8. ANALYSING AND PRESENTING RESULTS

Edited by *Jon Deeks, Julian Higgins* and *Doug Altman* on behalf of the *Cochrane Statistical Methods Group*.

Do not start here! Please consult Sections 2 to 6 before reading this Section. It can be tempting to jump prematurely into a statistical analysis when undertaking a systematic review. The production of a diamond at the bottom of a plot is an exciting moment for many reviewers, but results of meta-analyses can be very misleading if suitable attention has not been given to formulating the review question; specifying inclusion criteria; identifying, selecting and critically appraising studies; collecting appropriate data; and deciding what would be meaningful to analyse.

This version of section 8 contains references to subsections that are not yet complete. Where this happens, the name of the subsection is given, together with the number 8.X as we do not yet know what the numbering of these will be. We hope that these subsections will be completed and published during 2004.

Within this section 'RevMan' is used to refer to the Cochrane Collaboration's Review Manager software including its statistical component, which is now called RevMan Analyses. Previous versions of RevMan used a statistical program called MetaView, which is currently one option for viewing graphs in The Cochrane Library. Thus people reading a review may see a slightly different output to that the reviewer sees in RevMan.

8.1 Planning the analysis

While in primary studies the investigators select and collect data from individual patients, in systematic reviews the investigators select and collect data from primary studies. While primary studies include analyses of their patients, Cochrane Reviews contain analyses of the primary studies. Analyses may be narrative, such as a structured summary and discussion of the studies' characteristics and findings, or quantitative, that is involving statistical analysis. **Meta-analysis** – the statistical combination of results from two or more separate studies – is the most commonly used statistical technique. Cochrane Review writing software (RevMan) can perform a variety of meta-analyses, but it must be stressed that meta-analysis is not appropriate in all Cochrane Reviews. Issues to consider when deciding whether a meta-analysis is appropriate in your review are discussed in this section and in 8.2.2 When not to use meta-analysis in a review?

Studies comparing health care interventions, notably randomised trials, use the outcomes of participants to compare the effects of different interventions. Meta-analyses focus on pair wise comparisons of interventions, such as an experimental intervention versus a control intervention, or the comparison of two experimental interventions. The terminology used throughout this section of the Handbook (experimental versus control interventions) implies the former, but is intended to include the latter.

The contrast between the outcomes of two groups treated differently is known as the *effect* or the *treatment effect*. Whether analysis of included studies is narrative or quantitative, a general framework for synthesis may be provided by considering four questions:

(1) What is the direction of effect?
(2) What is the size of effect?
(3) Is the effect consistent across studies?
(4) What is the strength of evidence for the effect?

Meta-analysis provides a statistical method for (1)-(3). Assessment of (4) relies additionally on judgements based on assessments of study design and study quality, as well as statistical measures of uncertainty.

Narrative synthesis uses subjective (rather than statistical) methods to follow through stages (1)-(4) for reviews where meta-analysis is either not feasible or not sensible. In a narrative synthesis the method used for each stage should be pre-specified, justified and followed systematically. Bias may be introduced if the results of one study are inappropriately stressed over those of another.

The analysis plan follows from the scientific aim of the review. Reviews have different types of aims, and may therefore contain different approaches to analysis.

(1) The most straightforward Cochrane Review assembles studies that make one particular comparison between two treatment options, for example, comparing inhaled steroids with placebo for bronchiectasis. Meta-analysis and related techniques can be used if there is a consistent outcome measure to:
    (i) establish whether there is evidence of an effect;
    (ii) estimate the size of the effect and the uncertainty surrounding that size; and
    (iii) investigate whether the effect is consistent across studies.
(2) Some reviews may have a broader focus than a single comparison. The first is where the intention is to identify and collate all studies in a particular field. An example of such a review is that of topical treatments for fungal infections of the skin and nails of the foot, which included studies of any topical treatment. The second, related aim is that of identifying a 'best' intervention. A review of interventions for emergency contraception sought that which was most effective (while also considering potential adverse effects). Such reviews may include multiple comparisons and meta-analyses between all possible pairs of treatments, and require care when it comes to planning analyses – see 8.1.4 Which comparisons should be made?
(3) Occasionally review comparisons have particularly wide scopes that make the use of meta-analysis problematic. For example, a review of media-based behavioural treatments for behavioural disorders in children covers diverse media-based treatments (including written material and film) and diverse behavioural problems (including Attention Deficit/Hyperactivity Disorder and enuresis). When reviews contain very diverse studies a meta-analysis might be useful to answer the overall question of whether there is evidence that, for example, media-based treatments can work (but see 8.1.2 When not to use meta-analysis in a review). But use of meta-analysis to describe the size of effect may not be meaningful if the implementations are so diverse that an effect estimate cannot be interpreted in any specific context.
(4) An aim of some reviews is to investigate the relationship between the size of an effect and some characteristic(s) of the studies. This is uncommon as a primary aim in Cochrane Reviews, but may be a secondary aim. For example, in the review of inhaled steroids for bronchiectasis, there was interest in whether the

administered dose of steroid affected its efficacy. Such investigations of **heterogeneity** need to be undertaken with care: see 8.8 Investigating heterogeneity.

8.1.1 Why perform a meta-analysis in a review?

The value a meta-analysis can add to a review depends on the context in which it is used, as described in 8.1 Planning the analysis. Reasons for considering including a meta-analysis in a review are:

(1) To increase power. Power is the chance of detecting a real effect as statistically significant if it exists. Many individual studies are too small to detect small effects, but when several are combined there is a higher chance of detecting an effect.
(2) To improve precision. The estimation of a treatment effect can be improved when it is based on more information.
(3) To answer questions not posed by the individual studies. Primary studies often involve a specific type of patient and explicitly defined interventions. A selection of studies in which these characteristics differ can allow investigation of the consistency of effect and, if relevant, allow reasons for differences in effect estimates to be investigated.
(4) To settle controversies arising from apparently conflicting studies or to generate new hypotheses. Statistical analysis of findings allows the degree of conflict to be formally assessed, and reasons for different results to be explored and quantified.

Of course, the use of statistical methods does not guarantee that the results of a review are valid, any more than it does for a primary study. Moreover, like any tool, statistical methods can be misused.

8.1.2 When not to use meta-analysis in a review

If used appropriately, meta-analysis is a powerful tool for deriving meaningful conclusions from data and can help prevent errors in interpretation. However, there are situations in which a meta-analysis can be more of a hindrance than a help. A common criticism of meta-analyses is that they 'combine apples with oranges'. If studies are clinically diverse then a meta-analysis may be meaningless, and genuine differences in effects may be obscured. A particularly important type of diversity is in the comparisons being made by the primary studies. Often it is nonsensical to combine all included studies in a single meta-analysis: sometimes there is a mix of comparisons of different treatments with different comparators, each combination of which may need to be considered separately. Further, it is important not to combine outcomes that are too diverse.

Decisions concerning what should and should not be combined are inevitably subjective, and are not amenable to statistical solutions but require discussion and clinical judgement. In some cases consensus may be hard to reach.

Meta-analyses of poor quality studies may be seriously misleading. If bias is present in each (or some) of the individual studies, meta-analysis will simply compound the errors, and produce a 'wrong' result that may be interpreted as having more credibility.

Finally, meta-analyses in the presence of serious publication and/or reporting biases may produce an inappropriate summary.

8.1.3 What does a meta-analysis entail?

While the use of statistical methods in reviews can be extremely helpful, the most essential element of an analysis is a thoughtful approach, to both its narrative and quantitative elements. This entails consideration of the following questions:

(1) Which comparisons should be made?
(2) Which study results should be used in each comparison?
(3) What is the best summary of effect for each comparison?
(4) Are the results of studies similar within each comparison?
(5) How reliable are those summaries?

The first step in addressing these questions is to decide which comparisons to make (8.1.4.Which comparisons should be made?). The next step is to prepare tabular summaries of the characteristics and results of the studies that are included in each comparison (8.2 Types of data and effect measures, 8.4 Extraction of study results). It is then possible to derive estimates of effect across studies in a systematic way (8.6 Summarising effects across studies), to measure and investigate differences among studies (8.7 Heterogeneity) and to interpret the findings and conclude how much confidence should be placed in them (8.X Issues in interpretation).

8.1.4 Which comparisons should be made?

The first and most important step in planning the analysis is to specify the pair wise comparisons that will be made. The comparisons addressed in the review should relate clearly and directly to the questions or hypotheses that are posed when the review is formulated (see Section 4). It should be possible to specify in the protocol of a review the main comparisons that will be made. However, it will often be necessary to modify comparisons and add new ones in light of the data that are collected. For example, important variations in the intervention may only be discovered after data are collected.

Decisions about which studies are similar enough for their results to be grouped together require an understanding of the problem that the review addresses, and judgement by the reviewer and the user. The formulation of the questions that a review addresses is discussed in Section 4. Essentially the same considerations apply to deciding which comparisons to make, which outcomes to combine and which key characteristics (of study design, participants, interventions and outcomes) to consider when investigating variation in effects (heterogeneity). These considerations must be addressed when setting up the Table of Comparisons in RevMan and in deciding what information to put in the table of Characteristics of Included Studies.

8.1.5 Writing the analysis section of the protocol

The analysis section of a Cochrane Review protocol may be more susceptible to change than other protocol sections (such as criteria for including studies and how methodological quality will be assessed). It is rarely possible to anticipate all the statistical issues that may arise, for example, finding outcomes that are similar but not the

same as each other; outcomes measured at multiple or varying time-points; and use of concomitant treatments

However the protocol should provide a strong indication as to how the reviewer will approach the statistical evaluation of studies' findings. At least one member of the review team should be familiar with the majority of the contents of Section 8 when the protocol is written. As a guideline we recommend that the following be addressed (more details of all the issues may be found in the rest of Section 8):

(1) ensure that the analysis strategy firmly addresses the stated objectives of the review (8.1 Planning the analysis);
(2) consider which types of study design would be appropriate for the review. Parallel group trials are the norm, but other randomized designs may be appropriate to the topic (e.g. cross-over trials, cluster randomized trials, factorial trials). Decide how such studies will be addressed in the analysis (8.X Other types of study);
(3) decide whether a meta-analysis is intended and consider how the decision as to whether a meta-analysis is appropriate will be made (8.1.1 Why perform a meta-analysis in a review? 8.1.2 When not to use meta-analysis in a review);
(4) determine the likely nature of outcome data (e.g. dichotomous, continuous etc) (8.2 Types of data and effect measures);
(5) consider whether it is possible to specify in advance what treatment effect measures will be used (e.g. risk ratio, odds ratio or risk difference for dichotomous outcomes, mean difference or standardised mean difference for continuous outcomes) (8.6.3.4 Which measure for dichotomous outcomes? 8.6.4.1 Which measure for continuous outcomes?);
(6) decide how statistical heterogeneity will be identified (8.7.2 Identifying and measuring heterogeneity);
(7) decide whether random effects meta-analyses, fixed effect meta-analyses or both methods will be used for each planned meta-analysis (8.7.4 Incorporating heterogeneity into random effects models);
(8) consider how clinical and methodological diversity (heterogeneity) will be assessed and whether (and how) these will be incorporated into the analysis strategy (8.7 Heterogeneity and 8.8 Investigating heterogeneity);
(9) decide how quality of included studies will be assessed and addressed in the analysis (Section 6, Assessing trial quality);
(10) pre-specify characteristics of the studies that may be examined as potential causes of heterogeneity. (8.8.4 Selection of study characteristics for subgroup analyses and meta-regression);
(11) consider how missing data will be handled (e.g. imputing data for intention-to-treat analyses) (8.X Missing data);
(12) decide whether (and how) evidence of possible publication and/or reporting biases will be sought (8.X Investigating and dealing with bias).

It may become apparent when writing the protocol that additional expertise is likely to be required: see 8.X Where to go for help.

8.2 Types of data and effect measures

The starting point of all meta-analyses of studies of effectiveness involves the identification of the **data type** for the outcome measurements.

Through Section 8 we consider outcome data to be of five different types:

(1) Dichotomous (or binary) data, where each individual's outcome is one of only two possible categorical responses;
(2) Continuous data, where each individual's outcome is a measurement of a numerical quantity;
(3) Ordinal data (including measurement scales), where the outcome is one of several ordered categories, or generated by scoring and summing categorical responses;
(4) Counts and rates calculated from counting the number of events that each individual experiences;
(5) Time-to-event (typically survival) data that analyse the time until an event occurs, but where not all individuals in the study experience the event (censored data).

The ways in which the effect of a treatment can be measured depends on the nature of the data being collected. In this section we briefly examine the types of outcome data that might be encountered in systematic reviews of clinical trials, and review definitions, properties and interpretation of standard measures for treatment effect. In Section 8.6.3.4 Which measure for dichotomous outcomes? and Section 8.6.4.1 Which measure for continuous outcomes? we discuss issues in the selection of one of these measures for a particular meta-analysis.

8.2.1 Effect measures for dichotomous outcomes

Dichotomous outcome data arise when the outcome for every participant is one of two possibilities, for example, dead or alive, or clinical improvement or no clinical improvement. This section considers the possible summary statistics when the outcome of interest has such a binary form. The most commonly encountered effect measures used in clinical trials with dichotomous data are:

- the risk ratio (RR) (also called the relative risk);
- the odds ratio (OR);
- the risk difference (RD) (also called the absolute risk reduction, ARR);
- the number needed to treat (NNT).

Details of the calculations of the first three of these measures are given in Box 8.2.1. Numbers needed to treat are discussed in detail in 8.X Re-expressing meta-analysis results as NNTs.

Aside: As events may occasionally be desirable rather than undesirable, it would be preferable to use a more neutral term than risk (such as probability), but for the sake of convention we use the terms risk ratio and risk difference throughout. We also use the term 'risk ratio' in preference to 'relative risk' for consistency with other terminology. The two are interchangeable and both conveniently abbreviate to 'RR'. Note also that we have been careful with the use of the words 'risk' and 'rates'. These words are often treated synonymously. However, we have tried to reserve use of the word 'rate' for the data type 'counts and rates' where it describes the frequency of events in a measured period of time.

Box 8.2.1 Calculation of RR, OR and RD from a 2×2 Table
The results of a clinical trial can be displayed as a 2×2 table:

	Event	No event	Total
Intervention	a	b	a+b
Control	c	d	c+d

where a, b, c and d are the numbers of participants with each outcome in each group. The following summary statistics can be calculated:

$$\text{risk ratio} = \frac{\text{risk of event in intervention group}}{\text{risk of event in control group}} = \frac{a/(a+b)}{c/(c+d)}$$

$$\text{odds ratio} = \frac{\text{odds of event in intervention group}}{\text{odds of event in control group}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$\text{risk difference} = \text{risk of event in intervention group} - \text{risk of event in control group}$$

$$= \frac{a}{a+b} - \frac{c}{c+d}$$

8.2.1.1 Risk and odds

In general conversation the terms 'risk' and 'odds' are used interchangeably (as are the terms 'chance', 'probability' and 'likelihood') as if they describe the same quantity. In statistics, however, risk and odds have particular meanings and are calculated in different ways. When the difference between them is ignored the results of a systematic review may be misinterpreted.

Risk is the concept more familiar to patients and health professionals. Risk describes the probability with which a health outcome (usually an adverse event) will occur. In research, risk is commonly expressed as a decimal number between 0 and 1, although it is occasionally converted into a percentage. It is simple to grasp the relationship between a risk and the likely occurrence of events: in a sample of 100 people the number of events observed will on average be the risk multiplied by 100. For example, when the risk is 0.1, about ten people out of every 100 will have the event, when the risk is 0.5, about 50 people out of every 100 will have the event.

Odds is a concept that is more familiar to gamblers. The odds is the ratio of the probability that a particular event will occur to the probability that it will not occur, and can be any number between zero and infinity. In gambling, the odds describes the ratio of the size of the potential winnings to the gambling stake; in health care it is the ratio of the number of people with the event to the number without. It is commonly expressed as a ratio of two integers. For example, an odds of 0.01 is often written as 1:100, odds of 0.33 as 1:3, and odds of 3 as 3:1. Odds can be converted to risks, and risks to odds, using the formulae:

$$risk = \frac{odds}{1+odds}; \qquad odds = \frac{risk}{1-risk}.$$

The interpretation of an odds is more complicated than for a risk. The simplest way to ensure that the interpretation is correct is to first convert the odds into a risk. For example, when the odds are 1:10, or 0.1, one person will have the event for every 10 who do not,

and, using the above formula, the risk of the event is 0.1/(1+0.1) = 0.091. In a sample of one hundred, about nine individuals will have the event and 91 will not. When the odds is equal to one, one person will have the event for every one who does not, so in a sample of 100, 100 × 1/(1+1) = 50 will have the event and 50 will not.

The difference between odds and risk is small when the event is rare (as illustrated in the first example above where a risk of 0.091 was seen to be similar to an odds of 0.1). When events are common, as is often the case in clinical trials, the differences between odds and risks are large. For example, a risk of 0.5 is equivalent to an odds of 1; and a risk of 0.95 is equivalent to odds of 19.

Measures of effect for clinical trials with dichotomous outcomes involve comparing either risks or odds from two treatment groups. To compare them we can look at their ratio (risk ratio or odds ratio) or their difference in risk (risk difference).

8.2.1.2 Measures of relative effect: the risk ratio and odds ratio

Measures of relative effect express the outcome in one group relative to that in the other. The risk ratio (relative risk) is the ratio of the risk of an event in the two groups whereas the odds ratio is the ratio of the odds of an event (Box 8.2.1). For both measures a value of one indicates that the estimated effects are the same for both treatments.

Neither the risk ratio nor the odds ratio can be calculated for a trial if there are no events in the control group. This is because, as can be seen from the formulae in box 8.2.1, we would be trying to divide by zero. The odds ratio also cannot be calculated if everybody in the intervention group experiences an event. In these situations, and others where standard errors cannot be computed, it is customary to add ½ to each cell of the 2×2 table (RevMan automatically makes this correction when necessary). In the case where no events (or all events) are observed in *both* groups the trial provides no information about relative probability of the event and is automatically omitted from the meta-analysis. This is entirely appropriate. Zeros arise particularly when the event of interest is rare – such events are often unintended adverse outcomes. For further discussion of choice of effect measures for such sparse data (often with lots of zeros) see 8.X Rare events (including zero frequencies).

Risk ratios describe the multiplication of the risk that occurs with use of the intervention. For example, a risk ratio of 3 implies that events with treatment are three times more likely than events without treatment. Alternatively we can say that treatment increases the risk of events by 100 × (*RR* – 1)% = 200%. Similarly a risk ratio of 0.25 is interpreted as the probability of an event with treatment being one-quarter of that without treatment. This may be expressed alternatively by saying that treatment decreases the risk of events by 100 × (1 –*RR*)% = 75%. This is known as the relative risk reduction. The interpretation of the clinical importance of a given risk ratio cannot be made without knowledge of the typical risk of events without treatment: a risk ratio of 0.75 could correspond to a clinically important reduction in events from 80% to 60%, or a small, less clinically important reduction from 4% to 3%.

The numerical value of the observed risk ratio must always be between 0 and 1/ CGR, where CGR (abbreviation of 'control group risk', sometimes referred to as the CER or control event rate) is the observed risk of the event in the control group (expressed as a number between 0 and 1). This means that for common events large values of risk ratio

are impossible. For example, when the observed risk of events in the control group is 0.66 (or 66%) then the observed risk ratio cannot exceed 1.5. This problem applies only for increases in risk, and causes problems only when the results are extrapolated to risks above those observed in the trial

Odds ratios, like odds, are more difficult to interpret (Sackett 1996, Sinclair 1994). Odds ratios describe the multiplication of the odds of the outcome that occur with use of the intervention. To understand what an odds ratio means in terms of changes in numbers of events it is simplest to first convert it into a risk ratio, and then interpret the risk ratio in the context of a typical baseline risk (BR) without treatment, as outlined above. Formulae for converting an odds ratio to a risk ratio, and vice versa, are:

$$\mathrm{RR} = \frac{\mathrm{OR}}{1-(\mathrm{BR}\times(1-\mathrm{OR}))}; \qquad \mathrm{OR} = \frac{\mathrm{RR}(1-\mathrm{BR})}{1-(\mathrm{BR}\times\mathrm{RR})},$$

where BR is the typical risk of an event without treatment (as a number between 0 and 1). Please note that this conversion requires specification of a value of BR. Often the value of CGR is used, but use of different values of baseline risk will give different answers when the conversion is made. Sometimes it may be sensible to calculate the RR for more than one value of the BR.

8.2.1.3 Warning: OR and RR are not the same

Because risk and odds are different when events are common, the risk ratio and the odds ratio also differ when events are common. The non-equivalence of the risk ratio and odds ratio does not indicate that either is wrong: both are entirely valid ways of describing a treatment effect. Problems may arise, however, if the odds ratio is misinterpreted as a risk ratio. For treatments that increase the chances of events, the odds ratio will be larger than the risk ratio, so the misinterpretation will tend to overestimate the treatment effect, especially when events are common (with, say, risks of events more than 20%). For treatments that reduce the chances of events, the odds ratio will be smaller than the risk ratio, so that again misinterpretation overestimates the effect of treatment. This error in interpretation is unfortunately quite common in published reports of individual studies and systematic reviews.

