Cochrane Handbook for Systematic Reviews of Interventions

8 Assessing risk of bias in included studies

Editors: Julian PT Higgins, Douglas G Altman

PRELIMINARY DRAFT, 12 October 2007

This document is posted for consultation and does not yet constitute formal Handbook advice

Table of Contents

8.1 Introduction

The extent to which a Cochrane review can draw conclusions about the effects of an intervention depends on whether the findings of the included studies are valid. In particular, a meta-analysis of invalid studies may produce a misleading result, yielding a narrow confidence interval around the wrong intervention effect estimate. The evaluation of the validity of the included studies is therefore an essential component of a Cochrane review, and should influence the analysis, interpretation and conclusions of the review.

The validity of a study may be considered to have two dimensions. The first dimension is whether the study is asking an appropriate research question. This is often described as 'external validity', and its assessment depends on the purpose for which the study is to be used. External validity is closely connected with the generalizability or applicability of a study's findings, and is addressed in Chapter 12.

The second dimension of a study's validity relates to whether it answers its research question 'correctly', that is, in a manner free from bias. This is often described as 'internal validity', and it is this aspect of validity that we address in this chapter. As most Cochrane reviews focus on randomized controlled trials, we concentrate on how to appraise the validity of this type of study. Chapter 13 addresses further issues in the assessment of non-randomized studies. Assessments of internal validity are frequently referred to as 'quality assessments'. However, we will avoid the term quality, for reasons explained below. In the next section we define 'bias' and distinguish it from related concepts of random error, and quality.

8.2 What is bias?

8.2.1 'Bias' and 'risk of bias'

A **bias** is a systematic error, or deviation from the truth, in results or inferences. Biases can operate in either direction: they can lead to underestimation or overestimation of the true intervention effect. Biases can vary in magnitude: some are small (and trivial compared with the observed effect) and some are substantial (so that an observed effect may be entirely due to bias). Biases can also vary in direction: bias due to a particular design flaw (e.g. lack of allocation concealment) may lead to underestimation of an effect in one study but overestimation in another study. It is usually impossible to know to what extent biases have affected the results of a particular study, although there is good empirical evidence that particular flaws in the design, conduct and analysis of randomized clinical trials lead to bias (see Section 8.2.3). Because the results of a study may in fact be unbiased despite a methodological flaw, it is more appropriate to consider **risk of bias**.

Differences in risks of bias can help explain variation in the results of the studies included in a systematic review (i.e. can explain heterogeneity of results). More rigorous studies are more likely to yield results that are close to the truth. Meta-analysis of results from studies of variable validity can result in false positive conclusions (erroneously concluding an intervention is effective) if the less rigorous studies are biased toward overestimating an intervention's effect. They might also come to false negative conclusions (erroneously concluding no effect) if the less rigorous studies are biased towards underestimating an intervention's effect (Detsky 1992).

It is important to assess risk of bias in all studies in a review irrespective of the anticipated variability in either the results or the validity of the included studies. For instance, the results may be consistent among studies but all the studies may be flawed. In this case, the review's conclusions should not be as strong as if a series of rigorous studies yielded consistent results about an intervention's effect. In a Cochrane review, this appraisal process is described as the *assessment of risk of bias in included*

studies. A tool that has been developed and implemented in RevMan for this purpose is described in Section 8.5. The rest of this chapter provides the rationale for this tool as well as explaining how bias assessments should be summarized and incorporated in analyses. Sections 8.9 to 8.14 provide background considerations to assist review authors in using the tool

Bias should not be confused with **precision**. Bias refers to *systematic error*, meaning that multiple replications of the same study would reach the wrong answer on average. Precision refers to *random error*, meaning that multiple replications of the same study will produce different effect estimates because of sampling variation even if they would give the right answer on average. The results of smaller studies are subject to greater sampling variation and hence are less precise. Precision is reflected in the confidence interval around the intervention effect estimate from each study and in the weight given to the results of each study in a meta-analysis. More precise results are given more weight.

8.2.2 'Risk of bias' and 'quality'

Bias may be distinguished from **quality**. The term 'assessment of methodological quality' has been used extensively in the context of systematic review methods to refer to the critical appraisal of included studies. The term suggests an investigation of the extent to which study authors conducted their research to the highest possible standards. This Handbook draws a distinction between assessment of methodological quality and assessment of risk of bias, and recommends a focus on the latter. The reasons for this distinction include:

1. The key consideration in a Cochrane review is the extent to which results of included studies should be *believed*. Assessing risk of bias targets this question squarely.

2. A study may be performed to the highest possible standards yet still be at an important risk of bias. For example, in many situations it is impractical or impossible to blind participants or study personnel to intervention group. It is inappropriately judgmental to describe such studies as of 'low quality', but that does not mean they are free of bias resulting from knowledge of intervention status.

3. Some markers of quality in medical research, such as obtaining ethical approval, performing a sample size calculation and reporting a study in line with the CONSORT Statement, are unlikely to have direct implications for risk of bias.

4. An emphasis on risk of bias overcomes ambiguity between the quality of reporting and the quality of the underlying research (although does not overcome the problem of having to rely on reports to assess the underlying research).

Notwithstanding these concerns about the term 'quality', the term 'quality of evidence' is used in 'Summary of findings' tables in Cochrane reviews to describe the extent to which one can be confident that an estimate of effect is near the true value for an outcome, across studies, as described in Chapter 11 (Section 11.5) and Chapter 12 (Section 12.2). The risk of bias in the results of each study contributing to an estimate of effect is one of several factors that must be considered when judging the quality of a body of evidence, as defined in this context.

8.2.3 Establishing empirical evidence of biases

Biases associated with particular characteristics of studies may be examined using a technique often known as **meta-epidemiology** (Naylor 1997, Sterne 2002). A meta-epidemiological study analyses a collection of meta-analyses, in each of which the component studies has been classified according to some study-level characteristic. An early example was the study of clinical trials with dichotomous outcomes included in meta-analyses from the Cochrane Pregnancy and Childbirth Database (Schulz 1995b). This study demonstrated that trials in which randomization was inadequately concealed or inadequately reported yielded exaggerated estimates of treatment effect compared with trials reporting adequate concealment, and found a similar (but smaller) effect for trials that were not described as double-blind.

