, J., Meijer, G. *et al.* (2003). Evidence b(i)ased medicine—selective reporting from studies sponsored by pharmaceutical industry: Review of studies in new drug applications. *BMJ*, **326**, 1171-3.

Mitchell, J.R.A.(1981). Timolol after myocardial infarction: an answer or a new set of questions? *BMJ*, **282**, 1565-70.

Moher, D., Tetzlaff, J., Tricco, A.C. *et al.* (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine*, **4**, e78.

Montori, V.M., Devereaux, P.J., Adhikari, N.K. *et al.* (2005). Randomized trials stopped early for benefit: a systematic review. *JAMA*, **294**, 2203-9.

Mulrow, C.D. (1987). The medical review article: State of the science. *Annals of Internal Medicine*, **106**, 485-8.

Multicentre International Study: Supplementary report (1977). Reduction in mortality after myocardial infarction with long-term betaadrenoceptor blockade. *BMJ*, **2**, 419-21.

Nurmohamed, M.T., Rosendaal, F.R., Bueller, H.R. *et al.* (1992). Lowmolecular-weight heparin versus standard heparin in general and orthopaedic surgery: A meta-analysis. *Lancet*, **340**, 152-6.

O'Farrell, N. and Egger, M. (2000). Circumcision in men and the prevalence of HIV infection: a meta-analysis revisited. *International Journal of STD & AIDS*, **11**, 137-42.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *BMJ*, **3**, 1243-6.

Peters, J.L., Sutton, A.J., Jones, D.R. *et al*. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, **295**, 676-80.

Phillips, A.N. and Davey Smith, G. (1991). How independent are ‘independent’ effects? Relative risk estimation when correlated exposures are measured imprecisely. *Journal of Clinical Epidemiology*, **44**, 1223-31.

Rembold, C.M. (1998). Number needed to screen: Development of a statistic for disease screening. *BMJ*, **317**, 307-12.

Reynolds, J.L. and Whitlock, R.M.L. (1972). Effects of a beta-adrenergic receptor blocker in myocardial infarctation treated for one year from onset. *British Heart Journal*, **34**, 252-9.

Rosenthal, R. (1979). The ‘file drawer problem’ and tolerance for null results. *Psychological Bulletin*, **86**, 638-41.

Sampson, M., Barrowman, N.J., Moher, D. *et al*. (2003). Should metaanalysts search Embase in addition to Medline? *Journal of Clinical Epidemiology*, **56**, 943-55.

Schulz KF (1996). Randomised trials, human nature, and reporting guidelines. *Lancet*, **348**, 596-8.

Shapiro, S. (1994). Meta-analysis/Shmeta-analysis. *American Journal of Epidemiology*, **140**, 771-8.

Song, F. (1999). Exploring heterogeneity in meta-analysis: Is the L'Abbé Plot useful? *Journal of Clinical Epidemiology*, **52**, 725-30.

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, **54**, 30-4.

Sterne J.A.C., Gavaghan D.J., and Egger M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, **53**, 1119-29.

Sterne, J.A.C. Egger, M., and Sutton, A.J. (2001). Meta-analysis software. In *Systematic Reviews in Health Care: Meta-Analysis in Context* (eds. M. Egger, D.G. Smith, and D.G. Altman), p. 336. BMJ Books, London.

The Norwegian Multicenter Study Group (1981). Timolol-induced reduction in mortality and reinfarction in patients surviving acute myocardial infarction. *The New England Journal of Medicine*, **304**, 801-7.

Thornley, B. and Adams, C. (1998). Content and quality of 2 000 controlled trials in schizophrenia over 50 years. *BMJ*, **317**, 1181-4.

Tramèr, M.R., Reynolds, D.J.M., Moore, R.A. *et al*. (1997). Impact of covert duplicate publication on meta-analysis: a case study. *BMJ*, **315**, 635-40.

Vandenbroucke, J.P., von Elm, E., Altman, D.G. *et al*. (2004). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Medicine*, **4**, e297.

von Elm, E., Costanza, M.C., Walder, B. *et al*. (2003). More insight into the fate of biomedical meeting abstracts: a systematic review. *BMC Medical Research Methodology*, **3**, 12.

von Elm, E., Poglia, G., Walder, B. *et al*. (2004). Different patterns of duplicate publication: An analysis of articles used in systematic reviews. *JAMA*, **291**, 974-80.

von Elm, E., Röllin, A., Blümle, A. *et al*. Publication and non-publication of clinical trials: Longitudinal study of applications submitted to a research ethics committee. *Swiss Medical Weekly*, 2008; **138**:197-203.

Wood, L., Egger, M., Gluud, L.L. *et al*. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *BMJ*, 2008; **336**:601-5.