8.2.1.4 Measure of absolute effect: the risk difference

The risk difference is the difference between the observed risks (proportions of individuals with the outcome of interest) in the two groups (Box 8.2.1). The risk difference can be calculated for any trial, even when there are no events in either group. The risk difference is straightforward to interpret: it describes the actual difference in the risk of events that was observed with treatment and with control; for an individual it describes the estimated difference in the probability of experiencing the event. However, the clinical importance of a risk difference may depend on the underlying risk of events. For example, a risk difference of 0.02 (or 2%) may represent a small, clinically insignificant change from a risk of 58% to 60% or a proportionally much larger and potentially important change from 1% to 3%. Although there are arguments that the risk difference provides more complete information than relative measures (Sackett 1997, Laupacis 1988) it is still important to be aware of the underlying risk of events and consequences of the events when interpreting a risk difference.

The risk difference is naturally constrained (like the risk ratio), which may create difficulties when applying results to other patient groups and settings. For example, if a

trial or meta-analysis estimates a risk difference of –0.1 (or –10%), then for a group with an initial risk of, say, 7% the outcome will have an impossible estimated negative probability of –3%. Similar scenarios for increases in risk occur at the other end of the scale. Such problems can arise only when the results are applied to patients with different risks from those observed in the trial(s).

The number needed to treat is obtained from the risk difference. Although it is often used to summarise results of clinical trials, NNTs cannot be combined in a meta-analysis (see 8.6.3.4 Which measure for dichotomous outcomes?).

8.2.1.5 What is the event?

In the context of dichotomous outcomes, health care interventions are intended either to reduce the risk of occurrence of an adverse outcome or increase the chance of a good outcome. All of the effect measures described above apply equally to both scenarios.

In many situations it is natural to talk about one of the outcome states as being an event. For example, when participants have particular symptoms at the start of the trial the event of interest is usually recovery or cure. If participants are well or alternatively at risk of some adverse outcome at the beginning of the trial, then the event is the onset of disease or occurrence of the adverse outcome. Because the focus is usually on the experimental intervention group, a trial in which the experimental intervention reduces the occurrence of an adverse outcome will have an odds ratio and risk ratio less than one, and a negative risk difference. A trial in which the experimental intervention increases the occurrence of a good outcome will have an odds ratio and risk ratio greater than one, and a positive risk difference (see Box 8.2.1).

However, it is possible to switch events and non-events and consider instead the proportion of patients not recovering or not experiencing the event. For meta-analyses using risk differences or odds ratios the impact of this switch is of no great consequence: the switch simply changes the sign of a risk difference, whilst for odds ratios the new odds ratio is the reciprocal ($1/x$) of the original odds ratio.

By contrast, switching the outcome can make a substantial difference for risk ratios, affecting the effect estimate, its significance, and the consistency of treatment effects across studies. This is because the precision of a risk ratio estimate differs markedly between situations with low risks of events and situations with high risks of events. In a meta-analysis the effect of this reversal cannot easily be predicted. The identification, before data analysis, of which risk ratio is more likely to be the most relevant summary statistic is therefore important and discussed further in 8.5.3.4 Which measure for dichotomous outcomes?.

8.2.2 Effect measures for continuous outcomes

The term 'continuous' in statistics conventionally refers to data that can take any value in a specified range. When dealing with numerical data, this means that any number may be measured and reported to arbitrarily many decimal places. Examples of truly continuous data are weight, area, volume and blood concentrations. In practice, in Cochrane Reviews we can use the same statistical methods for other types of data, most commonly measurement scales and counts of large numbers of events (see 8.2.3 Effect measures for ordinal outcomes (including measurement scales)).

Two summary statistics are commonly used for meta-analysis of continuous data: the mean difference and the standardised mean difference. These can be calculated whether the data from each individual are single assessments or change from baseline measures. It is also possible to measure effects by taking ratios of means, or by comparing statistics other than means (e.g. medians). However, methods for these are under development and are not addressed here.

8.2.2.1 The mean difference (and 'WMD')

The 'difference in means' is a standard statistic that measures the absolute difference between the mean value in the two groups in a clinical trial. It estimates the amount by which the treatment changes the outcome on average. It can be used as a summary statistic in meta-analysis when outcome measurements in all trials are made on the same scale. Analyses based on this effect measure are termed **weighted mean difference** (WMD) analyses in RevMan and the Cochrane Database of Systematic Reviews (CDSR). This name is potentially confusing. This is for three reasons. First, the measure is a difference in means and not a mean of differences. Second, although the meta-analysis computes a weighted average of these differences in means, no weighting is involved in calculation of a statistical summary of a single trial. Third, all meta-analyses involve a weighted combination of estimates, yet we don't use the word 'weighted' when referring to other methods.

8.2.2.2 The standardised mean difference

The **standardised mean difference** is used as a summary statistic in meta-analysis when the trials all assess the same outcome, but measure it in a variety of ways (for example, all trials measure depression but they use different psychometric scales). In this circumstance it is necessary to standardise the results of the trials to a uniform scale before they can be combined. The standardised mean difference expresses the size of the treatment effect in each trial relative to the variability observed in that trial. (Again in reality the treatment effect is a difference in means and not a mean of differences.):

$$\text{SMD} = \frac{\text{Difference in mean outcome between groups}}{\text{Standard deviation of outcome among participants}}$$

Thus trials for which the difference in means is the same proportion of the standard deviation will have the same SMD, regardless of the actual scales used to make the measurements.

However, the method assumes that the differences in standard deviations among trials reflect differences in measurement scales and not real differences in variability among trial populations. This assumption may be problematic in some circumstances where we expect real differences in variability between the participants in different trials. For example, where pragmatic and explanatory trials are combined in the same review, pragmatic trials may include a wider range of participants and may consequently have higher standard deviations. The overall treatment effect can also be difficult to interpret as it is reported in units of standard deviation rather than in units of any of the measurement scales used in the review, but in some circumstances it is possible to transform the effect back to the units used in a specific trial (see Section 8.X Re-expressing standardised mean differences).

The term 'effect size' is frequently used in the social sciences, particularly in the context of meta-analysis. Effect sizes typically, though not always, refer to versions of the standardised mean difference. It is recommended that the term 'standardised mean difference' be used in Cochrane Reviews in preference to 'effect size' to avoid confusion with the more general medical use of the latter term as a synonym for 'treatment effect' or 'effect estimate'. The particular definition of standardised mean difference used in Cochrane Reviews is the effect size known in social science as Hedges' (adjusted) *g*.

It should be noted that the SMD method does not correct for differences in the direction of the scale. If some scales increase with disease severity whilst others decrease it is essential to multiply the mean values from one set of trials by –1 (or alternatively to subtract the mean from the maximum possible value for the scale) to ensure that all the scales point in the same direction. Any such adjustment should be described in the statistical methods section of the review. The standard deviation does not need to be modified.

8.2.3 Effect measures for ordinal outcomes (including measurement scales)

Ordinal outcome data arise when each participant is classified in a category and when the categories have a natural order. For example, a 'trichotomous' outcome with an ordering to the categories, such as the classification of disease severity into 'mild', 'moderate' or 'severe' is of ordinal type. As the number of categories increases, ordinal outcomes acquire properties similar to continuous outcomes, and probably will have been analysed as such in a clinical trial.

Measurement scales are one particular type of ordinal outcome frequently used to measure conditions that are difficult to quantify, such as behaviour, depression, and cognitive abilities. Measurement scales typically involve a series of questions or tasks, each of which is scored, and the scores then summed to yield a total 'score'. If the items are not considered of equal importance a weighted sum may be used. See Box 8.4 for an example.

It is important to know whether scales have been validated: that is, that they have been proven to measure the conditions that they claim to measure. When a scale is used to assess an outcome in a clinical trial the cited reference to the scale should be studied in order to understand the objective, the target population and the assessment questionnaire. As investigators often adapt scales to suit their own purpose by adding, changing or dropping questions, check whether an original or adapted questionnaire is being used. This is particularly important when pooling outcomes for a meta-analysis. Clinical trials may appear to use the same rating scale, but closer examination may reveal differences that must be taken into account. It is possible that modifications to a scale were made in the light of the results of a trial, in order to highlight components that appear to benefit from an experimental intervention.

Specialist methods are available for analysing ordinal outcome data that describe effects in terms of **proportional odds ratios**, but they are not available in RevMan, and become unwieldy (and unnecessary) when the number of categories is large. In practice longer ordinal scales are often analysed in meta-analyses as continuous data, whilst shorter ordinal scales are often made into binary data by combining adjacent categories together. Scales may sometimes be analysed as dichotomous data if an established defensible cut-

point is available. Inappropriate choice of a cut-point can induce bias, particularly if it is chosen to maximise the difference between two intervention arms in a clinical trial.

Where ordinal scales are summarised using methods for binary data, one of the two sets of grouped categories is defined to be the event and treatment effects are described using risk ratios, odds ratios or risk differences (see 8.2.1 Effect measures for dichotomous outcomes). When ordinal scales are summarised using methods for continuous data, the treatment effect is expressed as a difference in means or standardised difference in means (see 8.2.2 Effect measures for continuous outcomes). Difficulties will be encountered if trials have summarised their results using medians (see 8. 5.2 Data extraction for continuous data).

Unless individual patient data are available, the analyses reported by the investigators in the clinical trials typically determine the approach that is used in the meta-analysis.

Box 8.2.3
An example of a scale is the Clinical Dementia Rating (CDR) (Berg 1988). The CDR is a quantitative global assessment of the severity of dementia. The clinician rates the patient's cognitive function in each of six categories: memory, orientation, judgement and problem solving, function in community affairs, function in home and hobbies, and function in personal care. Impairment is rated in each category on a five point scale (none=0, questionable=0.5, mild=1, moderate=2, severe=3). From these six ratings the CDR is established from a simple algorithm that is slightly more complex than an average. The result is a rating of no dementia (CDR=0), questionable (CDR=0.5), mild (CDR=1), moderate (CDR=2) and severe dementia (CDR=3). A second scale is formed by summing the category scores with equal weights. This is called the CDR sum of boxes and it has a range of 0 - 18.

8.2.4 Effect measures for counts and rates

Some types of event can happen to a person more than once, for example, a myocardial infarction, fracture, an adverse reaction or a hospitalisation. It may be preferable, or necessary, to address the number of times these events occur rather than simply whether each person experienced any event (that is, rather than treating them as dichotomous data). We refer to this type of data as **count data**. For practical purposes, count data may be conveniently divided into counts of rare events and counts of common events.

Counts of rare events are often referred to as 'Poisson data' in statistics. Analyses of rare events often focus on *rates*. Rates relate the counts to the amount of time during which they could have happened. For example, the result of one arm of a clinical trial could be that 18 myocardial infarctions (MIs) were experienced, across all participants in that arm, during a period of 314 person-years of follow-up, the rate is 0.057 per person year or 5.7 per 100 person years. The summary statistic used in meta-analysis is the *rate ratio* (also abbreviated to RR), which compares the rate of events in the two groups by dividing one by the other. It is also possible to use a difference in rates as a summary statistic, although this is much less common.

Counts of more common events, such as counts of decayed, missing or filled teeth, may often be treated in the same way as continuous outcome data. The treatment effect used will be the mean difference which will compare the difference in the mean number of

events (possibly standardised to a unit time period) experienced by participants in the intervention group compared to participants in the control group.

8.2.4.1 Warning: counting events or counting participants?

A common error is to attempt to treat count data as dichotomous data. Suppose that in the example just presented, the 314 person-years arose from 157 patients observed on average for 2 years. One may be tempted to quote the results as 18/157. This is inappropriate if multiple MIs from the same patient could have contributed to the total of 18 (say if the 18 arose through 12 patients having single MIs and 3 patients each having 2 MIs). It is also possible that the total number of events could theoretically exceed the number of patients, making the results nonsensical. For example, over the course of one year, 35 epileptic participants in a trial may experience 63 seizures among them.

8.2.5 Effect measures for time-to-event (survival) outcomes

Time-to-event data arise when interest is focused on the time elapsing before an event is experienced. They are known generically as **survival data** in statistics, since death is often the event of interest, particularly in cancer and heart disease. Time-to-event data consist of pairs of observations for each individual: (i) a length of time during which no event was observed, and (ii) an indicator of whether the end of that time period corresponds to an event or just the end of observation. Participants who contribute some period of time that does not end in an event are said to be 'censored'. Their event-free time contributes information and they are included in the analysis. Time-to-event data may be based on events other than death, such as recurrence of a disease event (for example, time to the end of a period free of epileptic fits) or discharge from hospital.

Time-to-event data can sometimes be analysed as dichotomous data. This requires the status of all patients in a trial to be known at a fixed time-point. For example, if all patients have been followed for at least 12 months, and the proportion who have incurred the event before 12 months is known for both groups, then a 2x2 table can be constructed (see Box 8.3) and treatment effects expressed as risk ratios, odds ratios or risk differences.

It is not appropriate to analyse time-to-event data using methods for continuous outcomes (e.g. using mean times-to-event) as the relevant times are only known for the subset of participants who have had the event. Censored participants must be excluded, which may well introduce bias.

The most appropriate way of summarising time-to-event data is to use methods of survival analysis and express the treatment effect as a **hazard ratio**. Hazard is similar in notion to risk, but is subtly different in that it measures instantaneous risk and may change continuously (for example, your hazard of death changes as you cross a busy road). A hazard ratio is interpreted in a similar way to a risk ratio, as it describes how many times more (or less) likely a participant is to suffer the event at a particular point in time if they receive the experimental rather than the control intervention. When comparing treatments in a trial or meta-analysis a simplifying assumption is often made that the hazard ratio is constant across the follow-up period, even though hazards themselves may vary continuously. This is known as the proportional hazards assumption.

8.2.6 Expressing treatment effects on log scales

The values of ratio treatment effects (such as the odds ratio, risk ratio, rate ratio and hazard ratio) undergo log transformations before being analysed, and they may occasionally be referred to in terms of their log transformed values. Typically the natural log (log base e) transformation is used.

Ratio summary statistics all have the common feature that the lowest value that they can take is 0, that the value 1 corresponds with no treatment effect, and the highest value that an odds ratio can ever take is infinity. This number scale is not symmetric. For example, whilst an odds ratio of 0.5 (a halving) and an OR of 2 (a doubling) are opposites such that they should average to no effect, the average of 0.5 and 2 is not an OR of 1 but an OR of 1.25. The log transformation makes the scale symmetric: the log of zero is minus infinity, the log of one is zero, and the log of infinity is infinity. In the example, the log of the OR of 0.5 is -0.69 and the log of the OR of 2 is 0.69. The average of -0.69 and 0.69 is 0 which is the log transformed value of an OR of 1, correctly implying no average treatment effect.

Graphics for ratio scale meta-analysis usually use a log scale. This has the effect of making the confidence intervals appear symmetric for the same reasons.

8.3 Study designs and identifying the unit of analysis

An important principle in clinical trials is that the analysis must take into account the level at which randomization occurred. In most circumstances the number of observations in the analysis should match the number of 'units' that were randomized. In a simple parallel group design for a clinical trial, participants are individually randomized to one of two intervention groups, and a single measurement for each outcome from each participant is collected and analysed. However, there are numerous variations on this design. Reviewers should consider whether in each trial

- groups of individuals were randomized together to the same intervention (i.e. cluster randomized trials);
- individuals undergo more than one intervention (e.g. in a cross-over trial, or simultaneous treatment of multiple sites on each individual);
- there are multiple observations for the same outcome (e.g. repeated measurements, recurring events, measurements on different body parts).

There follows a more detailed list of situations in which unit-of-analysis issues commonly arise, together with directions to relevant discussions elsewhere in the Handbook.

8.3.1 Cluster randomized trials

In cluster randomized trials, groups of participants are randomized to different interventions. For example, the groups may be schools, villages, medical practices, patients of a single doctor or families. See 8.X Cluster randomized trials.

8.3.2 Cross-over trials

In a cross-over trial all participants receive all interventions in sequence – they are randomized to an ordering of interventions, and participants act as their own control. See Section 8.X Cross-over trials.

8.3.3 Repeated observations on participants

In studies of long duration, results may be presented for several periods of follow-up (for example, at 6 months, 1 year and 2 years). Results from more than one time point for each trial cannot be combined in a standard meta-analysis without a unit of analysis error. Some options are:

- to obtain individual patient data and perform an analysis (such as time-to-event analysis) that uses the whole follow up for each participant. Alternatively, compute an effect measure for each individual participant which incorporates all time points, such as total number of events, an overall mean, or a trend over time. Occasionally, such analyses are available in published reports;
- to define several different outcomes, based on different periods of follow-up, and to perform separate analyses. For example, time frames might be defined to reflect short-term, medium-term and long-term follow-up;
- to select a single time point and analyse only data at this time for trials in which it is presented. Ideally this should be a clinically important time point. Sometimes it might be chosen to maximise the data available, although reviewers should be aware of the possibility of reporting biases;
- to select the longest follow-up from each trial. This may induce a lack of consistency across studies that gives rise to heterogeneity.

8.3.4 Events that may re-occur

If the outcome of interest is an event that can occur more than once, then care must be taken to avoid a unit-of-analysis error. Count data should not be treated as if they are dichotomous data. See 8.2.4 Effect measures for counts and rates.

8.3.5 Multiple treatment attempts

Similarly, multiple treatment attempts per participant can cause a unit of analysis error. Care must be taken to ensure that the number of participants randomized, and not the number of treatment attempts, is used to calculate confidence intervals. For example, in subfertility studies, women may undergo multiple cycles, and authors might erroneously use cycles as the denominator rather than women. This is similar to the situation in cluster randomized trials, except that each participant is the 'cluster'. See methods described in 8.X Cluster randomized trials.

8.3.6 Multiple body parts I: body parts receive the same treatment

In some trials, whole people are randomized, but multiple parts of the body receive the same treatment and the number of body parts is used as the denominator in the analysis. For example, eyes may be mistakenly used as the denominator without adjustment for the non-independence between eyes. This is similar to the situation in cluster randomized trials, except that participants are the 'clusters'. See methods described in 8.X Cluster randomized trials.

8.3.7 Multiple body parts II: body parts receive different treatments

A different situation is that in which different parts of the body are randomized to *different* treatments. 'Split-mouth' designs in oral health are of this sort, in which different

areas of the mouth are assigned different interventions. These are similar to cross-over trials. See methods described in Section 8.X Cross-over trials. It is important to distinguish these studies from those in which participants receive multiple versions of the *same* treatment.

8.3.8 Multiple intervention groups

Trials that compare more than two intervention groups need to be treated with care. A serious unit of analysis problem arises if the same group of participants is included twice in the same meta-analysis (for example, if 'Dose 1 vs Placebo' and 'Dose 2 vs Placebo' are both included in the same meta-analysis, with the same placebo patients in both comparisons). See 8.X Trials with more than two treatment groups.

8.4 Intention to treat issues

From the emphasis given to proper randomisation it follows that analysis of a randomised trial should ideally compare the groups exactly as randomised. Often some participants are excluded, either because they were lost to follow up and no outcome was obtained, or for some deviation from the protocol, such as receiving the wrong treatment or no treatment, lack of compliance, or ineligibility. Alternatively, it may be impossible to measure certain outcomes for all participants because their availability depends on another outcome (see 8.4.4 Identifying conditional outcomes only available for subsets of participants).

8.4.1 What are intention-to-treat analyses?

An estimated treatment effect may be biased if some randomised participants are excluded from the analysis. Imbalances in such omissions between groups may be especially indicative of bias. Intention-to-treat (ITT) analysis aims to include all participants randomized into a trial irrespective of what happened subsequently (Lewis 1993, Newell 1992). ITT analyses are generally preferred as they are unbiased, and also because they address a more pragmatic and clinically relevant question.

The simple idea of an ITT analysis, to include all randomised patients, is not always easy to implement, and there are confusions about terminology. There are two criteria for an ITT analysis:

1. Trial participants should be analysed in the groups to which they were randomised regardless of which (or how much) treatment they actually received, and regardless of other protocol irregularities, such as ineligibility.

2. All participants should be included regardless of whether their outcomes were actually collected.

There is no clear consensus on whether both criteria should be applied (Hollis 1999). While the first is widely agreed, the second is contentious, since to include participants whose outcomes are unknown (mainly through loss to follow up) involves 'filling-in' ('imputing') missing data.

Many trials report having undertaken ITT analyses when they have met only the first of the two criteria, the second being impossible to achieve when contact is lost with the trial participants. An analysis in which data are analysed for every participant for whom the outcome was obtained is more properly called an **available case analysis**. Some trial reports present analyses of the results of only those participants who completed the trial *and* who complied with (or received some of) their allocated treatment. Some authors incorrectly call these ITT analyses, but they are in fact **per-protocol** or **treatment-received** analyses. Here we interpret the term ITT to mean that both of the above criteria are fulfilled. Reviewers should critically consider and report which type of analysis each trial has presented. Reviewers should avoid using the terms 'intention-to-treat' and 'ITT' without explicitly defining them.

8.4.1.1 Available case analyses

In most situations reviewers should attempt to extract from papers the data to enable at least an **available case analysis**. Avoidable exclusions should be reinstated if possible. The proportion of participants in each study arm who do not provide outcome data should be noted in the Study Characteristics table.

Three types of exclusions deserve specific mention. First, some trial participants may legitimately be excluded (i.e. without introducing bias) if their reason for exclusion was specified in the protocol and relates only to information collected before randomisation. For example, a condition may be defined by delayed blood tests on samples taken before randomization. Such exclusions are generally unwise, however, as the results do not then relate to the real clinical situation.