A simple analysis of a meta-epidemiological study is to calculate the 'ratio of odds ratios' within each meta-analysis (for example, the odds ratio in trials with inadequate/unclear allocation concealment divided by the odds ratio in trials with adequate allocation concealment). These ratios of odds ratios are then combined across meta-analyses, in a meta-analysis. Thus, such analyses are also known as meta-meta-analyses. In subsequent sections of this chapter, empirical evidence of bias from meta-epidemiological studies is cited where available as part of the rationale for assessing each domain of potential bias.

8.3 Tools for assessing quality and risk of bias

8.3.1 Types of tools

Many tools have been proposed for assessing the quality of studies for use in the context of a systematic review and elsewhere. Most tools are **scales**, in which various components of quality are scored and combined to give a summary score; or **checklists**, in which specific questions are asked (Jüni 2001).

In 1995, Moher and colleagues identified 25 scales and 9 checklists that had been used to assess the validity or 'quality' of randomized controlled trials (Moher 1995, Moher 1996). These scales and checklists included between 3 and 57 items and were found to take from 10 to 45 minutes to complete for each study. Almost all of the items in the instruments were based on suggested or 'generally accepted' criteria that are mentioned in clinical trial textbooks. Many also contained items that were not directly related to internal validity, such as whether a power calculation was done (an item that relates more to the precision of the results) or whether the inclusion and exclusion criteria were clearly described (an item that relates more to applicability than validity). Scales were more likely than checklists to include criteria that do not directly measure internal validity.

The Collaboration's recommended tool for assessing risk of bias is neither a scale nor a checklist. It is a **domain-based evaluation**,, in which critical assessments are made separately for different domains, described in Section 8.5. It was developed between 2005 and 2007 by a working group of methodologists, editors and review authors. Because it is impossible to know the extent of bias (or even the true risk of bias) in a given study, the possibility of validating any proposed tool is limited. The most realistic assessment of the validity of a study may involve subjectivity: for example an assessment of whether lack of blinding of patients might plausibly have affected recurrence of a serious condition such as cancer.

8.3.2 Reporting versus conduct

A key difficulty in the assessment of risk of bias or quality is the obstacle provided by incomplete reporting. While the emphasis should be on the risk of bias in the actual design and conduct of a study, it can be tempting to resort to assessing the adequacy of reporting. Many of the tools reviewed by Moher et al. were liable to confuse these separate issues (Moher 1995). Moreover, scoring in scales was often based on whether something was reported (such as stating how participants were allocated) rather than whether it was done appropriately in the study.

8.3.3 Quality scales and Cochrane reviews

The use of scales for assessing quality or risk of bias is explicitly discouraged in Cochrane reviews. While the approach offers appealing simplicity, it is not supported by empirical evidence (Emerson 1990, Schulz 1995b). Calculating a summary score inevitably involves assigning 'weights' to different items in the scale, and it is difficult to justify the weights assigned. Furthermore, scales have been

shown to be unreliable assessments of validity (Jüni 1999) and they are less likely to be transparent to users of the review. It is preferable to use simple approaches for assessing validity that can be fully reported (i.e. how each trial was rated on each criterion).

One commonly-used scale was developed by Jadad and colleagues for randomized trials in pain research (Jadad 1996). The use of this scale is explicitly discouraged. As well as suffering from the generic problems of scales, it has a strong emphasis on reporting rather than conduct, and does not cover one of the most important potential biases in randomized trials, namely allocation concealment (see Section 8.10.1).

8.3.4 Collecting information for assessments of risk of bias

Despite the limitations of reports, information about the design and conduct of studies will often be obtained from published reports, including journal papers, book chapters, dissertations, conference abstracts and web sites (including trials registries). The extraction of information from such reports is discussed in Chapter 7. Data extraction forms should include space to extract sufficient details to allow implementation of the Collaboration's 'Risk of bias' tool (Section 8.5). When extracting this information, it is particularly desirable to record the exact source of each piece of information (including the precise location within a document). It is helpful to test data extraction forms and assessments of risk of bias within a review team on a pilot sample of articles to ensure that criteria are applied consistently, and that consensus can be reached. Three to six papers that, if possible, span a range from low to high risk of bias might provide a suitable sample for this.

Authors must also decide whether those assessing risk of bias will be blinded to the names of the authors, institutions, journal and results of a study when they assess its methods. One study suggested that blind assessment of reports might produce lower and more consistent ratings than open assessments (Jadad 1996), whereas another suggested little benefit from blind assessments (Berlin 1997). However, blinded assessments are very time consuming, and not all domains of bias can be assessed independently of the outcome data. Furthermore, knowledge of who undertook a study can sometimes allow reasonable assumptions to be made about how the study was conducted (although such assumptions must be reported by the review author). Authors must weigh the potential benefits against the costs involved when deciding whether or not to blind review authors to certain information in study reports.

Review authors with different levels of methodological training and experience may extract different sources of evidence and reach different judgments about risk of bias. Although experts in content areas may have pre-formed opinions that can influence their assessments (Oxman 1993), they may nonetheless give more consistent assessments of the validity of studies than people without content expertise (Jadad 1996). Content experts may have valuable insights into the magnitudes of biases, and experienced methodologists may have valuable insights into biases that are not at first apparent. It is desirable to include both content experts and methodologists and to ensure that all have an adequate understanding of the relevant methodological issues.

Attempts to assess risk of bias are often hampered by incomplete reporting of what happened during the conduct of the study. One option for collecting missing information is to contact the study investigators. Unfortunately, contacting authors of trial reports may lead to overly positive answers. In a survey of 104 trialists, using direct questions about blinding with named categories of trial personnel, 43% responded that the data analysts in their double blind trials were blinded, and 19% responded that the manuscript writers were blinded (Haahr 2006). This is unlikely to be true, given that such procedures were reported in only 3% and 0% of the corresponding published articles, and that they are very rarely described in other trial reports.

