

Chapter 16: Special topics in statistics

Editors: Julian PT Higgins, Jonathan J Deeks 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 16.10. 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

- When missing data prevent a study from being included in a meta-analysis (and attempts to obtain the data from the original investigators have been unsuccessful), any strategies for imputing them should be described and assessed in sensitivity analyses.
- Non-standard designs, such as cluster-randomized trials and cross-over trials, should be analysed using methods appropriate to the design. Even if study authors fail to account for correlations among outcome data, approximate methods can often be applied by review authors.
- To include a study with more than two intervention groups in a meta-analysis, the recommended approach is usually to combine relevant groups to create a single pair-wise comparison.
- Indirect comparisons of interventions may be misleading, but methods are available that exploit randomization, including extensions into ‘multiple-treatments meta-analysis’.
- To reduce misleading conclusions resulting from multiple statistical analyses, review authors should state in the protocol which analyses they will perform, keep the number of these to a minimum, and interpret statistically significant findings in the context of how many analyses were undertaken.
- Bayesian approaches and hierarchical (or multilevel) models allow more complex meta-analyses to be performed, and can offer some technical and interpretative advantages over the standard methods implemented in RevMan.
- Studies with no events contribute no information about the risk ratio or odds ratio. For rare events, the Peto method has been observed to be less biased and more powerful than other methods.

16.1 Missing data

16.1.1 Types of missing data

There are many potential sources of missing data in a systematic review or meta-analysis (see [Table 16.1.a](#)). For example, a whole study may be missing from the review, an outcome may be missing from a study, summary data may be missing for an outcome, and individual participants may be missing from the summary data. Here we discuss a variety of potential sources of missing data, highlighting where more detailed discussions are available elsewhere in the *Handbook*.

Whole **studies** may be missing from a review because they are never published, are published in obscure places, are rarely cited, or are inappropriately indexed in databases. Thus review authors should always be aware of the possibility that they have failed to identify relevant studies. There is a strong possibility that such studies are missing because of their ‘uninteresting’ or ‘unwelcome’ findings (that is, in the presence of publication bias). This problem is discussed at length in Chapter 10. Details of comprehensive search methods are provided in Chapter 6.

Some studies might not report any information on **outcomes** of interest to the review. For example, there may be no information on quality of life, or on serious adverse effects. It is often difficult to determine whether this is because the outcome was not measured or because the outcome was not reported. Furthermore, failure to report that outcomes were measured may be dependent on the unreported results (selective outcome reporting bias; see Chapter 8, Section 8.13). Similarly, **summary data** for an outcome, in a form that can be included in a meta-analysis, may be missing. A common example is missing standard deviations for continuous outcomes. This is often a problem when change-from-baseline outcomes are sought. We discuss imputation of missing standard deviations in Section 16.1.3. Other examples of missing summary data are missing sample sizes (particularly those for each intervention group separately), numbers of events, standard errors, follow-up times for calculating rates, and sufficient details of time-to-event outcomes. Inappropriate analyses of studies, for example of cluster-randomized and cross-over trials, can lead to missing summary data. It is sometimes possible to approximate the correct analyses of such studies, for example by imputing correlation coefficients or standard deviations, as discussed in Section 16.3 for cluster-randomized studies and Section 16.4 for cross-over trials. As a general rule, most methodologists believe that missing summary data (e.g. “no usable data”) should not be used as a reason to exclude a study from a systematic review. It is more appropriate to include the study in the review, and to discuss the potential implications of its absence from a meta-analysis.

It is likely that in some, if not all, included studies, there will be **individuals** missing from the reported results. Analyses of randomized trials that do not include all randomized participants are not intention-to-treat (ITT) analyses. It is sometimes possible to perform ITT analyses, even if the original investigators did not. We provide a detailed discussion of ITT issues in Section 16.2.

Missing data can also affect subgroup analyses. If subgroup analyses or meta-regressions are planned (see Chapter 9, Section 9.6), they require details of the **study-level characteristics** that distinguish studies from one another. If these are not available for all studies, review authors should consider asking the study authors for more information.

Table 16.1.a: Types of missing data in a meta-analysis

Type of missing data	Some possible reasons for missing data
Missing studies.	Publication bias;

	Search not sufficiently comprehensive.
Missing outcomes.	Outcome not measured; Selective reporting bias.
Missing summary data.	Selective reporting bias; Incomplete reporting.
Missing individuals.	Lack of intention-to-treat analysis; Attrition from the study; Selective reporting bias.
Missing study-level characteristics (for subgroup analysis or meta-regression).	Characteristic not measured; Incomplete reporting.

16.1.2 General principles for dealing with missing data

There is a large literature of statistical methods for dealing with missing data. Here we briefly review some key concepts and make some general recommendations for Cochrane review authors. It is important to think *why* data may be missing. Statisticians often use the terms ‘missing at random’ and ‘not missing at random’ to represent different scenarios.

Data are said to be ‘missing at random’ if the fact that they are missing is unrelated to actual values of the missing data. For instance, if some quality-of-life questionnaires were lost in the postal system, this would be unlikely to be related to the quality of life of the trial participants who completed the forms. In some circumstances, statisticians distinguish between data ‘missing at random’ and data ‘missing completely at random’, although in the context of a systematic review the distinction is unlikely to be important. Data that are missing at random may not be important. Analyses based on the available data will tend to be unbiased, although based on a smaller sample size than the original data set.

Data are said to be ‘not missing at random’ if the fact that they are missing is related to the actual missing data. For instance, in a depression trial, participants who had a relapse of depression might be less likely to attend the final follow-up interview, and more likely to have missing outcome data. Such data are ‘non-ignorable’ in the sense that an analysis of the available data alone will typically be biased. Publication bias and selective reporting bias lead by definition to data that are ‘not missing at random’, and attrition and exclusions of individuals within studies often do as well.

The principal options for dealing with missing data are:

1. analysing only the available data (i.e. ignoring the missing data);
2. imputing the missing data with replacement values, and treating these as if they were observed (e.g. last observation carried forward, imputing an assumed outcome such as assuming all were poor outcomes, imputing the mean, imputing based on predicted values from a regression analysis);
3. imputing the missing data and accounting for the fact that these were imputed with uncertainty (e.g. multiple imputation, simple imputation methods (as point 2) with adjustment to the standard error); and
4. using statistical models to allow for missing data, making assumptions about their relationships with the available data.

Option 1 may be appropriate when data can be assumed to be missing at random. Options 2 to 4 are attempts to address data not missing at random. Option 2 is practical in most circumstances and very commonly used in systematic reviews. However, it fails to acknowledge uncertainty in the imputed values and results, typically, in confidence intervals that are too narrow. Options 3 and 4 would require involvement of a knowledgeable statistician.

Four general recommendations for dealing with missing data in Cochrane reviews are as follows.

- Whenever possible, contact the original investigators to request missing data.
- Make explicit the assumptions of any methods used to cope with missing data: for example, that the data are assumed missing at random, or that missing values were assumed to have a particular value such as a poor outcome.
- Perform sensitivity analyses to assess how sensitive results are to reasonable changes in the assumptions that are made (see Chapter 9, Section 9.7).
- Address the potential impact of missing data on the findings of the review in the Discussion section.

16.1.3 Missing standard deviations

16.1.3.1 Imputing standard deviations

Missing standard deviations are a common feature of meta-analyses of continuous outcome data. One approach to this problem is to impute standard deviations. Before imputing missing standard deviations however, authors should look carefully for statistics that allow calculation or estimation of the standard deviation (e.g. confidence intervals, standard errors, t values, P values, F values), as discussed in Chapter 7 (Section 7.7.3).

The simplest imputation is of a particular value borrowed from one or more other studies. Furukawa et al. found that imputing standard deviations either from other studies in the same meta-analysis, or from studies in another meta-analysis, yielded approximately correct results in two case studies (Furukawa 2006). If several candidate standard deviations are available, review authors would have to decide whether to use their average, the highest, a 'reasonably high' value, or some other strategy. For meta-analyses of mean differences, choosing a higher standard deviation down-weights a study and yields a wider confidence interval. However, for standardized mean difference meta-analyses, choice of an overly large standard deviation will bias the result towards a lack of effect. More complicated alternatives are available for making use of multiple candidate standard deviations. For example, Marinho et al. implemented a linear regression of $\log(\text{standard deviation})$ on $\log(\text{mean})$, because of a strong linear relationship between the two (Marinho 2003).

All imputation techniques involve making assumptions about unknown statistics, and it is best to avoid using them wherever possible. If the majority of studies in a meta-analysis have missing standard deviations, these values should not be imputed. However, imputation may be reasonable for a small proportion of studies comprising a small proportion of the data if it enables them to be combined with other studies for which full data are available. Sensitivity analyses should be used to assess the impact of changing the assumptions made.

16.1.3.2 Imputing standard deviations for changes from baseline

A special case of missing standard deviations is for changes from baseline. Often, only the following information is available:

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

Note that the mean change in each group can always be obtained by subtracting the final mean from the baseline mean even if it is not presented explicitly. However, the information in this table does *not* allow us to calculate the standard deviation of the changes. We cannot know whether the changes were very consistent or very variable. Some other information in a paper may help us determine the standard deviation of the changes. If statistical analyses comparing the changes themselves are presented (e.g. confidence intervals, standard errors, t values, P values, F values) then the techniques described in Chapter 7 (Section 7.7.3) may be used.

When there is not enough information available to calculate the standard deviations for the changes, they can be imputed. When change-from-baseline standard deviations for the same outcome measure are available from other studies in the review, it may be reasonable to use these in place of the missing standard deviations. However, the appropriateness of using a standard deviation from another study relies on whether the studies used the same measurement scale, had the same degree of measurement error and had the same time periods (between baseline and final value measurement).

The following alternative technique may be used for imputing missing standard deviations for changes from baseline (Follmann 1992, Abrams 2005). A typically unreported number known as the correlation coefficient describes how similar the baseline and final measurements were across participants. Here we describe (1) how to calculate the correlation coefficient from a study that is reported in considerable detail and (2) how to impute a change-from-baseline standard deviation in another study, making use of an imputed correlation coefficient. Note that the methods in (2) are applicable both to correlation coefficients obtained using (1) and to correlation coefficients obtained in other ways (for example, by reasoned argument). These methods should be used sparingly, because one can never be sure that an imputed correlation is appropriate (correlations between baseline and final values will, for example, decrease with increasing time between baseline and final measurements, as well as depending on the outcomes and characteristics of the participants). An alternative to these methods is simply to use a comparison of final measurements, which in a randomized trial in theory estimates the same quantity as the comparison of changes from baseline.

(1) Calculating a correlation coefficient from a study reported in considerable detail

Suppose a study is available that presents means and standard deviations for change as well as for baseline and final measurements, for example:

	Baseline	Final	Change
Experimental intervention (sample size 129)	mean=15.2 SD=6.4	mean=16.2 SD=7.1	mean=1.0 SD=4.5
Control intervention (sample size 135)	mean=15.7 SD=7.0	mean=17.2 SD=6.9	mean=1.5 SD=4.2

An analysis of change from baseline is available from this study, using only the data in the final column. However, we can use the other data from the study to calculate two correlation coefficients, one for each intervention group. Let us use the following notation:

	Baseline	Final	Change
Experimental intervention (sample size N_E)	$M_{E,baseline}, SD_{E,baseline}$	$M_{E,final}, SD_{E,final}$	$M_{E,change}, SD_{E,change}$
Control intervention (sample size N_C)	$M_{C,baseline}, SD_{C,baseline}$	$M_{C,final}, SD_{C,final}$	$M_{C,change}, SD_{C,change}$

The correlation coefficient in the experimental group, $Corr_E$, can be calculated as:

$$Corr_E = \frac{SD_{E,baseline}^2 + SD_{E,final}^2 - SD_{E,change}^2}{2 \times SD_{E,baseline} \times SD_{E,final}};$$

and similarly for the control intervention, to obtain $Corr_C$. In the example, these turn out to be

$$Corr_E = \frac{6.4^2 + 7.1^2 - 4.5^2}{2 \times 6.4 \times 7.1} = 0.78,$$

$$Corr_C = \frac{7.0^2 + 6.9^2 - 4.2^2}{2 \times 7.0 \times 6.9} = 0.82.$$

Where either the baseline or final standard deviation is unavailable, then it may be substituted by the other, providing it is reasonable to assume that the intervention does not alter the variability of the outcome measure. Correlation coefficients lie between -1 and 1 . If a value less than 0.5 is obtained, then there is no value in using change from baseline and an analysis of final values will be more precise. Assuming the correlation coefficients from the two intervention groups are similar, a simple average will provide a reasonable measure of the similarity of baseline and final measurements across all individuals in the study (the average of 0.78 and 0.82 for the example is 0.80). If the correlation coefficients differ, then either the sample sizes are too small for reliable estimation, the intervention is affecting the variability in outcome measures, or the intervention effect depends on baseline level, and the use of imputation is best avoided. Before imputation is undertaken it is recommended that correlation coefficients are computed for many (if not all) studies in the meta-analysis and it is noted whether or not they are consistent. Imputation should be done only as a very tentative analysis if correlations are inconsistent.

