2 Principles of and procedures for systematic reviews

MATTHIAS EGGER, GEORGE DAVEY SMITH

Summary points

- Reviews and meta-analyses should be as carefully planned as any other research project, with a detailed written protocol prepared in advance.
- The formulation of the review question, the a priori definition of eligibility criteria for trials to be included, a comprehensive search for such trials and an assessment of their methodological quality, are central to high quality reviews.
- The graphical display of results from individual studies on a common scale ("Forest plot") is an important step, which allows a visual examination of the degree of heterogeneity between studies.
- There are different statistical methods for combining the data in meta-analysis but there is no single "correct" method. A thorough sensitivity analysis should always be performed to assess the robustness of combined estimates to different assumptions, methods and inclusion criteria and to investigate the possible influence of bias.
- When interpreting results, reviewers should consider the importance of beneficial and harmful effects of interventions in absolute and relative terms and address economic implications and implications for future research.

Systematic reviews allow a more objective appraisal of the evidence than traditional narrative reviews and may thus contribute to resolve uncertainty when original research, reviews and editorials disagree. Systematic reviews are also important to identify questions to be addressed in future studies. As will be discussed in the subsequent chapter, ill conducted

reviews and meta-analyses may, however, be biased due to exclusion of relevant studies, the inclusion of inadequate studies or the inappropriate statistical combination of studies. Such bias can be minimised if a few basic principles are observed. Here we will introduce these principles and give an overview of the practical steps involved in performing systematic reviews. We will focus on systematic reviews of controlled trials but the basic principles are applicable to reviews of any type of study (see Chapters 12–14 for a discussion of systematic reviews of observational studies). Also, we assume here that the review is based on summary information obtained from published papers, or from the authors. Systematic reviews and meta-analyses based on individual patient data are discussed in Chapter 6. We stress that the present chapter can only serve as an elementary introduction. Readers who want to perform systematic reviews should consult the ensuing chapters and consider joining forces with the Cochrane Collaboration (see Chapters 25 and 26).

Developing a review protocol

Systematic reviews should be viewed as observational studies of the evidence. The steps involved, summarised in Box 2.1, 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 are likely to produce spurious results.[1,2] 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

Box 2.1 Steps in conducting a systematic review*

1 Formulate review question

2 Define inclusion and exclusion criteria
- participants
- interventions and comparisons
- outcomes
- study designs and methodological quality

3 Locate studies (see also Chapter 4)

Develop search strategy considering the following sources:
- *The Cochrane Controlled Trials Register* (CCTR)
- electronic databases and trials registers not covered by CCTR
- checking of reference lists
- handsearching of key journals
- personal communication with experts in the field

4 Select studies
- have eligibility checked by more than one observer
- develop strategy to resolve disagreements
- keep log of excluded studies, with reasons for exclusions

5 Assess study quality (see also Chapter 5)
- consider assessment by more than one 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

6 Extract data
- design and pilot data extraction form
- consider data extraction by more than one observer
- consider blinding of observers to authors, institutions and journals

7 Analyse and present results (see also Chapters 8–11, 15, 16)
- tabulate results from individual studies
- examine forest plot
- explore possible sources of heterogeneity
- 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

8 Interpret results (see also Chapters 19–24)
- 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

* Points 1–7 should be addressed in the review protocol.

the combinability of patients, treatments, outcomes and lengths of follow-up. As discussed in detail in Chapter 5, quality and design features of clinical trials can influence the results.[3-5] Ideally, only controlled trials with proper patient randomisation 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.[6] Formulating assessments regarding study quality can be a subjective process, however, especially since the information reported is often inadequate for this purpose.[7-10] 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. As discussed in Chapter 4, identifying controlled trials for systematic reviews has become more straightforward in recent years. Appropriate terms to index randomised 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 because the overlap in journals covered by the two databases is only about 34%.[11] The majority of journals indexed in MEDLINE are published in the US whereas EMBASE has better coverage of European journals (see Box 4.1 in Chapter 4 for a detailed comparison of MEDLINE and EMBASE). Re-tagging continues in MEDLINE and EMBASE and projects to cover other databases are ongoing or planned. 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 *The Cochrane Controlled Trials Register* which is available in the Cochrane Library on CD ROM or online (see Chapter 25). This register currently includes over 250 000 records and is clearly 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 (see the search strategy described in Chapter 4). Specialised databases, conference proceedings and the bibliographies of review articles, monographs and the located studies should be scrutinised as well. Finally, the searching by hand of key journals should be considered, keeping in mind that many journals are already being searched by the Cochrane Collaboration.