To reduce the risk of overly positive answers, review authors should use open-ended questions when asking trial authors for information about study design and conduct. For example, to obtain information about blinding, a request of the following form might be appropriate: "Please describe all measures used, if any, to ensure blinding of trial participants and key trial personnel from knowledge of which intervention a participant had received." To obtain information about the randomization process, a request of the following form might be appropriate: "How did you decide which treatment the next patient should get?". More focussed questions can then be asked to clarify remaining uncertainties.

8.4 Introduction to sources of bias in clinical trials

The reliability of the results of a randomized controlled trial (RCT) depends on the extent to which potential sources of bias have been avoided. A key part of a review is to consider the risk of bias in the results of each of the eligible studies. We introduce six issues to consider briefly here, then describe a tool for assessing them in Section 8.5. We provide more detailed consideration of each issue in Sections 8.9 to 8.14.

The unique strength of randomization is that, if successfully accomplished, it prevents selection bias in allocating interventions to participants. Its success in this respect depends on fulfilling several interrelated processes. A rule for allocating interventions to participants must be specified, based on some chance (random) process. We call this **sequence generation**. Furthermore, steps must be taken to secure strict implementation of that schedule of random assignments by preventing fore-knowledge of the forthcoming allocations, a process termed **allocation concealment**.

After enrolment into the study, **blinding** (or masking) of study participants and personnel may reduce the risk that knowledge of which intervention was received, rather than the intervention itself, affects outcomes and outcome measurements. Blinding can be especially important for assessment of subjective outcomes, such as degree of postoperative pain. Effective blinding can also ensure that the compared groups receive a similar amount of attention, ancillary treatment and diagnostic investigations. Blinding is not always possible, however. For example, it can be impossible to blind people to whether or not major surgery has been undertaken.

Incomplete outcome data raise the possibility that effect estimates are biased. There are two reasons for incomplete (or missing) outcome data in clinical trials. *Exclusions* refer to situations in which some participants are omitted from reports of analyses, despite outcome data being available. *Attrition* refers to situations in which outcome data are not available to the trialists.

Within a published report those analyses with statistically significant differences between treatment groups are more likely to be reported than non-significant differences. This sort of 'within-study publication bias' is usually known as outcome reporting bias or **selective reporting** bias, and may be one of the most substantial biases affecting results from individual studies (Chan 2005).

In addition there are **other sources of bias** that are relevant only in certain circumstances. Some can be found only in particular trial designs (e.g. carry-over in cross-over trials and recruitment bias in cluster randomized trials); some can be found across a broad spectrum of trials, but only for specific circumstances (e.g. bias due to early stopping); and there may be sources of bias that are only found in a particular clinical setting. There are also some complex interrelationships between elements of allocation and elements of blinding in terms of whether bias may be introduced. For example, one approach to sequence generation is through 'blocking', whereby a set number of experimental group

and a set group of control group allocations are randomly ordered within a 'block' of sequentially recruited participants. If there is a lack of blinding after enrolment, such that allocations are revealed to the clinician recruiting to the trial, then it may be possible for some future allocations to be predicted, thus compromising the assignment process.

For all potential sources of bias it is important to consider the likely magnitude and the likely direction of the bias. If all methodological limitations of studies were expected to bias the results towards the null, and the treatment is shown to be effective, then one can still conclude that the treatment is effective.

A useful classification of biases is into selection bias, performance bias, attrition bias, detection bias and reporting bias. Table 8.4.a describes each of these and shows how the domains of assessment in the Collaboration's 'Risk of bias' tool fit with these categories.

Table 8.4.a: A common classification scheme for bias

Type of bias	Description	Relevant domains in the Collaboration's 'Risk of bias' tool
Selection bias	Systematic differences between the groups that are compared.	• Sequence generation; • Allocation concealment.
Performance bias	Systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest.	• Blinding of participants, personnel and outcome assessors; • Topic-specific, design-specific or other potential threats to validity.
Attrition bias	Systematic differences between groups in withdrawals from a study.	• Incomplete outcome data.
Detection bias	Systematic differences between groups in how outcomes are determined.	• Blinding of participants, personnel and outcome assessors; • Topic-specific, design-specific or other potential threats to validity.
Reporting bias	Systematic differences between reported and unreported findings.	• Selective outcome reporting; • (see also Chapter 10).

8.5 The Cochrane Collaboration's tool for assessing risk of bias

8.5.1 Overview

This section describes the recommended approach for assessing risk of bias in studies included in Cochrane reviews. It is a two-part tool, addressing the six specific domains discussed in Sections 8.9 to 8.14 (namely sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting and 'other issues'). The tool is summarized in Table 8.5.a. Each domain includes one or more specific items. Within each item the first part of the tool involves describing

what was reported to have happened in the study. The second part of the tool involves assigning a judgment relating to the risk of bias for each item. This is achieved by answering a pre-specified question about the adequacy of the study in relation to the item, such that a judgement of 'Yes' indicates low risk of bias, 'No' indicates high risk of bias, and 'Unclear' indicates unclear or unknown risk of bias.

The domains of sequence generation, allocation concealment and selective outcome reporting should each be addressed in the tool by a single item for each study. For blinding and for incomplete outcome data, two or more items may be used because assessments generally need to be made separately for different outcomes (or for the same outcome at different time points). Review authors should limit the number of items used by grouping outcomes, for example, as 'subjective, or 'objective' outcomes for the purposes of assessing blinding; or as 'patient-reported at 6 months' or 'patient-reported at 12 months' for incomplete outcome data. The final domain ('other sources of bias') can be assessed as a single item for studies as a whole (the default in RevMan). It is recommended, however, that multiple, pre-specified, items be used to address specific other risks of bias. Such author-specified items may be for studies as a whole or for individual (or grouped) outcomes within every study. Adding new items involves specifying a question that should be answerable as 'Yes' to indicate a low risk of bias.

