

Chapter 9: Analysing data and undertaking meta-analyses

Editors: Jonathan J Deeks, Julian PT Higgins and Douglas G Altman on behalf of the Cochrane Statistical Methods Group.

Copyright © 2008 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd under “The Cochrane Book Series” Imprint.

This extract is made available solely for use in the authoring, editing or refereeing of Cochrane reviews, or for training in these processes by representatives of formal entities of The Cochrane Collaboration. Other than for the purposes just stated, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the copyright holders.

Permission to translate part or all of this document must be obtained from the publishers.

This extract is from *Handbook* version 5.0.1. For guidance on how to cite it, see Section 9.8. The material is also published in Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on www.wiley.com.

Key points

- Meta-analysis is the statistical combination of results from two or more separate studies.
- Potential advantages of meta-analyses include an increase in power, an improvement in precision, the ability to answer questions not posed by individual studies, and the opportunity to settle controversies arising from conflicting claims. However, they also have the potential to mislead seriously, particularly if specific study designs, within-study biases, variation across studies, and reporting biases are not carefully considered.
- It is important to be familiar with the type of data (e.g. dichotomous, continuous) that result from measurement of an outcome in an individual study, and to choose suitable effect measures for comparing intervention groups.
- Most meta-analysis methods are variations on a weighted average of the effect estimates from the different studies.
- Variation across studies (heterogeneity) must be considered, although most Cochrane reviews do not have enough studies to allow the reliable investigation of the reasons for it. Random-effects meta-analyses allow for heterogeneity by assuming that underlying effects follow a normal distribution.
- Many judgements are required in the process of preparing a Cochrane review or meta-analysis. Sensitivity analyses should be used to examine whether overall findings are robust to potentially influential decisions.

9.1 Introduction

9.1.1 Do not start here!

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 authors, but results of meta-analyses can be very misleading if suitable attention has not been given to formulating the review question; specifying eligibility criteria; identifying, selecting and critically appraising studies; collecting appropriate data; and deciding what would be meaningful to analyse. Review authors should consult the chapters that precede this one before a meta-analysis is undertaken.

9.1.2 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 participants, 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 a review are discussed in this section and in Section 9.1.4.

Studies comparing healthcare interventions, notably randomized 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 here (experimental versus control interventions) implies the former, although the methods apply equally to the latter.

The contrast between the outcomes of two groups treated differently is known as the 'effect', the 'treatment effect' or the 'intervention 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 questions 1 to 3. Assessment of question 4 relies additionally on judgements based on assessments of study design and risk of bias, as well as statistical measures of uncertainty.

Narrative synthesis uses subjective (rather than statistical) methods to follow through questions 1 to 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 kava extract versus placebo for treating

anxiety (Pittler 2003). Meta-analysis and related techniques can be used if there is a consistent outcome measure to:

- Establish whether there is evidence of an effect;
 - Estimate the size of the effect and the uncertainty surrounding that size; and
 - 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 studies of numerous interventions for the same disease or condition. 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 (Crawford 2007). 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 Section 9.1.6 and Chapter 16, Section 16.6).
 3. Occasionally review comparisons have particularly wide scopes that make the use of meta-analysis problematic. For example, a review of workplace interventions for smoking cessation covered diverse types of interventions (Moher 2005). 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, work-based interventions can work (but see Section 9.1.4). 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 a review of beclomethasone versus placebo for chronic asthma, there was interest in whether the administered dose of beclomethasone affected its efficacy (Adams 2005). Such investigations of heterogeneity need to be undertaken with care (see Section 9.6).

9.1.3 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 Section 9.1.2. The following are reasons for considering including a meta-analysis in a review.

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 an intervention 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.

9.1.4 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 studies that are at risk of bias 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 are likely to produce an inappropriate summary.

9.1.5 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 (see Section 9.1.6) and what sorts of data are appropriate for the outcomes of interest (see Section 9.2). The next step is to prepare tabular summaries of the characteristics and results of the studies that are included in each comparison (extraction of data and conversion to the desired format is discussed in Chapter 7, Section 7.7). It is then possible to derive estimates of effect across studies in a systematic way (Section 9.4), to measure and investigate differences among studies (Sections 9.5 and 9.6) and to interpret the findings and conclude how much confidence should be placed in them (see Chapter 12).

9.1.6 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 Chapter 5). 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 author and the user. The formulation of the questions that a review addresses is discussed in Chapter 5. 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 ‘Data and analyses’ tables in RevMan and in deciding what information to put in the table of ‘Characteristics of included studies’.

9.1.7 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 author 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 this chapter when the protocol is written. As a guideline we recommend that the following be addressed:

1. Ensure that the analysis strategy firmly addresses the stated objectives of the review (see Section 9.1.2).
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 (see Section 9.3).
3. Decide whether a meta-analysis is intended and consider how the decision as to whether a meta-analysis is appropriate will be made (see Sections 9.1.3 and 9.1.4).
4. Determine the likely nature of outcome data (e.g. dichotomous, continuous etc) (see Section 9.2).
5. Consider whether it is possible to specify in advance what intervention effect measures will be used (e.g. risk ratio, odds ratio or risk difference for dichotomous outcomes, mean difference or standardized mean difference for continuous outcomes) (see Sections 9.4.4.4 and 9.4.5.1).
6. Decide how statistical heterogeneity will be identified or quantified (see Section 9.5.2).
7. Decide whether random-effects meta-analyses, fixed-effect meta-analyses or both methods will be used for each planned meta-analysis (see Section 9.5.4).
8. Consider how clinical and methodological diversity (heterogeneity) will be assessed and whether (and how) these will be incorporated into the analysis strategy (see Sections 9.5 and 9.6).
9. Decide how the risk of bias in included studies will be assessed and addressed in the analysis (see Chapter 8).
10. Pre-specify characteristics of the studies that may be examined as potential causes of heterogeneity (see Section 9.6.5).
11. Consider how missing data will be handled (e.g. imputing data for intention-to-treat analyses) (see Chapter 16, Sections 16.1 and 16.2).
12. Decide whether (and how) evidence of possible publication and/or reporting biases will be sought (see Chapter 10).

It may become apparent when writing the protocol that additional expertise is likely to be required, and if so, a statistician should be sought to join the review team.

9.2 Types of data and effect measures

9.2.1 Types of data

The starting point of all meta-analyses of studies of effectiveness involves the identification of the data type for the outcome measurements. Throughout this chapter 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; and
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 an intervention can be measured depend 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 of intervention effect. In Sections 9.4.4.4 and 9.4.5.1 we discuss issues in the selection of one of these measures for a particular meta-analysis.