Second, and by contrast, exclusions immediately post-randomisation (and perhaps before treatment) may introduce bias, as they could be related to the treatment allocation.

Third, if dropout is very high or is different across treatment groups then the systematic review's protocol may dictate that a study be given a low quality rating and perhaps excluded from a meta-analysis (though usually not from the systematic review).

Many (but not all) people consider that available case and ITT analyses are not appropriate when assessing unintended (adverse) effects, as it is wrong to attribute these to a treatment that somebody did not receive. As ITT analyses tend to bias the results towards no difference they may not be the most appropriate when attempting to establish equivalence or non-inferiority of a treatment.

8.4.1.2 Full intention-to-treat analyses

In some rare situations it is possible to create a genuine ITT analysis from information presented in the text and tables of the paper, or by obtaining extra information from the author about participants who were followed up but excluded from the trial report. If this is possible without imputing study results it should be done.

Otherwise an intention to treat analysis can only be produced by using imputation. This involves making assumptions about the outcomes of participants for whom no outcome was recorded, and making up data for these participants. Some statistical techniques exist for imputing data but, ultimately, assessing the results of trials in the presence of more than minimal amounts of missing data is a matter of judgement. Statistical analysis cannot reliably compensate for missing data (Unnebrink 2001). No assumption is likely

adequately to reflect the truth, and the impact of any assumption should be assessed by trying more than one method as a sensitivity analysis (see 8.X Sensitivity analyses).

In the next two sections we consider some ways to take account of missing observations for dichotomous or continuous outcomes. Although imputation is possible, at present a sensible decision in most cases is to include data for only those participants whose results are known, and discuss the potential impact of the missing data. Where imputation is used the methods and assumptions for imputing data for dropouts should be described in the Methods section of the protocol and review.

8.4.2 ITT issues for dichotomous data

Percentages of participants for whom no outcome data were obtained should always be collected and reported in the Characteristics of Included Studies table; note that the percentages may vary by outcome. However, there is no consensus on the best way to handle these participants in an analysis. There are two basic options, and it may be wise to plan to undertake both and compare their results in a sensitivity analysis (see 8.X Sensitivity Analyses).

- **Available case analysis:** Include data on only those whose results are known, using as a denominator the total number of people who completed the trial for the particular outcome in question. The potential impact of the missing data on the results should be considered in the interpretation of the results of the review. This will depend on the degree of 'missingness', the frequency of the events and the size of the pooled effect estimate. Variation in the degree of missing data across studies may also be considered as a potential source of heterogeneity.

- **ITT analysis using imputation:** Base an analysis on the total number of randomized participants, irrespective of how the original trialists analysed the data. This will involve 'imputing' (a formal term for 'making up') outcomes for the missing patients. Studies with imputed data will be given more weight than they warrant if entered as dichotomous data into RevMan. It is possible to determine more appropriate weights; consultation with a statistician is recommended.

  There are several approaches to imputing dichotomous outcome data. One common approach is to assume either that all missing participants experienced the event, or that all missing participants did not experience the event. The choice among these assumptions should be based on clinical judgement as to what would be the most likely outcome. An alternative approach is to impute data according to the event rate observed in the control group, or according to event rates among completers in the separate groups. None of these assumptions is likely to reflect the truth, and the latter achieves little other than an unwarranted inflation of the precision of effect estimates. Thus this approach is generally not recommended. The impact of any assumptions can be tested by undertaking sensitivity analyses where first it is assumed that all missing participants in the first group incurred the event and those in the second group did not, and then assuming the opposite. When missing data are common, these worst-case/best-case scenarios will cover a very wide range of possible treatment effects and thus the analysis will not be very informative. However, when missing data are not common and this procedure is done across all trials in the review with little impact on

the results, it can be concluded that the missing data could not affect the outcome of the review.

8.4.3 ITT issues for continuous data

In full ITT analyses, all participants who did not receive the assigned intervention according to the protocol as well as those who were lost to follow-up are included in the analysis. Inclusion of these in an analysis requires that means and standard deviations for all randomized participants are available. As for dichotomous data, dropout rates should always be collected and reported in the Characteristics of Included Studies table. There are two basic options, and it may be wise to plan to undertake both and formally compare their results in a sensitivity analysis (see 8.X Sensitivity Analyses).

- **Available case analysis:** Include data only on those whose results are known. The potential impact of the missing data on the results should be considered in the interpretation of the results of the review. This will depend on the degree of 'missingness', the pooled estimate of the treatment effect and the variability of the outcomes. Variation in the degree of missing data may also be considered as a potential source of heterogeneity.

- **ITT analysis using imputation:** Base an analysis on the total number of randomized participants, irrespective of how the original trialists analysed the data. This will involve imputing outcomes for the missing patients. Approaches to imputing missing continuous data in the context of a meta-analysis have received little attention in the methodological literature. In some situations it may be possible to exploit standard (although often questionable) approaches such as 'last observation carried forward', or, for change from baseline outcomes, to assume that no change took place, but such approaches generally require access to the raw patient data. Inflating the sample size of the available data up to the total numbers of randomized participants is based on an assumption that those dropping out from the study were a random sample of all those included, and is not recommended as it will artificially inflate the precision of the effect estimate

8.4.4 Identifying conditional outcomes only available for subsets of participants

Some trial outcomes may only be applicable to a proportion of participants. For example, in subfertility trials the proportion of clinical pregnancies that miscarry following treatment is often reported. By definition this outcome excludes participants who do not achieve an interim state (clinical pregnancy), so the comparison is not of all participants randomized. As a general rule it is better to re-define such outcomes so that the analysis includes all randomized participants. In this example, the outcome could be whether the woman has a 'successful pregnancy' (becoming pregnant and reaching, say, 24 weeks or term).

Another example is a morbidity outcome measured in the medium or long term (e.g. development of chronic lung disease), when there is a distinct possibility of a death preventing assessment of the morbidity. A convenient way to deal with such situations is to combine the outcomes, for example as 'death or chronic lung disease'.

Some intractable problems arise when a continuous outcome (say a measure of functional ability or quality of life following stroke) is measured only on those who survive to the end of follow-up. Two unsatisfactory alternatives exist: (a) imputing zero functional ability scores for those who die (which may not appropriately represent the death state and will make the outcome severely skewed), and (b) analysing the available data (which must be interpreted as a non-randomised comparison applicable only to survivors).

8.5 Extraction of study results

This section outlines the data that need to be extracted from trial reports for analyses of each of the data types described in 8.2 Types of data and effect measures. For many studies the required data will be presented clearly. However, sometimes the required data may be obtained only indirectly, and the relevant results may not be obvious. This section provides some useful tips and techniques to deal with these situations.

The section concludes with some important considerations that despite being mentioned last must be considered before starting the data extraction process. First, a common error when extracting data is to fail to recognise what the unit of analysis should be. A 'unit of analysis error' may arise when results entered into an analysis do not suitably reflect the design of the study. It is important to recognise such situations. Second, intention-to-treat analyses may require collection of data from different parts of a paper.

8.5.1 Data extraction for dichotomous outcomes

Dichotomous data are described in 8.2.1 Effect measures for dichotomous outcomes. The only data required for a dichotomous outcome are the numbers in each of the two categories in each of the intervention groups – the numbers needed to fill in the four boxes *a*, *b*, *c* and *d* in Box 8.2.1. The data are often available as the number assessed and the number incurring the event of interest in each group. Difficulties may be experienced in clearly identifying the numbers actually assessed for each outcome due to poor reporting, and occasionally the numbers incurring the event need to be derived from percentages (although it is not always clear which denominator to use, and rounded percentages may be compatible with more than one numerator).

See also 8.6.3 Meta-analysis of dichotomous outcomes.

8.5.1.1 Extracting effect estimates calculated from dichotomous outcomes

Sometimes the numbers of participants and numbers of events are not available, but results calculated from them are. For example, an estimate of an odds ratios or a risk ratio may be present in an abstract, while the full text of the paper cannot be obtained so further data are unavailable. Such data may be included in meta-analyses only if they are accompanied by measures of uncertainty such as a 95% confidence interval or an exact *P*-value. The numbers then must be analysed using the generic inverse variance method in RevMan (see 8.6.2 A generic inverse variance approach to meta-analysis). This requires the reviewer to enter an estimate and a standard error for each study. The process of obtaining a suitable estimate and standard error from a confidence interval or *P*-value is described in 8.5.6 Obtaining standard errors from confidence intervals and *P*-values

A limitation of this approach is that estimates and standard errors of the same effect measure must be calculated for all the other studies in the same meta-analysis, even if they provide the original numbers of participants and events. If the numbers of events and participants are known the necessary summary statistics may be obtained from RevMan (entering the data as dichotomous data), and copied manually into the data entry window for the generic inverse variance outcome. The confidence intervals estimated in RevMan will need to be converted into standard errors.

When extracting data from non-randomized studies, and from some randomized studies, adjusted odds ratios may be available from logistic regression analyses. The process of data extraction, and analysis using the generic inverse variance method, is the same as for unadjusted estimates.

8.5.2 Data extraction for continuous outcomes

Continuous data are described in 8.2.2 Effect measures for continuous outcomes. To perform a meta-analysis of continuous data using either mean differences or standardised mean differences one needs to extract the mean values of the outcomes, the standard deviations of the outcomes, and the number of participants on whom the outcome was assessed in each of the two groups.

In many cases the relevant information can be extracted directly from trial reports in a straightforward way. However, due to poor and variable reporting occasionally it is difficult or impossible to obtain the necessary information from the data summaries presented. Trials vary in the statistics they use to summarise average (sometimes using medians rather than means) and variation (sometimes using standard errors, confidence intervals, interquartile ranges and ranges rather than standard deviations).

When needed, missing information and clarification about the statistics presented should always be sought from the authors. However, for several of the measures of variation there is an approximate or direct algebraic relationship with standard deviations, so it may be possible to obtain the required statistic even if it is not published directly in the paper as is explained in the subsections that follow. For more details and examples see (Deeks 1997a, Deeks 1997b).

A particularly misleading error is to misinterpret a standard error as a standard deviation. Unfortunately it is not always clear what is being reported and some intelligent reasoning may be required. Standard deviations and standard errors are occasionally confused by authors of trial reports, and the terminology is used inconsistently.

See also 8.6.4 Meta-analysis of continuous outcomes.

8.5.2.1 Medians

The median is very similar to the mean when the distribution of the data is symmetrical, and so occasionally can be used directly in meta-analyses. However, means and medians can be very different from each other if the data are skewed, and medians are often the summary statistic of choice when data are skewed (see 8.5.2.11 Skewed data).

8.5.2.2 Standard errors of group means

Standard deviations are obtained by multiplying standard errors of means by the square-root of the sample size:

$$SD = SE \times \sqrt{N}$$

When making this transformation ensure that standard errors are standard errors of means calculated from *within* a treatment group and not standard errors of the difference in means computed *between* treatment groups.

8.5.2.3 Confidence intervals for group means

Confidence intervals for means can also be used to calculate standard deviations via calculation of the standard error of the mean. The following applies to confidence intervals for mean values calculated *within* treatment group results and *not* from comparisons of treatments. Most confidence intervals are 95% confidence intervals. If the sample size is large (say bigger than 100), the 95% confidence interval is 3.92 (2 × 1.96) standard errors wide. The standard deviation for each group is obtained by dividing the length of the confidence interval by 3.92, and then multiplying by the square root of the sample size:

$$SD = \sqrt{N} \times \text{(upper limit – lower limit)}/3.92$$

For 90% confidence intervals divide by 3.29 rather than 3.92, for 99% confidence intervals divide by 5.15.

If the sample size is smaller than 60 then confidence intervals should have been calculated using a value from a *t*-distribution. The numbers 3.92, 3.29 and 5.15 need to be replaced with slightly larger numbers specific to both the *t*-distribution and the sample size which can be obtained from tables of the *t*-distribution with degrees of freedom equal to the group sample size minus 1. (Relevant details of the *t*-distribution are available as appendices of many statistical textbooks, or using standard computer spreadsheet packages. For example the *t*-value for a 95% confidence interval from a sample size of 27 can be obtained by typing **=tinv(1-0.95,27-1)** in a cell in a Microsoft Excel spreadsheet.)

As an example, consider data presented as follows:

Group	*Sample size*	*Mean*	*95% CI*
Experimental intervention	25	32.1	(30.0, 34.2)
Control intervention	22	28.3	(26.5, 30.1)

The confidence intervals should have been based on *t*-distributions with 24 and 21 degrees of freedom respectively. The relevant numbers for the divisor are then 2 × 2.06 = 4.12 and 2 × 2.08 = 4.16. The standard deviations for the two groups are $\sqrt{25} \times (34.2 – 30.0)/4.12 = 5.10$ and $\sqrt{22} \times (30.1 – 26.5)/4.16 = 4.06$.

It is important to check that the confidence interval is symmetrical about the mean (the distance between the lower limit and the mean is the same as the distance between the mean and the upper limit). If this is not the case the confidence interval may have been calculated on transformed values (see Section 8.5.2.11 Skewed data below).

8.5.2.4 *t*-values, standard errors and confidence intervals for differences in means

The same ingredients of means, standard deviations and sample sizes are involved in *t*-tests used to compute the statistical significance of differences in means. The methods do not actually estimate the two standard deviations observed in the two groups but estimate

the average of their values. This simplification does not matter for the purpose of meta-analysis.

The *t*-value is the ratio of the difference in means to the standard error of the difference in means. Computing the standard deviation first involves computing the standard error of the difference in means by dividing the difference in means (MD) by the *t*-value:

$$\text{standard error of difference in means} = \frac{\text{MD}}{t}$$

If a 95% confidence interval is available for the difference in means, then the same standard error can be calculated as:

$$\text{SE} = (\text{upper limit} - \text{lower limit})/3.92$$

as long as the trial is large. For 90% confidence intervals divide by 3.29 rather than 3.92, for 99% confidence intervals divide by 5.15. If the sample size is small then confidence intervals should have been calculated using a *t*-distribution. The numbers 3.92, 3.29 and 5.15 need to be replaced with larger numbers specific to both the *t*-distribution and the sample size, and can be obtained from tables of the *t*-distribution with degrees of freedom equal to $N_E + N_C - 2$, where N_E and N_C are the sample sizes in the two groups. (Relevant details of the *t*-distribution are available as appendices of many statistical textbooks, or using standard computer spreadsheet packages. For example the *t*-value for a 95% confidence interval from a comparison of a sample size of 27 with a sample size of 24 can be obtained by typing **=tinv(1-0.95,27+24-2)** in a cell in a Microsoft Excel spreadsheet).

The standard deviation can then be obtained from the standard error of the difference in means using the following formula:

$$\text{standard deviation} = \frac{\text{standard error of difference in means}}{\sqrt{\left(\frac{1}{N_E} + \frac{1}{N_C}\right)}}.$$

See below (Section 8.5.2.5 *P*-values) for an example. This standard deviation must be entered into RevMan for both intervention groups.

Related methods can be used to derive standard deviations from certain F-statistics, although methods are somewhat complex and advice of a knowledgeable statistician is recommended.

8.5.2.5 P-values

Where actual *P*-values obtained from *t*-tests are quoted, it is possible to extract standard deviations by first obtaining the corresponding *t*-value from a table of the *t*-distribution (noting that the degrees of freedom are given by $N_E + N_C - 2$), and then transforming the *t*-value into a standard deviation as described in 8.5.2.4 *t*-values, standard errors and confidence intervals for differences in means.

As an example, consider a trial of an experimental intervention ($N_E = 25$) versus a control intervention ($N_C = 22$), where the difference in means was MD = 3.8. It is noted that the *P*-value for the comparison was *P* = 0.008 obtained using a two-sample *t*-test.

The *t*-statistic that corresponds with a *P*-value of 0.008 and 25+22-2=45 degrees of freedom is *t* = 2.78. This can be obtained from a table of the t-distribution with 45 degrees of freedom or a computer (for example, by entering **=tinv(0.008, 45)** into any cell in a Microsoft Excel spreadsheet).

The standard error of the difference in means is obtained by dividing the MD (3.8) by the t-value (2.78), which gives 1.37. To calculate the standard deviation from the *t*-statistic we use

$$\text{standard deviation} = \frac{1.37}{\sqrt{\left(\frac{1}{25} + \frac{1}{22}\right)}} = 4.69 .$$

Note that this standard deviation is the average of the standard deviations of the experimental and control arms, and must be entered into RevMan for both groups.

Difficulties are encountered when levels of significance are reported (such as *P*<0.05 or even *P*=NS which usually implies *P*>0.05) rather than exact *P*-values. A conservative approach would be to take the *P*-value at the upper limit (e.g. for *P*<0.05 take *P*=0.05, for *P*<0.01 take *P*=0.01 and for *P*<0.001 take *P*=0.001). However, this is not a solution for results which are reported as *P*=NS. It may be preferable to impute a value for the standard deviation for studies that report *P*=NS from those observed in other studies rather than inevitably introducing bias by excluding them from the meta-analysis (see 8.X Missing Data).

8.5.2.6 Interquartile ranges

Interquartile ranges describe where the central 50% of participants' outcomes lie. When sample sizes are reasonably large and the distribution of the outcome is similar to the normal distribution, the width of the interquartile range will be approximately 1.35 standard deviations. In other situations, and especially when the outcome's distribution is skewed, it is not possible to estimate a standard deviation from an interquartile range. Note that the use of interquartile ranges rather than standard deviations can often be taken as an indicator that the outcome's distribution is skewed.

8.5.2.7 Ranges

Ranges are very unstable and, unlike other measures of variation, increase when the sample size increases. They describe the extremes of observed outcomes rather than the average variation. It is not possible to reliably estimate a standard deviation from a range. One common approach has been to make use of the fact that, with normally distributed data, 95% of values will lie within 2×SD either side of the mean. The SD may therefore be estimated to be approximately one quarter of the 'typical' range of data values. This method is not robust and is discouraged.

8.5.2.8 No information on variability

If none of the above methods allow calculation of the standard deviation(s) from the trial report (and the information is not available directly from the trialists) then, in order to

perform a meta-analysis, a reviewer is forced either to exclude the study and risk introducing bias, or to impute missing data (see 8.X Missing data) and risk making a different type of error. Alternatively a narrative approach to synthesis may be used. It is valuable to tabulate available results for all studies included in the systematic review, even if they cannot be included in a formal meta-analysis.

8.5.2.9 Change from baseline

A common feature of continuous data (and also possible with ordinal data) is that a measurement used to assess the outcome of each participant is also measured at baseline, that is at or before randomization into the trial. This gives rise to the possibility of using differences in **changes from baseline** (also called a **change score**) as the primary outcome. Reviewers are advised not to focus on change from baseline unless this method of analysis was used in some of the trial reports.

When addressing change from baseline, a single measurement is created for each participant, obtained either by subtracting the final measurement from the baseline measurement or by subtracting the baseline measurement from the final measurement. Analyses then proceed as for any other type of continuous outcome variable using the changes rather than the final measurements.

The principal difficulty associated with change from baseline analyses is the availability of data from published reports. It is very common for standard deviations of the changes to be unavailable. A common situation is that the following data are available:

	Baseline	*Final*	*Change*
Experimental intervention (sample size n_1)	mean, SD	mean, SD	mean
Control intervention (sample size n_2)	mean, SD	mean, SD	mean

Note that the mean change in each group can always be obtained by subtracting the final mean from the baseline mean even if it is not presented explicitly. However, the information in this table does *not* allow us to calculate the standard deviation of the changes. We cannot know whether the changes were very similar or very variable. Some other information in a paper may help us determine the standard deviation of the changes. If statistical analyses comparing the changes themselves are presented (e.g. confidence intervals, *t*-values or *P*-values) then the techniques described above (see Sections 8.5.2.3 to 8.5.2.5) may be used.

In other situations it is possible to impute standard deviations for the changes. Follmann (Follmann 1992) discusses techniques for imputing missing standard deviations, some of which are described in Section 8.5.2.10 Imputing standard deviations for changes from baseline. However, all imputation techniques involve making assumptions about unknown statistics, and it is best to avoid using them wherever possible If they are used the impact of the imputations should be tested in planned sensitivity analyses (see 8.X Sensitivity analyses). Imputed standard deviations should not be used for a majority of studies in a meta-analysis, but may be reasonable for a small proportion of studies comprising a small proportion of the data if it enables them to be combined with other studies for which full data are available.

Reviewers are advised to extract data on both change from baseline and final value outcomes if the required means and standard deviations are available. Commonly a reviewer will find that they end up with a mixture of changes from baseline and final values for trials included in a review. Some trials will report both; others will report only change scores or only final values. As explained in Section 8.6.4.2 Meta-analysis of change scores, both final values and change scores can often be combined in the same analysis so this is not necessarily a problem.

A final problem with using change from baseline measures is that often baseline and final measurements will be reported for different numbers of participants due to missed visits and study withdrawals. It may be difficult to identify the subset of participants who report both baseline and final value measurements for whom change scores can be computed.