Table 8.5.a: The Cochrane Collaboration's tool for assessing risk of bias

Domain	Description	Review authors' judgment
Sequence generation	Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups.	Was the allocation sequence adequately generated?
Allocation concealment	Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen in advance of, or during, recruitment.	Was allocation adequately concealed?
Blinding of participants, personnel and outcome assessors *Assessments should be made for each main outcome (or class of outcomes)*	Describe all measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.	Was knowledge of the allocated intervention adequately prevented during the study?
Incomplete outcome data *Assessments should be made for each main outcome (or class of outcomes)*	Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. State whether attrition and exclusions were reported, the numbers (compared with total randomized participants), reasons for attrition/exclusions where reported, and any re-inclusions in analyses performed by the review authors.	Were incomplete outcome data adequately addressed?
Selective outcome reporting	State how the possibility of selective outcome reporting was examined by the review authors, and what was found.	Are reports of the study free of suggestion of selective outcome reporting?
Other sources of bias	State any important concerns about bias not addressed in the other items in the tool. If particular questions/items were pre-specified in the review's protocol, responses should be provided for each question/item.	Was the study apparently free of other problems that could put it at a high risk of bias?

8.5.2 The description

The description provides a succinct summary from which judgments of risk of bias can be made. For a specific study, information for the description will often come from a single published study report, but may be obtained from a mixture of study reports, protocols, published comments on the study and contacts with the investigators. Where appropriate, the description should include verbatim quotes from reports or correspondence. Alternatively, or in addition, it may include a summary of known facts, or a comment from the review authors. In particular, it should include other information that influences any judgments made (such as knowledge of other studies performed by the same investigators). A helpful construction to supplement an ambiguous quote is to state 'Probably done' or 'Probably not done', providing the rationale for such assertions. When no information is available from which to make a judgment, this should be stated explicitly. Examples of proposed formatting for the description are provided in Table 8.5.b.

Table 8.5.b: Examples of summary descriptions for Sequence generation item

Sequence generation	Comment: no information provided.
Sequence generation	Quote: "patients were randomly allocated.
Sequence generation	Quote: "patients were randomly allocated". Comment: Probably done, since earlier reports from the same investigators clearly describe use of random sequences (Cartwright 1980).
Sequence generation	Quote: "patients were randomly allocated". Comment: Probably not done, as a similar trial by these investigators included the same phrase yet used alternate allocation (Winrow 1983).
Sequence generation	Quote (from report): "patients were randomly allocated". Quote (from correspondence): "Randomization was performed according to day of treatment". Comment: Not randomized.

8.5.3 The judgment

Review authors' judgments involve answering a specific question for each item. In all cases**, an answer 'Yes' indicates a low risk of bias,** and **an answer 'No' indicates high risk of bias**.

Table 8.5.c provides criteria for making judgments about risk of bias from each of the six domains in the tool. If insufficient detail is reported of what happened in the study, the judgment will usually be 'Unclear' risk of bias. An 'Unclear' judgment should also be made if (i) what happened in the study is known, but the risk of bias in unknown; or (ii) an item is not relevant to the study at hand (particularly for assessing blinding and incomplete outcome data, when outcomes may not have been measured in the study).

Table 8.5.c: Criteria for judging risk of bias in the 'Risk of bias' assessment tool

SEQUENCE GENERATION **Was the allocation sequence adequately generated? [Short form: *Adequate sequence generation*?]**	
Criteria for a judgment of 'YES' (i.e. low risk of bias)	The investigators describe a random component in the sequence generation process such as: • Referring to a random number table; • Using a computer random number generator; • Coin tossing; • Shuffling cards or envelopes; • Throwing dice; • Drawing of lots; • Minimization*. *Minimisation may be implemented without a random element, and this is considered to be equivalent to being random.
Criteria for the judgment of 'NO' (i.e. high risk of bias)	The investigators describe a non-random component in the sequence generation process. Usually, the description would involve some systematic, non-random approach, for example: • Sequence generated by odd or even date of birth; • Sequence generated by some algorithm based on date (or day) of admission; • Sequence generated by some algorithm based on hospital or clinic record number. Other non-random approaches happen much less frequently than the systematic approaches mentioned above and tend to be obvious. They usually involve judgment or some method of non-random categorization of participants, for example: • Allocation by judgment of the clinician;

	• Allocation by preference of the participant; • Allocation based on the results of a laboratory test or a series of tests; • Allocation by availability of the intervention.
Criteria for the judgment of 'UNCLEAR' (uncertain risk of bias)	Insufficient information about the sequence generation process to permit judgment of 'Yes' or 'No'.
ALLOCATION CONCEALMENT **Was allocation adequately concealed? [Short form: *Allocation concealment*?]**	
Criteria for a judgment of 'YES' (i.e. low risk of bias)	Participants and investigators enrolling participants could not foresee assignment because one of the following, or an equivalent method, was used to conceal allocation: • Sequentially numbered drug containers of identical appearance; • Central allocation (including web-based, and pharmacy-controlled, randomization); • Sequentially numbered, opaque, sealed envelopes.
Criteria for the judgment of 'NO' (i.e. high risk of bias)	Participants or investigators enrolling participants could possibly foresee assignments and thus introduce selection bias, such as allocation based on: • Using an open random allocation schedule; • Assignment envelopes were used without appropriate safeguards (for example if envelopes were unsealed or non-opaque or not sequentially numbered); • Alternation or rotation; • Date of birth; • Case record number; • Any other explicitly unconcealed procedure.
Criteria for the judgment of 'UNCLEAR' (uncertain risk of	Insufficient information to permit judgment of 'Yes' or 'No'. This is usually the case if the method of concealment is not described or not described in sufficient detail to allow a definite judgment – for example if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially

bias)	numbered, opaque and sealed.
BLINDING OF PARTICIPANTS, PERSONNEL AND OUTCOME ASSESSORS **Was knowledge of the allocated interventions adequately prevented during the study? [Short form: *Blinding*?]**	
Criteria for a judgment of 'YES' (i.e. low risk of bias)	Any one of the following: • No blinding, but the review authors judge that the outcome and the outcome measurement are not likely to be influenced by lack of blinding; • Blinding of participants and key study personnel ensured, and unlikely that the blinding could have been broken; • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the non-blinding of others unlikely to introduce bias.
Criteria for the judgment of 'NO' (i.e. high risk of bias)	Any one of the following: • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding; • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken; • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias.
Criteria for the judgment of 'UNCLEAR' (uncertain risk of bias)	Any one of the following: • Insufficient information to permit judgment of 'Yes' or 'No'; • The study did not address this outcome.
INCOMPLETE OUTCOME DATA **Were incomplete outcome data adequately addressed? [Short form: *Incomplete outcome data addressed*?]**	
Criteria for a judgment of 'YES'	Any one of the following:

(i.e. low risk of bias)	• No missing outcome data; • Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring unlikely to be introducing bias); • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups; • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to impact to any clinically relevant extent on the intervention effect estimate; • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to impact to any clinically relevant extent on observed effect size.
Criteria for the judgment of 'NO' (i.e. high risk of bias)	Any one of the following: • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups; • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce clinically relevant bias in intervention effect estimate; • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size; • 'As-treated' analysis with substantial departure of the intervention received from that assigned at randomization; • Potentially inappropriate application of simple imputation.
Criteria for the judgment of 'UNCLEAR' (uncertain risk of bias)	Any one of the following: • Insufficient reporting of attrition/exclusions to permit judgment of 'Yes' or 'No' (e.g. number randomized not stated, no reasons for missing data provided); • The study did not address this outcome.

SELECTIVE OUTCOME REPORTING

Are reports of the study free of suggestion of selective outcome reporting? [Short form: *Free of selective reporting*?]

Criteria for a judgment of 'YES' (i.e. low risk of bias)	Any of the following: • The study protocol is available and all of the studies' pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way; • The study protocol is not available but it is clear that the published reports include all of the study's pre-specified outcomes and all expected outcomes that are of interest in the review (convincing text of this nature may be uncommon).
Criteria for the judgment of 'NO' (i.e. high risk of bias)	Any one of the following: • Not all of the study's pre-specified primary outcomes have been reported; • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data that were not pre-specified; • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect); • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis; • The study report fails to include results for a key outcome that would be expected to have been reported for such a study.
Criteria for the judgment of 'UNCLEAR' (uncertain risk of bias)	Insufficient information to permit judgment of 'Yes' or 'No'. It is likely that the majority of studies will fall into this category.

TOPIC-SPECIFIC, DESIGN-SPECIFIC OR OTHER POTENTIAL THREATS TO VALIDITY

Was the study apparently free of other problems that could put it at a risk of bias? [Short form: *Free of other bias*?]

Criteria for a judgment of 'YES' (i.e. low risk of bias)	The study appears to be free of other sources of bias.
Criteria for the judgment of 'NO' (i.e. high risk of bias)	There is at least one important risk of bias. For example, the study: • Had a potential source of bias related to the specific study design used; or

	• Stopped early due to some data-dependent process (including a formal-stopping rule); or • Had extreme baseline imbalance; or • Has been claimed to have been fraudulent; or • Had some other problem.
Criteria for the judgment of 'UNCLEAR' (uncertain risk of bias)	There may be a risk of bias, but there is either • Insufficient information to assess whether an important risk of bias exists; or • Insufficient rationale or evidence that an identified problem will introduce bias.

8.6 Presentation of assessments of risk of bias

A 'Risk of bias' table is available in a Cochrane review as part of the 'Table of characteristics of included studies'. For each question-based item, the judgment ('Yes' for low risk of bias; 'No' for high risk of bias, or 'Unclear') is followed by a text box providing a description of the design, conduct or observations that underlie the judgment. Figure 8.6.a provides an example of how it might look.

Figure 8.6.a: Example of a 'Risk of bias' table for a single study (fictional)

Item	Judgment	Description
Adequate sequence generation?	Yes	Quote: "patients were randomly allocated." Comment: Probably done done, since earlier reports from the same investigators clearly describe use of random sequences (Cartwright 1980).
Allocation concealment?	No	Quote: "...using a table of random numbers." Comment: Probably not done.
Blinding? (Patient-reported outcomes)	Yes	Quote: "double blind, double dummy"; "High and low dose tablets or capsules were indistinguishable in all aspects of their outward appearance. For each drug an identically matched placebo was available (the success of blinding was evaluated by examining the drugs before distribution)." Comment: Probably done.
Blinding? (Mortality)	Yes	Obtained from medical records; review authors do not believe this will introduce bias.
Incomplete outcome data addressed? (Short-term outcomes (2-6 wks))	No	4 weeks: 17/110 missing from intervention group (9 due to 'lack of efficacy'); 7/113 missing from control group (2 due to 'lack of efficacy').
Incomplete outcome data addressed? (Longer-term outcomes (>6 wks))	No	12 weeks: 31/110 missing from intervention group; 18/113 missing from control group. Reasons differ across groups.
Free of selective reporting?	No	Three rating scales for cognition listed in Methods, but only one reported.
Free of other bias?	No	Trial stopped early due to apparent benefit.

Considerations for presentation of 'Risk of bias' assessments in the review text are discussed in Chapter 4 (Section 4.5) (under the Results sub-heading 'Risk of bias in included studies' and the Discussion recommended sub-heading 'Quality of the evidence').

Two Figures may be generated using RevMan for inclusion in a published review. A 'Risk of bias graph' Figure illustrates the proportion of studies with each of the judgments ('Yes', 'No', 'Unclear') across for each item in the tool (see Figure 8.6.b). A 'Risk of bias summary' Figure presents all of the judgments in a cross-tabulation of study by item (see Figure 8.6.c).