9.2.2 Effect measures for dichotomous outcomes

Dichotomous (binary) 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); and
- the number needed to treat (NNT).

Details of the calculations of the first three of these measures are given in [Box 9.2.a](#). Numbers needed to treat are discussed in detail in Chapter 12 (Section 12.5).

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 9.2.a: Calculation of risk ratio (RR), odds ratio (OR) and risk difference (RD) from a 2×2 table

The results of a clinical trial can be displayed as a 2×2 table:

	Event	No event	Total
--	-------	----------	-------

	(‘Success’)	(‘Fail’)	
Experimental intervention	S _E	F _E	N _E
Control intervention	S _C	F _C	N _C

where S_E, S_C, F_E and F_C are the numbers of participants with each outcome (‘S’ or ‘F’) in each group (‘E’ or ‘C’). The following summary statistics can be calculated:

$$RR = \frac{\text{risk of event in experimental group}}{\text{risk of event in control group}} = \frac{S_E/N_E}{S_C/N_C}$$

$$OR = \frac{\text{odds of event in experimental group}}{\text{odds of event in control group}} = \frac{S_E/F_E}{S_C/F_C} = \frac{S_E F_C}{F_E S_C}$$

$$RD = \text{risk of event in experimental group} - \text{risk of event in control group}$$

$$= \frac{S_E}{N_E} - \frac{S_C}{N_C}$$

9.2.2.1 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. In ‘Summary of findings’ tables in Cochrane reviews, it is often expressed as a number of individuals per 1000 (see Chapter 11, Section 11.5). 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 10 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. In a sample of 1000 people, these numbers are 100 and 500 respectively.

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:

$$\text{risk} = \frac{\text{odds}}{1 + \text{odds}} ; \quad \text{odds} = \frac{\text{risk}}{1 - \text{risk}}$$

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 formula, the risk of the event is $0.1/(1+0.1) = 0.091$. In a sample of 100, about 9 individuals will have the event and 91 will not. When the odds is equal to 1, one person will have the event for every one who does not, so in a sample of 100, $100 \times 1/(1+1) = 50$ 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 intervention groups. To compare them we can look at their ratio (risk ratio or odds ratio) or their difference in risk (risk difference).

9.2.2.2 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** (or 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 (see [Box 9.2.a](#)). For both measures a value of 1 indicates that the estimated effects are the same for both interventions.

Neither the risk ratio nor the odds ratio can be calculated for a study if there are no events in the control group. This is because, as can be seen from the formulae in [Box 9.2.a](#), 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 $\frac{1}{2}$ 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 study 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 Chapter 16 (Section 16.9).

Risk ratios describe the multiplication of the risk that occurs with use of the experimental intervention. For example, a risk ratio of 3 for a treatment 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 $100 \times (RR - 1)\% = 200\%$. 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 $100 \times (1 - RR)\% = 75\%$. This is known as the relative risk reduction (see also Chapter 12, Section 12.5.1). 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 $1/\text{CGR}$, where CGR (abbreviation of ‘control group risk’, sometimes referred to as the 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 study.

Odds ratios, like odds, are more difficult to interpret (Sinclair 1994, Sackett 1996). 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 control group risk, as outlined above. The formula for converting an odds ratio to a risk ratio is provided in Chapter 12 (Section 12.5.4.4). Sometimes it may be sensible to calculate the RR for more than one assumed control group risk.

9.2.2.3 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 an intervention effect. Problems may arise, however, if the odds ratio is misinterpreted as a risk ratio. For interventions that increase the chances of events, the odds ratio will be larger than the risk ratio, so the misinterpretation will tend to overestimate the intervention effect, especially when events are common (with, say, risks of events more than 20%). For interventions that reduce the chances of events, the odds ratio will be smaller than the risk ratio, so that again misinterpretation overestimates the effect of the intervention. This error in interpretation is unfortunately quite common in published reports of individual studies and systematic reviews.

9.2.2.4 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 (see [Box 9.2.a](#)). The risk difference can be calculated for any study, even when there are no events in either group. The risk difference is straightforward to interpret: it describes the actual difference in the observed risk of events between experimental and control interventions; 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 the risk difference provides more directly relevant information than relative measures (Laupacis 1988, Sackett 1997) it is still important to be aware of the underlying risk of events and consequences of the events when interpreting a risk difference. Absolute measures, such as the risk difference, are particularly useful when considering trade-offs between likely benefits and likely harms of an intervention.

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 study 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 studies.

The number needed to treat is obtained from the risk difference. Although it is often used to summarize results of clinical trials, NNTs cannot be combined in a meta-analysis (see [Section 9.4.4.4](#)). However, odds ratios, risk ratios and risk differences may be usefully converted to NNTs and used when interpreting the results of a meta-analysis as discussed in Chapter 12 (Section 12.5).

9.2.2.5 What is the event?

In the context of dichotomous outcomes, healthcare 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 in Section 9.2.2 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 study 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 study, 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 study in which the experimental intervention reduces the occurrence of an adverse outcome will have an odds ratio and risk ratio less than 1, and a negative risk difference. A study in which the experimental intervention increases the occurrence of a good outcome will have an odds ratio and risk ratio greater than 1, and a positive risk difference (see Box 9.2.a).

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 intervention effects across studies. This is because the precision of a risk ratio estimate differs markedly between situations where risks are low and situations where risks are high. 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 Section 9.4.4.4.

9.2.3 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 and volume. 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 Section 9.2.4).

Two summary statistics are commonly used for meta-analysis of continuous data: the mean difference and the standardized 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 not addressed here.

9.2.3.1 The mean difference (or difference in means)

The **mean difference** (more correctly, ‘difference in means’) is a standard statistic that measures the absolute difference between the mean value in two groups in a clinical trial. It estimates the amount by which the experimental intervention changes the outcome on average compared with the control. It can be used as a summary statistic in meta-analysis when outcome measurements in all studies are made on the same scale.

Aside: Analyses based on this effect measure have historically been termed weighted mean difference (WMD) analyses in the Cochrane Database of Systematic Reviews (CDSR). This name is potentially

confusing: 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 study. Furthermore, all meta-analyses involve a weighted combination of estimates, yet we do not use the word 'weighted' when referring to other methods.