8.5.2.10 Imputing standard deviations for changes from baseline

A 'hidden' number known as the correlation coefficient describes how similar the baseline and final measurements were across participants. Here we describe (1) how to estimate the correlation coefficient from a study that is reported in considerable detail and (2) how to impute a change from baseline standard deviation in another study, making use of an imputed correlation coefficient. Note that the methods in (2) are applicable both to correlation coefficients obtained using (1) and to correlation coefficients obtained in other ways (for example, by reasoned argument). These methods should be used sparingly, if at all. This is partly because one can never be sure that an imputed correlation is appropriate (correlations between baseline and final values will, for example, decrease with increasing time between baseline and final measurements, as well as depending on the outcomes and characteristics of the participants). A further reason is that a comparison of final measurements in a randomised trial in theory estimates the same quantity as the comparison of changes from baseline, so imputation is often not necessary to enable trials to be included in the analysis.

(1) Suppose a study is available that presents the following information:

	Baseline	*Final*	*Change*
Experimental intervention (sample size n_1)	$mean_1(B)$, $SD_1(B)$	$mean_1(F)$,$SD_1(F)$	$mean_1(C)$,$SD_1(C)$
Control intervention (sample size n_2)	$mean_2(B)$, $SD_2(B)$	$mean_2(F)$, $SD_2(F)$	$mean_2(C)$, $SD_2(C)$

An analysis of change from baseline is available from this study, using only the data in the final column. We can use the other data from the study to estimate the correlation coefficient in the experimental intervention, r_1, as follows:

$$r_1 = \frac{SD_1(B)^2 + SD_1(F)^2 - SD_1(C)^2}{2 \times SD_1(B) \times SD_1(F)},$$

and similarly for the control intervention, r_2. Where either *SD(F)* or *SD(B)* are unavailable, then it may be substituted by the other if it is reasonable to assume that the intervention does not alter the variability of the outcome measure. Correlation coefficients lie between –1 and 1. If zero or a negative number is obtained, then there is no value in using change from baseline and an analysis of final values should be performed.

Assuming the correlation coefficients from the two intervention groups are similar, a simple average will provide a reasonable measure of the similarity of baseline and final measurements across individuals. If the correlation coefficients differ, then either the sample sizes are too small for reliable estimation, or the intervention is affecting the variability in outcome measures, and the use of imputation is best avoided. Before imputation is undertaken it is recommended that correlation coefficients are computed for many (if not all) studies in the meta-analysis and it is noted whether or not they are consistent. Imputation should be done only as a very tentative analysis if correlations are inconsistent.

(2) To impute the standard deviation of a change from baseline, when baseline and final standard deviations are known, we use an imputed value R_I for the correlation coefficient. The value R_I might be imputed from another study in the meta-analysis (using the method in (1) above), it might be imputed from elsewhere, or it might be hypothesised based on reasoned argument. In all of these situations, a sensitivity analysis should be undertaken, trying different values of R_I, to determine whether the overall result of the analysis is robust to the use of imputed correlation coefficients.

To obtain a standard deviation of the change from baseline for the experimental intervention, use

$$SD_1(C) = \sqrt{SD_1(B)^2 + SD_1(F)^2 - (2 \times R_I \times SD_1(B) \times SD_1(F))},$$

and similarly for the control intervention. Again, if either *SD(F)* or *SD(B)* are unavailable, then one may be substituted by the other if it is reasonable to assume that the intervention does not alter the variability of the outcome measure.

As an example, given the following data:

	Baseline	*Final*	*Change*
Experimental intervention (sample size 35)	mean=12.4 SD=4.2	mean=15.2 SD=3.8	mean=2.8
Control intervention (sample size 38)	mean=10.7 SD=4.0	mean=13.8 SD=4.4	mean=3.1

and using an imputed correlation coefficient of 0.5, we can impute the standard deviation for the change score in the control group as:

$$SD_2(C) = \sqrt{4.0^2 + 4.4^2 - (2 \times 0.5 \times 4.0 \times 4.0)} = 4.21.$$

8.5.2.11 Skewed data

Analyses based on means or standardised means are appropriate for data that are at least approximately normally distributed, and for data from very large trials. If the true distribution of outcomes is asymmetrical then the data are said to be skewed. Methods for meta-analysing skewed data are lacking at present, though they are the subject of current research.

Transformation of the original outcome data may substantially reduce skewness. Reports of trials may present results on a transformed scale, usually a log scale. More often they

do not. Collection of appropriate data summaries from the trialists, or acquisition of individual patient data, is currently the approach of choice. Appropriate data summaries and analysis strategies for the individual patient data will depend on the situation. Consultation with a knowledgeable statistician is advised.

With the more common positive skewness, presentation of a 'geometric mean' with its 95% confidence interval is equivalent to an analysis of a log transformation of the data. The difference in means of the log transformed data may be obtained from a ratio of geometric means (geometric mean ratio, GMR) as log(GMR), and the standard error of this difference as [log(lower confidence limit for GMR) – log(upper confidence limit for GMR)]/3.92. The standard deviation of the log transformed data may be determined from the standard error as described above (see Sections 8.5.2.2 to 8.5.2.5). This approach depends on being able to obtain transformed data for all trials. Log-transformed and untransformed data can not be mixed in a meta-analysis.

Skewness can sometimes be diagnosed from the means and standard deviations of the outcomes. A rough 'check' is available, but it is only valid if a lowest or highest possible value for an outcome is known to exist. Thus the check may be used for outcomes such as weight, volume and blood concentrations, which have lowest possible values of 0, or for scale outcomes that may have lowest and highest possible values. The check is not appropriate for change from baseline measures. The check involves calculating the observed mean minus the lowest possible value (or the highest possible value minus the observed mean), and dividing this by the standard deviation. A ratio less than 2 suggests skewness. If the ratio is less than 1 there is strong evidence of a skewed distribution (Altman 1996).

It should be noted that skewness is not necessarily a problem for meta-analyses in RevMan if the sample sizes in the individual studies are large.

8.5.2.12 Extracting effect estimates calculated from continuous data

Sometimes only effect estimates (estimates of a mean difference or standardized mean difference) are available with a standard error or confidence interval. If this is the case, the analysis should be performed using the generic inverse variance method in RevMan (8.6.2 A generic inverse variance approach to meta-analysis). This requires the reviewer to enter the estimate and standard error for each study. The process of obtaining a suitable standard error from a confidence interval for a mean difference is described in 8.5.2.4 *t*-values, standard errors and confidence intervals for differences in means. For standardized mean differences, see 8.5.6 Obtaining standard errors from confidence intervals and *P*-values.

A limitation of this approach is that all other studies in the same meta-analysis must provide estimates and standard errors of the same effect measure, even if they provide the six numbers usually required to analyse continuous data. However, the necessary numbers may be obtained from RevMan (entering the data as continuous data), and copied manually into the data entry window for a generic inverse variance outcome, converting the confidence interval into a standard error.

When extracting data from non-randomized studies, and from some randomized studies, adjusted estimates of mean differences may be available from multiple regression

analyses and analyses of covariance. The process of data extraction and analysis using the generic inverse variance method is the same as for unadjusted estimates.

8.5.3 Data extraction for ordinal outcomes and measurement scales

Ordinal data and measurement scales are described in 8.2.3 Effect measures for ordinal outcomes (including measurement scales). The data that need to be extracted for ordinal outcomes depend on whether the ordinal scale will be dichotomised for analysis (see 8.5.1 Data extraction for dichotomous data), treated as a continuous outcome (see 8.5.2 Data extraction for continuous data) or analysed directly as ordinal data. This decision, in turn, will be influenced by the way in which authors of the trials analysed their data. Thus it may be impossible to pre-specify whether data extraction will involve calculation of numbers of participants above and below a defined threshold, or mean values and standard deviations. In practice, it is wise to extract data in all forms in which they are given as it will not be clear which is the most common until all trials have been reviewed, and in some circumstances more than one form of analysis may justifiably be included in a review.

Where ordinal data are being dichotomised and there are several options for selecting a cutpoint (or the choice of cutpoint is arbitrary) it is sensible to plan from the outset to investigate the impact of choice of cutpoint in a sensitivity analysis (see Section 8.X Sensitivity Analyses). To do this it is necessary to collect the data that would be used for each alternative dichotomisation. Hence it is preferable to record the numbers in each category of short ordinal scales to avoid having to extract data from a paper multiple times. This approach of recording all categorisations is also sensible when trials use slightly different short ordinal scales, and it is not clear whether there will be a cutpoint that is common across all the trials which can be used for dichotomisation.

It is also necessary to record the numbers in each category of the ordinal scale for each treatment group if the proportional odds ratio method (see 8.2.3 Effect measures for ordinal outcomes (including measurement scales)) will be used.

8.5.4 Data extraction for counts and rates

Counts and rates are described in 8.2.4 Effect measures for counts and rates. Data that are inherently counts may be analysed in several ways. The essential decision is whether to make the outcome of interest dichotomous, continuous, time-to-an-event or a rate. A common error is to treat counts directly as dichotomous data, using as sample sizes either the total number of participants or the total number of, say, person-years of follow-up. Neither of these approaches is appropriate for an event that may occur more than once for each participant. This becomes obvious when the total number of events exceeds the sample size, leading to nonsensical results. Although it is preferable to decide how count data will be analysed in advance, the choice is often determined by the format of the available data, and thus cannot be decided until the majority of studies have been reviewed.

8.5.4.1 Extracting counts as dichotomous data

To consider the outcome as a dichotomous outcome, the reviewer must determine the number of participants in each intervention group, and the number of participants in each intervention group who experience *at least one* event (or some other appropriate criterion

which classified all participants into one of two possible groups). Any time element in the data is lost through this approach, though it may be possible to create a series of dichotomous outcomes, for example 'at least one stroke during the first year of follow-up', 'at least one stroke during the first two years of follow-up', and so on. Such data may be hard to derive from published reports. See also 8.6.3 Meta-analysis of dichotomous outcomes.

8.5.4.2 Extracting counts as continuous data

To extract counts as continuous data, guidance in 8.5.2 Data extraction for continuous outcomes should be followed, although particular attention should be paid to the likelihood that the data will be highly skewed. See also 8.6.4 Meta-analysis of continuous outcomes.

8.5.4.3 Extracting counts as time-to-event data

For rare events that can happen more than once, a reviewer may be faced with studies that treat the data as time-to-*first*-event. To extract counts as time-to-event data, guidance in 8.5.5 Data extraction for time-to-event outcomes should be followed. See also 8.6.8 Meta-analysis of time-to-event outcomes.

8.5.4.4 Extracting counts as rate data

To analyse rate data a reviewer should extract the total number of events in each group, and the total amount of person-time at risk in each group. Unlike for dichotomous data, the total number of events may include multiple events for some participants, and may even exceed the total number of participants. Note that the total number of participants is not required for an analysis of rate data but you will probably wish to record it as part of the trial description. See also 8.6.7 Meta-analysis of counts and rates.

8.5.4.5 Extracting effect estimates calculated from rate data

Sometimes detailed data on events and person-years at risk are not available, but results calculated from them are. For example, an estimate of a rate ratio or rate difference may be present in an abstract, while the full text of the paper unavailable. Such data may be included in meta-analyses only if they are accompanied by measures of uncertainty such as a 95% confidence interval. See 8.5.6 Obtaining standard errors from confidence intervals and *P*-values. When extracting data from non-randomized studies, and from some randomized studies, adjusted rate ratios may be available from Poisson regression analyses. Data extraction is the same as for unadjusted rate ratios.

8.5.5 Data extraction for time-to-event outcomes

Meta-analysis of time-to-event data commonly involves obtaining individual patient data from the trialists, re-analysing the data to obtain estimates of the log hazard ratio and its standard error, and then performing a meta-analysis. Conducting a meta-analysis using summary information from published papers or trial reports is often problematic as the most appropriate summary statistics are typically not explicitly presented.

Two approaches can be used to obtain estimates of log hazard ratios regardless of whether individual patient data or aggregate data are being used.

In the first approach an estimate of the log hazard ratio can be obtained from statistics computed during a logrank analysis. Collaboration with a knowledgeable statistician is advised if this approach is followed. The log hazard ratio (experimental relative to control) is estimated by (O – E)/V, which has standard error $1/\sqrt{V}$, where O is the

observed number of events on the experimental intervention, E is the logrank expected number of events on the experimental intervention, (O – E) is the logrank statistic and V is the variance of the logrank statistic. It is therefore necessary to obtain values of O – E and V for each study.

These statistics are easily computed if individual patient data are available, and can sometimes be extracted from quoted statistics and survival curves as discussed by Parmar, Torri and Stewart (Parmar 1998). Alternatively, use can sometimes be made of aggregated data for each treatment group in each trial. For example, suppose that the data comprise the number of participants who have the event during the first year, second year, etc., and the number of participants who are event free and still being followed up at the end of each year. A logrank analysis can be performed on these data, to provide the (O – E) and V values, although careful thought needs to be given to the handling of censored times. Because of the coarse grouping the log hazard ratio is estimated only approximately, and in some reviews has been referred to as a log odds ratio (Early Breast Cancer Trialists' Collaborative Group 1990). If the time intervals are large, a more appropriate approach is one based on interval-censored survival (Collett 1994).

The second approach can be used if trialists have analysed the data using a Cox proportional hazards model, or if a Cox model is fitted to individual patient data. Cox models produce direct estimates of the log hazard ratio and its standard error. If the hazard ratio is quoted in a report together with a confidence interval or *P*-value, estimates of standard error can be obtained as described in 8.5.6 Obtaining standard errors from confidence intervals and *P*-values.

8.5.6 Obtaining standard errors from confidence intervals and *P*-values

Estimates of an effect measure of interest are typically presented along with a confidence interval or a *P*-value. On occasion, the data contributing to the estimate (for example, numbers of events and participants, or means and standard deviations) cannot be extracted. In such situations it may still be possible to include the data in a meta-analysis using the generic inverse variance method, which requires only an estimate and a standard error from each study (See 8.6.2 A generic inverse variance approach to meta-analysis). This section describes how to obtain a standard error from a confidence interval or a *P*-value. If extracting data concerning a mean from one treatment arm, or the difference between two means, then section 8.5.2 Data extraction for continuous data should be followed instead.

The procedure for obtaining a standard error depends on whether the effect measure is a ratio measure (e.g. odds ratio, risk ratio, hazard ratio, rate ratio) or an absolute measure (e.g. mean difference, standardized mean difference, risk difference).

8.5.6.1 Standard error for absolute (difference) measures

If a 95% confidence interval is available for an absolute measure of treatment effect, then the standard error can be calculated as

$$SE = (\text{upper limit} - \text{lower limit})/3.92.$$

For 90% confidence intervals divide by 3.29 rather than 3.92; for 99% confidence intervals divide by 5.15.

Where exact *P*-values are quoted alongside estimates of treatment effect, it is possible to estimate standard errors. While all tests of statistical significance produce *P*-values, different tests use different mathematical approaches to obtain a *P*-value. The method here assumes *P*-values have been obtained through a particular simple approach known as a Wald test. Where significance tests have used other mathematical approaches the estimated standard errors may not coincide exactly with the true standard errors.

The first step is to obtain the *Z*-value corresponding to the reported *P*-value from a table of the standard normal distribution. A standard error may then be calculated as

$$\text{SE} = \text{treatment effect estimate} / Z .$$

As an example, suppose a conference abstract presents an estimate of a risk difference of 0.03 ($P = 0.008$). The *Z*-statistic that corresponds with a *P*-value of 0.008 is $Z = 2.652$. This can be obtained from a table of the standard normal distribution or a computer (for example, by entering **=abs(normsinv(0.008/2))** into any cell in a Microsoft Excel spreadsheet). The standard error of the risk difference is obtained by dividing the risk difference (0.03) by the *Z*-value (2.652), which gives 0.011.

8.5.6.2 Standard error for ratio measures

The process of obtaining standard errors for ratio measures is similar to that for absolute measures, but with an additional first step. Analyses of ratio measures are performed on the log scale (see 8.2.6 Expressing treatment effects on log scales). For a ratio measure *R*, such as an odds ratio or hazard ratio, first calculate

lower limit = log(lower confidence limit given for *R*)
upper limit = log(upper confidence limit given for *R*)
treatment effect estimate = log(*R*)

Then the formulae in Section 8.5.6.1 Standard error for absolute (difference) measures can be used. Note that the standard error refers to the log of the ratio measure. When using the generic inverse variance method in RevMan, the data should be entered on the log scale, that is as log(*R*) and the standard error of log(*R*), as calculated here (see 8.6.2 A generic inverse variance approach to meta-analysis).

8.6 Summarising effects across studies

An important step in a systematic review is the thoughtful consideration of whether it is appropriate to combine the numerical results of all, or perhaps some, of the studies. Such a 'meta-analysis' yields an overall statistic (together with its confidence interval) that summarises the effectiveness of the experimental intervention compared with a control intervention (see 8.1 Planning the analysis). This section describes the principles and methods used to carry out a meta-analysis for the main types of data encountered.

Formulae for all the methods described and a much longer discussion of the issues discussed in this section appears in Deeks et al (Deeks 2001a) and Deeks and Altman (Deeks 2001b).

8.6.1 Principles of meta-analysis

All commonly used methods for meta-analysis follow the following basic principles.

(1) Meta-analysis is typically a two-stage process. In the first stage, a summary statistic is calculated for each study. For controlled trials, these values describe the treatment effects observed in each individual trial. For example, the summary statistic may be a risk ratio if the data are dichotomous or a difference between means if the data are continuous.

(2) In the second stage, a summary (pooled) treatment effect estimate is calculated as a weighted average of the treatment effects estimated in the individual studies. A weighted average is defined as

$$\text{weighted average} = \frac{\text{sum of}\left(\text{estimate} \times \text{weight}\right)}{\text{sum of weights}} = \frac{\sum T_i W_i}{\sum W_i}$$

where T_i is the treatment effect estimated in study i, W_i is the weight given to study i and the summation is across all studies. Note that if all the weights are the same then the weighted average is equal to the mean treatment effect. The bigger the weight given to study i the more it will contribute to the weighted average. The weights are therefore chosen to reflect the amount of information that each trial contains. For ratio measures (OR, RR, etc.) T_i is the logarithm of the measure.

3. The combination of treatment effect estimates across studies may optionally incorporate an assumption that the studies are not all estimating the same treatment effect, but estimate treatment effects that follow a distribution across studies. This is the basis of a **random effects meta-analysis** (see Section 8.7.4 Incorporating heterogeneity in random effects models). Alternatively, if it is assumed that each study is estimating exactly the same quantity a **fixed effect meta-analysis** is performed.

4. The standard error of the summary (pooled) treatment effect can be used to derive a confidence interval which communicates the precision (or uncertainty) of the summary estimate, and to derive a *P*-value (significance level) which communicates the strength of the evidence against the null hypothesis of no treatment effect.

5. As well as yielding a summary quantification of the pooled effect, all methods of meta-analysis can incorporate an assessment of whether the variation among the results of the separate studies is compatible with random variation, or whether it is large enough to indicate inconsistency of treatment effects across studies (see 8.7 Heterogeneity).

8.6.2 A generic inverse variance approach to meta-analysis

A very common and simple version of the meta-analysis procedure is commonly referred to as the **inverse variance method**. This approach was implemented in its most basic form in RevMan version 4.2, although iit has been used behind the scenes in certain meta-analyses of both dichotomous and continuous data.

The inverse variance method is so named because the weight given to each study is chosen to be the inverse of the variance of the effect estimate (i.e. one over the square of

its standard error). Thus larger studies, which have smaller standard errors, are given more weight than smaller studies, which have larger standard errors. This choice of weight minimises the imprecision (uncertainty) of the pooled effect estimate.

A fixed effect meta-analysis using the inverse variance method calculates a weighted average as

$$\text{generic inverse variance weighted average} = \frac{\sum\left(T_i / S_i^2\right)}{\sum\left(1/S_i^2\right)}$$

where T_i is the treatment effect estimated in study i, S_i is the standard error of that estimate and the summation is across all studies. The basic data required for the analysis are therefore an estimate of the treatment effect and its standard error from each study.

8.6.2.1 Random effects (DerSimonian and Laird) method for meta-analysis

A variation on the inverse variance method is to incorporate an assumption that the different studies are estimating different, yet related, treatment effects. This produces a random effects meta-analysis, and the simplest version is known as the DerSimonian and Laird method (DerSimonian 1986). Random effects meta-analysis is discussed in 8.7.4 Incorporating heterogeneity into random effects models. To undertake a random effects meta-analysis, the standard errors of the study-specific estimates (S_i above) are adjusted to incorporate a measure of the extent of variation, or heterogeneity, among the treatment effects observed in different studies. The size of this adjustment can be estimated from the treatment effects and standard errors of the studies included in the meta-analysis.

8.6.2.2 The generic inverse variance outcome type in RevMan 4.2

Estimates and standard errors may be entered directly into RevMan 4.2 (and subsequent versions) under the 'Generic inverse variance' outcome. The software will undertake fixed effect meta-analyses and random effects (DerSimonian and Laird) meta-analyses, along with assessments of heterogeneity. For ratio measures of treatment effect, the data should be entered as logarithms (for example as a log odds ratio and the standard error of the log odds ratio). However, it is straightforward to instruct the software to display results on the original (e.g. odds ratio) scale. Rather than displaying summary data separately for the treatment groups, the forest plot will display the estimates and standard errors as they were entered beside the study identifiers. It is possible to supplement or replace this with a column providing the sample sizes in the two groups.