Figure 8.6.b: Example of a 'Risk of bias graph' Figure

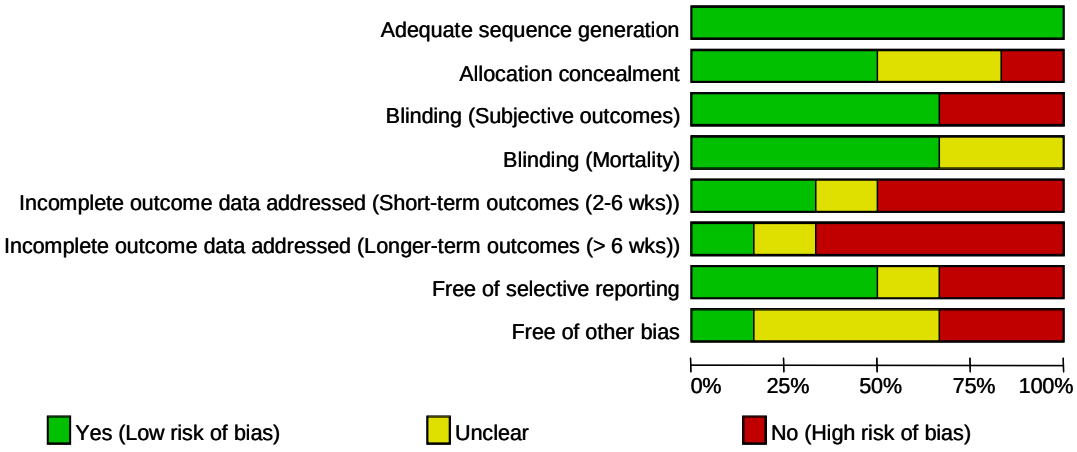
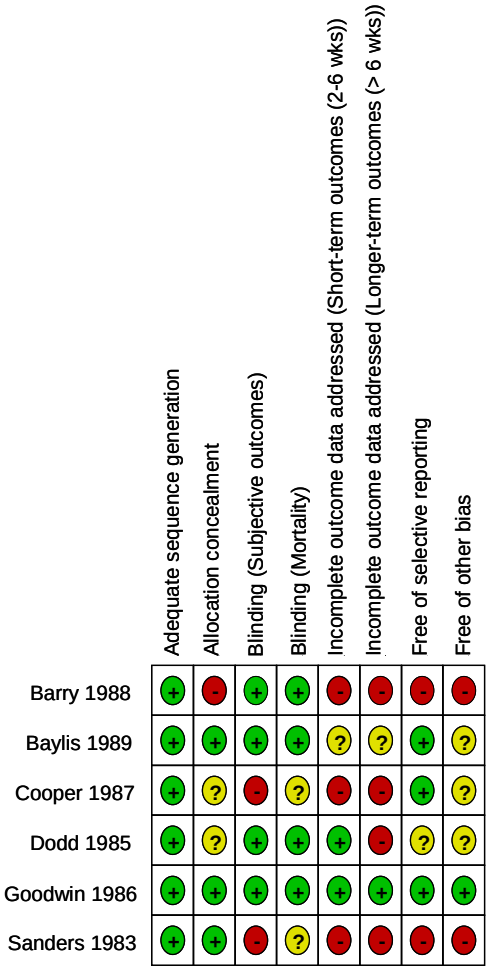


Figure 8.6.c: Example of a 'Risk of bias summary' Figure



8.7 Summary assessments of risk of bias

The Collaboration's recommended tool for assessing risk of bias in included studies involves the assessment and presentation of individual items, such as allocation concealment and blinding. To draw conclusions about the overall risk of bias for an outcome it is necessary to summarize these. The use of scales (in which scores for multiple items are added up to produce a total) is discouraged for reasons outlined in Section 8.3.1.

Nonetheless, any assessment of the overall risk of bias involves consideration of the relative importance of different items. A review author will have to make judgments about which items are most important in the current review. For example, for highly subjective outcomes such as pain, authors may decide that blinding of participants is critical. Such judgments should be explicit and they should be informed by:

- **Empirical evidence of bias**: Sections 8.5 to 8.14 summarize empirical evidence of the association between items such as allocation concealment and blinding and estimated magnitudes of effect. However, the evidence base remains incomplete.
- **Likely direction of bias**: The available empirical evidence suggests that failure to meet most items, such as adequate allocation concealment, is associated with overestimates of effect. If the likely direction of bias for an item is such that effects will be underestimated (biased towards the null), then, providing the review demonstrates an important effect of the intervention, such an item may be of less concern.
- **Likely magnitude of bias**: The likely magnitude of bias associated with any item may vary. For example, the magnitude of bias associated with inadequate blinding of participants is likely to be greater for more subjective outcomes. Some indication of the likely magnitude of bias may be provided by the empirical evidence base (see above), but this does not yet provide clear information on the particular scenarios in which biases may be large or small. It may, however, be possible to consider the likely magnitude of bias relative to the estimated magnitude of effect. For example, inadequate allocation concealment and a small estimate of effect might substantially reduce one's confidence in the estimate, whereas minor inadequacies in how incomplete outcome data were addressed might not substantially reduce one's confidence in a large estimate of effect.

Summary assessment of risk of bias might be considered at four levels:

- **Summarizing risk of bias for a study across outcomes**: Some items affect the risk of bias across outcomes in a study; e.g. sequence generation and allocation concealment. Other items, such as blinding and incomplete outcome data, may have different risks of bias for different outcomes within a study. Thus, review authors should be cautious about assuming that the risk of bias is the same for all outcomes in a study. Moreover, a summary assessment of the risk of bias across all outcomes for a study is generally of little interest.
- **Summarizing risk of bias for an outcome within a study (across items)**: This is the recommended level at which to summarize the risk of bias in a study, because some risks of bias may be different for different outcomes. A summary assessment of the risk of bias for an outcome should include all of the items relevant to that outcome; i.e. both study-level items, such as allocation concealment, and outcome specific items, such as blinding.
- **Summarizing risk of bias for an outcome across studies (e.g. for a meta-analysis)**: These are the main summary assessments that will be made by review authors and incorporated into judgements about the 'quality of evidence' in 'Summary of findings' tables, as described in Chapter 11 (Section 11.5).
- **Summarizing risk of bias for a review as a whole (across studies and outcomes)**: It may be tempting to summarize the overall risk of bias in a review, but this should be avoided for two reasons. First, this requires value judgements about which outcomes are critical to a decision and, therefore, should be included in this assessment. Frequently no data are available from the studies included in a review for some outcomes that may be critical, such as adverse effects, and the risk of bias is rarely the same across all of the outcomes that are critical to such an assessment. Second, judgements about which outcomes are critical to a decision may vary from setting to setting, both due to differences in values and due to differences in other factors, such as baseline risk. Thus, judgements about the overall risk of bias of evidence across studies and outcomes should be made in a specific context, for example in the context of clinical practice guidelines, and not in the context of systematic reviews that are intended to inform decisions across a variety of settings.