The search should be extended to include unpublished studies, as their results may systematically differ from published trials. As discussed in Chapter 3, a systematic review which is restricted to published evidence may produce distorted results due to publication bias. Registration of trials at the time they are established (and before their results become known) would eliminate the risk of publication bias.[12] A number of such registers have been set up in recent years and access to these has improved, for example through the Cochrane Collaboration's *Register of Registers* or the internet-based *meta*Register of Controlled Trials which has been established by the publisher Current Science (see Chapters 4 and 24). 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.

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.

Randomised controlled trials provide the best evidence of the efficacy of medical interventions but they are not immune to bias. Studies relating methodological features of trials to their results have shown that trial quality influences effect sizes.[4,5,13] Inadequate concealment of treatment allocation, resulting, for example, from the use of open random number tables, is on average associated with larger treatment effects.[4,5,13] Larger effects were also found if trials were not double-blind.[4] In some instances effects may also be overestimated if some participants, for example, those not adhering to study medications, were excluded from the analysis.[14-16] Although widely recommended, the assessment of the methodological quality of clinical trials is a matter of ongoing debate.[7] This is reflected by the large number of different quality scales and checklists that are available.[10,17] Empirical evidence[10] and theoretical considerations[18] suggests 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. Rather, as discussed in Chapter 5, 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 should also be considered as this may lead to more consistent assessments.[19] Blinding involves photocopying of papers removing the title page and concealing journal identifications and other characteristics with a black marker, or scanning the text of papers into a computer and preparing standardised formats.[20,21] This is time consuming and potential benefits may not always justify the additional costs.[22]

It is important that two independent observers extract the data, so errors can be avoided. A standardised record form is needed for this purpose. Data extraction forms should be carefully designed, piloted and revised if necessary. Electronic data collection 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.[23]

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. Table 2.1 shows the characteristics of the long term trials that were included in a systematic review[24] of the effect of beta blockade in secondary prevention after myocardial infarction (we mentioned this example in Chapter 1 and will return to it later in this chapter). Freemantle *et al.*[24] included all parallel group randomised trials that examined the effectiveness of beta blockers versus placebo or alternative treatment in patients who had had a myocardial infarction. The authors searched 11 bibliographic databases, including dissertation abstracts and grey literature databases, examined existing reviews and checked the reference lists of each identified study. Freemantle *et al.* identified 31 trials of at least six months' duration which contributed 33 comparisons of beta blocker with control groups (Table 2.1).

Standardised outcome measure

Individual results have to be expressed in a standardised format to allow for comparison between studies. If the endpoint is binary (e.g. 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

Table 2.1 Characteristics of long term trials comparing beta blockers with control. Adapted from Freemantle *et al.*[24]

Author	Year	Drug	Study duration (years)	Concealment of treatment allocation	Double-blind	Mortality (No./total no.)	
						Beta blocker	Control
Barber	1967	Practolol	2	Unclear	Unclear	33/ 207	38/ 213
Reynolds	1972	Alprenolol	1	Yes	Yes	3/ 38	3/ 39
Ahlmark	1974	Alprenolol	2	Unclear	Unclear	5/ 69	11/ 93
Wilhelmsson	1974	Alprenolol	2	Unclear	Yes	7/ 114	14/ 116
Multicentre International	1975	Practolol	2	Unclear	Yes	102/ 1533	127/ 1520
Yusuf	1979	Atenolol	1	Unclear	Yes	1/ 11	1/ 11
Andersen	1979	Alprenolol	1	Unclear	Yes	61/238	62/242
Rehnqvist	1980	Metroprolol	1	Unclear	Unclear	4/59	6/52
Baber	1980	Propranolol	0·75	Unclear	Yes	28/355	27/365
Wilcox (Atenolol)	1980	Atenolol	1	Yes	Yes	17/132	19/129
Wilcox (Propanolol)	1980	Propranolol	1	Yes	Yes	19/ 127	19/ 129
Hjalmarson	1981	Metoprolol	2	Unclear	No	40/ 698	62/ 697
Norwegian Multicentre	1981	Timolol	1·4	Unclear	Yes	98/ 945	152/ 939
Hansteen	1982	Propranolol	1	Unclear	Yes	25/ 278	37/ 282
Julian	1982	Sotalol	1	Yes	Yes	64/ 873	52/ 583
BHAT	1982	Propranolol	2·1	Yes	Yes	138/ 1916	188/ 1921
Taylor	1982	Oxprenolol	4	Done	Yes	60/ 632	48/ 471
Manger Cats	1983	Metoprolol	1	Unclear	Yes	9/ 273	16/ 280
Rehnqvist	1983	Metroprolol	3	Unclear	Yes	25/ 154	31/ 147
Australian-Swedish	1983	Pindolol	2	Unclear	Yes	45/ 263	47/ 266
Mazur	1984	Propranolol	1·5	Unclear	No	5/ 101	11/ 103
EIS	1984	Oxprenolol	1	Unclear	Yes	57/ 853	45/ 883
Salathia	1985	Metoprolol	1	Unclear	Yes	49/ 416	52/ 348
Roque	1987	Timolol	2	Unclear	Yes	7/ 102	12/ 98
LIT	1987	Metoprolol	1·5	Unclear	Yes	86/ 1195	93/ 1200
Kaul	1988	Propranolol	0·5	Unclear	Yes	3/ 25	3/ 25
ASPI	1990	Aceburtolol	0·87	Yes	Yes	17/ 298	34/ 309
Schwartz (high risk)	1992	Oxprenolol	1·8	Unclear	No	2/ 48	12/ 56
Schwartz (low risk)	1992	Oxprenolol	1·8	Unclear	Yes	15/ 437	27/ 432
SSSD	1993	Metoprolol	3	Unclear	No	17/ 130	9/ 123
Darasz	1995	Xamoterol	0·5	Unclear	Yes	3/ 23	1/ 24
Basu	1997	Carvedilol	0·5	Unclear	Yes	2/ 75	3/ 71
Aronow	1997	Propranolol	1	Unclear	Unclear	44/ 79	60/ 79