(2) Imputing a change-from-baseline standard deviation using a correlation coefficient

Now consider a study for which the standard deviation of changes from baseline is missing. When baseline and final standard deviations are known, we can impute the missing standard deviation using an imputed value, $Corr$, for the correlation coefficient. The value $Corr$ might be imputed from another study in the meta-analysis (using the method in (1) above), it might be imputed from elsewhere, or it might be hypothesized based on reasoned argument. In all of these situations, a sensitivity analysis should be undertaken, trying different values of $Corr$, to determine whether the overall result of the analysis is robust to the use of imputed correlation coefficients.

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

$$SD_{E,change} = \sqrt{SD_{E,baseline}^2 + SD_{E,final}^2 - (2 \times Corr \times SD_{E,baseline} \times SD_{E,final})},$$

and similarly for the control intervention. Again, if either of the standard deviations (at baseline and final) are unavailable, then one may be substituted by the other if it is reasonable to assume that the intervention does not alter the variability of the outcome measure.

As an example, given the following data:

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

and using an imputed correlation coefficient of 0.80, we can impute the change-from-baseline standard deviation in the control group as:

$$SD_{C,\text{change}} = \sqrt{4.0^2 + 4.4^2 - (2 \times 0.80 \times 4.0 \times 4.4)} = 2.68 .$$

16.2 Intention-to-treat issues

16.2.1 Introduction

Often some participants are excluded from analyses of randomized trials, either because they were lost to follow-up and no outcome was obtained, or because there was some deviation from the protocol, such as receiving the wrong (or no) treatment, lack of compliance, or ineligibility. Alternatively, it may be impossible to measure certain outcomes for all participants because their availability depends on another outcome (see Section 16.2.4). As discussed in detail in Chapter 8 (Section 8.12), an estimated intervention effect may be biased if some randomized participants are excluded from the analysis. Intention-to-treat (ITT) analysis aims to include all participants randomized into a trial irrespective of what happened subsequently (Newell 1992, Lewis 1993). ITT analyses are generally preferred as they are unbiased, and also because they address a more pragmatic and clinically relevant question.

The following principles of ITT analyses are described in Chapter 8 (Section 8.12).

1. Keep participants in the intervention groups to which they were randomized, regardless of the intervention they actually received.
2. Measure outcome data on all participants.
3. Include all randomized participants in the analysis.

There is no clear consensus on whether all criteria should be applied (Hollis 1999). While the first is widely agreed, the second is often impossible and the third is contentious, since to include participants whose outcomes are unknown (mainly through loss to follow-up) involves imputing ('filling-in') the missing data (see Section 16.1.2).

An analysis in which data are analysed for every participant for whom the outcome was obtained is often described as an **available case analysis**. Some trial reports present analyses of the results of only those participants who completed the trial *and* who complied with (or received some of) their allocated intervention. Some authors incorrectly call this an ITT analysis, but it is in fact a **per-**

protocol analysis. Furthermore, some authors analyse participants only according to the actual interventions received, irrespective of the randomized allocations (**treatment-received analysis**). It is generally unwise to accept study authors' description of an analysis as ITT; such a judgement should be based on the detailed information provided.

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

In most situations, authors should attempt to extract from papers the data to enable at least an **available case analysis**. Avoidable exclusions should be 're-included' if possible. In some rare situations it is possible to create a genuine ITT analysis from information presented in the text and tables of the paper, or by obtaining extra information from the author about participants who were followed up but excluded from the trial report. If this is possible without imputing study results, it should be done.

Otherwise, it may appear that an intention-to-treat analysis can be produced by using imputation. This involves making assumptions about the outcomes of participants for whom no outcome was recorded. However, many imputation analyses differ from available case analyses only in having an unwarranted inflation in apparent precision. Assessing the results of studies in the presence of more than minimal amounts of missing data is ultimately a matter of judgement, as discussed in Chapter 8 (Section 8.12). Statistical analysis cannot reliably compensate for missing data (Unnebrink 2001). No assumption is likely adequately to reflect the truth, and the impact of any assumption should be assessed by trying more than one method as a sensitivity analysis (see Chapter 9, Section 9.7).

In the next two sections we consider some ways to take account of missing observations for dichotomous or continuous outcomes. Although imputation is possible, at present a sensible decision in most cases is to include data for only those participants whose results are known, and address the potential impact of the missing data in the assessment of risk of bias (Chapter 8, Section 8.12). Where imputation is used the methods and assumptions for imputing data for drop-outs should be described in the Methods section of the protocol and review.

If individual participant data are available, then detailed sensitivity analyses can be considered. Review authors in this position are referred to the extensive literature on dealing with missing data in clinical trials (Little 2004). Participants excluded from analyses in published reports should typically be re-included when possible, as is the case when individual participant data are available (Stewart 1995). Information should be requested from the trial authors when sufficient details are not available in published reports to re-include exclude participants in analyses.

16.2.2 Intention-to-treat issues for dichotomous data

Proportions of participants for whom no outcome data were obtained should always be collected and reported in a 'Risk of bias' table; note that the proportions may vary by outcome and by randomized group. However, there is no consensus on the best way to handle these participants in an analysis. There are two basic options, and a plausible option should be used both as a main analysis and as a basis for sensitivity analysis (see below and Chapter 9, Section 9.7).

- Available case analysis: Include data on only those whose results are known, using as a denominator the total number of people who had data recorded for the particular outcome in question. Variation in the degree of missing data across studies may be considered as a potential source of heterogeneity.

- ITT analysis using imputation: Base an analysis on the total number of randomized participants, irrespective of how the original study authors analysed the data. This will involve imputing outcomes for the missing participants. There are several approaches to imputing dichotomous outcome data. One common approach is to assume either that all missing participants experienced the event, or that all missing participants did not experience the event. An alternative approach is to impute data according to the event rate observed in the control group, or according to event rates among completers in the separate groups (the latter provides the same estimate of intervention effect but results in unwarranted inflation of the precision of effect estimates). The choice among these assumptions should be based on clinical judgement. Studies with imputed data may be given more weight than they warrant if entered as dichotomous data into RevMan. It is possible to determine more appropriate weights (Higgins 2008); consultation with a statistician is recommended. However, none of these assumptions is likely to reflect the truth, except for imputing ‘failures’ in some settings such as smoking cessation trials, so an imputation approach is generally not recommended.

The potential impact of the missing data on the results should be considered in the interpretation of the results of the review. This will depend on the degree of ‘missingness’, the frequency of the events and the size of the pooled effect estimate. Gamble and Hollis suggest a sensitivity analysis for dichotomous outcomes based on consideration of ‘best-case’ and ‘worst-case’ scenarios (Gamble 2005). The ‘best-case’ scenario is that all participants with missing outcomes in the experimental intervention group had good outcomes, and all those with missing outcomes in the control intervention group had poor outcomes; the ‘worst-case’ scenario is the converse. The sensitivity analysis down-weights studies in which the discrepancy between ‘best-case’ and ‘worst-case’ scenarios is high, although the down-weighting may be too extreme.

A more plausible sensitivity analysis explicitly considers what the event rates might have been in the missing data. For example, suppose an available case analysis has been used, and a particular study has 20% risk in the intervention arm and 15% risk in the control arm. An available case analysis implicitly assumes that the same fractions apply in the missing data, so three suitable sensitivity analyses to compare with this analysis might consider the risk in the missing data to be 15% in both arms, or 15% and 10% in the experimental and control arms respectively, or 20% and 10% respectively. Alternatively, suppose that in the main analysis, all missing values have been imputed as events. A sensitivity analysis to compare with this analysis could consider the case that, say, 10% of missing participants experienced the event, or 10% in the intervention arm and 5% in the control arm. Graphical approaches to sensitivity analysis have been considered (Hollis 2002).

Higgins et al. suggest an alternative approach that can incorporate specific reasons for missing data, which considers plausible event risks among missing participants in relation to risks among those observed (Higgins 2008). Bayesian approaches, which automatically down-weight studies with more missing data, are considered by White et al. (White 2008a, White 2008b).

16.2.3 Intention-to-treat issues for continuous data

In full ITT analyses, all participants who did not receive the assigned intervention according to the protocol as well as those who were lost to follow-up are included in the analysis. Inclusion of these in an analysis requires that means and standard deviations of the outcome for all randomized participants are available. As for dichotomous data, dropout rates should always be collected and reported in a ‘Risk of bias’ table. Again, there are two basic options, and in either case a sensitivity analysis should be performed (see Chapter 9, Section 9.7).

- Available case analysis: Include data only on those whose results are known. The potential impact of the missing data on the results should be considered in the interpretation of the results of the review. This will depend on the degree of ‘missingness’, the pooled estimate of the treatment

effect and the variability of the outcomes. Variation in the degree of missing data may also be considered as a potential source of heterogeneity.

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

A simple way to conduct a sensitivity analysis for continuous data is to assume a fixed difference between the actual mean for the missing data and the mean assumed by the analysis. For example, after an analysis of available cases, one could consider how the results would have differed if the missing data in the intervention arm had averaged 2 units *greater* than the observed data in the intervention arm, and the missing data in the control arm had averaged 2 units *less* than the observed data in the control arm. A Bayesian approach, which automatically down-weights studies with more missing data, has been considered (White 2007).

16.2.4 Conditional outcomes only available for subsets of participants

Some study outcomes may only be applicable to a proportion of participants. For example, in subfertility trials the proportion of clinical pregnancies that miscarry following treatment is often reported. By definition this outcome excludes participants who do not achieve an interim state (clinical pregnancy), so the comparison is not of all participants randomized. As a general rule it is better to re-define such outcomes so that the analysis includes all randomized participants. In this example, the outcome could be whether the woman has a ‘successful pregnancy’ (becoming pregnant and reaching, say, 24 weeks or term). Another example is provided by a morbidity outcome measured in the medium or long term (e.g. development of chronic lung disease), when there is a distinct possibility of a death preventing assessment of the morbidity. A convenient way to deal with such situations is to combine the outcomes, for example as ‘death or chronic lung disease’.

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

16.3 Cluster-randomized trials

16.3.1 Introduction

In **cluster-randomized trials**, groups of individuals rather than individuals are randomized to different interventions. Cluster-randomized trials are also known as group-randomized trials. We say the ‘unit of allocation’ is the cluster, or the group. The groups may be, for example, schools, villages, medical practices or families. Such trials may be done for one of several reasons. It may be to evaluate the group effect of an intervention, for example herd-immunity of a vaccine. It may be to avoid ‘contamination’ across interventions when trial participants are managed within the same setting, for

example in a trial evaluating a dietary intervention, families rather than individuals may be randomized. A cluster-randomized design may be used simply for convenience.

One of the main consequences of a cluster design is that participants within any one cluster often tend to respond in a similar manner, and thus their data can no longer be assumed to be independent of one another. Many of these studies, however, are incorrectly analysed as though the unit of allocation had been the individual participants. This is often referred to as a 'unit-of-analysis error' (Whiting-O'Keefe 1984) because the unit of analysis is different from the unit of allocation. If the clustering is ignored and cluster trials are analysed as if individuals had been randomized, resulting P values will be artificially small. This can result in false positive conclusions that the intervention had an effect. In the context of a meta-analysis, studies in which clustering has been ignored will have overly narrow confidence intervals and will receive more weight than is appropriate in a meta-analysis. This situation can also arise if participants are allocated to interventions that are then applied to parts of them (for example, to both eyes or to several teeth), or if repeated observations are made on a participant. If the analysis is by the individual units (for example, each tooth or each observation) without taking into account that the data are clustered within participants, then a unit-of-analysis error can occur.