9.2.3.2 The standardized mean difference

The **standardized mean difference** is used as a summary statistic in meta-analysis when the studies all assess the same outcome but measure it in a variety of ways (for example, all studies measure depression but they use different psychometric scales). In this circumstance it is necessary to standardize the results of the studies to a uniform scale before they can be combined. The standardized mean difference expresses the size of the intervention effect in each study relative to the variability observed in that study. (Again in reality the intervention 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 studies 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 studies reflect differences in measurement scales and not real differences in variability among study populations. This assumption may be problematic in some circumstances where we expect real differences in variability between the participants in different studies. 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 intervention 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 study (see Chapter 12, Section 12.6).

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 standardized mean difference. It is recommended that the term 'standardized 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 'intervention effect' or 'effect estimate'. The particular definition of standardized 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 studies 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.

9.2.4 Effect measures for ordinal outcomes and 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.

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, review authors should 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 study, 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 dichotomous data by combining adjacent categories together. The latter is especially appropriate if an established, defensible cut-point is available. Inappropriate choice of a cut-point can induce bias, particularly if it is chosen to maximize the difference between two intervention arms in a clinical trial.

Where ordinal scales are summarized using methods for dichotomous data, one of the two sets of grouped categories is defined to be the event and intervention effects are described using risk ratios, odds ratios or risk differences (see Section 9.2.2). When ordinal scales are summarized using methods for continuous data, the intervention effect is expressed as a difference in means or standardized difference in means (see Section 9.2.3). Difficulties will be encountered if studies have summarized their results using medians (see Chapter 7, Section 7.7.3.5).

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.

9.2.5 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 hospitalization. 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 usually 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 intervention effect used will be the mean difference which will compare the difference in the mean number of events (possibly standardized to a unit time period) experienced by participants in the intervention group compared with participants in the control group.

9.2.5.1 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). 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 study may experience 63 seizures among them.

9.2.6 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 study 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 2×2 table can be constructed (see [Box 9.2.a](#)) and intervention 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 almost certainly will introduce bias.

The most appropriate way of summarizing time-to-event data is to use methods of survival analysis and express the intervention 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 interventions in a study 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.

9.2.7 Expressing intervention effects on log scales

The values of ratio intervention effects (such as the odds ratio, risk ratio, rate ratio and hazard ratio) usually undergo log transformations before being analysed, and they may occasionally be referred to in terms of their log transformed values. Typically the *natural* log transformation (log base e , written ‘ln’) 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 intervention 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 0 is minus infinity, the log of 1 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 intervention effect.

Graphical displays for meta-analysis performed on ratio scales usually use a log scale. This has the effect of making the confidence intervals appear symmetric, for the same reasons.

9.3 Study designs and identifying the unit of analysis

9.3.1 Unit-of-analysis issues

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. Authors should consider whether in each study:

- 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); and
- 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*.

9.3.2 Cluster-randomized trials

In a cluster-randomized trial, 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 Chapter 16 (Section 16.3).

9.3.3 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 Chapter 16 (Section 16.4).

9.3.4 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 study cannot be combined in a standard meta-analysis without a unit-of-analysis error. Some options are as follows.

- 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.
- 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.
- Select a single time-point and analyse only data at this time for studies in which it is presented. Ideally this should be a clinically important time-point. Sometimes it might be chosen to maximize the data available, although authors should be aware of the possibility of reporting biases.
- Select the longest follow-up from each study. This may induce a lack of consistency across studies, giving rise to heterogeneity.

9.3.5 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 Section [9.2.5](#).

9.3.6 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 Chapter 16 (Section 16.3).

9.3.7 Multiple body parts I: body parts receive the same intervention

In some studies, people are randomized, but multiple parts (or sites) of the body receive the same intervention, a separate outcome judgement being made for each body part, 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 Chapter 16 (Section 16.3).

9.3.8 Multiple body parts II: body parts receive different interventions

A different situation is that in which different parts of the body are randomized to *different* interventions. ‘Split-mouth’ designs in oral health are of this sort, in which different areas of the mouth are assigned different interventions. These trials have similarities to cross-over trials: whereas in cross-over trials individuals receive multiple treatments at different times, in these trials they receive multiple treatments at different sites. See methods described in Chapter 16 (Section 16.4). It is important to distinguish these studies from those in which participants receive the same intervention at multiple sites (Section [9.3.7](#)).

9.3.9 Multiple intervention groups

Studies that compare more than two intervention groups need to be treated with care. Such studies are often included in meta-analysis by making multiple pair-wise comparisons between all possible pairs of intervention groups. 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 Chapter 16 (Section 16.5).

9.4 Summarizing effects across studies

9.4.1 Meta-analysis

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 summarizes the effectiveness of the experimental intervention compared with a control intervention (see Section 9.1.2). 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 are provided in a supplementary document Statistical algorithms in Review Manager 5 (available on the *Handbook* web site), and a longer discussion of the issues discussed in this section appear in Deeks et al. (Deeks 2001).

9.4.2 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, to describe the observed intervention effect. 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) intervention effect estimate is calculated as a weighted average of the intervention effects estimated in the individual studies. A weighted average is defined as

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

where Y_i is the intervention effect estimated in the i th study, W_i is the weight given to the i th study, 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 intervention effect. The bigger the weight given to the i th study, the more it will contribute to the weighted average. The weights are therefore chosen to reflect the amount of information that each study contains. For ratio measures (OR, RR, etc), Y_i is the natural logarithm of the measure.

3. The combination of intervention effect estimates across studies may optionally incorporate an assumption that the studies are not all estimating the same intervention effect, but estimate intervention effects that follow a distribution across studies. This is the basis of a **random-effects meta-analysis** (see Section 9.5.4). 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) intervention 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, which communicates the strength of the evidence against the null hypothesis of no intervention 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 intervention effects across studies (see Section 9.5).

9.4.3 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 is implemented in its most basic form in RevMan, and is 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 minimizes the imprecision (uncertainty) of the pooled effect estimate.

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

$$\text{generic inverse-variance weighted average} = \frac{\sum Y_i (1/SE_i^2)}{\sum (1/SE_i^2)},$$

where Y_i is the intervention effect estimated in the i th study, SE_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 intervention effect and its standard error from each study.

9.4.3.1 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, intervention 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 Section 9.5.4. To undertake a random-effects meta-analysis, the standard errors of the study-specific estimates (SE_i above) are adjusted to incorporate a measure of the extent of variation, or heterogeneity, among the intervention effects observed in different studies (this variation is often referred to as tau-squared (τ^2 , or Tau²)). The amount of variation, and hence the adjustment, can be estimated from the intervention effects and standard errors of the studies included in the meta-analysis.

9.4.3.2 The generic inverse variance outcome type in RevMan

Estimates and their standard errors may be entered directly into RevMan 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 intervention effect, the data should be entered as natural 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-randomized 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 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 author to calculate effect estimates, and allows the use of methods targeted specifically at different types of data (see Sections 9.4.4 and 9.4.5). Also, it is helpful for the readers of the review to see the summary statistics for each intervention group in each study.