the testing of the overall effect for statistical significance, but, as discussed in Box 2.2, the odds ratio will differ from the relative risk if the outcome is common. Relative risks should probably be prefered over odds ratios because they are more intuitively comprehensible to most people.[25,26] Absolute measures such as the absolute risk reduction or the number of patients needed to be treated for one person to benefit[27] are more helpful when applying results in clinical practice (see below). If the outcome is continuous and measurements are made on the same scale (e.g. blood pressure measured in mm Hg) the mean difference between the treatment and control groups is used. If trials measured outcomes in different ways, differences may be presented in standard deviation units, rather than as absolute differences. For example, the efficacy of non-steroidal antiinflammatory drugs for reducing pain in patients with rheumatoid arthritis was measured using different scales.[28] The choice and calculation of appropriate summary statistics is covered in detail in Chapters 15 and 16.

Graphical display

Results from each trial are usefully graphically 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 2.1 represents the forest plot for the trials of beta-blockers in secondary prevention after myocardial infarction which we mentioned in Chapter 1.[24] Each study is represented by a black square and a horizontal line which correspond to the point estimate and the 95% confidence intervals of the relative risk. The 95% confidence intervals would contain the true underlying effect in 95% of the occasions, if the study was repeated again and again. 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 ($P > 0{\cdot}05$). The confidence interval of most studies cross this line. The area of the black squares reflects the weight of the study in the meta-analysis (see below).

A logarithmic scale was used for plotting the relative risk in Figure 2.2. There are a number of reasons why ratio measures are best plotted on logarithmic scales.[29] 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.

Box 2.2 Odds ratio or relative risk?

Odds ratios are often used in order to bring the results of different trials into a standardised format. What is an odds ratio and how does it relate to the relative risk? The *odds* is defined as the number of patients who fulfill the criteria for a given endpoint divided by the number of patients who do not. For example, the odds of diarrhoea during treatment with an antibiotic in a group of 10 patients may be 4 to 6 (4 with diarrhoea divided by 6 without, 0·66), as compared to 1 to 9 (0·11) in a control group. A bookmaker (a person who takes bets, especially on horse-races, calculates odds, and pays out winnings) would, of course, refer to this as nine to one. The *odds ratio* of treatment to control group in this example is 6 (0·66 divided by 0·11). The risk, on the other hand, is calculated as the number of patients with diarrhoea divided by all patients. It would be 4 in 10 in the treatment group and 1 in 10 in the control group, for a risk ratio, or a *relative risk*, of 4 (0·4 divided by 0·1). As shown in Figure 2.1, the odds ratio will be close to the relative risk if the endpoint occurs relatively infrequently, say in less than 15%. If the outcome is more common, as in the diarrhoea example, then the odds ratio will differ increasingly from the relative risk. The choice of binary outcome measures is discussed in detail in Chapter 16.