There are several useful sources of information on cluster-randomized trials (Murray 1995, Donner 2000). A detailed discussion of incorporating cluster-randomized trials in a meta-analysis is available (Donner 2002), as is a more technical treatment of the problem (Donner 2001). Special considerations for analysis of standardized mean differences from cluster-randomized trials are discussed by White and Thomas (White 2005).

16.3.2 Assessing risk of bias in cluster-randomized trials

In cluster-randomized trials, particular biases to consider include: (i) recruitment bias; (ii) baseline imbalance; (iii) loss of clusters; (iv) incorrect analysis; and (v) comparability with individually randomized trials.

(i) Recruitment bias can occur when individuals are recruited to the trial after the clusters have been randomized, as the knowledge of whether each cluster is an 'intervention' or 'control' cluster could affect the types of participants recruited. Farrin et al. showed differential participant recruitment in a trial of low back pain randomized by primary care practice; a greater number of less severe participants were recruited to the 'active management' practices (Farrin 2005). Puffer et al. reviewed 36 cluster-randomized trials, and found possible recruitment bias in 14 (39%) (Puffer 2003).

(ii) Cluster-randomized trials often randomize all clusters at once, so lack of concealment of an allocation sequence should not usually be an issue. However, because small numbers of clusters are randomized, there is a possibility of chance baseline imbalance between the randomized groups, in terms of either the clusters or the individuals. Although not a form of bias as such, the risk of baseline differences can be reduced by using stratified or pair-matched randomization of clusters. Reporting of the baseline comparability of clusters, or statistical adjustment for baseline characteristics, can help reduce concern about the effects of baseline imbalance.

(iii) Occasionally complete clusters are lost from a trial, and have to be omitted from the analysis. Just as for missing outcome data in individually randomized trials, this may lead to bias. In addition, missing outcomes for individuals within clusters may also lead to a risk of bias in cluster-randomized trials.

(iv) Many cluster-randomized trials are analysed by incorrect statistical methods, not taking the clustering into account. For example, Eldridge et al. reviewed 152 cluster-randomized trials in primary care of which 41% did not account for clustering in their analyses (Eldridge 2004). Such analyses create a 'unit of analysis error' and produce over-precise results (the standard error of the estimated intervention effect is too small) and P values that are too small. They do not lead to biased estimates of effect. However, if they remain uncorrected, they will receive too much weight in a meta-analysis. Approximate methods of correcting trial results that do not allow for clustering are suggested in Section 16.3.6. Some of these can be implemented by review authors.

(v) In a meta-analysis including both cluster and individually randomized trials, or including cluster-randomized trials with different types of clusters, possible differences between the intervention effects being estimated need to be considered. For example, in a vaccine trial of infectious diseases, a vaccine applied to all individuals in a community would be expected to be more effective than if the vaccine was applied to only half of the people. Another example is provided by Hahn et al., who discussed a Cochrane review of hip protectors (Hahn 2005). The cluster trials showed large positive effect whereas individually randomized trials did not show any clear benefit. One possibility is that there was a 'herd effect' in the cluster-randomized trials (which were often performed in nursing homes, where compliance with using the protectors may have been enhanced). In general, such 'contamination' would lead to underestimates of effect. Thus, if an intervention effect is still demonstrated despite contamination in those trials that were not cluster-randomized, a confident conclusion about the presence of an effect can be drawn. However, the size of the effect is likely to be underestimated. Contamination and 'herd effects' may be different for different types of cluster.

16.3.3 Methods of analysis for cluster-randomized trials

One way to avoid unit-of-analysis errors in cluster-randomized trials is to conduct the analysis at the same level as the allocation, using a summary measurement from each cluster. Then the sample size is the number of clusters and analysis proceeds as if the trial was individually randomized (though the clusters become the individuals). However, this might considerably, and unnecessarily, reduce the power of the study, depending on the number and size of the clusters.

Alternatively, statistical methods now exist that allow analysis at the level of the individual while accounting for the clustering in the data. The ideal information to extract from a cluster-randomized trial is a direct estimate of the required effect measure (for example, an odds ratio with its confidence interval) from an analysis that properly accounts for the cluster design. Such an analysis might be based on a 'multilevel model', a 'variance components analysis' or may use 'generalized estimating equations (GEEs)', among other techniques. Statistical advice is recommended to determine whether the method used is appropriate. Effect estimates and their standard errors from correct analyses of cluster-randomized trials may be meta-analysed using the generic inverse-variance method in RevMan.

16.3.4 Approximate analyses of cluster-randomized trials for a meta-analysis: effective sample sizes

Unfortunately, many cluster-randomized trials have in the past failed to report appropriate analyses. They are commonly analysed as if the randomization was performed on the individuals rather than the clusters. If this is the situation, approximately correct analyses may be performed if the following information can be extracted:

- the number of clusters (or groups) randomized to each intervention group; or the average (mean) size of each cluster;
- the outcome data ignoring the cluster design for the total number of individuals (for example, number or proportion of individuals with events, or means and standard deviations);

- an estimate of the intracluster (or intraclass) correlation coefficient (ICC).

The ICC is an estimate of the relative variability within and between clusters (Donner 1980). It describes the ‘similarity’ of individuals within the same cluster. In fact this is seldom available in published reports. A common approach is to use external estimates obtained from similar studies, and several resources are available that provide examples of ICCs (Ukoununne 1999, Campbell 2000, Health Services Research Unit 2004). ICCs may appear small compared with other types of correlations: values lower than 0.05 are typical. However, even small values can have a substantial impact on confidence interval widths (and hence weights in a meta-analysis), particularly if cluster sizes are large. Empirical research has observed that larger cluster sizes are associated with smaller ICCs (Ukoununne 1999).

An approximately correct analysis proceeds as follows. The idea is to reduce the size of each trial to its ‘effective sample size’ (Rao 1992). The effective sample size of a single intervention group in a cluster-randomized trial is its original sample size divided by a quantity called the ‘design effect’. The design effect is

$$1 + (M - 1) \text{ ICC},$$

where M is the average cluster size and ICC is the intracluster correlation coefficient. A common design effect is usually assumed across intervention groups. For dichotomous data both the number of participants and the number experiencing the event should be divided by the same design effect. Since the resulting data must be rounded to whole numbers for entry into RevMan this approach may be unsuitable for small trials. For continuous data only the sample size need be reduced; means and standard deviations should remain unchanged.

16.3.5 Example of incorporating a cluster-randomized trial

As an example, consider a cluster-randomized trial that randomized 10 school classrooms with 295 children into an intervention group and 11 classrooms with 330 children into a control group. The numbers of successes among the children, ignoring the clustering, are

Intervention: 63/295

Control: 84/330.

Imagine an intracluster correlation coefficient of 0.02 has been obtained from a reliable external source. The average cluster size in the trial is $(295+330)/(10+11) = 29.8$. The design effect for the trial as a whole is then $1 + (M - 1) \text{ ICC} = 1 + (29.8 - 1) \times 0.02 = 1.576$. The effective sample size in the intervention group is $295 / 1.576 = 187.2$ and for the control group is $330 / 1.576 = 209.4$.

Applying the design effects also to the numbers of events produces the following results:

Intervention: 40.0/187.2

Control: 53.3/209.4.

Once trials have been reduced to their effective sample size, the data may be entered into RevMan as, for example, dichotomous outcomes or continuous outcomes. Results from the example trial may be entered as

Intervention: 40/187

Control: 53/209.

16.3.6 Approximate analyses of cluster-randomized trials for a meta-analysis: inflating standard errors

A clear disadvantage of the method described in Section 16.3.4 is the need to round the effective sample sizes to whole numbers. A slightly more flexible approach, which is equivalent to calculating

effective sample sizes, is to multiply the standard error of the effect estimate (from an analysis ignoring clustering) by the square root of the design effect. The standard error may be calculated from a confidence interval (see Chapter 7, Section 7.7.7). Standard analyses of dichotomous or continuous outcomes may be used to obtain these confidence intervals using RevMan. The meta-analysis using the inflated variances may be performed using RevMan and the generic inverse-variance method.

As an example, the odds ratio (OR) from a study with the results

Intervention: 63/295

Control: 84/330

is $OR = 0.795$ (95% CI 0.548 to 1.154). Using methods described in Chapter 7 (Section 7.7.7.3), we can determine from these results that the log odds ratio is $\ln OR = -0.23$ with standard error 0.19. Using the same design effect of 1.576 as in Section 16.3.5, an inflated standard error that accounts for clustering is given by $0.19 \times \sqrt{1.576} = 0.24$. The log odds ratio (-0.23) and this inflated standard error (0.24) may be entered into RevMan under a generic inverse-variance outcome.

16.3.7 Issues in the incorporation of cluster-randomized trials

Cluster-randomized trials may, in principle, be combined with individually randomized trials in the same meta-analysis. Consideration should be given to the possibility of important differences in the effects being evaluated between the different types of trial. There are often good reasons for performing cluster-randomized trials and these should be examined. For example, in the treatment of infectious diseases an intervention applied to all individuals in a community may be more effective than treatment applied to select (randomized) individuals within the community since it may reduce the possibility of re-infection.

Authors should always identify any cluster-randomized trials in a review and explicitly state how they have dealt with the data. They should conduct sensitivity analyses to investigate the robustness of their conclusions, especially when ICCs have been borrowed from external sources (see Chapter 9, Section 9.7). Statistical support is recommended.

16.3.8 Individually randomized trials with clustering

Issues related to clustering can also occur in individually randomized trials. This can happen when the same health professional (for example doctor, surgeon, nurse or therapist) delivers the intervention to a number of participants in the intervention group. This type of clustering is discussed by Lee and Thompson, and raises issues similar to those in cluster-randomized trials (Lee 2005a).

16.4 Cross-over trials

16.4.1 Introduction

Parallel group trials allocate each participant to a single intervention for comparison with one or more alternative interventions. In contrast, **cross-over trials** allocate each participant to a sequence of interventions. A simple randomized cross-over design is an 'AB/BA' design in which participants are randomized initially to intervention A or intervention B, and then 'cross over' to intervention B or intervention A, respectively. It can be seen that data from the first period of a cross-over trial represent a parallel group trial, a feature referred to in Section 16.4.5. In keeping with the rest of the *Handbook*, we will use E and C to refer to interventions, rather than A and B.

Cross-over designs offer a number of possible advantages over parallel group trials. Among these are (i) that each participant acts as his or her own control, eliminating among-participant variation; (ii)

that, consequently, fewer participants are required to obtain the same power; and (iii) that every participant receives every intervention, which allows the determination of the best intervention or preference for an individual participant. A readable introduction to cross-over trials is given by Senn (Senn 2002). More detailed discussion of meta-analyses involving cross-over trials is provided by Elbourne et al. (Elbourne 2002), and some empirical evidence on their inclusion in systematic reviews by Lathyris et al. (Lathyris 2007).

16.4.2 Assessing suitability of cross-over trials

Cross-over trials are suitable for evaluating interventions with a temporary effect in the treatment of stable, chronic conditions. They are employed, for example, in the study of interventions to relieve asthma and epilepsy. They are not appropriate when an intervention can have a lasting effect that compromises entry to subsequent periods of the trial, or when a disease has a rapid evolution. The advantages of cross-over trials must be weighed against their disadvantages. The principal problem associated with cross-over trials is that of carry-over (a type of period-by-intervention interaction). Carry-over is the situation in which the effects of an intervention given in one period persist into a subsequent period, thus interfering with the effects of a different subsequent intervention. Many cross-over trials include a period between interventions known as a washout period as a means of reducing carry-over. If a primary outcome is irreversible (for example mortality, or pregnancy in a subfertility study) then a cross-over study is generally considered to be inappropriate. Another problem with cross-over trials is the risk of drop-out due to their longer duration compared with comparable parallel group trials. The analysis techniques for cross-over trials with missing observations are limited. The assessment of the risk of bias in cross-over trials is discussed in Section 16.4.3.

In considering the inclusion of cross-over trials in meta-analysis, authors should first address the question of whether a cross-over trial is a suitable method for the condition and intervention in question. For example, although they are frequently employed in the field, one group of authors decided cross-over trials were inappropriate for studies in Alzheimer's disease due to the degenerative nature of the condition, and included only data from the first period (Qizilbash 1998). The second question to be addressed is whether there is a likelihood of serious carry-over, which relies largely on judgement since the statistical techniques to demonstrate carry-over are far from satisfactory. The nature of the interventions and the length of any washout period are important considerations.