Note that the ability to enter estimates and standard errors directly into RevMan creates a high degree of flexibility in meta-analysis. For example, it facilitates the analysis of properly analysed cross-over trials, cluster randomised trials and non-randomized studies, as well as outcome data that are ordinal, time-to-event or rates. However, in most situations for analyses of continuous and dichotomous outcome data it is still preferable to enter more detailed data into RevMan (i.e. specifically as simple summaries of dichotomous or continuous data for each group). This avoids the need for the reviewer to calculate effect estimates, and allows the use of methods targeted specifically at different types of data (see 8.5.3 Meta-analysis of dichotomous outcomes and 8.5.4 Meta-analysis of continuous outcomes). Also, it is helpful for the readers of the review to see the summary statistics for each treatment group in each trial.

8.6.3 Meta-analysis of dichotomous outcomes

There are four widely used methods of meta-analysis for dichotomous outcomes, three fixed effect methods (Mantel-Haenszel, Peto and Inverse Variance) and one random effects method (DerSimonian and Laird). The Mantel-Haenszel, Peto and DerSimonian and Laird methods are available as options in RevMan analyses for dichotomous data, and the inverse variance analysis can be performed by using the generic inverse variance outcome data method (see 8.6.2.2 The generic inverse variance outcome type in RevMan 4.2). The Peto method can only pool odds ratios whilst the other three methods can pool odds ratios, risk ratios and risk differences. Formulae for all of the meta-analysis methods are given by Deeks et al (Deeks 2001a).

Note that zero cells (e.g. no events in one group) cause problems with computation of estimates and standard errors with some methods. The RevMan software automatically adds 0.5 to each cell of the 2×2 table for any such study.

8.6.3.1 Mantel-Haenszel methods

The Mantel-Haenszel methods (Mantel 1959, Greenland 1985) are the default fixed effect methods of meta-analysis programmed in RevMan. When data are sparse, either in terms of event rates being low or trial size being small, the estimates of the standard errors of the effect estimates that are used in the inverse variance methods may be poor. Mantel-Haenszel methods use a different weighting scheme that depends upon which effect measure (e.g. risk ratio, odds ratio, risk difference) is being used. They have been shown to have better statistical properties when there are few events. As this is a common situation in Cochrane Reviews, the Mantel-Haenszel method is generally preferable to the inverse variance method. In other situations the two methods give similar estimates.

8.6.3.2 Peto odds ratio method

Peto's method (Yusuf 1985) can only be used to pool odds ratios. It uses an inverse variance approach but utilises an approximate method of estimating the log odds ratio, and uses different weights. An alternative way of viewing the Peto method is as a sum of 'O – E' statistics. Here, O is the observed number of events and E is an expected number of events in the experimental intervention group of each trial.

The approximation used in the computation of the log odds ratio works well when treatment effects are small (odds ratios are close to one), events are not particularly common and the trials have similar numbers in experimental and control groups. In other situations it has been shown to give biased answers. As these criteria are not always fulfilled, Peto's method is not recommended as a default approach for meta-analysis.

Corrections for zero cell counts are not necessary when using Peto's method. Perhaps for this reason, this method performs well when events are very rare (Deeks 1998a) (see 8.X Rare events (including zero frequencies)). Also, Peto's method can be used to combine dichotomous outcome data with data from time-to-event analyses where log-rank tests have been used (see 8.6.8 Meta-analysis of time-to-event outcomes).

8.6.3.3 DerSimonian and Laird random effects method

The DerSimonian and Laird random effects method (DerSimonian 1986) incorporates an assumption that the different studies are estimating different, yet related, treatment effects. As described in 8.6.2.1 Random effects (DerSimonian and Laird) method for meta-analysis the method is based on the inverse variance approach, making an adjustment to the study weights according to the extent of variation, or heterogeneity, among the

varying treatment effects. The DerSimonian and Laird method and the inverse variance method will give identical results when there is no heterogeneity among the studies (and thus also gives results similar to the Mantel-Haenszel method in many situations). Where there is heterogeneity, confidence intervals for the average treatment effect will be wider if the DerSimonian and Laird method is used rather than a fixed effect method, and corresponding claims of statistical significance will be more conservative. It is also possible that the central estimate of the treatment effect will change if there are relationships between observed treatment effects and sample sizes. See 8.7.4 Incorporating heterogeneity into random effects models for further discussion of these issues.

8.6.3.4 Which measure for dichotomous outcomes?

Summary statistics for dichotomous data are described in 8.2.1 Effect measures for dichotomous outcomes. The effect of treatment can be expressed as either a relative or an absolute effect. The risk ratio (relative risk) and odds ratio are relative measures, while the risk difference and number needed to treat are absolute measures. A further complication is that there are in fact two risk ratios. We can calculate the risk ratio of an event occurring or the risk ratio of no event occurring. These give different pooled results in a meta-analysis, sometimes dramatically so.

The selection of a summary statistic for use in meta-analysis depends on balancing three criteria (Deeks 2002). First, we desire a summary statistic that gives values that are similar for all the trials in the meta-analysis and subdivisions of the population to which the treatment will be applied. The more consistent the summary statistic the greater is the justification for expressing the effect of treatment as a single summary number. Second, the summary statistic must have the mathematical properties required for performing a valid meta-analysis. Third, the summary statistic should be easily understood and applied by those using the review. It should present a summary of the effect of the intervention in a way that helps readers to interpret and apply the results appropriately. Among effect measures for dichotomous data, no single measure is uniformly best, so the choice inevitably involves a compromise.

Consistency: Empirical evidence suggests that relative effect measures are, on average, more consistent than absolute measures. For this reason it is wise to avoid performing meta-analyses of risk differences, unless there is a clear reason to suspect that risk differences will be consistent in a particular clinical situation. On average there is little difference between the odds ratio and risk ratio in this regard (Deeks 2002). When the trial aims to reduce the incidence of an adverse outcome (see 8.2.1.5 What is the event?) there is empirical evidence that risk ratios of the adverse outcome are more consistent than risk ratios of the non-event (Deeks 2002). Selecting an effect measure on the basis of what is the most consistent in a *particular* situation is not a generally recommended strategy, since it may lead to a selection that spuriously minimises the precision of a meta-analysis estimate.

Mathematical properties: The most important mathematical criterion is the availability of a reliable variance estimate. The number needed to treat does not have a simple variance estimator and cannot easily be used directly in meta-analysis, although it can be computed from the other summary statistics (see 8.X Re-expressing meta-analysis results as NNTs). There is no consensus as to the importance of two other often cited mathematical properties: the fact that the behaviour of the odds ratio and the risk difference do not rely

on which of the two outcome states is coded as the event, and the odds ratio being the only statistic which is unbounded (see 8.2.1 Effect measures for dichotomous outcomes).

Ease of interpretation: The odds ratio is the hardest summary statistic to understand and to apply in practice, and many practising clinicians report difficulties in using them. There are many published examples where authors have misinterpreted odds ratios from meta-analyses as if they were risk ratios. There must be some concern that routine presentation of the results of systematic reviews as odds ratios will lead to frequent overestimation of the benefits and harms of treatments when the results are applied in clinical practice. Absolute measures of effect are also thought to be more easily interpreted by clinicians than relative effects (Sinclair 1994), although they are less likely to be generalisable.

It seems important to avoid using summary statistics for which there is empirical evidence that they are unlikely to give consistent estimates of treatment effects (the risk difference) and it is impossible to use statistics for which meta-analysis cannot be performed (the number needed to treat). Thus it is generally recommended that analysis proceeds using risk ratios (taking care to make a sensible choice over which category of outcome is classified as the event) or odds ratios. It may be wise to plan to undertake a sensitivity analysis to investigate whether choice of summary statistic (and selection of the event category) is critical to the conclusions of the meta-analysis (see 8.X Sensitivity Analyses).

It is often sensible to use one statistic for meta-analysis and re-express the results using a second, more easily interpretable statistic. For example, meta-analysis may often be best performed using relative effect measures (risk ratios or odds ratio) and the results re-expressed using absolute effect measures (risk differences or numbers needed to treat – see 8.X Re-expressing meta-analysis results as NNTs). If odds ratios are used for meta-analysis they can also be re-expressed as risk ratios (see 8.2.1 Effect measures for dichotomous outcomes). In all cases the same formulae can be used to convert upper and lower confidence limits. However, it is important to note that all of these transformations require specification of a value of baseline risk indicating the likely risk of the outcome in the population to which the results will be applied. Where the chosen value for baseline risk is close to the average of the control group event rates across the trials the same estimates of NNT will be obtained regardless of whether odds ratios or risk ratios are used for meta-analysis. Where the chosen baseline risk differs from the average control group event rate, the predictions of absolute benefit will differ according to which summary statistic was used for meta-analysis.

8.6.4 Meta-analysis of continuous outcomes

Two methods of analysis are available in RevMan for meta-analysis of continuous data, one fixed effect method and one random effects method. The default fixed effect method uses the inverse variance approach whilst the random effects method uses the DerSimonian and Laird random effects approach. The methods will give exactly the same answers when there is no heterogeneity. Where there is heterogeneity, confidence intervals for the average treatment effect will be wider if the DerSimonian and Laird method is used rather than a fixed effect method, and corresponding P-values will be less significant. It is also possible that the central estimate of the treatment effect will change if there are relationships between observed treatment effects and sample sizes. See 8.7.4 Incorporating heterogeneity into random effects models for further discussion of these issues.

Reviewers should be aware that an assumption underlying methods for meta-analysis of continuous data is that the outcomes have a normal distribution in each treatment arm in each study. This assumption may not always be met, although it is unimportant in very large studies. It is useful to consider the possibility of skewed data (see 8. 5.2.11 Skewed data).

8.6.4.1 Which measure for continuous outcomes?

There are two summary statistics used for meta-analysis of continuous data, the mean difference (MD) and the standardised mean difference (SMD) (see 8.2.2 Effect measures for continuous outcomes). Selection of summary statistics for continuous data is principally determined by whether trials all report the outcome using the same scale (when the mean difference can be used) or using different scales (when the standardised mean difference has to be used).

It is important to note the different roles played in the two approaches by the standard deviations of outcomes observed in the two groups.

For the mean difference method the standard deviations are used together with the sample sizes to compute the weight given to each study. Studies with small standard deviations are given relatively higher weight whilst studies with larger standard deviations are given relatively smaller weights. This is appropriate if variation in standard deviations between studies reflects differences in the reliability of outcome measurements, but is probably not appropriate if the differences in standard deviation reflect real differences in the variability of outcomes in the study populations.

For the standardised mean difference approach the standard deviation is used to standardise the mean differences to a single scale (see 8.2.2.2 The standardised mean difference), as well as in the computation of study weights. It is assumed that variation between standard deviations reflects only differences in measurement scales and not differences in the reliability of outcome measures or variability among trial populations.

These limitations of the methods should be borne in mind where unexpected variation of standard deviations across studies is observed.

8.6.4.2 Meta-analysis of change scores

In some circumstances an analysis based on changes from baseline will be more efficient and powerful than comparison of final values as it removes a component of between person variability from the analysis. However, calculation of a change score requires measurement of the outcome twice and in practice may be less efficient for outcomes which are unstable or difficult to measure precisely, where the measurement error may be larger than true between person baseline variability. Change from baseline outcomes may also be preferred if they have a less skewed distribution than final measurement outcomes. Although sometimes used as a device to 'correct' for unlucky randomization, this practice is not recommended.

In practice a reviewer is likely to discover that the trials included in a review may include a mixture of change from baseline and final value scores. However, mixing of outcomes is not a problem when it comes to meta-analysis. There is no statistical reason why trials with change from baseline outcomes should not be combined in a meta-analysis with trials with final measurement outcomes when using the weighted mean difference method

in RevMan. In a randomized trial, mean differences based on changes from baseline can usually be assumed to be addressing exactly the same underlying treatment effects as analyses based on final measurements. That is to say, the difference in mean final values will on average be the same as the difference in mean change scores. If the use of change scores does increase precision, the studies presenting change scores will appropriately be given higher weights in the analysis than they would have received if final values had been used, as they will have smaller standard deviations.

When combining the data reviewers must be careful to use the appropriate means and standard deviations (either of final measurements or of changes from baseline) for each trial. Since the mean values and standard deviations for the two types of outcome may differ substantially it may be advisable to place them in separate subgroups to avoid confusion for the reader, but the results of the subgroups can legitimately be pooled together.

However, final value and change scores should not be combined together as standardised mean differences, since the difference in standard deviation reflects not differences in measurement scale, but differences in the reliability of the measurements.

8.6.5 Combining dichotomous and continuous outcomes

Occasionally reviewers encounter a situation where data for the same outcome are presented in some studies as dichotomous data and in other studies as continuous data. For example, scores on depression scales can be reported as means or as the percentage of patients who were depressed at some point after an intervention (i.e. with a score above a specified cut-point). This type of information is often easier to understand and more helpful when it is dichotomised. However, deciding on a cut-point may be arbitrary and information is lost when continuous data are transformed to dichotomous data.

There are several options for handling combinations of dichotomous and continuous data. Generally, it is useful to summarise results from all the relevant, valid studies in a similar way, but this is not always possible. It may be possible to collect missing data from investigators so that this can be done. If not, it may be useful to summarise the data in three ways: by placing the continuous data in a Continuous Data Table, dichotomous data in a Dichotomous Data Table and all of the data in an Other Data Table.

There are statistical approaches available which will re-express odds ratios as standardised mean differences (and vice versa) which allow dichotomous and continuous data to be pooled together, subject to making particular distributional assumptions. Based on an assumption that the underlying distribution of the continuous measurement in each treatment group follows a logistic distribution (which is a symmetrical distribution similar in shape to the normal distribution but with more data in the distributional tails), and that the variability of the outcomes is the same in both treated and control participants, the odds ratios can be re-expressed as a standardised mean difference according to the following simple formula (Chinn 2000):

$$\mathrm{SMD} = \frac{\sqrt{3}}{\pi} \log \mathrm{OR} .$$

The standard error of the log odds ratio can be converted to the standard error of a standardised mean difference by multiplying by the same constant (0.5513). Alternatively standardised mean differences can be re-expressed as log odds ratios by multiplying by $\pi/\sqrt{3} = 1.8140$.

Once standardised mean differences and standard errors have been computed for all trials in the meta-analysis they can be combined using the generic inverse variance method in RevMan (version 4.2 or later). Standard errors will first need to be computed for all trials by entering the data in RevMan as dichotomous and continuous outcome type data as appropriate, and converting the confidence intervals for the resulting log odds ratios and standardised mean differences into standard errors (see 8.5.6 Obtaining standard errors from confidence intervals and P-values).

8.6.6 Meta-analysis of ordinal and measurement scale outcomes

Ordinal and measurement scale outcomes are most commonly meta-analysed as dichotomous data (if so see Section 8.6.3) or continuous data (if so see Section 8.6.4) depending on the way that the trialists performed the original analyses.

Occasionally it is possible to analyse the data using proportional odds models where ordinal scales have a small number of categories, the numbers falling into each category for each treatment group can be obtained, and the same ordinal scale has been used in all trials. This approach may make more efficient use of all available data than dichotomisation, but requires access to advanced statistical software and results in a summary statistic for which it is challenging to find a clinical meaning.

The proportional odds model uses the proportional odds ratio as the measure of treatment difference (Agresti 1996). Suppose that there are 3 categories, which are ordered in terms of desirability such that 1 is the best and 3 the worst. The data could be dichotomised in 2 ways. That is, category 1 constitutes a success and categories 2-3 a failure, or categories 1-2 constitute a success and category 3 a failure. A proportional odds model would assume that there is an equal odds ratio for both dichotomies of the data. Therefore, the odds ratio calculated from the proportional odds model can be interpreted as the odds of success on the experimental intervention relative to control, irrespective of how the ordered categories might be divided into success or failure. Methods (specifically polychotomous logistic regression models) are available for calculating trial estimates of the log odds ratio and its standard error and for conducting a meta-analysis in advanced statistical software packages (Whitehead 1994).

Estimates of log odds ratios and their standard errors from a proportional odds model may be meta-analysed using the generic inverse variance method in RevMan version 4.2 or later (see 8.5.2.2 The generic inverse variance outcome type in RevMan 4.2). Both fixed effect and random effects methods of analysis are available. If the same ordinal scale has been used in all studies, but has in some reports been presented as a dichotomous outcome, it may still be possible to include all studies in the meta-analysis. In the context of the 3 category model, this might mean that for some studies category 1 constitutes a success, while for others both categories 1 and 2 constitute a success. Methods for dealing with this, and for combining data from scales which are related but have different definitions for their categories are available (Whitehead 1994).

8.6.7 Meta-analysis of counts and rates

Results may be expressed as **count data** when each participant may experience an event, and may experience it more than once. For example, 'number of strokes', or 'number of hospital visits' are counts. These events may not happen at all, but if they do happen there is no theoretical maximum number of occurrences for an individual.

As described in 8.5.4 Data extraction for counts and rates, count data may be analysed using methods for dichotomous (if so see Section 8.6.3), continuous (if so see Section 8.6.4) and time-to-event data (if so see Section 8.6.8) as well as being analysed as rate data.

Rate data occur if counts are measured for each participant along with the time over which they are observed. This is particularly appropriate when the events being counted are rare. For example, a woman may experience two strokes during a follow-up period of two years. Her **rate** of strokes is one per year of follow up (or, equivalently 0.083 per month of follow-up). Rates are conventionally summarised at the group level. For example, participants in the control group of a trial may experience 85 strokes during a total of 2836 person-years of follow-up. An underlying assumption associated with the use of rates is that the risk of an event is constant across participants and over time. This assumption should be carefully considered for each situation. For example, in contraception studies, rates have been used (known as Pearl indices) to describe the number of pregnancies per 100 women-years of follow-up. This is now considered inappropriate since couples have different risks of conception, and the risk for each woman changes over time. Pregnancies are now analysed more often using life tables or time to event methods that investigate the time elapsing before the first pregnancy.

Analysing count data as rates is not always the most appropriate approach and is uncommon in practice. This is because:

(1) the assumption of a constant underlying risk may not be suitable; and
(2) statistical methods are not as well developed as they are for other types of data.

The results of a trial may be expressed as a **rate ratio**, that is the ratio of the rate in the intervention group to the rate in the control group. Suppose A events occurred during X participant-years of follow-up in the intervention group, and C events during Y participant-years in the control group. The rate ratio is $(A/X)/(C/Y) = AY/CX$.

The (natural) logarithms of the rate ratios may be combined across trials using the generic inverse variance method (see 8.6.2.2 The generic inverse variance outcome type in RevMan 4.2). An approximate standard error of the log rate ratio is given by $\sqrt{(1/A + 1/C)}$. A correction of 0.5 may be added to each count in the case of zero events. Note that the choice of time unit (i.e. patient-months, women-years, etc) is irrelevant since it is cancelled out of the rate ratio and does not figure in the standard error. However the units should still be displayed when presenting the study results. An alternative means of estimating the rate ratio is through the approach of Whitehead and Whitehead (Whitehead 1991).

In a randomized trial rate ratios may often be very similar to relative risks obtained after dichotomising the participants, since the average period of follow-up should be similar in all intervention groups. Rate ratios and relative risks will differ, however, if an intervention affects the likelihood of some participants experiencing multiple events.

It is possible also to focus attention on the rate difference, $(A/X) - (C/Y)$. An approximate standard error for the rate difference is $\sqrt{(A/X^2 + C/Y^2)}$. The analysis again requires use of the generic inverse variance method in RevMan. One of the only discussions of meta-analysis of rates, which is still rather short, is that by Hasselblad and McCrory (Hasselblad 1995).

8.6.8 Meta-analysis of time-to-event outcomes

Two approaches to meta-analysis of time-to-event outcomes are available in RevMan. Which is used will depend on what data have been extracted from the primary studies, or obtained from reanalysis of individual patient data.

If logrank 'O – E' and 'V' statistics have been obtained, either through re-analysis of individual patient data or from aggregate statistics presented in the study reports, trial results can be combined using a modified version of the Peto method for dichotomous data (available as the only analysis option for the Individual Patient Data outcome type in RevMan). In the output 'Odds Ratio' will actually mean 'Hazard Ratio'. This is a fixed effect analysis – no equivalent random effects analysis is available in RevMan.

Alternatively if estimates of log hazard ratios and standard errors have been obtained from results of Cox proportional hazards regression models trial results can be combined using the generic inverse variance method (available in RevMan 4.2 and later), see 8.6.2.2 The generic inverse variance outcome type in RevMan 4.2. Both fixed effect and random (DerSimonian and Laird) effects analyses are available.

If a mixture of logrank and Cox model estimates are obtained from the trials, all results can be combined using the generic inverse variance method as the logrank estimates can be converted into log hazard ratios and standard errors using the formulae given in 8.5.5 Data extraction for time-to-event data.

8.6.9 A summary of meta-analysis methods available in RevMan

RevMan includes the following options for statistical analysis:

TYPE OF DATA	SUMMARY STATISTIC	METHOD (F:fixed, R:random)
Dichotomous	odds ratio	Mantel-Haenszel (F) Peto (F) DerSimonian and Laird (R)
	risk ratio	Mantel-Haenszel (F) DerSimonian and Laird (R)
	risk difference	Mantel-Haenszel (F) DerSimonian and Laird (R)
Continuous	(weighted) mean difference	inverse variance (F) DerSimonian and Laird (R)
	standardised mean difference	inverse variance (F) DerSimonian and Laird (R)

Time to event (IPD)	odds/hazard ratio	Peto (F)
Generic inverse variance*	defined by reviewer	inverse variance (F) DerSimonian and Laird (R)

*only available since RevMan 4.2

RevMan requires the reviewer to select one preferred method for each outcome. If these are not specified then the software defaults to the fixed effect Mantel-Haenszel odds ratio for dichotomous outcomes, the fixed effect weighted mean difference for continuous outcomes and the fixed effect model for generic inverse variance outcomes. It is important that reviewers make it clear which method they are using when results are presented in the text of a review, since it cannot be guaranteed that a meta-analysis displayed to the user will coincide with the selected preferred method.