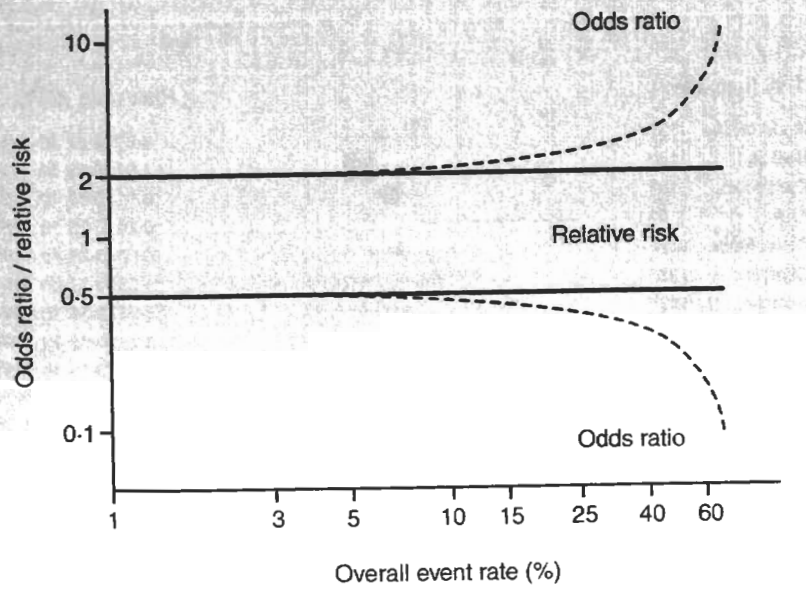

Figure 2.1

Heterogeneity between study results

The thoughtful consideration of heterogeneity between study results is an important aspect of systematic reviews.[30,31] As mentioned 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 (Figure 2.2). In contrast, trials of BCG vaccination for prevention of tuberculosis[32] (Figure 2.3) are clearly heterogeneous. The findings of the UK trial, which indicate substantial benefit of BCG vaccination are not compatible with those from the Madras or Puerto Rico trials which suggest little effect or only a modest benefit. There is no overlap in the confidence intervals of the three trials. Other graphical representations, discussed elsewhere, are particularly useful to detect and investigate heterogeneity. These include Galbraith plots[29] (see Chapter 9), L'Abbé plots[33] (see Chapters 8, 10 and 16) and funnel plots[34] (see Chapter 11).

Statistical tests of homogeneity (also called tests for heterogeneity) assess whether the individual study results are likely to reflect a single underlying effect, as opposed to a distribution of effects. If this test fails

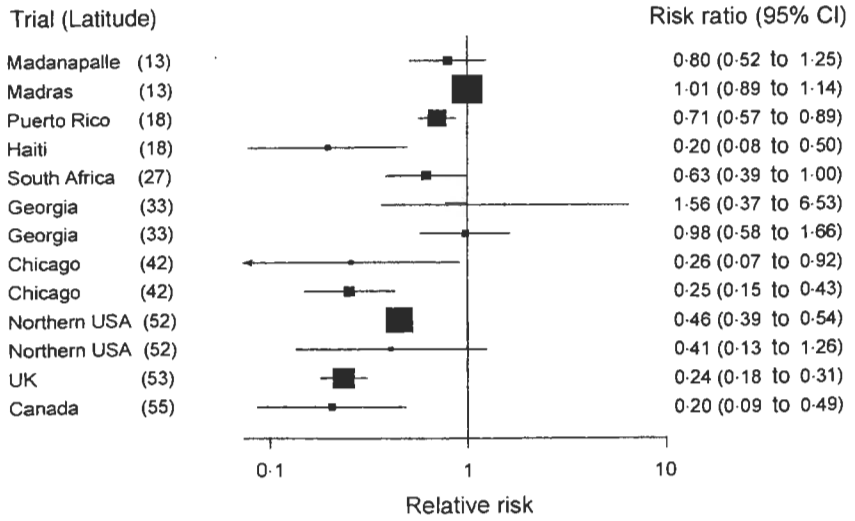

Figure 2.3 Forest plot of trials of BCG vaccine to prevent tuberculosis. Trials are ordered according to the latitude of the study location, expressed as degrees from the equator. No meta-analysis is shown. Adapated from Colditz *et al.*[32]