It is only justifiable to exclude cross-over trials from a systematic review if the design is inappropriate to the clinical context. Very often, however, it is difficult or impossible to extract suitable data from a cross-over trial. In Section 16.4.5 we outline some considerations and suggestions for including cross-over trials in a meta-analysis. First we discuss how the 'Risk of bias' tool described in Chapter 8 can be extended to address questions specific to cross-over trials.

16.4.3 Assessing risk of bias in cross-over trials

The main concerns over risk of bias in cross-over trials are: (i) whether the cross-over design is suitable; (ii) whether there is a carry-over effect; (iii) whether only first period data are available; (iv) incorrect analysis; and (v) comparability of results with those from parallel-group trials.

(i) The cross-over design is suitable to study a condition that is (reasonably) stable (e.g. asthma), and where long-term follow-up is not required. The first issue to consider therefore is whether the cross-over design is suitable for the condition being studied.

(ii) Of particular concern is the possibility of a 'carry over' of treatment effect from one period to the next. A carry-over effect means that the observed difference between the treatments depends upon the

order in which they were received; hence the estimated overall treatment effect will be affected (usually underestimated, leading to a bias towards the null).

The use of the cross-over design should thus be restricted to situations in which there is unlikely to be carry-over of treatment effect across periods. Support for this notion may not be available, however, before the trial is done. Review authors should seek information in trial reports about the evaluation of the carry-over effect. However, in an unpublished review of 116 published cross-over trials from 2000 (Mills 2005), 30% of the studies discussed carry-over but only 12% reported the analysis.

(iii) In the presence of carry-over, a common strategy is to base the analysis on only the first period. Although the first period of a cross-over trial is in effect a parallel group comparison, use of data from only the first period will be biased if, as is likely, the decision to do so is based on a test of carry-over. Such a 'two stage analysis' has been discredited (Freeman 1989) but is still used. Also, use of the first period only removes the main strength of the cross-over design, the ability to compare treatments within individuals.

Cross-over trials for which only first period data are available should be considered to be at risk of bias, especially when the investigators explicitly used the two-stage strategy.

(iv) The analysis of a cross-over trial should take advantage of the within-person design, and use some form of paired analysis (Elbourne 2002). Although trial authors may have analysed paired data, poor presentation may make it impossible for review authors to extract paired data. Unpaired data may be available and will generally be unrelated to the estimated treatment effect or statistical significance. So it is not a source of bias, but rather will usually lead to a trial getting (much) less than its due weight in a meta-analysis.

In the review above (Mills 2005), only 38% of 116 cross-over trials performed an analysis of paired data.

(v) In the absence of carry-over, cross-over trials should estimate the same treatment effect as parallel group trials. Although one study reported a difference in the treatment effect found in cross-over trials compared with parallel group trials (Khan 1996), they had looked at treatments for infertility, an area notorious for the inappropriateness of the cross-over design, and a careful re-analysis did not support the original findings (te Velde 1998).

Other issues to consider for risk of bias in cross-over trials include the following.

- Participants may drop out after the first treatment, and not receive the second treatment. Such participants are usually dropped from the analysis.
- There may be a systematic difference between the two periods of the trial. A period effect is not too serious, as it applies equally to both treatments, although it may suggest that the condition being studied is not stable.
- It may not be clear how many treatments or periods were used. Lee could not identify the design for 12/64 published cross-over trials (Lee 2005b).
- It should not be assumed that the order of treatments was randomized in a cross-over trial. Occasionally a study may be encountered in which it is clear that all participants had the treatments in the same order. Such a trial does not provide a valid comparison of the treatments, since there may be a trend in outcomes over time in addition to the change in treatments.

- Reporting of drop-outs may be poor, especially for those participants who completed one treatment period. The number of participants who dropped out was specified in only nine of the 64 trials in Lee's review (Lee 2005b).

Some suggested questions for assessing risk of bias in cross-over trials are as follows:

- Was use of a cross-over design appropriate?
- Is it clear that the order of receiving treatments was randomized?
- Can it be assumed that the trial was not biased from carry-over effects?
- Are unbiased data available?

16.4.4 Methods of analysis for cross-over trials

If neither carry-over nor period effects are thought to be a problem, then an appropriate analysis of continuous data from a two-period, two-intervention cross-over trial is a paired t-test. This evaluates the value of 'measurement on experimental intervention (E)' minus 'measurement on control intervention (C)' separately for each participant. The mean and standard error of these difference measures are the building blocks of an effect estimate and a statistical test. The effect estimate may be included in a meta-analysis using the generic inverse-variance method in RevMan.

A paired analysis is possible if the data in any one of the following bullet points is available:

- individual participant data from the paper or by correspondence with the trialist;
- the mean and standard deviation (or standard error) of the participant-specific differences between experimental intervention (E) and control intervention (C) measurements;
- the mean difference and one of the following: (i) a t-statistic from a paired t-test; (ii) a P value from a paired t-test; (iii) a confidence interval from a paired analysis;
- a graph of measurements on experimental intervention (E) and control intervention (C) from which individual data values can be extracted, as long as matched measurements for each individual can be identified as such.

For details see Elbourne et al. (Elbourne 2002).

If results are available broken by the particular sequence each participant received, then analyses that adjust for period effects are straightforward (e.g. as outlined in Chapter 3 of Senn (Senn 2002)).

16.4.5 Methods for incorporating cross-over trials into a meta-analysis

Unfortunately, the reporting of cross-over trials has been very variable, and the data required to include a paired analysis in a meta-analysis are often not published. A common situation is that means and standard deviations (or standard errors) are available only for measurements on E and C separately. A simple approach to incorporating cross-over trials in a meta-analysis is thus to take all measurements from intervention E periods and all measurements from intervention C periods and analyse these as if the trial were a parallel group trial of E versus C. This approach gives rise to a unit-of-analysis error (see Chapter 9, Section 9.3) and should be avoided unless it can be demonstrated that the results approximate those from a paired analysis, as described in Section 16.4.4. The reason for this is that confidence intervals are likely to be too wide, and the trial will receive too little weight, with the possible consequence of disguising clinically important heterogeneity. Nevertheless, this incorrect analysis is conservative, in that studies are under-weighted rather than over-weighted. While some argue against the inclusion of cross-over trials in this way, the unit-of-analysis error might be regarded as less serious than some other types of unit-of-analysis error.

A second approach to incorporating cross-over trials is to include only data from the first period. This might be appropriate if carry-over is thought to be a problem, or if a cross-over design is considered inappropriate for other reasons. However, it is possible that available data from first periods constitute a biased subset of all first period data. This is because reporting of first period data may be dependent on the trialists having found statistically significant carry-over.

A third approach to incorporating inappropriately reported cross-over trials is to attempt to approximate a paired analysis, by imputing missing standard deviations. We address this approach in detail in Section 16.4.6.

Cross-over trials with dichotomous outcomes require more complicated methods and consultation with a statistician is recommended (Elbourne 2002).

16.4.6 Approximate analyses of cross-over trials for a meta-analysis

Table 16.4.a presents some results that might be available from a report of a cross-over trial, and presents the notation we will use in the subsequent sections. We review straightforward methods for approximating appropriate analyses of cross-over trials to obtain mean differences or standardized mean differences for use in meta-analysis. Review authors should consider whether imputing missing data is preferable to excluding cross-over trials completely from a meta-analysis. The trade-off will depend on the confidence that can be placed on the imputed numbers, and in the robustness of the meta-analysis result to a range of plausible imputed results.

Table 16.4.a: Some possible data available from the report of a cross-over trial

Data relate to	Core statistics	Related, commonly-reported statistics
Intervention E	N, M_E, SD_E	Standard error of M_E .
Intervention C	N, M_C, SD_C	Standard error of M_C .
Difference between E and C	N, MD, SD_{diff}	Standard error of MD; Confidence interval for MD; Paired t-statistic; P value from paired t-test.

16.4.6.1 Mean differences

The point estimate of mean difference for a paired analysis is usually available, since it is the same as for a parallel group analysis (the mean of the differences is equal to the difference in means):

$$MD = M_E - M_C.$$

The standard error of the mean difference is obtained as

$$SE(MD) = \frac{SD_{diff}}{\sqrt{N}}.$$

where N is the number of participants in the trial, and SD_{diff} is the standard deviation of *within-participant differences between E and C measurements*. As indicated in Section 16.4.4, the standard error can also be obtained directly from a confidence interval for MD, from a paired t-statistic, or from

the P value from a paired t-test. The quantities MD and SE(MD) may be entered into RevMan under the generic inverse-variance outcome type.

When the standard error is not available directly and the standard deviation of the differences is not presented, a simple approach is to impute the standard deviation, as is commonly done for other missing standard deviations (see Section 16.1.3). Other studies in the meta-analysis may present standard deviations of differences, and as long as the studies use the same measurement scale, it may be reasonable to borrow these from one study to another. As with all imputations, sensitivity analyses should be undertaken to assess the impact of the imputed data on the findings of the meta-analysis (see Section 16.1 and Chapter 9, Section 9.7).

If no information is available from any study on the standard deviations of the differences, imputation of standard deviations can be achieved by assuming a particular correlation coefficient. The correlation coefficient describes how similar the measurements on interventions E and C are within a participant, and is a number between -1 and 1. It may be expected to lie between 0 and 1 in the context of a cross-over trial, since a higher than average outcome for a participant while on E will tend to be associated with a higher than average outcome while on C. If the correlation coefficient is zero or negative, then there is no statistical benefit of using a cross-over design over using a parallel group design.

A common way of presenting results of a cross-over trial is as if the trial had been a parallel group trial, with standard deviations for each intervention separately (SD_E and SD_C ; see Table 16.4.a). The desired standard deviation of the differences can be estimated using these intervention-specific standard deviations and an imputed correlation coefficient (Corr):

$$SD_{\text{diff}} = \sqrt{SD_E^2 + SD_C^2 - (2 \times \text{Corr} \times SD_E \times SD_C)} .$$

16.4.6.2 Standardized mean difference

The most appropriate standardized mean difference (SMD) from a cross-over trial divides the mean difference by the standard deviation of measurements (and not by the standard deviation of the differences). A SMD can be calculated by pooled intervention-specific standard deviations as follows:

$$SMD = \frac{MD}{SD_{\text{pooled}}} ,$$

where

$$SD_{\text{pooled}} = \sqrt{\frac{SD_E^2 + SD_C^2}{2}} .$$

A correlation coefficient is required for the standard error of the SMD:

$$SE(SMD) = \sqrt{\frac{1}{N} + \frac{SMD^2}{2N}} \times \sqrt{2(1 - \text{Corr})} .$$

Alternatively, the SMD can be calculated from the MD and its standard error, using an imputed correlation:

$$SMD = \frac{MD}{SE(MD) \times \sqrt{\frac{N}{2(1 - \text{Corr})}}}$$

In this case, the imputed correlation impacts on the magnitude of the SMD effect estimate itself (rather than just on the standard error, as is the case for MD analyses in Section 16.4.6.1). Imputed correlations should therefore be used with great caution for estimation of SMDs.

16.4.6.3 Imputing correlation coefficients

The value for a correlation coefficient might be imputed from another study in the meta-analysis (see below), it might be imputed from a source outside of the meta-analysis, or it might be hypothesized based on reasoned argument. In all of these situations, a sensitivity analysis should be undertaken, trying different values of Corr, to determine whether the overall result of the analysis is robust to the use of imputed correlation coefficients.

Estimation of a correlation coefficient is possible from another study in the meta-analysis if that study presents all three standard deviations in Table 16.4.a. The calculation assumes that the mean and standard deviation of measurements for intervention E is the same when it is given in the first period as when it is given in the second period (and similarly for intervention C).

$$\text{Corr} = \frac{SD_E^2 + SD_C^2 - SD_{\text{diff}}^2}{2 \times SD_E \times SD_C}.$$

Before imputation is undertaken it is recommended that correlation coefficients are computed for as many studies as possible and compared. If these correlations vary substantially then sensitivity analyses are particularly important.

16.4.6.4 Example

As an example, suppose a cross-over trial reports the following data:

Intervention E (sample size 10)	$M_E = 7.0,$ $SD_E = 2.38$
Intervention C (sample size 10)	$M_C = 6.5,$ $SD_C = 2.21$

Mean difference, imputing SD of differences (SD_{diff})

The estimate of the mean difference is $MD = 7.0 - 6.5 = 0.5$. Suppose that a typical standard deviation of differences had been observed from other trials to be 2. Then we can estimate the standard error of MD as

$$SE(MD) = \frac{SD_{\text{diff}}}{\sqrt{N}} = \frac{2}{\sqrt{10}} = 0.632.$$

The numbers 0.5 and 0.632 may be entered into RevMan as the estimate and standard error of a mean difference, under a generic inverse-variance outcome.