Review authors should make explicit judgements about the risk of bias for important outcomes both within and across studies. This requires identifying the most important items ('key items') that feed into these summary assessments. Table 8.7.a provides a possible approach to making summary assessments of the risk of bias for important outcomes within and across studies.

Table 8.7.a: Possible approach for *summary assessments* of the risk of bias for each important outcome (across items) within and across studies

Risk of bias	Interpretation	Within a study	Across studies
Low risk of bias	Plausible bias unlikely to seriously alter the results.	Low risk of bias for all key items.	Most information is from studies at low risk of bias.
Unclear risk of bias	Plausible bias that raises some doubt about the results.	Unclear risk of bias for one or more key items.	Most information is from studies at low or unclear risk of bias.
High risk of bias	Plausible bias that seriously weakens confidence in the results.	High risk of bias for one or more key items.	The proportion of information from studies at high risk of bias is sufficient to affect the interpretation of results.

8.8 Incorporating assessments into analyses

8.8.1 Introduction

Statistical considerations often involve a trade-off between bias and precision. A meta-analysis that includes all eligible studies may produce a result with high precision (narrow confidence interval), but be seriously biased because of flaws in the conduct of the studies. On the other hand, including only the studies at low risk of bias in all domains assessed may produce a result that is unbiased but imprecise (if there are only a few high-quality studies).

When performing and presenting meta-analyses, review authors must address risk of bias in the results of included studies. It is not appropriate to present analyses and interpretations based on all studies, ignoring flaws identified during the assessment of risk of bias. The higher the proportion of studies assessed to be at high risk of bias, the more cautious should be the analysis and interpretation of their results.

8.8.2 Exploring the impact of risk of bias

8.8.2.1 Graphing results according to risk of bias

In the discussion that follows, we refer to comparisons of results according to individual bias domains. However, such comparisons can also be made according to risk of bias summarized at the study level (see Section 8.7).

Plots of intervention effect estimates (e.g. forest plots) stratified according to risk of bias are likely to be a useful way to begin examining the potential for bias to affect the results of a meta-analysis. Forest

plots ordered by judgments on each 'Risk of bias' item are available in RevMan 5. Such plots give a visual impression both of the relative contributions of the studies at low, unclear and high risk of bias, and also of the extent of differences in intervention effect estimates between studies at low, unclear and high risk of bias. It will usually be sensible to restrict such plots to key bias items (see Section 8.7).

8.8.2.2 Studies assessed as at unclear risk of bias

Studies are assessed as at unclear risk of bias when too few details are available to make a judgement of 'high' or 'low' risk, when the risk of bias is genuinely unknown despite sufficient information about the conduct, or when an item is not relevant to a study. When the first reason dominates, it is reasonable to assume that the average bias in results from such studies will be less than in studies at high risk of bias, because the conduct of some studies assessed as unclear will in fact have avoided bias. Limited evidence from empirical studies that examined the 'high' and 'unclear' categories separately confirms this: for example, the study of Schulz et al found that intervention odds ratios were exaggerated by 41% for inadequately concealed trials (high risk of bias) and by 30% for unclearly concealed trials (unclear risk of bias) (Schulz 1995b). However, most empirical studies have combined the 'high' and 'unclear' categories, which were then compared with the 'low' category.

It is recommended that review authors do not combine studies at 'low' and 'unclear' risk of bias in analyses, unless they provide specific reasons for believing that these studies are likely to have been conducted in a manner that avoided bias. In the rest of this section, we will assume that studies assessed as at low risk of bias will be treated as a separate category.

8.8.2.3 Meta-regression and comparisons of subgroups

Formal comparisons of intervention effects according to risk of bias can be done using meta-regression (see Chapter 9). For studies with dichotomous outcomes, results of meta-regression analyses are most usefully expressed as ratios of odds ratios (or risk ratios) comparing results of studies at high or unclear risk of bias with those of studies at low risk of bias.

$$\text{Ratio of odds ratios} = \frac{\text{Intervention odds ratio in studies at high or unclear risk of bias}}{\text{Intervention odds ratio in studies at low risk of bias}}$$

Alternatively, separate comparisons of high versus low and unclear versus low can be made. For studies with continuous outcomes (e.g. blood pressure), intervention effects are expressed as mean differences between intervention groups, and results of meta-regression analyses correspond to differences of mean differences.

If the estimated effect of the intervention is the same in studies at high and unclear risk of bias as in studies at low risk of bias then the ratio of odds ratios (or risk ratios) equals 1, while the difference between mean differences will equal zero. As explained in Section 8.2.3, empirical evidence from collections of meta-analyses assembled in meta-epidemiological studies suggests that, on average, intervention effect estimates tend to be more beneficial in studies at high or unclear risk of bias than in studies at low risk of bias.

When a meta-analysis includes sufficient studies, meta-regression analyses can include more than one item (e.g. both allocation concealment and blinding).

Results of meta-regression analyses include a confidence interval for the ratio of odds ratios, and a P value for the null hypothesis that there is no difference between the results of studies at high or unclear and low risk of bias. Because meta-analyses usually contain a small number of studies, the ratio of

odds ratios is usually imprecisely estimated. It is therefore important not to conclude, on the basis of a non-significant P value, that there is no difference between the results of studies at high or unclear and low risk of bias, and therefore no impact of bias on the results. Examining the confidence interval will often show that the difference between studies at high or unclear and low risk of bias is consistent with both no and a substantial effect of bias.