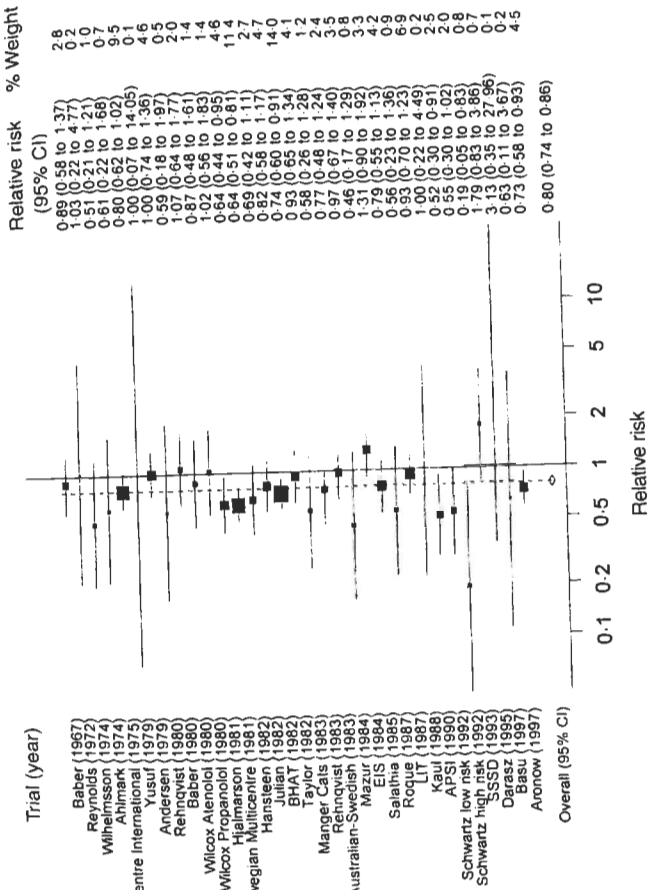

Figure 2.2 Forest plot showing total mortality from trials of beta-blockers in secondary prevention after myocardial infarction. The black square and horizontal line correspond to the relative risk and 95% confidence intervals. The area of the black squares reflects the weight each trial contributes to the meta-analysis. The diamond at the bottom of the graph represents the combined relative risk and its 95% confidence interval, indicating a 20% reduction in the risk of death. The solid vertical line corresponds to no effect of treatment (relative risk 1·0), the dotted vertical line to the combined relative risk (0·8). The relative risk, 95% confidence interval and weights are also given in tabular form. The graph was produced in STATA (see Chapter 18). Adapted from Freemantle *et al.*[34]

to detect heterogeneity among results, then it is assumed that the differences observed between individual studies are a consequence of sampling variation and simply due to chance. A chi-square test of homogeneity gives P = 0·25 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 inter-study differences exist. Reviewers should therefore not assume that a non-significant test of heterogeneity excludes important heterogeneity. 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 and in Chapters 8 and 9.

Methods for estimating a combined effect estimate

If, after careful consideration, a meta-analysis is deemed appropriate, the last step consists in estimating an overall effect by combining the data. Two principles are important. Firstly, simply pooling the data from different studies and treating them as one large study would fail to preserve the randomisation and introduce bias and confounding. For example, a recent review and "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.[35] 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.[36] Confounding by study thus led to a change in the direction of the association (a case of "Simpson's paradox" in epidemiological parlance[37]). The unit of the trial must therefore always be maintained when combining data.

Secondly, 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. Methods used for meta-analysis employ a weighted average of the results in which the larger trials generally have more influence than the smaller ones. There are a variety of statistical techniques available for this purpose (see Chapter 15), which can be broadly classified into two models.[38] The difference consists in the way the variability of the results *between* the studies is treated. The "*fixed effects*" model considers this variability as exclusively due to random variation and individual studies are simply weighted by their precision.[39] Therefore, if all the studies were infinitely

large they would give identical results. The main alternative, the "*random effects*" model,[40] assumes a different underlying effect for each study and takes this into consideration as an additional source of variation. Effects are assumed to be randomly distributed and the central point of this distribution is the focus of the combined effect estimate. The random effects model leads to relatively more weight being given to smaller studies and to wider confidence intervals than the fixed effects model. The use of random effects models has been advocated if there is heterogeneity between study results. This is problematic, however. Rather than simply ignoring it after applying some statistical model, the approach to heterogeneity should be to scrutinise, and attempt to explain it.[30,31]

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 2.2). Combining trials using a random effects model indicates that BCG vaccination halves the the risk of tuberculosis, whereas fixed effects analysis indicates that the risk is only reduced by 35%. This is essentially explained by the different weight given to the large Madras trial which showed no protective effect of vaccination (41% of the total weight with fixed effects model, 10% with random effects model, Table 2.2). Both analyses are probably misguided. As shown in Figure 2.2, BCG vaccination appears to be effective at higher latitudes but not in warmer regions, possibly because

Table 2.2 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. The raw data (from Colditz *et al.*[32]) are given in Chapter 18.