Mean difference, imputing correlation coefficient (Corr)

The estimate of the mean difference is again $MD = 0.5$. Suppose that a correlation coefficient of 0.68 has been imputed. Then we can impute the standard deviation of the differences as:

$$\begin{aligned} SD_{\text{diff}} &= \sqrt{SD_E^2 + SD_C^2 - (2 \times \text{Corr} \times SD_E \times SD_C)} \\ &= \sqrt{2.38^2 + 2.21^2 - (2 \times 0.68 \times 2.38 \times 2.21)} = 1.8426 \end{aligned}$$

The standard error of MD is then

$$SE(MD) = \frac{SD_{diff}}{\sqrt{N}} = \frac{1.8426}{\sqrt{10}} = 0.583 .$$

The numbers 0.5 and 0.583 may be entered into RevMan as the estimate and standard error of a mean difference, under a generic inverse-variance outcome. Correlation coefficients other than 0.68 should be used as part of a sensitivity analysis.

Standardized mean difference, imputing correlation coefficient (Corr)

The standardized mean difference can be estimated directly from the data:

$$SMD = \frac{MD}{SD_{pooled}} = \frac{MD}{\sqrt{\frac{SD_E^2 + SD_C^2}{2}}} = \frac{0.5}{\sqrt{\frac{2.38^2 + 2.21^2}{2}}} = 0.218 .$$

The standard error is obtained thus:

$$SE(SMD) = \sqrt{\frac{1}{N} + \frac{SMD^2}{2N}} \times \sqrt{2(1 - Corr)} = \sqrt{\frac{1}{10} + \frac{0.218^2}{20}} \times \sqrt{2(1 - 0.68)} = 0.256 .$$

The numbers 0.218 and 0.256 may be entered into RevMan as the estimate and standard error of a standardized mean difference, under a generic inverse-variance outcome.

We could also have obtained the SMD from the MD and its standard error:

$$SMD = \frac{MD}{SE(MD) \times \sqrt{\frac{N}{2(1 - Corr)}}} = \frac{0.5}{0.583 \times \sqrt{\frac{10}{2(1 - 0.68)}}} = 0.217$$

The minor discrepancy arises due to the slightly different ways in which the two formulae calculate a pooled standard deviation for the standardizing.

16.4.7 Issues in the incorporation of cross-over trials

Cross-over trials may, in principle, be combined with parallel group trials in the same meta-analysis. Consideration should be given to the possibility of important differences in other characteristics between the different types of trial. For example, cross-over trials may have shorter intervention periods or may include participants with less severe illness. It is generally advisable to meta-analyse parallel-group and cross-over trials separately irrespective of whether they are also combined together.

Authors should explicitly state how they have dealt with data from cross-over trials and should conduct sensitivity analyses to investigate the robustness of their conclusions, especially when correlation coefficients have been borrowed from external sources (see Chapter 9, Section 9.7). Statistical support is recommended.

16.5 Studies with more than two intervention groups

16.5.1 Introduction

It is not uncommon for clinical trials to randomize participants to one of several intervention groups. A review of randomized trials published in December 2000 found that a quarter had more than two intervention groups (Chan 2005). For example, there may be two or more experimental intervention groups with a common control group, or two control intervention groups such as a placebo group and a standard treatment group. We refer to these studies as ‘multi-arm’ studies. A special case is a

factorial trial, which addresses two or more simultaneous intervention comparisons using four or more intervention groups (see Section 16.5.6).

Although a systematic review may include several intervention comparisons (and hence several meta-analyses), almost all meta-analyses address pair-wise comparisons. There are three separate issues to consider when faced with a study with more than two intervention groups:

1. Determine which intervention groups are relevant to the systematic review.
2. Determine which intervention groups are relevant to a particular meta-analysis.
3. Determine how the study will be included in the meta-analysis if more than two groups are relevant.

16.5.2 Determining which intervention groups are relevant

For a particular multi-arm study, the intervention groups of relevance to a *systematic review* are all those that could be included in a pair-wise comparison of intervention groups that, if investigated alone, would meet the criteria for including studies in the review. For example, a review addressing only a comparison of ‘nicotine replacement therapy versus placebo’ for smoking cessation might identify a study comparing ‘nicotine gum versus behavioural therapy versus placebo gum’. Of the three possible pair-wise comparisons of interventions, only one (‘nicotine gum versus placebo gum’) addresses the review objective, and no comparison involving behavioural therapy does. Thus, the behavioural therapy group is not relevant to the review. However, if the study had compared ‘nicotine gum plus behavioural therapy versus behavioural therapy plus placebo gum versus placebo gum alone’, then a comparison of the first two interventions might be considered relevant and the placebo gum group not.

As an example of multiple control groups, a review addressing the comparison ‘acupuncture versus no acupuncture’ might identify a study comparing ‘acupuncture versus sham acupuncture versus no intervention’. The review authors would ask whether, on the one hand, a study of ‘acupuncture versus sham acupuncture’ would be included in the review and, on the other hand, a study of ‘acupuncture versus no intervention’ would be included. If both of them would, then all three intervention groups of the study are relevant to the review.

As a general rule, and to avoid any confusion for the reader over the identity and nature of each study, it is recommended that all intervention groups of a multi-intervention study be mentioned in the table of ‘Characteristics of included studies’, either in the ‘Interventions’ cell or the ‘Notes’ cell. However, it is necessary to provide detailed descriptions of only the intervention groups relevant to the review, and only these groups should be used in analyses.

The same considerations of relevance apply when determining which intervention groups of a study should be included in a particular *meta-analysis*. Each meta-analysis addresses only a single pair-wise comparison, so review authors should consider whether a study of each possible pair-wise comparison of interventions in the study would be eligible for the meta-analysis. To draw the distinction between the review-level decision and the meta-analysis-level decision consider a review of ‘nicotine therapy versus placebo or other comparators’. All intervention groups of a study of ‘nicotine gum versus behavioural therapy versus placebo gum’ might be relevant to the review. However, the presence of multiple interventions may not pose any problem for meta-analyses, since it is likely that ‘nicotine gum versus placebo gum’, and ‘nicotine gum versus behavioural therapy’ would be addressed in different meta-analyses. Conversely, all groups of the study of ‘acupuncture versus sham acupuncture versus no intervention’ might be considered eligible for the same meta-analysis, if the meta-analysis would include a study of ‘acupuncture versus sham acupuncture’ and a study of ‘acupuncture versus no intervention’. We describe methods for dealing with the latter situation in Section 16.5.4.

16.5.3 Assessing risk of bias in studies with more than two groups

Bias may be introduced in a multiple-intervention study if the decisions regarding data analysis are made after seeing the data. For example, groups receiving different doses of the same intervention may be combined only after seeing the results, including P values. Also, different outcomes may be presented when comparing different pairs of groups, again potentially in relation to the findings.

Juszczak et al. reviewed 60 multiple-intervention randomized trials, of which over a third had at least four intervention arms (Juszczak 2003). They found that only 64% reported the same comparisons of groups for all outcomes, suggesting selective reporting analogous to selective outcome reporting in a two-arm trial. Also, 20% reported combining groups in an analysis. However, if the summary data are provided for each intervention group, it does not matter how the groups had been combined in reported analyses; review authors do not need to analyse the data in the same way as the study authors.

Some suggested questions for assessing risk of bias in multiple-intervention studies are as follows:

- Are data presented for each of the groups to which participants were randomized?
- Are reports of the study free of suggestion of selective reporting of comparisons of intervention arms for some outcomes?

If the answer to the first question is ‘yes’, then the second question is unimportant (so could be answered also with a ‘yes’).

16.5.4 How to include multiple groups from one study

There are several possible approaches to including a study with multiple intervention groups in a particular meta-analysis. One approach that must be avoided is simply to enter several comparisons into the meta-analysis when these have one or more intervention groups in common. This ‘double-counts’ the participants in the ‘shared’ intervention group(s), and creates a unit-of-analysis error due to the unaddressed correlation between the estimated intervention effects from multiple comparisons (see Chapter 9, Section 9.3). An important distinction to make is between situations in which a study can contribute several *independent* comparisons (i.e. with no intervention group in common) and when several comparisons are *correlated* because they have intervention groups, and hence participants, in common. For example, consider a study that randomized participants to four groups: ‘nicotine gum’ versus ‘placebo gum’ versus ‘nicotine patch’ versus ‘placebo patch’. A meta-analysis that addresses the broad question of whether nicotine replacement therapy is effective might include the comparison ‘nicotine gum versus placebo gum’ as well as the independent comparison ‘nicotine patch versus placebo patch’. It is usually reasonable to include independent comparisons in a meta-analysis as if they were from different studies, although there are subtle complications with regard to random-effects analyses (see Section 16.5.5).

Approaches to overcoming a unit-of-analysis error for a study that could contribute multiple, correlated, comparisons include the following.

- Combine groups to create a single pair-wise comparison (recommended).
- Select one pair of interventions and exclude the others.
- Split the ‘shared’ group into two or more groups with smaller sample size, and include two or more (reasonably independent) comparisons.
- Include two or more correlated comparisons and account for the correlation.
- Undertake a *multiple-treatments meta-analysis* (see Section 16.6).

The recommended method in most situations is to combine all relevant experimental intervention groups of the study into a single group, and to combine all relevant control intervention groups into a single control group. As an example, suppose that a meta-analysis of ‘acupuncture versus no acupuncture’ would consider studies of either ‘acupuncture versus sham acupuncture’ or studies of ‘acupuncture versus no intervention’ to be eligible for inclusion. Then a study comparing ‘acupuncture versus sham acupuncture versus no intervention’ would be included in the meta-analysis by combining the participants in the ‘sham acupuncture’ group with participants in the ‘no intervention’ group. This combined control group would be compared with the ‘acupuncture’ group in the usual way. For dichotomous outcomes, both the sample sizes and the numbers of people with events can be summed across groups. For continuous outcomes, means and standard deviations can be combined using methods described in Chapter 7 (Section 7.7.3.8).

The alternative strategy of selecting a single pair of interventions (e.g. choosing either ‘sham acupuncture’ or ‘no intervention’ as the control) results in a loss of information and is open to results-related choices, so is not generally recommended.

A further possibility is to include each pair-wise comparison separately, but with shared intervention groups divided out approximately evenly among the comparisons. For example, if a trial compares 121 patients receiving acupuncture with 124 patients receiving sham acupuncture and 117 patients receiving no acupuncture, then two comparisons (of, say, 61 ‘acupuncture’ against 124 ‘sham acupuncture’, and of 60 ‘acupuncture’ against 117 ‘no intervention’) might be entered into the meta-analysis. For dichotomous outcomes, both the number of events and the total number of patients would be divided up. For continuous outcomes, only the total number of participants would be divided up and the means and standard deviations left unchanged. This method only partially overcomes the unit-of-analysis error (because the resulting comparisons remain correlated) so is not generally recommended. A potential advantage of this approach, however, would be that approximate investigations of heterogeneity across intervention arms are possible (for example, in the case of the example here, the difference between using sham acupuncture and no intervention as a control group).

Two final options, which would require statistical support, are to account for the correlation between correlated comparisons from the same study in the analysis, and to perform a multiple-treatments meta-analysis. The former involves calculating an average (or weighted average) of the relevant pair-wise comparisons from the study, and calculating a variance (and hence a weight) for the study, taking into account the correlation between the comparisons. It will typically yield a similar result to the recommended method of combining across experimental and control intervention groups. Multiple-treatments meta-analysis is discussed in more detail in Section 16.6.

16.5.5 Heterogeneity considerations with multiple-intervention studies

Two possibilities for addressing heterogeneity between studies are to allow for it in a random-effects meta-analysis, and to investigate it through subgroup analyses or meta-regression (Chapter 9, Section 9.6). Some complications arise when including multiple-intervention studies in such analyses. First, it will not be possible to investigate certain intervention-related sources of heterogeneity if intervention groups are combined as in the recommended approach in Section 16.5.4. For example, subgrouping according to ‘sham acupuncture’ or ‘no intervention’ as a control group is not possible if these two groups are combined prior to the meta-analysis. The simplest method for allowing an investigation of this difference, across studies, is to create two or more comparisons from the study (e.g. ‘acupuncture versus sham acupuncture’ and ‘acupuncture versus no intervention’). However, if these contain a common intervention group (here, acupuncture), then they are not independent and a unit-of-analysis error will occur, even if the sample size is reduced for the shared intervention group(s). Nevertheless, splitting up the sample size for the shared intervention group remains a practical means of performing approximate investigations of heterogeneity.