8.6.10 Use of vote counting for meta-analysis

Occasionally meta-analyses use "vote-counting" to compare the number of positive studies with the number of negative studies. Vote-counting is limited to answering the simple question "is there any evidence of an effect?" Two problems can occur with vote-counting, which suggest that it should be avoided whenever possible. Firstly, problems occur if subjective decisions or statistical significance are used to define "positive" and "negative" studies (Cooper 1980, Antman 1992). To undertake vote counting properly the number of studies showing harm should be compared with the number showing benefit, regardless of the statistical significance or size of their results. A sign test can be used to assess the significance of evidence for the existence of an effect in either direction (if there is no effect the studies will be distributed evenly around the null hypothesis of no difference). Secondly, vote-counting takes no account of the differential weights given to each study. Vote-counting might be considered as a last resort in situations when standard meta-analytical methods cannot be applied (such as when there is no consistent outcome measure).

8.7 Heterogeneity

8.7.1 What is heterogeneity?

Inevitably, studies brought together in a systematic review will differ. Any kind of variability among studies in a systematic review may be termed heterogeneity. It can be helpful to distinguish between different types of heterogeneity. Variability in the participants, interventions and outcomes studied may be described as **clinical diversity** (sometimes called clinical heterogeneity), and variability in trial design and quality may be described as **methodological diversity** (sometimes called methodological heterogeneity). Variability in the treatment effects being evaluated in the different trials is known as **statistical heterogeneity**, and is a consequence of clinical and/or methodological diversity among the studies. Statistical heterogeneity manifests itself in the observed treatment effects being more different from each other than one would expect due to random error (chance) alone. We will follow convention and refer to **statistical heterogeneity** simply as **heterogeneity**.

Clinical variation will lead to heterogeneity if the treatment effect is affected by the factors that vary across studies – most obviously, the specific interventions or patient characteristics. In other words, the true treatment effect will be different in different studies.

Differences between trials in terms of methodological factors, such as use of blinding and concealment of allocation, or if there are differences between trials in the way the outcomes are defined and measured, may be expected to lead to differences in the observed treatment effects. Significant statistical heterogeneity arising from methodological diversity or differences in outcome assessments suggests that the studies are not all estimating the same quantity, but does not necessarily suggest that the true treatment effect varies. In particular, heterogeneity associated solely with methodological diversity would indicate the studies suffer from different degrees of bias. Empirical evidence suggests that some aspects of design can affect the result of clinical trials, although this is not always the case. Further discussion appears in Section 6.

The scope of a review will largely determine the extent to which studies included in a review are diverse. Sometimes a review will include trials addressing a variety of questions, for example when several different interventions for the same condition are of interest. Trials of each intervention should be analysed and presented separately (see also 4.5 broad versus narrow questions). Meta-analysis should only be considered when a group of trials is sufficiently homogeneous in terms of participants, interventions and outcomes to provide a meaningful summary. It is often appropriate to take a broader perspective in a meta-analysis than in a single clinical trial. A common analogy is that systematic reviews bring together apples and oranges, and that combining these can yield a meaningless result. This is true if apples and oranges are of intrinsic interest on their own, but may not be if they are used to contribute to a wider question about fruit. For example, a meta-analysis may reasonably evaluate the average effect of a class of drugs by combining results from trials where each evaluates the effect of a different drug from the class.

There may be specific interest in a review in investigating how clinical and methodological aspects of trials relate to their results. Where possible these investigations should be specified a priori, i.e. in the systematic review protocol. It is legitimate for a systematic review to focus on examining the relationship between some clinical characteristic(s) of the studies and the size of treatment effect, rather than on obtaining a summary effect estimate across a series of trials (see 8.8 Investigating heterogeneity). Meta-regression may best be used for this purpose, although it is not implemented in RevMan (see 8.8.3 Meta-regression).

8.7.2 Identifying and measuring heterogeneity

It is important to consider to what extent the results of studies are consistent. If confidence intervals for the results of individual studies (generally depicted graphically using horizontal lines) have poor overlap, this generally indicates the presence of statistical heterogeneity. More formally, a statistical test for heterogeneity is available. This chi-squared test is included in the graphical output of Cochrane Reviews. It assesses whether observed differences in results are compatible with chance alone. A low p-value (or a large chi-squared statistic relative to its degree of freedom) provides evidence of heterogeneity of treatment effects (variation in effect estimates beyond chance).

Care must be taken in the interpretation of the chi-squared test, since it has low power in the (common) situation of a meta-analysis when trials have small sample size or are few in number. This means that while a statistically significant result may indicate a problem with heterogeneity, a non-significant result must not be taken as evidence of no heterogeneity. This is also why a *P*-value of 0.10, rather than the conventional level of 0.05, is sometimes used to determine statistical significance. A further problem with the test, which seldom occurs in Cochrane Reviews, is that when there are many studies in a meta-analysis, the test has high power to detect a small amount of heterogeneity that may be clinically unimportant.

Some argue that, since clinical and methodological diversity always occur in a meta-analysis, statistical heterogeneity is inevitable. Thus the test for heterogeneity is irrelevant to the choice of analysis; heterogeneity will always exist whether or not we happen to be able to detect it using a statistical test. Methods have been developed for quantifying inconsistency across studies that move the focus away from testing whether heterogeneity is present to assessing its impact on the meta-analysis. A useful statistic for quantifying inconsistency is $I^2 = [(Q - df)/Q] \times 100\%$, where Q is the chi-squared statistic and df is its degrees of freedom (Higgins 2003, Higgins 2002). This describes the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance). A value greater than 50% may be considered substantial heterogeneity.

8.7.3 Strategies for addressing heterogeneity

A number of options are available if (statistical) heterogeneity is identified among a group of trials that would otherwise be considered suitable for a meta-analysis.

1. Check again that the data are correct
Severe heterogeneity can indicate that data have been incorrectly extracted or entered into RevMan. For example, if standard errors have mistakenly been entered as standard deviations for continuous outcomes, this could manifest itself in overly narrow confidence intervals with poor overlap and hence substantial heterogeneity. Unit of analysis errors may also be causes of heterogeneity (see 8. 3 Study designs and identifying the unit of analysis).

2. Do not do a meta-analysis
A systematic review need not contain any meta-analyses (O'Rourke 1989). If there is considerable variation in results, and particularly if there is inconsistency in the direction of effect, it may be misleading to quote an average value for the treatment effect.

3. Explore heterogeneity
It is clearly of interest to determine the causes of heterogeneity among results of studies. This process is problematic since there are often many characteristics that vary across studies from which one may choose. Heterogeneity may be explored by conducting subgroup analyses (see 8.8.2 Undertaking subgroup analyses) or meta-regression (8.8.3 Meta-regression), though this latter method is not implemented in RevMan. Ideally, investigations of characteristics of trials that may be associated with heterogeneity should be pre-specified in the protocol of a review (see 8.1.5 Writing the analysis section of the protocol). Reliable conclusions can only be drawn from analyses that are truly pre-specified before inspecting the trials' results, and even these conclusions should be

interpreted with caution. In practice, reviewers will often be familiar with some trial results when writing the protocol, so true pre-specification is not possible. Explorations of heterogeneity that are devised after heterogeneity is identified can at best lead to the generation of hypotheses. They should be interpreted with even more caution and should generally not be listed among the conclusions of a review. Also, investigations of heterogeneity when there are very few studies are of questionable value.

4. Ignore heterogeneity
Fixed effect meta-analyses ignore heterogeneity. The pooled effect estimate from a fixed effect meta-analysis is normally interpreted as being the best estimate of the treatment effect. However, the existence of heterogeneity suggests that there may not be a single treatment effect but a distribution of treatment effects. Thus the pooled fixed effect estimate may be a treatment effect that does not actually exist in any population, and therefore have a confidence interval that is meaningless as well as being too narrow, (see 8.7.4 Incorporating heterogeneity into random effects models). The *P*-value obtained from a fixed effect meta-analysis does however provide a meaningful test of the null hypothesis that there is no effect in every study.

5. Perform a random effects meta-analysis
A random effects meta-analysis may be used to incorporate heterogeneity among trials. This is not a substitute for a thorough investigation of heterogeneity. It is intended primarily for heterogeneity that cannot be explained. An extended discussion of this option appears below (8.7.4 Incorporating heterogeneity into random effects models).

6. Change the effect measure
Heterogeneity may be an artificial consequence of an inappropriate choice of effect measure. For example, when trials collect continuous outcome data using different scales or different units, extreme heterogeneity may be apparent when using the mean difference but not when the more appropriate standardised mean difference is used. Furthermore, choice of effect measure for dichotomous outcomes (odds ratio, relative risk, or risk difference) may affect the degree of heterogeneity among results. In particular, when control group event rates vary, homogeneous odds ratios or risk ratios will necessarily lead to heterogeneous risk differences, and vice versa. However, it remains unclear whether homogeneity of treatment effect in a particular meta-analysis is a suitable criterion for choosing between these measures (see also 8.6.3.4 Which measure for dichotomous outcomes?).

7. Exclude studies
Heterogeneity may be due to the presence of one or two outlying trials with results that conflict with the rest of the trials. In general it is unwise to exclude studies from a meta-analysis on the basis of their results as this may introduce bias. However, if an obvious reason for the outlying result is apparent, the study might be removed with more confidence. Since usually at least one characteristic can be found for any trial in any meta-analysis which makes it different from the others, this criterion is unreliable because it is all too easy to fulfil. It is advisable to perform analyses both with and without outlying trials as part of a sensitivity analysis (see 8.10 Sensitivity analysis). Whenever possible, potential sources of clinical diversity that might lead to such situations should be specified in the protocol.

8.7.4 Incorporating heterogeneity into random effects models

A fixed effect meta-analysis provides a result that may be viewed as a 'typical treatment effect' from the studies included in the analysis. In order to calculate a confidence interval for a fixed effect meta-analysis the assumption is made that the true effect of treatment (in both magnitude and direction) is the same value in every study (that is, fixed across studies). This assumption implies that the observed differences among study results are due solely to the play of chance: i.e. that there is no statistical heterogeneity.

When there is heterogeneity that cannot readily be explained, one analytical approach is to incorporate it into a random effects model. A random effects meta-analysis model involves an assumption that the effects being estimated in the different studies are not identical, but follow some distribution. The model represents our lack of knowledge about why real, or apparent, treatment effects differ by considering the differences as if they were random. The centre of this symmetric distribution describes the average of the effects, while its width describes the degree of heterogeneity. The conventional choice of distribution is a normal distribution. It is difficult to establish the validity of any distributional assumption, and this is a common criticism of random effects meta-analyses. The importance of the particular assumed shape for this distribution is not known.

Note that a random effects model does not 'take account' of the heterogeneity, in the sense that it is no longer an issue. It is always advisable to explore possible causes of heterogeneity, although there may be too few studies to do this adequately (see 8.8 Investigating heterogeneity).

For random effects analyses in RevMan, the pooled estimate and confidence interval refer to the centre of the distribution of treatment effects, and do not describe the width of the distribution. Often the pooled estimate and its confidence interval are quoted in isolation as an alternative estimate of the quantity evaluated in a fixed effect meta-analysis, which is inappropriate. Note that the confidence interval from a random effects meta-analysis describes uncertainty in the location of the mean of systematically different effects in the different studies. It does not describe the degree of heterogeneity among studies as may be commonly believed. For example, when there are many studies in a meta-analysis, one may obtain a tight confidence interval around the random effects estimate of the mean effect even when there is a large amount of heterogeneity. The range of the treatment effects observed in the trials may be thought to give a rough idea of the spread of the distribution of true treatment effects, but in fact it will be slightly too wide as it also describes the random error in the observed effect estimates.

If variation in effects (statistical heterogeneity) is believed to be due to clinical diversity, the centre of the distribution should be interpreted differently from the fixed effect estimate since it relates to a different question. The random effects estimate and its confidence interval address the question 'what is the average treatment effect?' while the fixed effect estimate and its confidence interval addresses the question 'what is the best estimate of the treatment effect?' The answers to these questions coincide either when no heterogeneity is present, or when the distribution of the treatment effects is roughly symmetrical. When the answers do not coincide, the random effects estimate may not reflect the actual effect in any particular population being studied.

For any particular set of studies in which heterogeneity is present, a confidence interval around the random effects pooled estimate is wider than a confidence interval around a fixed effect pooled estimate. This will happen if the I^2 statistic is greater than zero, even if the heterogeneity is not detected by the chi-squared test for heterogeneity (Higgins 2003) (see 8.7.2 Identifying and measuring heterogeneity).

In a heterogeneous set of studies, a random effects meta-analysis will award relatively more weight to smaller studies than such studies would receive in a fixed effect meta-analysis. This is because small studies are more informative for learning about the distribution of effects across studies than for learning about an assumed common treatment effect. Care must be taken that random effects analyses are applied only when the idea of a 'random' distribution of treatment effects can be justified. In particular, if results of smaller studies are systematically different from results of larger ones, which can happen as a result of publication bias or low study quality bias, (Poole 1999) (Egger 1997b, Kjaergard 2001), then a random effects meta-analysis will exacerbate the effects of the bias. A fixed effect analysis will be affected less, although strictly it will also be inappropriate. In this situation it may be wise to present neither type of meta-analysis, or to perform a sensitivity analysis in which small studies are excluded.

Similarly, when there is little information, either because there are few trials or if the trials are small with few events, a random effects analysis will provide poor estimates of the width of the distribution of treatment effects. The Mantel-Haenszel method will provide more robust estimates of the average treatment effect, but at the cost of ignoring the observed heterogeneity.

RevMan implements a version of random effects meta-analysis that is described by DerSimonian and Laird (DerSimonian 1986). The attraction of this method is that the calculations are straightforward, but it has a theoretical disadvantage that the confidence intervals are slightly too narrow to encompass full uncertainty resulting from having estimated the degree of heterogeneity. Alternative methods exist that encompass full uncertainty, but they require advanced statistical software (see 8.X Bayesian meta-analysis, 8.X Hierarchical models). In practice, the difference in the results is likely to be small unless there are few studies.

8.8 Investigating heterogeneity

Does the treatment effect vary with different populations or treatment characteristics (such as dose or duration)? Such variation is known as interaction by statisticians and as effect modification by epidemiologists. Methods to search for such interactions include subgroup analyses and meta-regression. All methods have considerable pitfalls.

8.8.1 What are subgroup analyses?

Subgroup analyses involve splitting all the participant data into subgroups, often so as to make comparisons between them. Subgroup analyses may be done for subsets of participants (such as males and females), or for subsets of studies (such as different geographical locations). Subgroup analyses may be done as a means of investigating heterogeneous results, or to answer specific questions about particular patient groups, types of intervention or types of study.

Subgroup analyses of subsets of participants within trials are uncommon in systematic reviews of the literature because sufficient details to extract data about separate participant types are seldom published in reports. By contrast, such subsets of participants are easily analysed when individual patient data have been collected (see Appendix 11A).

Findings from multiple subgroup analyses may be misleading. Subgroup analyses are observational by nature and are not based on randomized comparisons. False negative and false positive significance tests increase in likelihood rapidly as more subgroup analyses are performed. If their findings are presented as definitive conclusions there is clearly a risk of patients being denied an effective intervention or treated with an ineffective (or even harmful) intervention. Subgroup analyses can also generate misleading recommendations about directions for future research that, if followed, would waste scarce resources.

It is useful to distinguish between the notions of 'qualitative interaction' and 'quantitative interaction' (Yusuf 1991). Qualitative interaction exists if the direction of effect is reversed, that is if an intervention is beneficial in one subgroup but is harmful in another. Qualitative interaction is rare. This may be used as an argument that the most appropriate result of a meta-analysis is the overall effect across all subgroups. Quantitative interaction exists when the size of the effect varies but not the direction, that is if an intervention is beneficial to different degrees in different subgroups.

Reviewers will find useful advice concerning subgroup analyses in Oxman and Guyatt (Oxman 1992) and Yusuf et al (Yusuf 1991). See also 8.8.5 Interpretation of subgroup analyses and meta-regressions.

8.8.2 Undertaking subgroup analyses

Subgroup analyses may be undertaken within RevMan. Meta-analyses within subgroups and meta-analyses that combine several subgroups are both permitted. It is tempting to compare effect estimates in different subgroups by considering the meta-analysis results from each subgroup separately. This should only be done informally by comparing the magnitudes of effect. Noting that either the effect or the test for heterogeneity in one subgroup is statistically significant whilst that in other subgroup is not statistically significant does not indicate that the subgroup factor explains heterogeneity. Since different subgroups are likely to contain different amounts of information and thus have different abilities to detect effects, it is extremely misleading simply to compare the statistical significance of the results.

8.8.2.1 Is the effect different in different subgroups?

Valid investigations of whether an intervention works differently in different subgroups involve comparing the subgroups with each other. No formal method is currently implemented in RevMan. When there are only two subgroups the overlap of the confidence intervals of the summary estimates in the two groups can be considered. Non-overlap of the confidence intervals indicates statistical significance, but note that the confidence intervals can overlap to a small degree and the difference still be statistically significant.

A simple approach for a significance test that can be used to investigate differences between two or more subgroups is described by Deeks et al, although some statistical help

may be required (Deeks 2001a). This method uses information given by RevMan when subgroups and totals are displayed. It is based on the test for heterogeneity chi-squared statistics that appear in the bottom left hand corner of the forest plots, and proceeds as follows. Suppose a chi-squared heterogeneity statistic, Q_{all}, is available for all of the trials, and that chi-squared heterogeneity statistics Q_1 up to Q_m are available for m subgroups (such that every trial is in one and only one subgroup). Then the new statistic $Q_{int} = Q_{all} - (Q_1 + \ldots + Q_m)$, compared with a chi-squared distribution with $m - 1$ degrees of freedom, tests for a difference among the subgroups. (Relevant details of the chi-squared distribution are available as appendices of many statistical textbooks, or using standard computer spreadsheet packages. For example typing **=chidist(5.2,2)** in any cell in a Microsoft Excel spreadsheet will give the P-value for a value of Q_{int} of 5.2 on 2 degrees of freedom). If the values of the heterogeneity chi-squared statistics are obtained from the continuous or generic inverse variance data types in RevMan then there are no problems in using this test. However, if the dichotomous data type is used, then the test will currently include a slight inaccuracy due to the way in which the heterogeneity chi-squared statistic is calculated in RevMan.

A more flexible alternative to testing for differences between subgroups is to use meta-regression techniques, in which residual heterogeneity (that is, heterogeneity not explained by the subgrouping) is allowed (see 8.8.3 Meta-regression).

8.8.3 Meta-regression

If studies are divided into subgroups (see 8.8.2 Subgroup analysis), this may be viewed as an investigation of how a categorical study characteristic is associated with the treatment effects in the meta-analysis. For example, studies in which allocation concealment was adequate may yield different results from those in which allocation concealment was inadequate. Here, allocation concealment, being either adequate or inadequate, is a categorical characteristic at the study level. Meta-regression is an extension to subgroup analyses that allows the effect of continuous, as well as categorical, characteristics to be investigated, and in principle allows the effects of multiple factors to be investigated simultaneously (although this is rarely possible due to inadequate numbers of trials) (Thompson 2002). Meta-regression should generally not be considered when there are fewer than 10 trials in a meta-analysis.

Meta-regressions are similar in essence to simple regressions, in which an **outcome variable** is predicted according to the values of one or more **explanatory variables**. In meta-regression, the outcome variable is the effect estimate (for example, a mean difference, a risk difference, a log odds ratio or a log risk ratio). The explanatory variables are characteristics of studies that might influence the size of treatment effect. These are often called 'potential effect modifiers' or covariates. Meta-regressions usually differ from simple regressions in two ways. First, larger studies have more influence on the relationship than smaller studies, since studies are weighted by the precision of their respective effect estimate. Second, it is wise to allow for the residual heterogeneity among treatment effects not modelled by the explanatory variables. This gives rise to the term 'random effects meta-regression', since the extra variability is incorporated in the same way as in a random effects meta-analysis (Thompson 1999).

The regression coefficient obtained from a meta-regression analysis will describe how the outcome variable (the treatment effect) changes with a unit increase in the explanatory

variable (the potential effect modifier). The statistical significance of the regression coefficient is a test of whether there is a linear relationship between treatment effect and the explanatory variable. If the treatment effect is a ratio measure, the log-transformed value of the treatment effect should always be used in the regression model (see 8.2.6 Expressing treatment effects on log scales), and the exponential of the regression coefficient will give an estimate of the relative change in treatment effect with a unit increase in the explanatory variable.

Meta-regression can also be used to investigate differences for categorical explanatory variables as done in subgroup analyses. If there are *m* subgroups membership of particular subgroups is indicated by using *m*−1 dummy variables (which can only take values of zero or one) in the meta-regression model (as in standard linear regression modelling). The regression coefficients will estimate how the treatment effect in each subgroup differs from a nominated reference subgroup. The *P*-value of each regression coefficient will indicate whether this difference is statistically significant.