Trial	Relative risk (95% CI)	Fixed effects weight (%)	Random effects weight (%)
Madanapalle	0·80 (0·52 to 1·25)	3·20	8·88
Madras	1·01 (0·89 to 1·14)	41·40	10·22
Puerto Rico	0·71 (0·57 to 0·89)	13·21	9·93
Haiti	0·20 (0·08 to 0·50)	0·73	6·00
South Africa	0·63 (0·39 to 1·00)	2·91	8·75
Georgia	0·98 (0·58 to 1·66)	0·31	3·80
Georgia	1·56 (0·37 to 6·53)	2·30	8·40
Chicago	0·26 (0·07 to 0·92)	0·40	4·40
Chicago	0·25 (0·15 to 0·43)	2·25	8·37
Northern USA	0·41 (0·13 to 1·26)	23·75	10·12
Northern USA	0·46 (0·39 to 0·54)	0·50	5·05
UK	0·24 (0·18 to 0·31)	8·20	9·71
Canada	0·20 (0·09 to 0·49)	0·84	6·34
Combined relative risks		0·65 (0·60 to 0·70)	0·49 (0·35 to 0·70)

exposure to certain environmental mycobacteria acts as a "natural" BCG inoculation in warmer regions.[41] 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 (see Chapter 18 for further analyses of the BCG trials).

Bayesian meta-analysis

There are other statistical approaches, which some feel are more appropriate than either of the above. One uses Bayes' theorem, named after the 18th century English clergyman Thomas Bayes.[42–44] 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.[45] 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% credible interval which covers 95% of the posterior probability distribution) will be slightly wider than that derived from using the conventional models.[46,47]

Bayesian methods allow probability statements to be made directly regarding, for example, the comparative effects of two treatments ("the probability that treatment A is better than B is 0·99").[48] 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 (see Chapter 10).[44,49] Finally, they provide a natural framework for cumulative meta-analysis.[49,50]

Bayesian approaches are, however, controversial because the definition of prior probability will often involve subjective assessments and opinion which runs against the principles of systematic review. Furthermore, analyses are complex to implement and time consuming. More methodological research is required to define the appropriate place of bayesian methods in systematic reviews and meta-analysis.[44,49]

Sensitivity analysis

There will often be diverging opinions on the correct method for performing a particular meta-analysis. The robustness of the findings to different assumptions should therefore always be examined in a thorough sensitivity analysis. This is illustrated in Figure 2.4 for the beta-blocker after myocardial infarction meta-analysis.[24] First, the overall

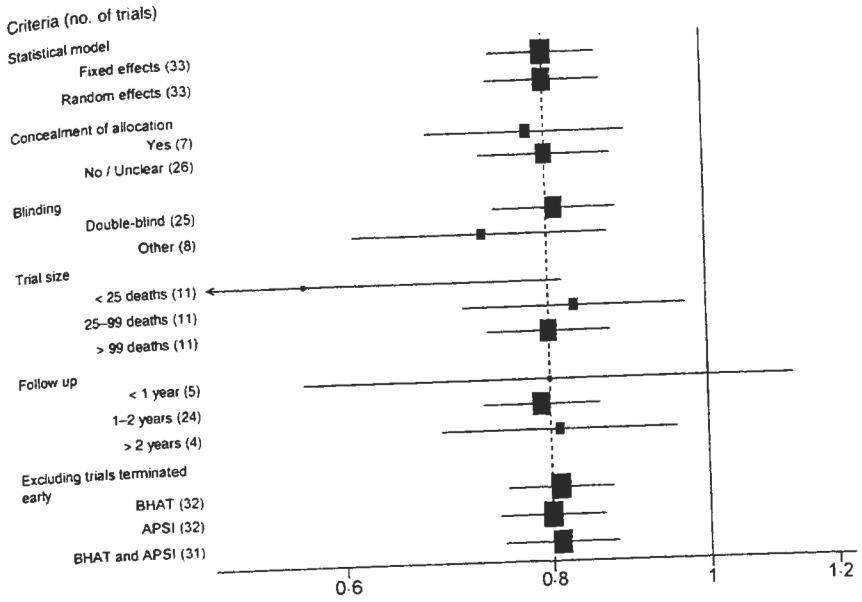

Figure 2.4 Sensitivity analyses examining the robustness of the effect on total mortality of beta-blockers in secondary prevention after myocardial infarction. The dotted vertical line corresponds to the combined relative risk from the fixed effects model (0·8).

effect was calculated by different statistical methods, both using 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 variation 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.[24] Figure 2.4 shows that the estimated treatment effect was similar for studies with and without concealment of treatment allocation. The eight studies that were not double-blind indicated more benefit than the 25 double-blind trials but confidence intervals overlap widely. Statistically significant results are more likely to get published than non-significant findings[51] and this can distort the findings of meta-analyses (see Chapter 3). Whether such publication bias is present can be examined by stratifying the analysis by study size. Smaller effects can be statistically significant in larger studies. If publication bias is present, it is expected that of published studies, the larger ones will report the smaller effects. The

figure shows that in the present example 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[52,53] 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 be biased away from the null value. Bias may thus be introduced in a meta-analysis which includes such trials.[54] Exclusion of these trials, however, again affects the overall estimate only marginally.