A more subtle problem occurs in random-effects meta-analyses if multiple comparisons are included from the same study. A random-effects meta-analysis allows for variation by assuming that the effects underlying the studies in the meta-analysis follow a distribution across studies. The intention is to allow for study-to-study variation. However, if two or more estimates come from the same study then the same variation is assumed across comparisons within the study and across studies. This is true whether the comparisons are independent or correlated (see Section 16.5.4). One way to overcome this is to perform a fixed-effect meta-analysis across comparisons within a study, and a random-effects meta-analysis across studies. Statistical support is recommended; in practice the difference between different analyses is likely to be trivial.

16.5.6 Factorial trials

In a factorial trial, two (or more) intervention comparisons are carried out simultaneously. Thus, for example, participants may be randomized to receive aspirin or placebo, and also randomized to receive a behavioural intervention or standard care. Most factorial trials have two ‘factors’ in this way, each of which has two levels; these are called 2×2 factorial trials. Occasionally 3×2 trials may be encountered, or trials that investigate three, four, or more interventions simultaneously. Often only one of the comparisons will be of relevance to any particular review. The following remarks focus on the 2×2 case but the principles extend to more complex designs.

In most factorial trials the intention is to achieve ‘two trials for the price of one’, and the assumption is made that the effects of the different active interventions are independent, that is, there is no interaction (synergy). Occasionally a trial may be carried out specifically to investigate whether there is an interaction between two treatments. That aspect may more often be explored in a trial comparing each of two active treatments on its own with both combined, without a placebo group. Such trials are not factorial trials.

The 2×2 factorial design can be displayed as a 2×2 table, with the rows indicating one comparison (e.g. aspirin versus placebo) and the columns the other (e.g. behavioural intervention versus standard care):

		Randomization of B	
		Behavioural intervention (B)	Standard care (not B)
Randomization of A	Aspirin (A)	A and B	A, not B
	Placebo (not A)	B, not A	Not A, not B

A 2×2 factorial trial can be seen as two trials addressing different questions. It is important that both parts of the trial are reported as if they were just a two-arm parallel group trial. Thus we expect to see the results for aspirin versus placebo, including all participants regardless of whether they had behavioural intervention or standard care, and likewise for the behavioural intervention. These results may be seen as relating to the margins of the 2×2 table. We would also wish to evaluate whether there may have been some interaction between the treatments (i.e. effect of A depends on whether B or ‘not B’ was received), for which we need to see the four cells within the table (McAlister 2003). It follows that the practice of publishing two separate reports, possibly in different journals, does not allow the full results to be seen.

McAlister et al. reviewed 44 published reports of factorial trials (McAlister 2003). They found that only 34% reported results for each cell of the factorial structure. However, it will usually be possible to derive the marginal results from the results for the four cells in the 2×2 structure. In the same review, 59% of the trial reports included the results of a test of interaction. On re-analysis, 2/44 trials (6%) had $P < 0.05$, which is close to expectation by chance (McAlister 2003). Thus, despite concerns about unrecognized interactions, it seems that investigators are appropriately restricting the use of the factorial design to those situations in which two (or more) treatments do not have the potential for substantive interaction. Unfortunately, many review authors do not take advantage of this fact and include only half of the available data in their meta-analysis (e.g. including only A versus not A among those that were *not* receiving B, and excluding the valid investigation of A among those that *were* receiving B).

A suggested question for assessing risk of bias in factorial trials is as follows:

- Are reports of the study free of suggestion of an important interaction between the effects of the different interventions?

16.6 Indirect comparisons and multiple-treatments meta-analysis

16.6.1 Introduction

Head-to-head comparisons of alternative interventions may be the focus of a Cochrane Intervention review, a secondary aim of a Cochrane Intervention review, or a key feature of a Cochrane Overview of reviews. Cochrane Overviews summarize multiple Cochrane Intervention reviews, typically of different interventions for the same condition (see Chapter 22). Ideally, direct head-to-head comparisons of alternative interventions would be made within randomized studies, but such studies are often not available. Indirect comparisons are comparisons that are made between competing interventions that have not been compared directly with each other: see Section 16.6.2. Multiple-treatments meta-analysis (MTM) is an extension to indirect comparisons that allows the combination of direct with indirect comparisons, and also the simultaneous analysis of the comparative effects of many interventions: see Section 16.6.3.

16.6.2 Indirect comparisons

Indirect comparisons are made between interventions in the absence of head-to-head randomized studies. For example, suppose that some trials have compared the effectiveness of ‘dietician versus doctor’ in providing dietary advice, and others have compared the effectiveness of ‘dietician versus nurse’, but no trials have compared the effectiveness of ‘doctor versus nurse’. We might then wish to learn about the relative effectiveness of ‘doctor versus nurse’ by making indirect comparisons. In fact, doctors and nurses can be compared indirectly by contrasting trials of ‘dietician versus doctor’ with trials of ‘dietician versus nurse’.

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

More appropriate methods for indirect comparisons are available, but the assumptions underlying the methods need to be considered carefully. A relatively simple method is to perform subgroup analyses, the different subgroups being defined by the different comparisons being made. For the particular case

of two subgroups (two comparisons; three interventions) the difference between the subgroups can be estimated, and the statistical significance determined, using a simple procedure described by Bucher (Bucher 1997). In the previous example, one subgroup would be the 'dietician versus doctor' trials, and the other subgroup the 'dietician versus nurse' trials. The difference between the summary effects in the two subgroups will provide an estimate of the desired comparison, 'doctor versus nurse'. The test can be performed using the test for differences between subgroups, as implemented in RevMan (see Chapter 9, Section 9.6.3.1). The validity of an indirect comparison relies on the different subgroups of trials being similar, on average, in all other factors that may affect outcome. More extensive discussions of indirect comparisons are available (Song 2003, Glenny 2005).

Indirect comparisons are not randomized comparisons, and cannot be interpreted as such. They are essentially observational findings across trials, and may suffer the biases of observational studies, for example due to confounding (see Chapter 9, Section 9.6.6). In situations when both direct and indirect comparisons are available in a review, then unless there are design flaws in the head-to-head trials, the two approaches should be considered separately and the direct comparisons should take precedence as a basis for forming conclusions.

16.6.3 Multiple-treatments meta-analysis

Methods are available for analysing, simultaneously, three or more different interventions in one meta-analysis. These are usually referred to as 'multiple-treatments meta-analysis' ('MTM'), 'network meta-analysis', or 'mixed treatment comparisons' ('MTC') meta-analysis. Multiple-treatments meta-analyses can be used to analyse studies with multiple intervention groups, and to synthesize studies making different comparisons of interventions. Caldwell et al. provide a readable introduction (Caldwell 2005); a more comprehensive discussion is provided by Salanti et al. (Salanti 2008). Note that multiple-treatments meta-analyses retain the identity of each intervention, allowing multiple intervention comparisons to be made. This is in contrast to the methods for dealing with a single study with multiple intervention groups that are described in Section 16.5, which focus on reducing the multiple groups to a single pair-wise comparison.

The simplest example of a multiple-treatments meta-analysis is the indirect comparison described in Section 16.6.2. With three interventions (e.g. advice from dietician, advice from doctor, advice from nurse), any two can be compared indirectly through comparisons with the third. For example, doctors and nurses can be compared indirectly by contrasting trials of 'dietician versus doctor' with trials of 'dietician versus nurse'. This analysis may be extended in various ways. For example, if there are also trials of the direct comparison 'doctor versus nurse', then these might be combined with the results of the indirect comparison. If there are more than three interventions, then there will be several direct and indirect comparisons, and it will be more convenient to analyse them simultaneously.

If each study compares exactly two interventions, then multiple-treatments meta-analysis can be performed using subgroup analyses, and the test for subgroup differences used as described in Chapter 9 (Section 9.6.3.1). However, it is preferable to use a random-effects model to allow for heterogeneity within each subgroup, and this can be achieved by using meta-regression instead (see Chapter 9, Section 9.6.4). When some studies include more than two intervention groups, the synthesis requires multivariate meta-analysis methods. Standard subgroup analysis and meta-regression methods can no longer be used, although the analysis can be performed in a Bayesian framework using WinBUGS: see Section 16.8.1. A particular advantage of using a Bayesian framework is that all interventions in the analysis can be ranked, using probabilistic, rather than crude, methods.

Multiple treatment meta-analyses are particularly suited to problems addressed by Overviews of reviews (Chapter 22). However, they rely on a strong assumption that studies of different comparisons are similar in all ways other than the interventions being compared. The indirect comparisons involved

are not randomized comparisons, and may suffer the biases of observational studies, for example due to confounding (see Chapter 9, Section 9.6.6). In situations when both direct and indirect comparisons are available in a review, any use of multiple-treatments meta-analyses should be to supplement, rather than to replace, the direct comparisons. Expert statistical support, as well as subject expertise, is required for a multiple-treatments meta-analysis.

16.7 Multiplicity and the play of chance

16.7.1 Introduction

A Cochrane review might include multiple analyses because of a choice of several outcome measures, outcomes measured at multiple time points, a desire to explore subgroup analyses, the inclusion of multiple intervention comparisons, or other reasons. The more analyses that are done, the more likely it is that some of them will be found to be ‘statistically significant’ by chance alone. Using the conventional significance level of 5%, it is expected that one in 20 tests will be statistically significant even when there is truly no difference between the interventions being compared. However, after 14 independent tests, it is more likely than not (probability greater than 0.5) that at least one test will be significant, even when there is no true effect. The probability of finding at least one statistically significant result increases with the number of tests performed. The likelihood of a spurious finding by chance is higher when the analyses are independent. For example, multiple analyses of different subgroups are usually more problematic in this regard than multiple analyses of various outcomes, since the latter involve the same participants so are not independent.

The problem of multiple significance tests occurs in clinical trials, epidemiology and public health research (Bauer 1991, Ottenbacher 1998) as well as in systematic reviews (Bender 2008). There is an extensive statistical literature about the multiplicity issue. Many statistical approaches have been developed to adjust for multiple testing in various situations (Bender 2001, Cook 2005, Dmitrienko 2006). However, there is no consensus about when multiplicity should be taken into account, or about which statistical approach should be used if an adjustment for multiple testing is made. For example, the use of adjustments appropriate for independent tests will lead to P values that are too large when the multiple tests are not independent. Adjustments for multiple testing are used in confirmatory clinical trials to protect against spuriously significant conclusions when multiple hypothesis tests are used (Koch 1996) and have been incorporated in corresponding statistical guidelines (CPMP Working Party on Efficacy of Medicinal Products 1995). In exploratory studies, in which there is no pre-specified key hypothesis, adjustments for multiple testing might not be required and are often not feasible (Bender 2001). Statistically significant results from exploratory studies should be thought of as ‘hypothesis generating’, regardless of whether adjustments for multiple testing have been performed.

16.7.2 Multiplicity in systematic reviews

Adjustments for multiple tests are not routinely used in systematic reviews, and we do not recommend their use in general. Nevertheless, issues of multiplicity apply just as much to systematic reviews as to other types of research. Review authors should remember that in a Cochrane review the emphasis should generally be on estimating intervention effects rather than testing for them. However, the general problem of multiple comparisons affects interval estimation just as much as hypothesis testing (Chen 2005, Bender 2008).

Some additional problems associated with multiplicity occur in systematic reviews. For instance, when the results of a study are presented, it is not always possible to know how many tests or analyses were done. It is likely that in some studies interesting findings were selected for presentation or publication in relation to statistical significance, and other ‘uninteresting’ findings omitted, leading to misleading

results and spurious conclusions. Such selective reporting is discussed in more detail in Chapter 8 (Section 8.13).

Adequate planning of the statistical testing of hypotheses (including any adjustments for multiple testing) should ideally be done at the design stage. Unfortunately, this can be difficult for systematic reviews when it might not be known, at the outset, which outcomes and which effect measures will be available from the included studies. This makes the *a priori* planning of multiple test procedures for systematic reviews more difficult or even impossible. Moreover, only some of the multiple comparison procedures developed for single studies can be used in meta-analyses of summary data. More research is required to develop adequate multiple comparison procedures for use in systematic reviews (Bender 2008).