9.4.4 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). All of these methods are available as analysis options in RevMan. 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 2001).

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.

9.4.4.1 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 study 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.

9.4.4.2 Peto odds ratio method

Peto's method (Yusuf 1985) can only be used to pool odds ratios. It uses an inverse variance approach but utilizes 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 study.

The approximation used in the computation of the log odds ratio works well when intervention effects are small (odds ratios are close to 1), events are not particularly common and the studies 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 (Bradburn 2007) (see Chapter 16, Section 16.9). Also, Peto's method can be used to combine studies with dichotomous outcome data with studies using time-to-event analyses where log-rank tests have been used (see Section 9.4.9).

9.4.4.3 Random-effects method

The random-effects method (DerSimonian 1986) incorporates an assumption that the different studies are estimating different, yet related, intervention effects. As described in Section 9.4.3.1, 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 intervention effects. The random-effects method and the fixed-effect method will give identical results when there is no heterogeneity among the studies. Where there is heterogeneity, confidence intervals for the average intervention effect will be wider if the random-effects 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 intervention effect will change if there are relationships between observed intervention effects and sample sizes. See Section 9.5.4 for further discussion of these issues.

RevMan implements two random-effects methods for dichotomous data: a Mantel-Haenszel method and an inverse-variance method. The difference between the two is subtle: the former estimates the amount of between-study variation by comparing each study's result with a Mantel-Haenszel fixed-effect meta-analysis result, whereas the latter estimates the amount of variation across studies by comparing each study's result with an inverse-variance fixed-effect meta-analysis result. In practice, the difference is likely to be trivial. The inverse-variance method was added in RevMan version 5.

9.4.4.4 Which measure for dichotomous outcomes?

Summary statistics for dichotomous data are described in Section 9.2.2. The effect of intervention 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 studies in the meta-analysis and subdivisions of the population to which the interventions will be applied. The more consistent the summary statistic the greater is the justification for expressing the intervention effect 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 (Engels 2000, Deeks 2002). 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 terms of consistency (Deeks 2002). When the study aims to reduce the incidence of an adverse outcome (see Section 9.2.2.5) 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 maximizes 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 Chapter 12, Section 12.5). 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 Section 9.2.2).

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), and allow trade-offs to be made between likely benefits and likely harms of interventions. However, they are less likely to be generalizable.

It seems important to avoid using summary statistics for which there is empirical evidence that they are unlikely to give consistent estimates of intervention 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 Section 9.7).

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 Chapter 12, Section 12.5). This is one of the key motivations for ‘Summary of findings’ tables in Cochrane reviews: see Chapter 11 (Section 11.5). If odds ratios are used for meta-analysis they can also be re-expressed as risk ratios (see Chapter 12, Section 12.5.4). 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 ‘control’ population to which the experimental intervention will be applied. Where the chosen value for this assumed control risk is close to the typical observed control group risks across the studies, similar estimates of absolute effect will be obtained regardless of whether odds ratios or risk ratios are used for meta-analysis. Where the assumed control risk differs from the typical observed control group risk, the predictions of absolute benefit will differ according to which summary statistic was used for meta-analysis.

9.4.5 Meta-analysis of continuous outcomes

Two methods of analysis are available in RevMan for meta-analysis of continuous data: the inverse-variance fixed-effect method and the inverse-variance random-effects method. The methods will give exactly the same answers when there is no heterogeneity. Where there is heterogeneity, confidence intervals for the average intervention effect will be wider if the random-effects 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 intervention effect will change if there are relationships between observed intervention effects and sample sizes. See Section 9.5.4 for further discussion of these issues.

Authors should be aware that an assumption underlying methods for meta-analysis of continuous data is that the outcomes have a normal distribution in each intervention 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 Section 9.4.5.3).

9.4.5.1 Which measure for continuous outcomes?

There are two summary statistics used for meta-analysis of continuous data: the mean difference (MD) and the standardized mean difference (SMD) (see Section 9.2.3). Selection of summary statistics for continuous data is principally determined by whether studies all report the outcome using the same scale (when the mean difference can be used) or using different scales (when the standardized mean difference has to be used).

The different roles played in the two approaches by the standard deviations of outcomes observed in the two groups should be understood.

- For the mean difference approach, 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 standardized mean difference approach, the standard deviations are used to standardize the mean differences to a single scale (see Section 9.2.3.2), as well as in the computation of study weights. It is assumed that between-study variation in standard deviations reflects only differences in measurement scales and not differences in the reliability of outcome measures or variability among study populations.

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

9.4.5.2 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.

The preferred statistical approach to accounting for baseline measurements of the outcome variable is to include the baseline outcome measurements as a covariate in a regression model or analysis of covariance (ANCOVA). These analyses produce an ‘adjusted’ estimate of the treatment effect together with its standard error. These analyses are the least frequently encountered, but as they give the most precise and least biased estimates of treatment effects they should be included in the analysis when they are available. However, they can only be included in a meta-analysis using the generic inverse-variance method, since means and standard deviations are not available for each intervention group separately.

In practice an author is likely to discover that the studies 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 of mean differences. There is no statistical reason why studies with change-from-baseline outcomes should not be combined in a meta-analysis with studies with final measurement outcomes when using the (unstandardized) 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 intervention 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 authors must be careful to use the appropriate means and standard deviations (either of final measurements or of changes from baseline) for each study. 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 standardized mean differences, since the difference in standard deviation reflects not differences in measurement scale, but differences in the reliability of the measurements.

A common practical problem associated with including change-from-baseline measures is that the standard deviation of changes is not reported. Imputations of standard deviations is discussed in Chapter 16 (Section 16.1.3).

9.4.5.3 Meta-analysis of skewed data

Analyses based on 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. Skew 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 with minimum or maximum scores, but it may not be 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 skew (Altman 1996). If the ratio is less than 1 there is strong evidence of a skewed distribution.

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

Where data have been analysed on a log scale, results are commonly presented as geometric means and ratios of geometric means. A meta-analysis may be then performed on the scale of the log-transformed data; an example of the calculation of the required means and standard deviation is given in Chapter 7 (Section 7.7.3.4). This approach depends on being able to obtain transformed data for all studies; methods for transforming from one scale to the other are available (Higgins 2008b). Log-transformed and untransformed data can not be mixed in a meta-analysis.