The sensitivity analysis thus shows that the results from this 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.

Relative or absolute measures of effect?

The relative risk of death associated with the use of beta-blockers after myocardial infarction is 0·80 (95% confidence interval 0·74 to 0·86) (Figure 2.2). The *relative risk reduction*, obtained by substracting the relative risk from 1 and expressing the result as a percentage, is 20% (95% confidence interval 14 to 26%). The relative measures ignore the underlying absolute risk. The risk of death among patients who have survived the acute phase of myocardial infarction, however, varies widely.[55] For example, among patients with three or more cardiac risk factors, the probability of death at two years after discharge ranged from 24 to 60%.[55] Conversely, two-year mortality among patients with no risk factors was less than three percent. 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 NNT(benefit).[27] The number of patients that need to be treated to harm one patient, denoted as NNH or, more appropriately, NNT(harm)[56] can also be calculated.

For a baseline risk of one per cent per year, the absolute risk difference indicates that two deaths are prevented per 1000 treated patients (Table 2.3). This corresponds to 500 patients (1 divided by 0·002) treated for one year to prevent one death. Conversely, if the risk is above 10%, 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

Table 2.3 Beta-blockade in secondary prevention after myocardial infarction. Absolute risk reductions and numbers-needed-to-treat for one year to prevent one death, NNT(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%.

have to be exposed to the adverse effects of beta-blockade to postpone 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[57] and confidence intervals can be calculated.[56]

Meta-analysis using absolute effect measures such as the risk difference may be useful to illustrate the range of absolute effects across studies. The *combined* risk difference (and the NNT calculated from it) will, however, be essentially determined by the number and size of trials in low, intermediate and high-risk patients. Combined results will thus be applicable only to patients at levels of risk corresponding to the average risk of the trial participants. It is therefore generally more meaningful to use relative effect measures when summarising the evidence while considering absolute measures when applying it to a specific clinical or public health situation. The use of numbers-needed-to-treat in meta-analysis is discussed in more detail in Chapter 20.

Conclusions

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. How likely is it that publication and related biases have been avoided? Is it sensible to combine the individual trials in meta-analysis or is there heterogeneity between individual study results which renders the calculation of an

overall estimate questionable? If meta-analysis was performed, how robust are the results to changes in assumptions? Finally, has the analysis contributed to the process of making rational health care decisions? These issues will be considered in more depth in the following chapters.

Acknowledgements

This chapter draws on material published earlier in the *BMJ*.[58]

We are grateful to Iain Chalmers for helpful comments on an earlier draft of this chapter.

1 Gelber RD, Goldhirsch A. From the overview to the patient: how to interpret meta-analysis data. *Recent Results Cancer Res* 1993;**127**:167–76.
2 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;**116**:78–84.
3 Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;**72**:233–40.
4 Schulz KF, Chalmers I, Hayes RJ, Altman D. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273**:408–12.
5 Moher D, Pham B, Jones A, *et al.* Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;**352**:609–13.
6 Prendiville W, Elbourne D, Chalmers I. The effects of routine oxytocic administration in the management of the third stage of labour: an overview of the evidence from controlled trials. *Br J Obstet Gynaecol* 1988;**95**:3–16.
7 Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. *Int J Technol Assess Hlth Care* 1996;**12**:195–208.
8 Begg C, Cho M, Eastwood S, *et al.* Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;**276**:637–9.
9 Schulz KF. Randomised trials, human nature, and reporting guidelines. *Lancet* 1996;**348**:596–8.
10 Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trial for meta-analysis. *JAMA* 1999;**282**:1054–60.
11 Smith BJ, Darzins PJ, Quinn M, Heller RF. Modern methods of searching the medical literature. *Med J Aust* 1992;**157**:603–11.
12 Dickersin K. Research registers. In: Cooper H, Hedges LV, eds. *The handbook of research synthesis*. New York: Russell Sage Foundation, 1994.
13 Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;**309**:1358–61.
14 Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. *N Engl J Med* 1979;**301**:1410–12.
15 Peduzzi P, Wittes J, Detre K, Holford T. Analysis as-randomized and the problem of non-adherence: an example from the veterans affairs randomized trial of coronary artery bypass surgery. *Stat Med* 1993;**12**:1185–95.
16 May GS, Demets DL, Friedman LM, Furberg C, Passamani E. The randomized clinical trial: bias in analysis. *Circulation* 1981;**64**:669–73.
17 Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clin Trials* 1995;**16**:62–73.
18 Greenland S. Quality scores are useless and potentially misleading. *Am J Epidemiol* 1994;**140**:300–2.
19 Jadad AR, Moore RA, Carrol D, *et al.* Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clin Trials* 1996;**17**:1–12.