In summary, there is no simple or completely satisfactory solution to the problem of multiple testing and multiple interval estimation in systematic reviews. However, the following general advice can be offered. More detailed advice can be found elsewhere (Bender 2008).

- In the protocol for the review, state which analyses and outcomes are of particular interest (the fewer the better). Outcomes should be classified in advance as primary and secondary outcomes, and main outcomes to appear in the 'Summary of findings' table should be pre-specified. If there is a clear key hypothesis, which could be tested by means of multiple significance tests, performing an adequate adjustment for multiple testing will lead to stronger confidence in any conclusions that are drawn.
- Although it is recommended that Cochrane reviews should seek to include all outcomes that are likely to be important to users of the review, overall conclusions are more difficult to draw if there are multiple analyses. Bear in mind, when drawing conclusions, that approximately one in 20 independent statistical tests will be statistically significant (at a 5% significance level) due to chance alone when there is no real difference between the groups.
- Do not select results for emphasis (e.g. in the abstract) on the basis of a statistically significant P value.
- If there is a choice of time-points for an outcome, attempts should be made to present a summary effect over all time points, or to choose one time-point that is the most appropriate one (although availability of suitable data from all trials may be a problem). Multiple testing of the effect at each of the time-points should be avoided.
- Keep subgroup analyses to a minimum and interpret them cautiously.
- Interpret cautiously any findings that were not hypothesized in advance, even when they are 'statistically significant'. Such findings should only be used to generate hypotheses, not to prove them.

16.8 Bayesian and hierarchical approaches to meta-analysis

16.8.1 Bayesian methods

Bayesian statistics is an approach to statistics based on a different philosophy from that which underlies significance tests and confidence intervals. It is essentially about updating of evidence. In a Bayesian analysis, initial uncertainty is expressed through a **prior distribution** about the quantities of interest. Current data and assumptions concerning how they were generated are summarized in the **likelihood**. The **posterior distribution** for the quantities of interest can then be obtained by combining the prior distribution and the likelihood. The posterior distribution may be summarized by point estimates and credible intervals, which look much like classical estimates and confidence

intervals. Bayesian analysis cannot be carried out in RevMan, but may be performed using WinBUGS software (Smith 1995, Lunn 2000).

In the context of a meta-analysis, the prior distribution will describe uncertainty regarding the particular effect measure being analysed, such as the odds ratio or the mean difference. This may be an expression of subjective belief about the size of the effect, or it may be from sources of evidence not included in the meta-analysis, such as information from non-randomized studies. The width of the prior distribution reflects the degree of uncertainty about the quantity. When there is little or no information, a 'non-informative' prior can be used, in which all values across the possible range are equally likely. The likelihood summarizes both the data from studies included in the meta-analysis (for example, 2×2 tables from randomized trials) and the meta-analysis model (for example, assuming a fixed effect or random effects).

The choice of prior distribution is a source of controversy in Bayesian statistics. Although it is possible to represent beliefs about effects as a prior distribution, it may seem strange to combine objective trial data with subjective opinion. A common practice in meta-analysis is therefore to use non-informative prior distributions to reflect a position of prior ignorance. This is particularly true for the main comparison. However, prior distributions may also be placed on other quantities in a meta-analysis, such as the extent of among-study variation in a random-effects analysis. It may be useful to bring in judgement, or external evidence, on some of these other parameters, particularly when there are few studies in the meta-analysis. It is important to carry out sensitivity analyses to investigate how the results depend on any assumptions made.

A difference between Bayesian analysis and classical meta-analysis is that the interpretation is directly in terms of belief: a 95% credible interval for an odds ratio is that region in which we believe the odds ratio to lie with probability 95%. This is how many practitioners actually interpret a classical confidence interval, but strictly in the classical framework the 95% refers to the long-term frequency with which 95% intervals contain the true value. The Bayesian framework also allows a review author to calculate the probability that the odds ratio has a particular range of values, which cannot be done in the classical framework. For example, we can determine the probability that the odds ratio is less than 1 (which might indicate a beneficial effect of an experimental intervention), or that it is no larger than 0.8 (which might indicate a clinically important effect). It should be noted that these probabilities are specific to the choice of the prior distribution. Different meta-analysts may analyse the same data using different prior distributions and obtain different results.

Bayesian methods offer some potential advantages over many classical methods for meta-analyses. For example, they can be used to:

- incorporate external evidence, such as on the effects of interventions or the likely extent of among-study variation;
- extend a meta-analysis to decision-making contexts, by incorporating the notion of the *utility* of various clinical outcome states;
- allow naturally for the imprecision in the estimated between-study variance estimate (see Chapter 9, Section 9.5.4);
- investigate the relationship between underlying risk and treatment benefit (see Chapter 9, Section 9.6.7);
- perform complex analyses (e.g. multiple-treatments meta-analysis), due to the flexibility of the WinBUGS software; and
- examine the extent to which data would change people's beliefs (Higgins 2002).

Statistical expertise is strongly recommended for review authors wishing to carry out Bayesian analyses. There are several good texts (Sutton 2000, Sutton 2001, Spiegelhalter 2004).

16.8.2 Hierarchical models

Some sophisticated techniques for meta-analysis exploit a statistical framework called hierarchical models, or multilevel models (Thompson 2001). This is because the information in a meta-analysis usually stems from two levels: studies at the higher level, and participants within studies at the lower level. Sometimes additional levels may be relevant, for example centres in a multicentre trial, or clusters in a cluster-randomized trial. A hierarchical framework is appropriate whether meta-analysis is of summary statistic information (for example, log odds ratios and their variances) or individual patient data (Turner 2000). Such a framework is particularly relevant when random effects are used to represent unexplained variation in effect estimates among studies (see Chapter 9, Section 9.5.4).

Hierarchical models rather than simpler methods of meta-analysis are useful in a number of contexts. For example, they can be used to:

- allow for the imprecision of the variance estimates of treatment effects within studies;
- allow for the imprecision in the estimated between-study variance estimate, tau-squared (see Chapter 9, Section 9.5.4);
- provide methods that explicitly model binary outcome data (rather than summary statistics);
- investigate the relationship between underlying risk and treatment benefit (see Chapter 9, Section 9.6.7); and
- extend methods to incorporate either study-level characteristics (see Chapter 9, Section 9.6.4) or individual-level characteristics (see Chapter 18).

Hierarchical models are particularly relevant where individual patient data (IPD) on both outcomes and covariates are available (Higgins 2001). However even using such methods, care still needs to be exercised to ensure that within- and between-study relationships are not confused.

Implementing hierarchical models needs sophisticated software, either using a classical statistical approach (e.g. SAS proc mixed, or MlwiN) or a Bayesian approach (e.g. WinBUGS). Much current methodological research in meta-analysis uses hierarchical model methods, often in a Bayesian implementation.

16.9 Rare events (including zero frequencies)

16.9.1 Meta-analysis of rare events

For rare outcomes, meta-analysis may be the only way to obtain reliable evidence of the effects of healthcare interventions. Individual studies are usually underpowered to detect differences in rare outcomes, but a meta-analysis of many studies may have adequate power to investigate whether interventions do impact on the incidence of the rare event. However, many methods of meta-analysis are based on large sample approximations, and are unsuitable when events are rare. Thus authors must take care when selecting a method of meta-analysis.

There is no single risk at which events are classified as ‘rare’. Certainly risks of 1 in 1000 constitute rare events, and many would classify risks of 1 in 100 the same way. However, the performance of methods when risks are as high as 1 in 10 may also be affected by the issues discussed in this section. What is typical is that a high proportion of the studies in the meta-analysis observe no events in one or more study arm.

16.9.2 Studies with zero-cell counts

Computational problems can occur when no events are observed in one or both groups in an individual study. Inverse variance meta-analytical methods (both the inverse-variance fixed-effect and DerSimonian and Laird random-effects methods) involve computing an intervention effect estimate and its standard error for each study. For studies where no events were observed in one or both arms, these computations often involve dividing by a zero count, which yields a computational error. Most meta-analytical software (including RevMan) automatically check for problematic zero counts, and add a fixed value (typically 0.5) to all cells of study results tables where the problems occur. The Mantel-Haenszel methods only require zero-cell corrections if the same cell is zero in all the included studies, and hence need to use the correction less often. However, in many software applications the same correction rules are applied for Mantel-Haenszel methods as for the inverse-variance methods. Odds ratio and risk ratio methods require zero cell corrections more often than difference methods, except for the Peto odds ratio method, which only encounters computation problems in the extreme situation of no events occurring in all arms of all studies.

Whilst the fixed correction meets the objective of avoiding computational errors, it usually has the undesirable effect of biasing study estimates towards no difference and overestimating variances of study estimates (consequently down-weighting inappropriately their contribution to the meta-analysis). Where the sizes of the study arms are unequal (which occurs more commonly in non-randomized studies than randomized trials), they will introduce a directional bias in the treatment effect. Alternative non-fixed zero-cell corrections have been explored by Sweeting et al., including a correction proportional to the reciprocal of the size of the contrasting study arm, which they found preferable to the fixed 0.5 correction when arm sizes were not balanced (Sweeting 2004).

16.9.3 Studies with no events

The standard practice in meta-analysis of odds ratios and risk ratios is to exclude studies from the meta-analysis where there are no events in both arms. This is because such studies do not provide any indication of either the direction or magnitude of the relative treatment effect. Whilst it may be clear that events are very rare on both the experimental intervention and the control intervention, no information is provided as to which group is likely to have the higher risk, or on whether the risks are of the same or different orders of magnitude (when risk are very low, they are compatible with very large or very small ratio measures). Whilst one might be tempted to infer that the risk would be lowest in the group with the larger sample size (as the upper limit of the confidence interval would be lower), this is not justified as the sample size allocation was determined by the study investigators and is not a measure of the incidence of the event.

Risk difference methods superficially appear to have an advantage over odds ratio methods in that the RD is defined (as zero) when no events occur in either arm. Such studies are therefore included in the estimation process. Bradburn et al. undertook simulation studies which revealed that all risk difference methods yield confidence intervals that are too wide when events are rare, and have associated poor statistical power, which make them unsuitable for meta-analysis of rare events (Bradburn 2007). This is especially relevant when outcomes that focus on treatment safety are being studied, as the ability to identify correctly (or attempt to refute) serious adverse events is a key issue in drug development.

It is likely that outcomes for which no events occur in either arm may not be mentioned in reports of many randomized trials, precluding their inclusion in a meta-analysis. It is unclear, though, when working with published results, whether failure to mention a particular adverse event means there *were* no such events, or simply that such events were not included as a measured endpoint. Whilst the results of risk difference meta-analyses will be affected by non-reporting of outcomes with no events,

odds and risk ratio based methods naturally exclude these data whether or not they are published, and are therefore unaffected.

16.9.4 Confidence intervals when no events are observed

It is possible to put upper confidence bounds on event risks when no events are observed, which may be useful when trying to ascertain possible risks for serious adverse events. A simple rule termed the 'rule of threes' has been proposed such that if no events are observed in a group, then the upper confidence interval limit for the number of events is three, and for the risk (in a sample of size N) is $3/N$ (Hanley 1983). The application of this rule has not directly been proposed or evaluated for systematic reviews. However, when looking at the incidence of a rare event that is not observed in any of the intervention groups in a series of studies (which randomized trials, non-randomized comparison or case series), it seems reasonable to apply it, taking N as the sum of the sample sizes of the arms receiving intervention. However, it will not provide any information about the relative incidence of the event between two groups.

The value 3 coincides with the upper limit of a one-tailed 95% confidence interval from the Poisson distribution (equivalent to a two-tailed 90% confidence interval). For the risk to be for a more standard one-tailed 97.5% confidence interval (equivalent to a two-tailed 95% confidence interval) then 3.7 should be used in all calculations in place of 3 (Newcombe 2000). An alternative recommendation which gives similar values is the 'rule of fours' which takes the upper limit of the risk to be $4/(N+4)$. Either of these options is recommended for use in Cochrane reviews. For example, if no events were observed out of 10, the upper limit of the confidence interval for the number of events is 3.7, and for the risk is 3.7 out of 10 (i.e. 0.37). If no events were observed out of 100, the upper limit on the number of events is still 3.7, but for the risk is 3.7 out of 100 (i.e. 0.037).