Meta-regression is currently best performed using the 'metareg' macro in the Stata statistical package (Sterne 2001).

8.8.4 Selection of study characteristics for subgroup analyses and meta-regression

Reviewers need to be cautious about undertaking subgroup analyses, and interpreting any that they do. Some considerations are outlined here for selecting characteristics (also called explanatory variables, potential effect modifiers or covariates) which will be investigated for their possible influence on the size of the treatment effect. These considerations apply similarly to subgroup analyses and to meta-regressions. Further details may be obtained from Oxman and Guyatt (Oxman 1992) and Berlin and Antman (Berlin 1994).

8.8.4.1 Ensure that there are adequate studies to justify subgroup analyses and meta-regressions

It is very unlikely that an investigation of heterogeneity will produce useful findings unless there is a substantial number of studies. It is worth noting the typical advice for undertaking simple regression analyses: that at least ten observations (i.e. ten studies in a meta-analysis) should be available for each characteristic modelled.

8.8.4.2 Specify characteristics in advance

Reviewers should, whenever possible, pre-specify characteristics in the protocol that later will be subject to subgroup analyses or meta-regression. Pre-specifying characteristics reduces the likelihood of spurious findings, first by limiting the number of subgroups investigated and second by preventing knowledge of the trials' results influencing which subgroups are analysed. True pre-specification is difficult in systematic reviews, because the results of some of the relevant trials are often known when the protocol is drafted. If a characteristic was overlooked in the protocol, but is clearly of major importance and justified by external evidence, then reviewers should not be reluctant to explore it. However, such post hoc analyses should be identified as such.

8.8.4.3 Select a small number of characteristics

The likelihood of a false positive result among subgroup analyses and meta-regression increases with the number of characteristics investigated. It is difficult to suggest a

maximum number of characteristics to look at, especially since the number of available studies is unknown in advance. If more than one or two characteristics are investigated it may be sensible to adjust the level of significance to account for making multiple comparisons. The help of a statistician is recommended (see 8.X Multiple comparisons and the play of chance).

8.8.4.4 Ensure there is scientific rationale for investigating each characteristic

Selection of characteristics should be motivated by biological and clinical hypotheses, ideally supported by evidence from sources other than the included studies. Subgroup analyses using characteristics that are implausible or clinically irrelevant are not likely to be useful and should be avoided. For example, a relationship between treatment effect and year of publication is seldom in itself clinically informative, and if statistically significant runs the risk of initiating a post-hoc data dredge of factors that may have changed over time.

Prognostic factors are those that predict the outcome of a disease or condition, whereas effect modifiers are factors that influence how well a treatment works in affecting the outcome. Confusion between prognostic factors and effect modifiers is common in planning subgroup analyses, especially at the protocol stage. Prognostic factors are not good candidates for subgroup analyses unless they are also believed to modify the effect of treatment. For example, being a smoker may be a strong predictor of mortality within the next ten years, but there may not be reason for it to influence the effect of a drug therapy on mortality (Deeks 1998b). Potential effect modifiers may include the precise interventions (dose of active treatment, choice of comparison treatment), how the study was done (length of follow-up) or methodology (design and quality).

8.8.4.5 Be aware that the effect of a characteristic may not always be identified

Many characteristics that might have important effects on how well an intervention works cannot be investigated using subgroup analysis or meta-regression. These are characteristics of participants that might vary substantially within studies, but which can only be summarised at the level of the study. An example is age. Consider a collection of clinical trials involving adults ranging from 18 to 60 years old. There may be a strong relationship between age and treatment effect that is apparent within each study. However, if the mean ages for the trials are similar, then no relationship will be apparent by looking at trial mean ages and trial-level effect estimates. The problem is one of aggregating individuals' results and is variously known as aggregation bias, ecological bias or the ecological fallacy (Morgenstern 1982, Greenland 1987, Berlin 2002). It is even possible for the differences between trials to display the opposite pattern to that observed within each trial.

8.8.4.6 Think about whether the characteristic is closely related to another characteristic (confounded)

The problem of 'confounding' complicates interpretation of subgroup analyses and meta-regressions and can lead to incorrect conclusions. Two characteristics are confounded if their influences on the treatment effect cannot be disentangled. For example, if those studies implementing an intensive version of a therapy happened to be the studies that involved patients with more severe disease, then one cannot tell which aspect is the cause of any difference in effect estimates between these studies and others. In meta-regression, co-linearity between potential effect modifiers leads to similar difficulties as is discussed by Berlin and Antman (Berlin 1994). Computing correlations between trial characteristics

will give some information about which trial characteristics may be confounded with each other.

8.8.5 Interpretation of subgroup analyses and meta-regressions

Appropriate interpretation of subgroup analyses and meta-regressions requires caution. For more detailed discussion see Oxman and Guyatt (Oxman 1992).

- Subgroup comparisons are observational

It must be remembered that subgroup analyses and meta-regressions are entirely observational in their nature. These analyses investigate differences between trials, and while individuals are randomised to one group or other within a trial, they are not randomised to go in one trial or another. Hence, subgroup analyses suffer the limitations of any observational investigation, including possible bias through confounding by other trial-level characteristics. Furthermore, even a genuine difference between subgroups is not necessarily due to the classification of the subgroups. As an example, a subgroup analysis of bone marrow transplantation for treating leukaemia might show a strong association between the age of a sibling donor and the success of the transplant. However, this probably does not mean that the age of donor is important. In fact, the age of the recipient is probably a key factor and the subgroup finding would simply be due to the strong association between the age of the recipient and the age of their sibling.

- Was the analysis pre-specified or post hoc?

Reviewers should state whether subgroup analyses were pre-specified or undertaken after the results of the studies had been compiled (post hoc). More reliance may be placed on a subgroup analysis if it was one of a small number of pre-specified analyses. Performing numerous post hoc subgroup analyses to explain heterogeneity is data dredging. Data dredging is condemned because it is usually possible to find an apparent, but false, explanation for heterogeneity by considering lots of different characteristics.

- Is there indirect evidence in support of the findings?

Differences between subgroups should be clinically plausible and supported by other external or indirect evidence, if they are to be convincing.

- Is the magnitude of the difference practically important?

If the magnitude of a difference between subgroups will not result in different recommendations for different subgroups, then it may be better to present only the overall analysis results.

- Is there a statistically significant difference between subgroups?

To establish whether there is a different effect of an intervention in different situations, the magnitudes of effects in different subgroups should be compared directly with each other. In particular, statistical significance of the results within separate subgroup analyses (as presented in RevMan) should not be compared. See 8.8.2 Undertaking subgroup analyses.

- Are analyses looking at within-study or between-study relationships?

For patient and intervention characteristics, differences in subgroups that are observed within studies are more reliable than analyses of subsets of studies. If such within-study relationships are replicated across studies then this adds confidence to the findings.

8.8.6 Investigating the effect of baseline risk

One potentially important source of heterogeneity among a series of studies is when the underlying average risk of the outcome event varies between the studies. The baseline risk of a particular event may be viewed as an aggregate measure of case-mix factors such as age or disease severity. It is generally measured as the observed risk of the event in the control group of each trial (the control group risk (CGR) or control event rate (CER)). The notion is controversial in its relevance to clinical practice since baseline risk represents a summary of both known and unknown risk factors. Problems also arise because baseline risk will depend on the length of follow-up, which often varies across studies. However, baseline risk has received particular attention in meta-analysis because the information is readily available once dichotomous data have been prepared for use in meta-analyses. A full discussion of the subject appears in Sharp (Sharp 2000).

Intuition would suggest that participants are more or less likely to benefit from an effective treatment according to their risk status. However, the relationship between baseline risk and treatment effect is a complicated issue. For example, suppose a treatment is equally beneficial in the sense that for all patients it reduces the risk of an event, say a stroke, to 80% of the baseline risk. Then it is not equally beneficial in terms of absolute differences in risk in the sense that it reduces a 50% stroke rate by 10 percentage points to 40% (number needed to treat = 10), but a 20% stroke rate by 4 percentage points to 16% (number needed to treat = 25).

Use of different summary statistics (risk ratio, odds ratio and risk difference) will demonstrate different relationships with baseline risk. Summary statistics that show close to no relationship with baseline risk are generally preferred for use in meta-analysis (see 8.6.3.4 Which measure for dichotomous outcomes?).

Investigating any relationship between effect estimates and the control group risk is also complicated by a technical phenomenon known as regression to the mean. This arises because the control group risk forms an integral part of the effect estimate. A high risk in a control group, observed entirely by chance, will on average give rise to a higher than expected effect estimate, and vice versa. This phenomenon results in a false correlation between effect estimates and control group risks. Methods are available, requiring sophisticated software, that correct for regression to the mean (McIntosh 1996, Thompson 1997). These should be used for such analyses and statistical expertise is recommended.

8.8.7 Dose-response analyses

The principles of meta-regression can be applied to the relationships between treatment effect and dose (commonly termed dose-response), treatment intensity or treatment duration (Greenland 1992, Berlin 1993). Conclusions about differences in effect due to differences in dose (or similar factors) are on strongest ground if participants are randomized to one dose or another within a study and a consistent relationship is found across similar studies. While reviewers should consider these effects, particularly as a possible explanation for heterogeneity, they should be cautious about drawing conclusions based on between-study differences. Reviewers should be particularly cautious about claiming that a dose-response relationship does not exist, given the low power of many meta-regression analyses to detect genuine relationships.

8.8.8 Indirect comparisons

Indirect comparisons are made between interventions in the absence of head-to-head randomized studies. Consider the situation in which some trials have compared the effectiveness of doctors versus dieticians in providing dietary advice, and others the effectiveness of nurses versus dieticians, but no trials have compared the effectiveness of doctors versus nurses. We might then wish to learn about the relative effectiveness of doctors versus nurses.

The problem can be considered as an investigation of a source of heterogeneity (different intervention) in a subgroup analysis. The trials should be considered in two separate subgroups, one of doctors versus dieticians and one of nurses versus dieticians. The difference between the summary effects in the two subgroups will be an estimate of the difference between doctors and nurses. The significance of this difference is best assessed by using meta-regression, although for this particular example the approach is equivalent to a simpler procedure described by Bucher (Bucher 1997). The validity of an indirect comparison relies on the two subgroups of trials being similar on average in other factors that may affect outcome.

One approach that should never be used is the direct comparison of the relevant single arms of the trials. For example, patients receiving advice from a nurse in the nurse versus dietician trials should not be compared directly with patients receiving advice from a doctor in the doctor versus dietician trials. This comparison ignores the potential benefits of randomization and suffers from the same (usually extreme) biases as a comparison of independent cohort studies.

Indirect comparisons are not randomized comparisons, and cannot be interpreted as such. They are essentially observational findings across trials, and may suffer the biases of observational studies, for example due to confounding (see 8.8.5 Interpretation of subgroup analyses and meta-regressions). In situations when both direct and indirect comparisons are available in a review, then unless there are design flaws in the head-to-head trials, the two approaches should always be considered separately and the direct comparisons should take precedence as a basis for forming conclusions.

8.9 Presenting, illustrating and tabulating results

Several types of figures and tables are available for the presentation of results in a Cochrane Review. This section reviews those available in RevMan, and describes how to incorporate results produced outside of RevMan. First we address some issues to consider when reporting results in the text of the review.

8.9.1 Presenting results in the text

The results of individual studies and meta-analyses in a Cochrane Review are displayed in Figures and Tables. Each Figure and Table should be referred to in the results section of the review text. The results section should summarise the findings in a clear and logical order, and should explicitly address the objectives of the review. The section should be organised to follow the order of comparisons and outcomes specified in the protocol, and used as the data structure in RevMan.

Findings for the most important comparisons and/or outcomes should be prominent in the text of the review, even when little relevant data were available. Answers to post hoc analyses and less important questions for which there happen to be plentiful data should not be overemphasised. *Post hoc* analyses should always be identified as such.

The analytic methods that are used in a review should be described in the methods section. The reviewer should also make clear in the results section the method of analysis used for each quoted result (in particular, the choice of effect measure, the direction of a beneficial effect and the meta-analysis model used). Results should always be accompanied by a measure of uncertainty, such as a 95% confidence interval.

Reviewers should consider presenting results in formats that are easy to interpret. For example, odds ratios and standardized mean differences do not lend themselves to direct application in clinical practice but can be re-expressed in more accessible forms. See 8.X Re-expressing standardised mean differences and 8.X Re-expressing meta-analysis results as NNTs.

The abstract should summarise findings for only the most important comparisons and outcomes, and not selectively report those with the most significant results. It is helpful also to indicate the amount of information (numbers of studies and participants) on which analyses were based.

Methods for meta-analysis allow quantification of direction of effect, size of effect and consistency of effect. If suitable numerical data are not available for meta-analysis, or if meta-analyses are considered inappropriate, then these domains may often still be examined to provide a systematic assessment of the evidence available (see 8.1 Planning the analysis).

8.9.2 Figures

Graphical displays provide a clear and systematic means of presenting results both from individual studies and from meta-analyses. However, reviews that contain large numbers of figures are often difficult to follow, especially when each figure contains very little information.

The standard graphic in Cochrane Reviews is the forest plot, which doubles as both a Table and a Figure. The graphical section of a forest plot displays effect estimates and confidence intervals for both individual studies and meta-analyses. Each study is represented by a block at the point estimate of treatment effect with a horizontal line extending either side of the block. The area of the block indicates the weight assigned to that study in the meta-analysis while the horizontal line depicts the confidence interval (usually with a 95% level of confidence). The area of the block and the confidence interval convey similar information, but both make different contributions to the graphic. The confidence interval depicts the range of treatment effects compatible with the study's result and indicates whether each was individually statistically significant. The size of the block draws the eye towards the studies with larger weight (narrower confidence intervals), which dominate the calculation of the pooled result.

8.9.2.1 Forest plots in RevMan

RevMan produces forest plots and similar plots are automatically incorporated into the published version of the Cochrane Review. The different options for analyses, including the choice between fixed and random effects meta-analyses are available as options when forest plot figures are viewed in RevMan. Default analyses are displayed unless options are overridden. The defaults are Mantel-Haenszel odds ratios for dichotomous data, fixed effect meta-analyses of (weighted) mean differences for continuous data, Peto odds ratios for IPD outcomes and (in RevMan 4.2 and later) fixed effect meta-analyses for generic inverse variance outcomes. The reviewer should override any default settings that do not correspond with results reported in the text when setting up or editing outcomes in RevMan. This ensures that the results displayed are consistent with what is described in the text. Note that some users of the Cochrane Database of Systematic Reviews will be able to select alternative summary statistics and meta-analysis models to those intended by the reviewer when they view the results.

A past convention in CDSR has been that dichotomous outcomes have focussed on unfavourable outcomes, so that risk ratios and odds ratios less than one (and risk differences less than zero) indicate that an experimental intervention is superior to a control intervention. This would result in effect estimates to the left of the vertical line in a forest plot implying a benefit of the experimental intervention. The convention is no longer encouraged since it is not universally appropriate, and a much superior approach is to make it transparent which side of the line indicates benefit of which intervention by labelling the directions on the axis on the forest plots. RevMan allows reviewers to specify the labels used for 'treatment' and 'control' groups in each outcome. These labels are then used in the CDSR. Thus it is essential to know which way round figures are constructed and should be interpreted. This is particularly important for measurement scale data where it is not always apparent to a reader which direction on a scale indicates worsening health.

Presentation of data as a forest plot is discouraged where no study or only a single study is found for a particular outcome, except in circumstances where a blank forest plot makes a particular point about the lack of available data for an important outcome. To display outcomes noted only in single studies a forest plot using a subgroup for each outcome can be used (ensuring that the option to pool the data is disabled). Otherwise results of single studies may more conveniently be presented in an Additional Table (see 8.9.3 Tables).

Forest plots for dichotomous outcomes and IPD outcomes illustrate, by default:

(1) The raw data (corresponding to the 2×2 tables) for each study;
(2) Point estimates and confidence intervals for the chosen effect measure, both as blocks and lines and as text;
(3) A meta-analysis for each subgroup using the chosen effect measure and chosen method (fixed or random effects), both as a diamond and as text;
(4) The total numbers of participants in the experimental intervention and control intervention groups;
(5) Heterogeneity statistics (the chi-squared test and the I^2 statistic);
(6) A test for overall effect (overall average effect for random effects meta-analyses);
(7) The total numbers of events in the experimental intervention and control intervention groups;
(8) Percent weights given to each study.

Note that 3-8 are not displayed unless data are pooled. RevMan 4.2 separates 7 from 4, whereas earlier versions presented them together. This led to some confusion since it wrongly suggested to some users that the meta-analysis had been computed by comparing the totals of participants and events between experimental and control groups. For IPD outcomes it is also possible to enable display of the O – E and V statistics.

Forest plots for continuous outcomes illustrate, by default:

(1) The raw data (means, standard deviations and sample sizes) for each arm in each study;
(2) Point estimates and confidence intervals for the chosen effect measure, both as blocks and lines and as text;
(3) A meta-analysis for each subgroup using the chosen effect measure and chosen method (fixed or random effects), both as a diamond and as text;
(4) The total numbers of participants in the experimental and control groups;
(5) Heterogeneity statistics (the chi-squared test and the I^2 statistic);
(6) A test for overall effect (overall average effect for random effects meta-analyses);
(7) Percent weights given to each study.

Note that 3-7 are not displayed unless the data are pooled.

Forest plots for the generic inverse variance method illustrate, by default:

(1) The summary data for each study, as entered by the reviewer (for ratio measures these will be on the log scale);
(2) Point estimates and confidence intervals, both as blocks and lines and as text (for ratio measures these will be on the natural scale rather than the log scale);
(3) A meta-analysis for each subgroup using the chosen method (fixed or random effects), both as a diamond and as text;
(4) Heterogeneity statistics (the χ^2 test and the I^2 statistic);
(5) A test for overall effect (overall average effect for random effects meta-analyses);
(6) Percent weights given to each study.

Note that 3-6 are not shown unless data are pooled. It is possible additionally to enter sample sizes for experimental and control groups. These should be entered as appropriate for the design of the study. The sample sizes are not involved in the analysis, but if entered are displayed as:

(7) Numbers of participants in the experimental and control group for each study;
(8) The total numbers of participants in the experimental and control groups.

8.9.2.2 Additional figures

Additional figures may be attached to reviews in RevMan 4.2 and later. Examples of figures that reviewers may wish to include in a review include:

(1) forest plots where each line represents a meta-analysis rather than a study (for example, to illustrate multiple subgroup analyses or sensitivity analyses);
(2) funnel plots;
(3) plots illustrating meta-regression analyses;

(4) L'Abbé plots

Note that although funnel plots may be drawn using RevMan, they may only be included in the published review by attaching them as additional figures. Additional figures should not often be required, and should not be used to draw forest plots that can currently be drawn using RevMan. Where possible graphics should be produced using specialist statistical software packages such as Stata, SAS, SPSS, S-Plus or specialised meta-analysis software which produce appropriate publication quality graphics. General purpose spreadsheet programs may not provide suitable flexibility nor produce output of adequate quality.

A separate document (Appendix 8.1) is available that provides extensive guidance on the content of additional figures that illustrate numerical data. The document includes descriptions and recommendations for the four plots listed above among others. Reviewers should refer to this document before submitting a review containing additional figures. All additional figures should be assessed by a statistical editor or advisor prior to submission of a Cochrane Review for publication. Reviewers should be aware that additional figures can often be large and take up valuable storage space on the Cochrane Library. Guidance on technical aspects of additional figures is available at http://www.cc-ims.net.

Important results from all additional figures should be overviewed in the Results section of the review text. Wherever numerical results taken from a Figure are reported in the text of the review their meaning and derivation should be clear, and a reference to the relevant Figure should be provided.

8.9.3 Tables

RevMan supports three types of tables of results that can be linked to the Results text of the review.

(1) Forest plots generated by RevMan present summary data and effect estimates alongside their graphical representation (see 8.9.2.1 Forest plots in RevMan).
(2) The Table of Comparisons allows an outcome type of 'Other data'. Results of individual trials may be entered here as plain text. This option is well suited for entering summary data such as median values which cannot be pooled in a meta-analysis
(3) A flexible way of creating tables is provided by the Additional Tables feature, allowing presentation of results of both trials and meta-analyses, and other meta-analytical investigations (such as meta-regression analyses).

For further information see the RevMan User Guide.

Note that descriptions of study characteristics (methods, participants, interventions and outcomes studied) should be presented in the Table of Characteristics of Included Studies. Study results should not be included in this table.

The ability to incorporate additional figures in RevMan 4.2 and later technically allows reviewers to attach further additional tables as graphics files. Reviewers are discouraged from doing this due to the high volume of storage space taken up by graphics files.

Reviewers are instead asked to use the Additional Tables function, which is provided for this purpose.

Important results from all tables should be discussed and summarised in the Results section of the review text. When numerical results are reported in the text of the review a reference to the relevant Table should be provided.