20 Chalmers TC. Problems induced by meta-analyses. *Stat Med* 1991;**10**:971–80.
21 Moher D, Fortin P, Jadad AR, *et al.* Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 1996;**347**:363–6.
22 Berlin JA, on behalf of University of Pennsylvania Meta-analysis Blinding Study Group. Does blinding of readers affect the results of meta-analyses? *Lancet* 1997;**350**:185–6.
23 *Cochrane Reviewer's Handbook* (updated July 1999). In: *The Cochrane Library* (database on disk and CD-ROM). *The Cochrane Collaboration*. Oxford: Update Software, 1999.
24 Freemantle N, Cleland J, Young P, Mason J, Harrison J. Beta blockade after myocardial infarction: systematic review and meta regression analysis. *BMJ* 1999;**318**:1730–7.
25 Sackett DL, Deeks JJ, Altman D. Down with odds ratios! *Evidence-Based Med* 1996;**1**:164–7.
26 Deeks J. When can odds ratios mislead? *BMJ* 1998;**317**:1155.
27 Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;**318**:1728–33.
28 Gøtzsche PC. Sensitivity of effect variables in rheumatoid arthritis: a meta-analysis of 130 placebo controlled NSAID trials. *J Clin Epidemiol* 1990;**43**:1313–18.
29 Galbraith R. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;7:889–94.
30 Bailey K. Inter-study differences: how should they influence the interpretation and analysis of results? *Stat Med* 1987;**6**:351–8.
31 Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;**309**:1351–5.
32 Colditz GA, Brewer TF, Berkley CS, *et al.* Efficacy of BCG vaccine in the prevention of tuberculosis. *JAMA* 1994;**271**:698–702.
33 L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med* 1987;**107**:224–33.
34 Light RJ, Pillemer DB. *Summing up. The science of reviewing research*. Cambridge, MA: Harvard University Press, 1984.
35 Van Howe RS. Circumcision and HIV infection: review of the literature and meta-analysis. *Int J STD AIDS* 1999;**10**:8–16.
36 O'Farrell N, Egger M. Circumcision in men and the prevalence of HIV infection: a meta-analysis revisited. *Int J STD AIDS* 2000; **11**:137–42.
37 Last JM. *A dictionary of epidemiology*. New York: Oxford University Press, 1995.
38 Berlin J, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Stat Med* 1989;**8**:141–51.
39 Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis* 1985;**17**:335–71.
40 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;7:177–88.
41 Fine PEM. Variation in protection by BCG: implications of and for heterologous immunity. *Lancet* 1995;**346**:1339–45.
42 Carlin JB. Meta-analysis for 2 × 2 tables: a Bayesian approach. *Stat Med* 1992;**11**:141–58.
43 Bland JM, Altman DG. Bayesians and frequentists. *BMJ* 1998;**317**:1151.
44 Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to bayesian methods in health technology assessment. *BMJ* 1999;**319**:508–12.
45 Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;**313**:603–7.
46 Su XY, Li Wan Po A. Combining event rates from clinical trials: comparison of bayesian and classical methods. *Ann Pharmacother* 1996;**30**:460–5.
47 Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* 1997;**16**:2741–58.
48 Fredman L. Bayesian statistical methods. *BMJ* 1996;**313**:569–70.
49 Song F, Abrams KR, Jones DR, Sheldon TA. Systematic reviews of trials and other studies. *Health Technol Assess* 1998;**2(19)**.

50 Eddy DM, Hasselblad V, Shachter R. *Meta-analysis by the confidence profile method. The statistical synthesis of evidence*. Boston: Academic Press, 1992.
51 Easterbrook PJ, Berlin J, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;**337**:867–72.
52 Anon. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *JAMA* 1982;**247**:1707–14.
53 Boissel JP, Leizorovicz A, Picolet H, Ducruet T. Efficacy of acebutolol after acute myocardial infarction (the APSI trial). The APSI Investigators. *Am J Cardiol* 1990;**66**:24C–31C.
54 Green S, Fleming TR, Emerson S. Effects on overviews of early stopping rules for clinical trials. *Stat Med* 1987;**6**:361–7.
55 The Multicenter Postinfarction Research Group. Risk stratification and survival after myocardial infarction. *N Engl J Med* 1983;**309**:331–6.
56 Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998;**317**:1309–12.
57 Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P. The number needed to treat: a clinically useful nomogram in its proper context. *BMJ* 1996;**312**:426–9.
58 Egger M, Davey Smith G, Phillips AN. Meta-analysis: principles and procedures. *BMJ* 1997;**315**:1533–7.