9.4.6 Combining dichotomous and continuous outcomes

Occasionally authors 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 dichotomized. 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 summarize 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 summarize the data in three ways: by placing entering the means and standard deviations as continuous outcomes, by entering the counts as dichotomous outcomes and by entering all of the data in text form as ‘Other data’ outcomes.

There are statistical approaches available which will re-express odds ratios as standardized mean differences (and *vice versa*), allowing dichotomous and continuous data to be pooled together. Based on an assumption that the underlying continuous measurements in each intervention group follow 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 standardized mean difference according to the following simple formula (Chinn 2000):

$$\text{SMD} = \frac{\sqrt{3}}{\pi} \ln \text{OR} .$$

The standard error of the log odds ratio can be converted to the standard error of a standardized mean difference by multiplying by the same constant ($\sqrt{3}/\pi = 0.5513$). Alternatively standardized mean differences can be re-expressed as log odds ratios by multiplying by $\pi/\sqrt{3} = 1.814$. Once standardized mean differences (or log odds ratios) and their standard errors have been computed for all studies in the meta-analysis, they can be combined using the generic inverse-variance method in RevMan. Standard errors can be computed for all studies 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 standardized mean differences into standard errors (see Chapter 7, Section 7.7.7.2).

9.4.7 Meta-analysis of ordinal outcomes and measurement scales

Ordinal and measurement scale outcomes are most commonly meta-analysed as dichotomous data (if so see Section 9.4.4) or continuous data (if so see Section 9.4.5) depending on the way that the study authors 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 intervention group can be obtained, and the same ordinal scale has been used in all studies. This approach may make more efficient use of all available data than dichotomization, but requires access to 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 intervention effect (Agresti 1996). Suppose that there are three categories, which are ordered in terms of desirability such that 1 is the best and 3 the worst. The data could be dichotomized in two 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 study 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 (see Section 9.4.3.2). 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 three-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 are available for dealing with this, and for combining data from scales that are related but have different definitions for their categories (Whitehead 1994).

9.4.8 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 (see Section 9.2.5). 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 Chapter 7 (Section 7.7.5), count data may be analysed using methods for dichotomous (see Section 9.4.4), continuous (see Section 9.4.5) and time-to-event data (see Section 9.4.9) 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 summarized at the group level. For example, participants in the control group of a clinical 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 study may be expressed as a **rate ratio**, that is the ratio of the rate in the experimental intervention group to the rate in the control group. Suppose E_E events occurred during T_E participant-years of follow-up in the experimental intervention group, and E_C events during T_C participant-years in the control intervention group. The rate ratio is

$$\text{rate ratio} = \frac{E_E/T_E}{E_C/T_C} = \frac{E_E T_C}{E_C T_E}.$$

The (natural) logarithms of the rate ratios may be combined across studies using the generic inverse-variance method (see Section 9.4.3.2). An approximate standard error of the log rate ratio is given by

$$\text{SE of ln rate ratio} = \sqrt{\frac{1}{E_E} + \frac{1}{E_C}}.$$

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 dichotomizing 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,

$$\text{rate difference} = \frac{E_E}{T_E} - \frac{E_C}{T_C}.$$

An approximate standard error for the rate difference is

$$\text{SE of rate difference} = \sqrt{\frac{E_E}{T_E^2} + \frac{E_C}{T_C^2}}.$$

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).

9.4.9 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 re-analysis of individual patient data.

If ‘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, then these statistics may be entered directly into RevMan using the ‘O – E and Variance’ outcome type. There are several ways to calculate ‘O – E’ and ‘V’ statistics. Peto’s method applied to dichotomous data (Section 9.4.4.2) gives rise to an odds ratio; a log-rank approach gives rise to a hazard ratio, and a variation of the Peto method for analysing time-to-event data gives rise to something in between. The appropriate effect measure should be specified in RevMan. Only fixed-effect meta-analysis methods are available in RevMan for ‘O – E and Variance’ outcomes.

Alternatively if estimates of log hazard ratios and standard errors have been obtained from results of Cox proportional hazards regression models, study results can be combined using the generic inverse-variance method (see Section 9.4.3.2). Both fixed-effect and random-effects analyses are available.

If a mixture of log-rank and Cox model estimates are obtained from the studies, all results can be combined using the generic inverse-variance method, as the log-rank estimates can be converted into log hazard ratios and standard errors using the formulae given in Chapter 7 (Section 7.7.6).

9.4.10 A summary of meta-analysis methods available in RevMan

Table 9.4.a lists the options for statistical analysis that are available in RevMan. RevMan requires the author 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 mean difference for continuous outcomes and the fixed-effect model for generic inverse-variance outcomes. It is important that authors 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.

Table 9.4.a: Summary of meta-analysis methods available in RevMan

Type of data	Effect measure	Fixed-effect methods	Random-effects methods
Dichotomous	Odds ratio (OR)	Mantel-Haenszel (M-H)	Mantel-Haenszel (M-H)
		Inverse variance (IV) Peto	Inverse variance (IV)
	Risk ratio (RR)	Mantel-Haenszel (M-H) Inverse variance (IV)	Mantel-Haenszel (M-H) Inverse variance (IV)
	Risk difference (RD)	Mantel-Haenszel (M-H) Inverse variance (IV)	Mantel-Haenszel (M-H) Inverse variance (IV)
Continuous	Mean difference (MD)	Inverse variance (IV)	Inverse variance (IV)
	Standardized mean difference (SMD)	Inverse variance (IV)	Inverse variance (IV)
O – E and Variance	<i>User-specified</i> (default ‘Peto odds ratio’)	Peto	<i>None</i>
Generic inverse variance	<i>User-specified</i>	Inverse variance (IV)	Inverse variance (IV)
Other data	<i>User-specified</i>	<i>None</i>	<i>None</i>

9.4.11 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).

9.5 Heterogeneity

9.5.1 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 study design and risk of bias may be described as **methodological diversity** (sometimes called methodological heterogeneity). Variability in the intervention effects being evaluated in the different studies is known as **statistical heterogeneity**, and is a consequence of clinical or methodological diversity, or both, among the studies. Statistical heterogeneity manifests itself in the observed intervention 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 intervention effect is affected by the factors that vary across studies; most obviously, the specific interventions or patient characteristics. In other words, the true intervention effect will be different in different studies.

Differences between studies in terms of methodological factors, such as use of blinding and concealment of allocation, or if there are differences between studies in the way the outcomes are defined and measured, may be expected to lead to differences in the observed intervention 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 intervention 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 Chapter 8.