8.10 Sensitivity analyses

Because there are different approaches to conducting a systematic review, reviewers should ask: How sensitive are the results of the analysis to changes in the way it was done? This provides reviewers with an approach to testing how robust the results of the review are relative to key decisions and assumptions that were made in the process of conducting the review. Each reviewer must identify how the key decisions and assumptions might conceivably have affected the results for a particular review. Generally, the types of decisions and assumptions that might be examined in sensitivity analyses include:

- changing the inclusion criteria for the types of study (e.g. using different methodological cut-points), participants, interventions or outcome measures
- including or excluding studies where there is some ambiguity as to whether they meet the inclusion criteria
- reanalysing the data using a reasonable range of results for studies where there may be some uncertainty about the results (e.g. because of inconsistencies in how the results are reported that cannot be resolved by contacting the investigators, or because of differences in how outcomes are defined or measured)
- reanalysing the data imputing a reasonable range of values for missing data
- reanalysing the data using different statistical approaches (e.g. using a random effects model instead of a fixed effect model, or *vice versa*)

The same cautions discussed for subgroup analyses apply to sensitivity analyses. In particular, since many sensitivity analyses involve between study subgroup comparisons, these findings need to be interpreted very carefully.

If the sensitivity analyses that are done do not materially change the results, it strengthens the confidence that can be placed in these results. If the results do change in a way that might lead to different conclusions, this indicates a need for greater caution in interpreting the results and drawing conclusions. Such differences might also enable reviewers to clarify the source of existing controversies about the effectiveness of an intervention, or lead them to hypothesise potentially important factors that might be related to the effectiveness of the intervention and warrant further investigation.

8.11 Special topics

8.11.1 Publication bias and funnel plots

As discussed in section 5, a particularly important component of a review is the identification of relevant studies. Publication bias has long been recognised as a problem

in this regard since it means that the likelihood of finding studies is related to the results of those studies (Begg 1988, Begg 1989, Easterbrook 1991, Dickersin 1992b). One way to investigate whether a review is subject to publication bias is to prepare a 'funnel plot' and examine this for signs of asymmetry. RevMan 4.0 includes a facility to produce such a graph. However, if there is asymmetry, reasons other than publication bias should also be considered.

Funnel plots were first used in educational research and psychology (Light 1984a). They are simple scatter plots of the treatment effects estimated from individual studies (on the x axis) against some measure of each study's sample size (y axis). The name 'funnel plot' arises from the fact that precision in the estimation of the true treatment effect increases as the sample size of the component studies increases. Effect estimates from small studies will therefore scatter more widely at the bottom of the graph, with the spread narrowing among larger studies. In the absence of bias the plot should resemble a symmetrical inverted funnel (see panel A of the figure).

Relative measures of treatment effect (such as relative risks and odds ratios) are plotted on a logarithmic scale. This ensures that effects of the same magnitude but opposite directions (for example relative risks of 0.5 and 2) are equidistant from 1.0 (Galbraith 1988). Treatment effects have generally been plotted against sample sizes. However, the statistical power of a trial is determined both by its total sample size and the number of participants experiencing the event of interest. For example, a study with 100,000 patients and 10 events is less likely to show a statistically significant treatment effect than a study with 1000 patients and 100 events. The standard error (SE) or the variance of the effect estimate, rather than total sample size, have therefore been increasingly used for the y axis in funnel plots. RevMan 4.0 uses 1/SE, plotted against the effect size calculated by the statistical method chosen by the reviewer for the particular outcome.

If there is bias, for example because smaller studies without statistically significant effects (shown as open circles in the figure) remain unpublished, this will lead to an asymmetrical appearance of the funnel plot with a gap in a bottom corner of the graph (panel B). In this situation the effect calculated in a meta-analysis will overestimate the treatment effect (Villar 1997, Egger 1997b). The more pronounced the asymmetry, the more likely it is that the amount of bias will be substantial.

Publication bias has long been associated with funnel plot asymmetry (Light 1984a). However the funnel plot should be seen as a generic means of examining whether the smaller studies in a meta-analysis tend to show larger treatment effects and this may be due to reasons other than publication bias (Egger 1997a, Egger 1998). Some of these are shown in the table:

Possible sources of asymmetry in funnel plots

1. Selection biases

Publication bias

Location biases

Language bias

Citation bias

Multiple publication bias

2. Poor methodological quality of 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
5. Chance

Even if a study has been published, the probability of finding it is also influenced by its results. For example, language bias (the preferential publication of studies without significant findings in languages other than English), makes it less likely that such 'negative' studies will be found (Grégoire 1995, Egger 1997c). Citation bias leads to 'negative' studies being referred to less often and they are therefore more likely to be missed when searching for relevant trials (Gotzsche 1987, Gotzcshe 1989, Ravnskov 1992). Conversely, results of 'positive' trials are sometimes reported more than once, increasing the probability that they will be located (multiple publication bias) (Gotzsche 1989, Huston 1996, Tramèr 1997).

Another source of asymmetry arises from differences in methodological quality. Smaller studies are, on average, conducted and analysed with less methodological rigour than larger studies. Trials of lower quality also tend to show larger treatment effects (Schulz 1995, Moher 1998). Trials which, if conducted and analysed properly, would have been 'negative' may thus become 'positive' (panel C).

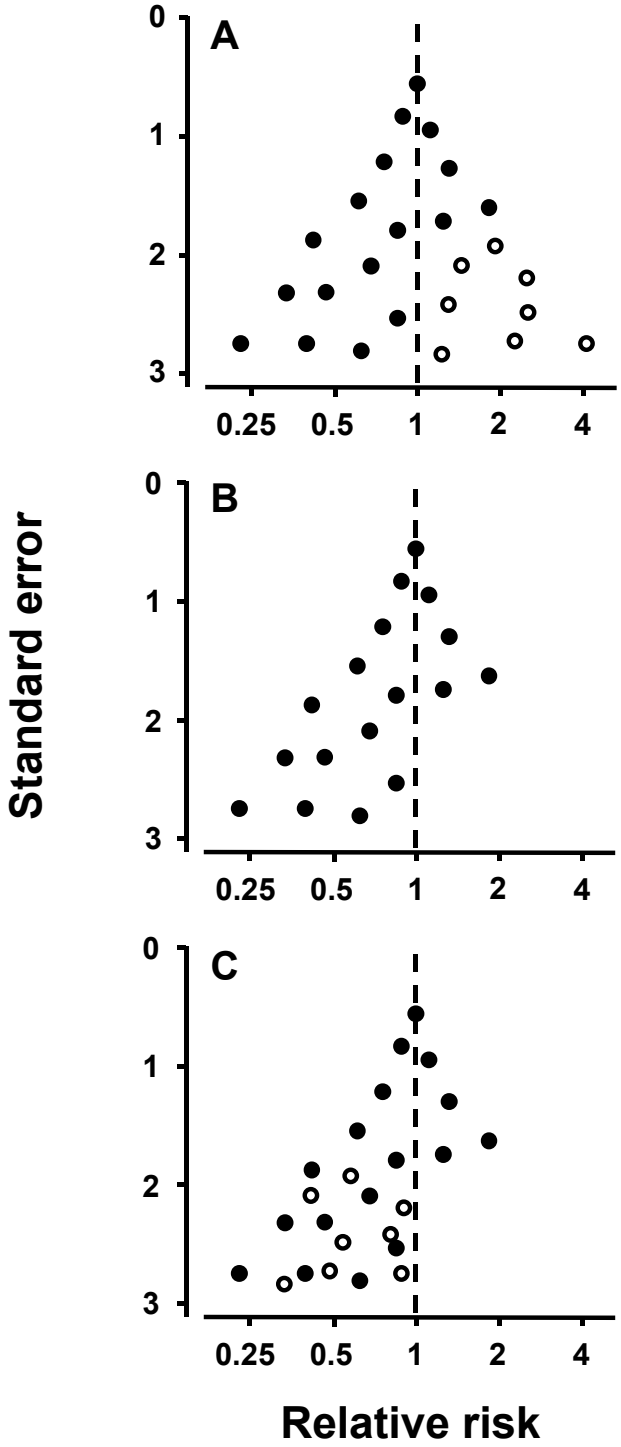
True heterogeneity in treatment effects may also lead to funnel plot asymmetry. For example, substantial benefit may be seen only in patients at high risk for the outcome which is affected by the intervention and these high risk patients are usually more likely to be included in early, small studies (Davey Smith 1994, Glasziou 1995). In addition, small trials are generally conducted before larger trials are established and in the intervening years standard treatment may have improved. Furthermore, some interventions may have been implemented less thoroughly in larger trials and, therefore, have resulted in smaller estimates of the treatment effect (Stuck 1998). It has also been argued that funnel plot asymmetry may be artefactual (Irwig 1998), but simulation studies have shown that this will occur infrequently, if the overall treatment effect is very large and the outcome of interest is rare (Sterne 2000). Finally, it is, of course, possible that an asymmetrical funnel plot arises merely by the play of chance.

Symmetry or asymmetry is generally defined informally, through visual examination, but the visual interpretation of funnel plots may vary between observers (Villar 1997). More formal statistical methods to examine associations between the study effects and size have been proposed (Begg 1994, Egger 1997b). At present there is debate regarding the statistical properties, potentials and limitations of these tests (Naylor 1997, Irwig 1998, Seagrott 1998, Egger 1998). No such tests are available in RevMan 4.0. Methodological work examining these issues is currently underway, but it is clear that both visual

examination and statistical analysis of funnel plots have limited power to detect bias if the number of studies is small.

Reviewers should look at the relevant funnel plot whenever they do a meta-analysis. If asymmetry is present, likely reasons should be explored. The power of this method is, however, at its most limited in those situations when bias is most likely to distort the results of the meta-analyses: when it comprises only a few small studies. Finally, it should be remembered that although funnel plots may alert reviewers to a problem which needs considering, they do not provide a solution to this problem. The only satisfactory way to address reporting bias and the inadequate quality of individual trials is through prospective registration of trials (Simes 1986, Dickersin 1988, Anonymous 1991, Dickersin 1992a) and improvements in the quality of the conduct, analysis and reporting of studies, meta-analyses and systematic reviews (Begg 1996, Moher 1995).

Legend to figure: Hypothetical funnel plots. Panel A: symmetrical plot in the absence of bias; Panel B: asymmetrical plot in the presence of reporting bias, Panel C: asymmetrical plot in the presence of bias due to low methodological quality of smaller studies.



8.12 Contributions

Contributing authors: Doug Altman, Deborah Ashby, Jacqueline Birks, Michael Borenstein, Marion Campbell, Jon Deeks, Matthias Egger, Julian Higgins, Joseph Lau, Keith O'Rourke, Rob Scholten, Jonathan Sterne, Simon Thompson, Anne Whitehead

Comments on drafts (statistical): Doug Altman, Deborah Ashby, Jesse Berlin, Joseph Beyene, Jacqueline Birks, Michael Bracken, Marion Campbell, Chris Cates, Mike Clarke, Albert Cobos, Francois Curtin, Roberto D'Amico, Keith Dear, Jon Deeks, Heather Dickinson, Diana Elbourne, Simon Gates, Paul Glasziou, Peter Herbison, Julian Higgins, Sally Hollis, David Jones, Steff Lewis, Nathan Pace, Craig Ramsey, Keith O'Rourke, Rob Scholten, Guido Schwarzer, Jonathan Sterne, Simon Thompson, Andy Vail, Clarine van Oel, Paula Williamson, Fred Wolf

Comments on drafts (non-statistical): Bodil Als-Nielsen, Wendong Chen, Esther Coren, Christian Gluud, Philippa Middleton, Jack Sinclair

8.13 References

Agresti 1996. Agresti A. An Introduction to Categorical Data Analysis. New York: Wiley, 1996.

Altman 1996. Altman DG, Bland JM. Detecting skewness from summary information. BMJ 1996; 313: 1200-1200.

Anonymous 1991. Anonymous. Making clinical trialists register. Lancet 1991; 338:244-5.

Antman 1992. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: Treatments for myocardial infarction. JAMA 1992; 268: 240-248.

Begg 1988. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. J Roy Stat Soc A 1988; 151:419-63.

Begg 1989. Begg CB, Berlin JA. Publication bias and dissemination of clinical research. J Natl Cancer Inst 1989; 81:107-15.

Begg 1994. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088-99.

Begg 1996. Begg CB, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637-639.

Berg 1988. Berg L. Clinical Dementia Rating (CDR). Psychopharm Bull 1988;24:637-639.

Berlin 1993. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. Epidemiology 1993; 4: 218-228.

Berlin 1994. Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trials data. Online J Curr Clin Trials 1994; Doc No 134.

Berlin 2002. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman KA. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat Med 2002; 21: 371-387.

Bucher 1997. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta- analysis of randomized controlled trials. J Clin Epidemiol 1997; 50: 683-691.

Chinn 2000. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med 2000; 19: 3127-3131.
Colletee 1994. Collett D. Modelling Survival Data in Medical Research. London: Chapman and Hall, 1994.

Cooper 1980. Cooper HM, Rosenthal R. Statistical versus traditional procedures for summarizing research findings. Psychol Bull 1980; 87: 442-449.

Davey Smith 1994. Davey Smith G, Egger M. Who benefits from medical interventions? Treating low risk patients can be a high risk strategy. Br Med J 1994;308:72-4.

Deeks 1997a. Deeks J. Are you sure that's a standard deviation? (part 1). Cochrane News 1997; Number 10; 11-12.

Deeks 1997b. Deeks J. Are you sure that's a standard deviation? (part 2). Cochrane News 1997; Number 11: 11-12.

Deeks 1998a. Deeks JJ, Bradburn MJ, Localio R, Berlin J. Much ado about nothing: Meta-analysis for rare events. 6th Cochrane Colloquium, Baltimore, 1998.

Deeks 1998b. Deeks JJ. Systematic reviews of published evidence: Miracles or minefields? Annals of Oncology 1998; 9: 703-709.

Deeks 2001a. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman DG (Eds). Systematic Reviews in Health Care: Meta-Analysis in Context (2nd edition). London: BMJ Publication Group, 2001.

Deeks 2001b. Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G, Altman DG (Eds). Systematic Reviews in Health Care: Meta-Analysis in Context (2nd edition). London: BMJ Publication Group, 2001.

Deeks 2002. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002; 21: 1575-1600.

DerSimonian 1986. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clin Trials 1986; 7: 177-188.

Dickersin 1988. Dickersin K. Report from the panel on the Case for Registers of Clinical Trials at the Eighth Annual Meeting of the Society for Clinical Trials. Controlled Clin Trials 1988; 9:76-81.

Dickersin 1992a. Dickersin K. Keeping posted. Why register clinical trials? - revisited. Controlled Clin Trials 1992; 13:170-7.

Dickersin 1992b. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. JAMA 1992; 263:374-8.

Early Breast Cancer Trialists' Collaborative Group 1990. Early Breast Cancer Trialists' Collaborative Group. Treatment of Early Breast Cancer. Volume 1: Worldwide Evidence 1985-1990 . Oxford: Oxford University Press, 1990.

Easterbrook 1991. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991; 337:867-72.

Egger 1997a. Egger M, Davey Smith G, Phillips AN. Meta-analysis: principles and procedures. Br Med J 1997;315:1533-7.

Egger 1997b. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315: 629-634.

Egger 1997c. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997;350:326-329.

Egger 1998. Egger M, Davey Smith G, Minder C. Authors' reply. Br Med J 1998;316:471.

Follman 1992. Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol 1992; 45: 769-773.

Galbraith 1988. Galbraith R. A note on graphical presentation of estimated odds ratios from several clinical trials. Stat Med 1988;7:889-94.

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

Gotzsche 1987. Gotzsche PC. Reference bias in reports of drug trials. Br Med J 1987;295:654-6.

Gotzsche 1989. Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis [published erratum appears in Controlled Clin. Trials 1989;50:356]. Controlled Clin Trials 1989; 10:31-56.

Greenland 1987. Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 1987; 9 : 1-30.

Greenland 1992. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. Am J Epidemiol 1992; 135: 1301-1309.

Greenland 1985. Greenland S, Robins J. Estimation of a common effect parameter from sparse follow-up data. Biometrics 1985; 41: 55-68.

Grégoire 1995. Grégoire G, Derderian F, LeLorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? J Clin Epidemiol 1995;48:159-63.

Hasselblad 1995. Hasselblad VIC, Mccrory DC. Meta-analytic tools for medical decision making: A practical guide. Med Decis Making 1995; 15: 81-96.

Higgins 2002. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21: 1539-1558.

Higgins 2003. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327: 557-560.

Hollis 1999. Hollis S, Campbell F. What is meant by intention to treat analysis? BMJ 1999; 319: 670-674.

Huston 1996. Huston P, Moher D. Redundancy, disaggregation, and the integrity of medical research. Lancet 1996;347:1024-6.

Irwig 1998. Irwig L, Macaskill P, Berry G. Graphical test is itself biased [letter]. Br Med J 1998;316:470.

Kjaergard 2001. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001; 135: 982-989.

Laupacis 1988. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. New Engl J Med 1988; 318: 1728-1733.

Lewis 1993. Lewis JA, Machin D. Intention to treat--who should use ITT? Br J Cancer 1993; 68: 647-650.

Light 1984. Light RJ, Pillemer DB. Summing up. The science of reviewing research. Cambridge, Massachusetts, and London, England: Harvard University Press, 1984.

Mantel 1959. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959; 22: 719-748.

McIntosh 1996. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. Stat Med 1996; 15: 1713-1728.

Moher 1998. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-13.

Morgenstern 1982. Morgenstern H. Uses of ecologic analysis in epidemiologic research. Am J Public Health 1982; 72: 1336-1344.

Naylor 1997. Naylor CD. Meta-analysis and the meta-epidemiology of clinical research. Br Med J 1997;315 :617-9.

Newell 1992. Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. Int J Epidemiol 1992; 21: 837-841.

O'Rourke 1989. O'Rourke K, Detsky AS. Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. J Clin Epidemiol 1989; 42: 1021-1026.

Oxman 1992. Oxman AD, Guyatt GH. A consumers guide to subgroup analyses. Ann Intern Med 1992; 116: 78-84.

Parmar 1998. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 1998; 17: 2815-2834.

Poole 1999. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. Am J Epidemiol 1999; 150: 469-475.

Ravnskov 1992. Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. Br Med J 1992;305:15-9.

Sackett 1996. Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! Evidence Based Medicine 1996; 1: 164-166.

Sackett 1997. Sackett DL, Richardson WS, Rosenberg W, Haynes BR. Evidence-Based Medicine: How to Practice and Teach EBM. Edinburgh: Churchill Livingstone, 1997.

Schulz 1995. 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.

Seagrott 1998. Seagrott V, Stratton I. Test had 10% false positive rate [letter]. Br Med J 1998;316:470.

Sharp 2000. Sharp SJ. Analysing the relationship between treatment benefit and underlying risk: precautions and practical recommendations. In: Egger M, Davey Smith G, Altman DG (Eds). Systematic Reviews in Health Care: Meta-Analysis in Context (2nd edition). London: BMJ Publications Group, 2000.

Simes 1986. Simes RJ. Publication bias: the case for an international registry of clinical trials. J Clin Oncol 1986; 4:1529-41.

Sinclair 1994. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. J Clin Epidemiol 1994; 47: 881-889.

Sterne 2000. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol 2000;53(11):1119-29.

Sterne 2001. Sterne JAC, Bradburn MJ, Egger M. Meta-analysis in Stata™. In: Egger M, Davey Smith G, Altman DG (Eds). Systematic Reviews in Health Care: Meta-Analysis in Context (2nd edition). London: BMJ Publication Group, 2001.

Stuck 1988. Stuck AE, Rubenstein LZ, Wieland D. Bias in meta-analysis detected by a simple, graphical test. Asymmetry detected in funnel plot was probably due to true heterogeneity [letter]. Br Med J 1998;316:469-71.

Thompson 1997. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med 1997; 16: 2741-2758.

Thompson 1999. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999; 18: 2693-2708.

Thompson 2002. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002; 21: 1559-1574.

Tramèr 1997. Tramèr MR, Reynolds DJM, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. Br Med J 1997;315:635-40.

Unnebrink 2001. Unnebrink K, Windeler J. Intention-to-treat methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Stat Med 2001; 20: 3931-3946.

Villar 1997. Villar J, Piaggio G, Carroli G, Donner A. Factors affecting the comparability of meta-analyses and largest trials results in perinatology. J Clin Epidemiol 1997;50:997-1002.

Whitehead 1991. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. Stat Med 1991; 10: 1665-1677.

Whitehead 1994. Whitehead A, Jones NMB. A meta-analysis of clinical trials involving different classifications of response into ordered categories. Stat Med 1994; 13: 2503-2515.

Yusuf 1995. Yusuf S, Peto R, Lewis J, Collins R, Sleight P, et al. Beta blockade during and after myocardial infarction: an overview of the randomised trials. Prog Cardiovasc Dis 1985; 27: 335-371.

Yusuf 1991. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991; 266: 93-98.

8.14 Sections under construction

You may have been directed to the following sections, which are currently under construction.

8.X Issues in interpretation
8.X Other types of study
8.X Missing data
8.X Investigating and dealing with bias
8.X Where to go for help
8.X Re-expressing meta-analysis results as NNTs
8.X Rare events (including zero frequencies)
8.X Re-expressing standardised mean differences
8.X Cluster randomized trials
8.X Cross-over trials
8.X Trials with more than two treatment groups
8.X Sensitivity analyses
8.X Bayesian meta-analysis
8.X Hierarchical models
8.X Multiple comparisons and the play of chance

The following parts of Section 8 are from an earlier version and will be replaced soon.

8.10 Sensitivity analyses
8.11.1 Publication bias and funnel plots