16.9.5 Validity of methods of meta-analysis for rare events

Simulation studies have revealed that many meta-analytical methods can give misleading results for rare events, which is unsurprising given their reliance on asymptotic statistical theory. Their performance has been judged suboptimal either through results being biased, confidence intervals being inappropriately wide, or statistical power being too low to detect substantial differences.

Below we consider the choice of statistical method for meta-analyses of odds ratios. Appropriate choices appear to depend on the control group risk, the likely size of the treatment effect and consideration of balance in the numbers of treated and control participants in the constituent studies. No research has evaluated risk ratio measures directly, but their performance is likely to be very similar to corresponding odds ratio measurement. When events are rare, estimates of odds and risks are near identical, and results of both can be interpreted as ratios of probabilities.

Bradburn et al. found that many of the most commonly used meta-analytical methods were biased when events were rare (Bradburn 2007). The bias was greatest in inverse variance and DerSimonian and Laird odds ratio and risk difference methods, and the Mantel-Haenszel odds ratio method using a 0.5 zero-cell correction. As already noted, risk difference meta-analytical methods tended to show conservative confidence interval coverage and low statistical power when risks of events were low.

At event rates below 1% the Peto one-step odds ratio method was found to be the least biased and most powerful method, and provided the best confidence interval coverage, provided there was no substantial imbalance between treatment and control group sizes within studies, and treatment effects were not exceptionally large. This finding was consistently observed across three different meta-analytical scenarios, and was also observed by Sweeting et al. (Sweeting 2004).

This finding was noted despite the method producing only an approximation to the odds ratio. For very large effects (e.g. risk ratio = 0.2) when the approximation is known to be poor, treatment effects were underestimated, but the Peto method still had the best performance of all the methods considered for event risks of 1 in 1000, and the bias was never more than 6% of the control group risk.

In other circumstances (i.e. event risks above 1%, very large effects at event risks around 1%, and meta-analyses where many studies were substantially imbalanced) the best performing methods were the Mantel-Haenszel OR without zero-cell corrections, logistic regression and an exact method. None of these methods is available in RevMan.

Methods that should be avoided with rare events are the inverse-variance methods (including the DerSimonian and Laird random-effects method). These directly incorporate the study's variance in the estimation of its contribution to the meta-analysis, but these are usually based on a large-sample variance approximation, which was not intended for use with rare events. The DerSimonian and Laird method is the only random-effects method commonly available in meta-analytic software. We would suggest that incorporation of heterogeneity into an estimate of a treatment effect should be a secondary consideration when attempting to produce estimates of effects from sparse data – the primary concern is to discern whether there is any signal of an effect in the data.

16.10 Chapter information

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

This chapter should be cited as: Higgins JPT, Deeks JJ, Altman DG (editors). Chapter 16: Special topics in statistics. 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, Ralf Bender, Catey Bunce, Marion Campbell, Mike Clarke, Jon Deeks, Simon Gates, Julian Higgins, Nathan Pace and Simon Thompson.

Acknowledgements: We particularly thank Joseph Beyene, Peter Gøtzsche, Steff Lewis, Georgia Salanti, Stephen Senn and Ian White for helpful comments on earlier drafts. For details of the Cochrane Statistical Methods Group, see Chapter 9 (Box 9.8.a).

16.11 References

Abrams 2005

Abrams KR, Gillies CL, Lambert PC. Meta-analysis of heterogeneously reported trials assessing change from baseline. *Statistics in Medicine* 2005; 24: 3823-3844.

Bauer 1991

Bauer P. Multiple testing in clinical trials. *Statistics in Medicine* 1991; 10: 871-889.

Bender 2001

Bender R, Lange S. Adjusting for multiple testing - when and how? *Journal of Clinical Epidemiology* 2001; 54: 343-349.

Bender 2008

Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, Thorlund K. Dealing with multiplicity in systematic reviews. *Journal of Clinical Epidemiology* 2008; 54: 343-349.

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.

Bucher 1997

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

Caldwell 2005

Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005; 331: 897-900.

Campbell 2000

Campbell M, Grimshaw J, Steen N. Sample size calculations for cluster randomised trials. Changing Professional Practice in Europe Group (EU BIOMED II Concerted Action). *Journal of Health Services Research and Policy* 2000; 5: 12-16.

Chan 2005

Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *The Lancet* 2005; 365: 1159-1162.

Chen 2005

Chen T, Hoppe FM. Simultaneous confidence intervals. In: Armitage P, Colton T (editors). *Encyclopedia of Biostatistics* (2nd edition). Chichester (UK): John Wiley & Sons, 2005.

Cook 2005

Cook RJ, Dunnett CW. Multiple comparisons. In: Armitage P, Colton T (editors). *Encyclopedia of Biostatistics* (2nd edition). Chichester (UK): John Wiley & Sons, 2005.

CPMP Working Party on Efficacy of Medicinal Products 1995

CPMP Working Party on Efficacy of Medicinal Products. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. *Statistics in Medicine* 1995; 14: 1659-1682.

Dmitrienko 2006

Dmitrienko A, Hsu JC. Multiple testing in clinical trials. In: Kotz S, Balakrishnan N, Read CB, Vidakovic B (editors). *Encyclopedia of Statistical Sciences* (2nd edition). Hoboken (NJ): John Wiley & Sons, 2006.

Donner 1980

Donner A, Koval JJ. The estimation of intraclass correlation in the analysis of family data. *Biometrics* 1980; 36: 19-25.

Donner 2000

Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London (UK): Arnold, 2000.

Donner 2001

Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomized trials. *Statistical Methods in Medical Research* 2001; 10: 325-338.

Donner 2002

Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. *Statistics in Medicine* 2002; 21: 2971-2980.

Elbourne 2002

Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vaillancourt JM. Meta-analyses involving cross-over trials: methodological issues. *International Journal of Epidemiology* 2002; 31: 140-149.

Eldridge 2004

Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials* 2004; 1: 80-90.

Farrin 2005

Farrin A, Russell I, Torgerson D, Underwood M, UK BEAM Trial Team. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. *Clinical Trials* 2005; 2: 119-124.

Follmann 1992

Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. *Journal of Clinical Epidemiology* 1992; 45: 769-773.

Freeman 1989

Freeman PR. The performance of the two-stage analysis of two-treatment, two-period cross-over trials. *Statistics in Medicine* 1989; 8: 1421-1432.

Furukawa 2006

Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of Clinical Epidemiology* 2006; 59: 7-10.

Gamble 2005

Gamble C, Hollis S. Uncertainty method improved on best-worst case analysis in a binary meta-analysis. *Journal of Clinical Epidemiology* 2005; 58: 579-588.

Glenny 2005

Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, Bradburn M, Eastwood AJ. Indirect comparisons of competing interventions. *Health Technology Assessment* 2005; 9: 26.

Hahn 2005

Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. *BMC Medical Research Methodology* 2005; 5: 10.

Hanley 1983

Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* 1983; 249: 1743-1745.

Health Services Research Unit 2004

Health Services Research Unit. Database of ICCs: Spreadsheet (Empirical estimates of ICCs from changing professional practice studies) [page last modified 11 Aug 2004]. Available from: <http://www.abdn.ac.uk/hsru/epp/cluster.shtml> (accessed 1 January 2008).

Higgins 2001

Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine* 2001; 20: 2219-2241.

Higgins 2002

Higgins JPT, Spiegelhalter DJ. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *International Journal of Epidemiology* 2002; 31: 96-104.

Higgins 2008

Higgins JPT, White IR, Wood AM. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials* 2008; 5: 225-239.

Hollis 1999

Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999; 319: 670-674.

Hollis 2002

Hollis S. A graphical sensitivity analysis for clinical trials with non-ignorable missing binary outcome. *Statistics in Medicine* 2002; 21: 3823-3834.

Juszczak 2003

Juszczak E, Altman D, Chan AW. A review of the methodology and reporting of multi-arm, parallel group, randomised clinical trials (RCTs). *3rd Joint Meeting of the International Society for Clinical Biostatistics and Society for Clinical Trials*, London (UK), 2003.

Khan 1996

Khan KS, Daya S, Collins JA, Walter SD. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertility and Sterility* 1996; 65: 939-945.

Koch 1996

Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. *Drug Information Journal* 1996; 30: 523-534.

Lathyris 2007

Lathyris DN, Trikalinos TA, Ioannidis JP. Evidence from crossover trials: empirical evaluation and comparison against parallel arm trials. *International Journal of Epidemiology* 2007; 36: 422-430.

Lee 2005a

Lee LJ, Thompson SG. Clustering by health professional in individually randomised trials. *BMJ* 2005; 330: 142-144.

Lee 2005b

Lee SHH. *Use of the two-stage procedure for analysis of cross-over trials in four aspects of medical statistics* (PhD thesis). University of London, 2005.

Lewis 1993

Lewis JA, Machin D. Intention to treat--who should use ITT? *British Journal of Cancer* 1993; 68: 647-650.

Little 2004

Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edition). Hoboken (NJ): John Wiley & Sons, 2004.

Lunn 2000

Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; 10: 325-337.

Marinho 2003

Marinho VCC, Higgins JPT, Logan S, Sheiham A. Fluoride toothpaste for preventing dental caries in children and adolescents. *Cochrane Database of Systematic Reviews* 2003, Issue 1. Art No: CD002278.

McAlister 2003

McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA* 2003; 289: 2545-2553.

Mills 2005

Mills EJ, Chan AW, Guyatt GH, Altman DG. Design, analysis, and presentation of cross-over trials. *5th Peer Review Congress*, Chicago (IL), 2005.

Murray 1995

Murray DM, Short B. Intraclass correlation among measures related to alcohol-use by young-adults - estimates, correlates and applications in intervention studies. *Journal of Studies on Alcohol* 1995; 56: 681-694.

Newcombe 2000

Newcombe RN, Altman DG. Proportions and their differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ (editors). *Statistics with Confidence* (2nd edition). London (UK): BMJ Books, 2000.

Newell 1992

Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. *International Journal of Epidemiology* 1992; 21: 837-841.

Ottenbacher 1998

Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. *American Journal of Epidemiology* 1998; 147: 615-619.

Puffer 2003

Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ* 2003; 327: 785-789.

Qizilbash 1998

Qizilbash N, Whitehead A, Higgins J, Wilcock G, Schneider L, Farlow M. Cholinesterase inhibition for Alzheimer disease: a meta-analysis of the tacrine trials. *JAMA* 1998; 280: 1777-1782.

Rao 1992

Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* 1992; 48: 577-585.

Salanti 2008

Salanti G, Higgins J, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* 2008; 17: 279-301.

Senn 2002

Senn S. *Cross-over Trials in Clinical Research* (2nd edition). Chichester (UK): John Wiley & Sons, 2002.

Smith 1995

Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* 1995; 14: 2685-2699.

Song 2003

Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ* 2003; 325: 472-475.

Spiegelhalter 2004

Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester (UK): John Wiley & Sons, 2004.

Stewart 1995

Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine* 1995; 14: 2057-2079.

Sutton 2000

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-analysis in Medical Research*. Chichester (UK): John Wiley & Sons, 2000.

Sutton 2001

Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 2001; 10: 277-303.

Sweeting 2004

Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; 23: 1351-1375.

te Velde 1998

te Velde ER, Cohlen BJ, Looman CW, Habbema JD. Crossover designs versus parallel studies in infertility research. *Fertility and Sterility* 1998; 69: 357-358.

Thompson 2001

Thompson SG, Turner RM, Warn DE. Multilevel models for meta-analysis, and their application to absolute risk differences. *Statistical Methods in Medical Research* 2001; 10: 375-392.

Turner 2000

Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2000; 19: 3417-3432.

Ukoumunne 1999

Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment* 1999; 3: 5.

Unnebrink 2001

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

White 2005

White IR, Thomas J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. *Clinical Trials* 2005; 2: 141-151.

White 2007

White IR, Carpenter J, Evans S, Schroter S. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials* 2007; 4: 125-139.

White 2008a

White IR, Higgins JPT, Wood A. Allowing for uncertainty due to missing data in meta-analysis. Part 1: Two-stage methods. *Statistics in Medicine* 2008; 27: 711-727.

White 2008b

White IR, Welton N, Wood A, Ades AE, Higgins JPT. Allowing for uncertainty due to missing data in meta-analysis. Part 2: Hierarchical models. *Statistics in Medicine* 2008; 27: 728-745.

Whiting-O'Keefe 1984

Whiting-O'Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in medical care experiments. *Medical Care* 1984; 22: 1101-1114.