The scope of a review will largely determine the extent to which studies included in a review are diverse. Sometimes a review will include studies addressing a variety of questions, for example when several different interventions for the same condition are of interest (see also Chapter 5, Section 5.6). Studies of each intervention should be analysed and presented separately. Meta-analysis should only be considered when a group of studies 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 studies 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 intervention effect, rather than on obtaining a summary effect estimate across a series of studies (see Section 9.6). Meta-regression may best be used for this purpose, although it is not implemented in RevMan (see Section 9.6.4).

9.5.2 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 (χ^2 , or Chi²) test is included in the forest plots in 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 intervention 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 studies 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 (Higgins 2003). 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

$$I^2 = \left(\frac{Q - df}{Q} \right) \times 100\%$$

where Q is the chi-squared statistic and df is its degrees of freedom (Higgins 2002, Higgins 2003). This describes the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance).

Thresholds for the interpretation of I^2 can be misleading, since the importance of inconsistency depends on several factors. A rough guide to interpretation is as follows:

- 0% to 40%: might not be important;
- 30% to 60%: may represent moderate heterogeneity*;
- 50% to 90%: may represent substantial heterogeneity*;
- 75% to 100%: considerable heterogeneity*.

*The importance of the observed value of I^2 depends on (i) magnitude and direction of effects and (ii) strength of evidence for heterogeneity (e.g. P value from the chi-squared test, or a confidence interval for I^2).

9.5.3 Strategies for addressing heterogeneity

A number of options are available if (statistical) heterogeneity is identified among a group of studies 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 Section 9.3).

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 intervention 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 Section 9.6.3) or meta-regression (see Section 9.6.4), though this latter method is not implemented in RevMan. Ideally, investigations of characteristics of studies that may be associated with heterogeneity should be pre-specified in the protocol of a review (see Section 9.1.7). Reliable conclusions can only be drawn from analyses that are truly pre-specified before inspecting the studies' results, and even these conclusions should be interpreted with caution. In practice, authors will often be familiar with some study 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 intervention effect. However, the existence of heterogeneity suggests that there may not be a single intervention effect but a distribution of intervention effects. Thus the pooled fixed-effect estimate may be an intervention 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 Section 9.5.4). 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 studies. 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 in Section 9.5.4.

6. Change the effect measure

Heterogeneity may be an artificial consequence of an inappropriate choice of effect measure. For example, when studies 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 standardized 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 risks vary, homogeneous odds ratios or risk ratios will necessarily lead to heterogeneous risk differences, and *vice versa*. However, it remains unclear whether homogeneity of intervention effect in a particular meta-analysis is a suitable criterion for choosing between these measures (see also Section 9.4.4.4).

7. Exclude studies

Heterogeneity may be due to the presence of one or two outlying studies with results that conflict with the rest of the studies. 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 study 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 studies as part of a sensitivity analysis (see Section 9.7). Whenever possible, potential sources of clinical diversity that might lead to such situations should be specified in the protocol.

9.5.4 Incorporating heterogeneity into random-effects models

A fixed-effect meta-analysis provides a result that may be viewed as a ‘typical intervention 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 intervention (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, intervention effects differ by considering the differences as if they were random. The centre of this 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 Section 9.6).

For random-effects analyses in RevMan, the pooled estimate and confidence interval refer to the centre of the distribution of intervention effects, but 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. 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.

In common with other meta-analysis software, RevMan presents an estimate of the between-study variance in a random-effects meta-analysis (known as tau-squared (τ^2 or Tau^2)). The square root of this number (i.e. tau) is the estimated standard deviation of underlying effects across studies. For absolute measures of effect (e.g. risk difference, mean difference, standardized mean difference), an approximate 95% range of underlying effects can be obtained by creating an interval from $2 \times \text{tau}$ below the random-effects pooled estimate, to $2 \times \text{tau}$ above it. For relative measures (e.g. odds ratio, risk ratio), the interval needs to be centred on the natural logarithm of the pooled estimate, and the limits anti-logged (exponentiated) to obtain an interval on the ratio scale. Alternative intervals, for the predicted effect in a new study, have been proposed (Higgins 2008a). The range of the intervention

effects observed in the studies may be thought to give a rough idea of the spread of the distribution of true intervention 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 random-effects pooled estimate 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 intervention effect?’ while the fixed-effect estimate and its confidence interval addresses the question ‘what is the best estimate of the intervention effect?’ The answers to these questions coincide either when no heterogeneity is present, or when the distribution of the intervention 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.

Methodological diversity creates heterogeneity through biases variably affecting the results of different studies. The random-effects pooled estimate will only estimate the average treatment effect if the biases are symmetrically distributed, leading to a mixture of over- and under-estimates of effect, which is unlikely to be the case. In practice it can be very difficult to distinguish whether heterogeneity results from clinical or methodological diversity, and in most cases it is likely to be due to both, so these distinctions in the interpretation are hard to draw.

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 Section 9.5.2). The choice between a fixed-effect and a random-effects meta-analysis should never be made on the basis of a statistical test for 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 intervention effect. Care must be taken that random-effects analyses are applied only when the idea of a ‘random’ distribution of intervention 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 within-study bias in smaller studies (Egger 1997, Poole 1999, Kjaergard 2001), then a random-effects meta-analysis will exacerbate the effects of the bias (see also Chapter 10, Section 10.4.4.1). 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 studies or if the studies are small with few events, a random-effects analysis will provide poor estimates of the width of the distribution of intervention effects. The Mantel-Haenszel method will provide more robust estimates of the average intervention 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 more advanced statistical software (see also Chapter 16, Section 16.8). In practice, the difference in the results is likely to be small unless there are few

studies. For dichotomous data, RevMan implements two versions of the DerSimonian and Laird random-effects model (see Section 9.4.4.3).

9.6 Investigating heterogeneity

9.6.1 Interaction and effect modification

Does the intervention effect vary with different populations or intervention 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.

9.6.2 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 studies 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 Chapter 18). The methods we describe in Section 9.6.3 are for subgroups of trials.

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.

Authors will find useful advice concerning subgroup analyses in Oxman and Guyatt (Oxman 1992) and Yusuf et al. (Yusuf 1991). See also Section 9.6.6.

9.6.3 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 the 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.

9.6.3.1 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. 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. (Deeks 2001). This method is implemented in RevMan for fixed-effect analyses based on the inverse-variance method. If Mantel-Haenszel methods for the dichotomous data type are used, then the test would include a slight inaccuracy due to the way in which the heterogeneity chi-squared statistic is calculated. The procedure 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_{tot} , is available for all of the studies, and that chi-squared heterogeneity statistics Q_1 up to Q_J are available for J subgroups (such that every study is in one and only one subgroup). Then the new statistic $Q_{int} = Q_{tot} - (Q_1 + \dots + Q_J)$, compared with a chi-squared distribution with $J - 1$ degrees of freedom, tests for a difference among the subgroups. 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 Section 9.6.4). This approach may be regarded as preferable due to the high risk of false-positive results when comparing subgroups in a fixed-effect model (Higgins 2004).

9.6.4 Meta-regression

If studies are divided into subgroups (see Section 9.6.2), this may be viewed as an investigation of how a categorical study characteristic is associated with the intervention effects in the meta-analysis. For example, studies in which allocation sequence concealment was adequate may yield different results from those in which it was inadequate. Here, allocation sequence 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 studies) (Thompson 2002). Meta-regression should generally not be considered when there are fewer than ten studies 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 intervention 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 intervention 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 intervention 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 intervention effect and the explanatory variable. If the intervention effect is a ratio measure, the log-transformed value of the intervention effect should always be used in the regression model (see Section 9.2.7), and the exponential of the regression coefficient will give an estimate of the relative change in intervention 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 J subgroups membership of particular subgroups is indicated by using $J - 1$ 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 intervention 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 may be performed using the ‘metareg’ macro available for the Stata statistical package.

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

Authors 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 intervention 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).

9.6.5.1 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. However, even this will be too few when the covariates are unevenly distributed.

9.6.5.2 Specify characteristics in advance

Authors 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 studies’ results influencing which subgroups are analysed. True pre-specification is difficult in systematic reviews, because the results of some of the relevant studies 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 authors should not be reluctant to explore it. However, such *post hoc* analyses should be identified as such.

9.6.5.3 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 Chapter 16, Section 16.7).

9.6.5.4 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 intervention 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 an intervention 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 intervention. 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 1998). 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).

9.6.5.5 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 summarized 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 intervention 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 studies to display the opposite pattern to that observed within each study.

9.6.5.6 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 intervention 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 study characteristics will give some information about which study characteristics may be confounded with each other.

9.6.6 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 studies. Even if individuals are randomized to one group or other within a clinical trial, they are not randomized to go in one trial or another. Hence, subgroup analyses suffer the limitations of any observational investigation, including possible bias through confounding by other study-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*?

Authors 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 should not be compared. See Section [9.6.3.1](#).

- 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.

9.6.7 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 study (the control group risk (CGR)). 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. Sharp provides a full discussion of the topic (Sharp 2000).

Intuition would suggest that participants are more or less likely to benefit from an effective intervention according to their risk status. However, the relationship between baseline risk and intervention effect is a complicated issue. For example, suppose an intervention 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 Section 9.4.4.4).

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.

9.6.8 Dose-response analyses

The principles of meta-regression can be applied to the relationships between intervention 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 authors should consider these effects, particularly as a possible explanation for heterogeneity, they should be cautious about drawing conclusions based on between-study differences. Authors 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.

9.7 Sensitivity analyses

The process of undertaking a systematic review involves a sequence of decisions. Whilst many of these decisions are clearly objective and non-contentious, some will be somewhat arbitrary or unclear. For instance, if inclusion criteria involve a numerical value, the choice of value is usually arbitrary: for example, defining groups of older people may reasonably have lower limits of 60, 65, 70 or 75 years, or any value in between. Other decisions may be unclear because a study report fails to include the required information. Some decisions are unclear because the included studies themselves never obtained the information required: for example, the outcomes of those who unfortunately were lost to follow-up. Further decisions are unclear because there is no consensus on the best statistical method to use for a particular problem.

It is desirable to prove that the findings from a systematic review are not dependent on such arbitrary or unclear decisions. A sensitivity analysis is a repeat of the primary analysis or meta-analysis, substituting alternative decisions or ranges of values for decisions that were arbitrary or unclear. For example, if the eligibility of some studies in the meta-analysis is dubious because they do not contain full details, sensitivity analysis may involve undertaking the meta-analysis twice: first, including all studies and second, only including those that are definitely known to be eligible. A sensitivity analysis asks the question, “Are the findings robust to the decisions made in the process of obtaining them?”.

There are many decision nodes within the systematic review process which can generate a need for a sensitivity analysis. Examples include:

Searching for studies:

- Should abstracts whose results cannot be confirmed in subsequent publications be included in the review?

Eligibility criteria:

- Characteristics of participants: where a majority but not all people in a study meet an age range, should the study be included?
- Characteristics of the intervention: what range of doses should be included in the meta-analysis?
- Characteristics of the comparator: what criteria are required to define usual care to be used as a comparator group?
- Characteristics of the outcome: what time-point or range of time-points are eligible for inclusion?
- Study design: should blinded and unblinded outcome assessment be included, or should study inclusion be restricted by other aspects of methodological criteria?

What data should be analysed?

- Time-to-event data: what assumptions of the distribution of censored data should be made?
- Continuous data: where standard deviations are missing, when and how should they be imputed? Should analyses be based on change scores or on final values?
- Ordinal scales: what cut-point should be used to dichotomize short ordinal scales into two groups?
- Cluster-randomized trials: what values of the intraclass correlation coefficient should be used when trial analyses have not been adjusted for clustering?
- Cross-over trials: what values of the within-subject correlation coefficient should be used when this is not available in primary reports?
- All analyses: what assumptions should be made about missing outcomes to facilitate intention-to-treat analyses? Should adjusted or unadjusted estimates of treatment effects used?

Analysis methods:

- Should fixed-effect or random-effects methods be used for the analysis?
- For dichotomous outcomes, should odds ratios, risk ratios or risk differences be used?
- And for continuous outcomes, where several scales have assessed the same dimension, should results be analysed as a standardized mean difference across all scales or as mean differences individually for each scale?

Some sensitivity analyses can be pre-specified in the study protocol, but many issues suitable for sensitivity analysis are only identified during the review process where the individual peculiarities of the studies under investigation are identified. When sensitivity analyses show that the overall result and conclusions are not affected by the different decisions that could be made during the review process, the results of the review can be regarded with a higher degree of certainty. Where sensitivity analyses identify particular decisions or missing information that greatly influence the findings of the review, greater resources can be deployed to try and resolve uncertainties and obtain extra information, possibly through contacting trial authors and obtaining individual patient data. If this cannot be achieved, the results must be interpreted with an appropriate degree of caution. Such findings may generate proposals for further investigations and future research.

Reporting of sensitivity analyses in a systematic review may best be done by producing a summary table. Rarely is it informative to produce individual forest plots for each sensitivity analysis undertaken.

Sensitivity analyses are sometimes confused with subgroup analysis. Although some sensitivity analyses involve restricting the analysis to a subset of the totality of studies, the two methods differ in two ways. First, sensitivity analyses do not attempt to estimate the effect of the intervention in the group of studies removed from the analysis, whereas in subgroup analyses, estimates are produced for each subgroup. Second, in sensitivity analyses, informal comparisons are made between different ways of estimating the same thing, whereas in subgroup analyses, formal statistical comparisons are made across the subgroups.

9.8 Chapter information

Editors: Jonathan J Deeks, Julian PT Higgins and Douglas G Altman on behalf of the Cochrane Statistical Methods Group.

This chapter should be cited as: Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

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 and Anne Whitehead.

Acknowledgements: We are grateful to the following for commenting helpfully on earlier drafts: Bodil Als-Nielsen, Doug Altman, Deborah Ashby, Jesse Berlin, Joseph Beyene, Jacqueline Birks, Michael Bracken, Marion Campbell, Chris Cates, Wendong Chen, Mike Clarke, Albert Cobos, Esther Coren, Francois Curtin, Roberto D'Amico, Keith Dear, Jon Deeks, Heather Dickinson, Diana Elbourne, Simon Gates, Paul Glasziou, Christian Gluud, Peter Herbison, Julian Higgins, Sally Hollis, David Jones, Steff Lewis, Philippa Middleton, Nathan Pace, Craig Ramsey, Keith O'Rourke, Rob Scholten, Guido Schwarzer, Jack Sinclair, Jonathan Sterne, Simon Thompson, Andy Vail, Clarine van Oel, Paula Williamson and Fred Wolf.

Box 9.8.a: The Cochrane Statistical Methods Group

Statistical issues are a core aspect of much of the work of the Cochrane Collaboration. The Statistical Methods Group (SMG) is a forum where all statistical issues related to the work of The Cochrane Collaboration are discussed. It has a broad scope, covering issues relating to statistical methods, training, software and research. It also attempts to ensure that adequate statistical and technical support is available to review groups.

The SMG dates back to 1993. Membership of the SMG is currently through membership of the group's email discussion list. The list is used for discussing all issues of importance for the group, whether research, training, software or administration. The group has over 130 members from over around 20 countries. All statisticians working with Cochrane Review Groups (CRGs) are strongly encouraged to join the SMG.

Specifically, the aims of the group are:

1. To develop general policy advice for the Collaboration on all statistical issues relevant to systematic reviews of healthcare interventions.
2. To take responsibility for statistics-orientated chapters of this *Handbook*.
3. To co-ordinate practical statistical support for CRGs.
4. To conduct training workshops and workshops on emerging topics as necessary.
5. To contribute to and review the statistical content of training materials provided within the Collaboration.
6. To develop and validate the statistical software used within the Collaboration.
7. To generate and keep up to date a list of the Statistical Methods Group, detailing their areas of interest and expertise, and maintain an email discussion list as a forum for discussing relevant methodological issues.
8. To maintain a research agenda dictated by issues important to the present and future functioning of the Collaboration, and to encourage research that tackles the agenda.

Web site: www.cochrane-smg.org

9.9 References

Adams 2005

Adams NP, Bestall JB, Malouf R, Lasserson TJ, Jones PW. Beclomethasone versus placebo for chronic asthma. *Cochrane Database of Systematic Reviews* 2005, Issue 1. Art No: CD002738.

Agresti 1996

Agresti A. *An introduction to categorical data analysis*. New York (NY): John Wiley & Sons, 1996.

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.

Altman 1996

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

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 Journal of Current Clinical Trials* 1994; Doc No 134.

Berlin 2002

Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman KA, Anti-Lymphocyte Antibody Induction

Therapy Study Group. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine* 2002; 21: 371-387.

Bradburn 2007

Bradburn MJ, Deeks JJ, Berlin JA, Russell LA. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine* 2007; 26: 53-77.

Chinn 2000

Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine* 2000; 19: 3127-3131.

Cooper 1980

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

Crawford 2007

Crawford F, Hollis S. Topical treatments for fungal infections of the skin and nails of the feet. *Cochrane Database of Systematic Reviews* 2007, Issue 3. Art No: CD001434.

Deeks 1998

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

Deeks 2001

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 (editors). *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd edition). London (UK): 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. *Statistics in Medicine* 2002; 21: 1575-1600.

DerSimonian 1986

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

Egger 1997

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

Engels 2000

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* 2000; 19: 1707-1728.

Greenland 1985

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

Greenland 1987

Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews* 1987; 9: 1-30.

Greenland 1992

Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology* 1992; 135: 1301-1309.

Hasselblad 1995

Hasselblad VIC, McCrory DC. Meta-analytic tools for medical decision making: A practical guide. *Medical Decision Making* 1995; 15: 81-96.

Higgins 2002

Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; 21: 1539-1558.

Higgins 2003

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

Higgins 2004

Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine* 2004; 23: 1663-1682.

Higgins 2008a

Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A* (in press, 2008).

Higgins 2008b

Higgins JPT, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. *Statistics in Medicine* (in press, 2008).

Kjaergard 2001

Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine* 2001; 135: 982-989.

Laupacis 1988

Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* 1988; 318: 1728-1733.

Mantel 1959

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; 22: 719-748.

McIntosh 1996

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

Moher 2005

Moher M, Hey K, Lancaster T. Workplace interventions for smoking cessation. *Cochrane Database of Systematic Reviews* 2005, Issue 2. Art No: CD003440.

Morgenstern 1982

Morgenstern H. Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health* 1982; 72: 1336-1344.

O'Rourke 1989

O'Rourke K, Detsky AS. Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. *Journal of Clinical Epidemiology* 1989; 42: 1021-1026.

Oxman 1992

Oxman AD, Guyatt GH. A consumers guide to subgroup analyses. *Annals of Internal Medicine* 1992; 116: 78-84.

Pittler 2003

Pittler MH, Ernst E. Kava extract versus placebo for treating anxiety. *Cochrane Database of Systematic Reviews* 2003, Issue 1. Art No: CD003383.

Poole 1999

Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology* 1999; 150: 469-475.

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 (UK): Churchill Livingstone, 1997.

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 (editors). *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd edition). London (UK): BMJ Publication Group, 2000.

Sinclair 1994

Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* 1994; 47: 881-889.

Thompson 1997

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

Thompson 1999

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

Thompson 2002

Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 2002; 21: 1559-1574.

Whitehead 1991

Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. *Statistics in Medicine* 1991; 10: 1665-1677.

Whitehead 1994

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

Yusuf 1985

Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Progress in Cardiovascular Diseases* 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.