A test for differences across subgroups provides an alternative to meta-regression for examination of a single item (e.g. comparing studies with adequate verses inadequate allocation concealment). Within a fixed-effect meta-analysis framework, such tests are available in RevMan 5. However, such P values are of limited use without corresponding confidence intervals, and they will in any case be too small in the presence of heterogeneity either within or between subgroups.

8.8.3 Including 'Risk of bias' assessments in analyses

Broadly speaking, studies at high or unclear risk of bias should be given reduced weight in meta-analyses, compared with studies at low risk of bias (Spiegelhalter 2003). However formal statistical methods to combine the results of studies at high and low risk of bias are not sufficiently well developed that they can currently be recommended for use in Cochrane reviews (see Section 8.8.4.2). Therefore, the major approach to incorporating risk of bias assessments in Cochrane reviews is to **restrict** meta-analyses to studies at low (or lower) risk of bias.

8.8.3.1 Possible analysis strategies

When risks of bias vary across studies in a meta-analysis, three broad strategies are available for choosing which result to present as the main finding for a particular outcome (for instance, in deciding which result to present in the Abstract).

1. Present all studies and provide a narrative discussion of risk of bias

The simplest approach to incorporating bias assessments in results is to present an estimated treatment effect based on all available studies, together with a description of the risk of bias in individual domains, or a description of the summary risk of bias, across studies. This is the only feasible option when all studies are at high risk, all are at unclear risk or all are at low risk of bias. However, when studies have different risks of bias, we discourage such an approach for two reasons. First, detailed descriptions of risk of bias in the results section, together with a cautious interpretation in the discussion section, will often be lost in the conclusions, abstract and summary of findings, so that the final interpretation ignores the risk of bias. Second, such an analysis fails to down-weight studies at high risk of bias and hence will lead to an overall treatment that is too precise as well as being potentially biased.

2. Primary analysis restricted to studies at low (or low and unclear) risk of bias

The second approach involves defining a threshold, based on key bias domains (see Section 8.7) such that only studies meeting specific criteria are included in the primary analysis. The threshold may be determined using the original review inclusion criteria, or using reasoned argument (which may draw on empirical evidence of bias from meta-epidemiological studies). If the primary analysis includes studies at unclear risk of bias, review authors must provide justification for this choice. Ideally the threshold, or the method for determining it, should be specified in the review protocol. Authors should keep in mind that all thresholds are arbitrary, and that studies may in theory lie anywhere on the spectrum from 'free of bias' to 'undoubtedly biased'. The higher the threshold, the more similar the studies will be in their risks of bias, but they may end up being few in numbers.

Having presented a restricted primary analysis, review authors are encouraged to perform **sensitivity analyses** showing how conclusions might be affected if more or all studies at high risk of bias were included in analyses. When analyses are presented that include studies judged to be at high risk of bias, review authors must present these judgments alongside their presentation of results.

3. Present multiple analyses

Two or more analyses incorporating different inclusion criteria might be presented with equal prominence, for example one including all studies and one including only those at low risk of bias. This avoids the need to make a difficult decision, but may be confusing for readers. In particular, people who need to make a decision usually require a single estimate of effect. Further, 'Summary of findings' tables will usually only present a single result for each outcome.

8.8.4 Other methods for addressing risk of bias

8.8.4.1 Direct weighting

Methods have been described for weighting studies in the meta-analysis according to their validity or risk of bias (Detsky 1992). The usual statistical methods for combining results of multiple studies weight studies by the amount of information they contribute (more specifically, by the inverse variances of their effect estimates). This gives studies with more precise results (narrower confidence intervals) more weight. It is also possible to weight studies according to validity, so that more valid studies have more influence on the summary result. A combination of inverse variances and validity assessments can be used. The main objection to this approach is that it requires a numerical summary of validity for each study, and there is no empirical basis for determining how much weight to assign to different domains of bias. Furthermore, the resulting weighted average will be biased if some of the studies are biased. Direct weighting of effect estimates by validity or assessments of risk of bias should be avoided (Greenland 2001).

8.8.4.2 Bayesian approaches

Bayesian analyses allow for the incorporation of external information or opinion on the nature of bias (see Chapter 16). Prior distributions for specific biases in intervention effect estimates might be based on empirical evidence of bias, on elicited prior opinion of experts, or on reasoned argument. Bayesian methods for adjusting meta-analyses for biases are a subject of current research; they are not currently sufficiently well developed for widespread adoption.

8.9 Sequence generation

8.9.1 Rationale for concern about bias

The starting point for an unbiased intervention study is the use of a mechanism that ensures that the same sorts of participants receive each intervention. Several interrelated processes need to be considered. First, an allocation rule must be used that, if perfectly implemented, would balance prognostic factors evenly across intervention groups. Randomization plays a fundamental role here. It can be argued that other assignment rules, such as alternation or rotation, can also fulfil this criterion (Hill 1990). However, a theoretically unbiased rule is insufficient to prevent bias in practice. If future assignments can be anticipated, either by predicting them or by knowing them, then selection bias can arise due to the selective enrolment of participants into a study in the light of the upcoming intervention assignment.

Future assignments can be anticipated for several reasons. These include (i) knowledge of a deterministic assignment rule, such as by alternation, date of birth or day of admission; (ii) knowledge of the sequence of assignments, whether randomized or not (e.g. if a sequence of random assignments

is posted on the wall); (iii) ability to predict assignments successfully, based on previous assignments (which may sometimes be possible when randomization methods are used that attempt to ensure an exact ratio of allocations to different interventions). Complex interrelationships between theoretical and practical aspects of allocation in intervention studies make the assessment of selection bias challenging. Perhaps the most important among the practical aspects is concealment of allocation, that is the use of mechanisms to prevent foreknowledge of the next assignment. This has historically been assessed in Cochrane reviews, with empirical justification. We address allocation concealment as a separate domain in the tool (see Section 8.10).

Under the domain of sequence generation, we address whether or not the study used a randomized sequence of assignments. Randomization allows for the sequence to be unpredictable. An unpredictable sequence, combined with allocation concealment, should be sufficient to prevent selection bias. However, selection bias may arise despite randomization if the random allocations are not concealed, and selection bias may (in theory at least) arise despite allocation concealment if the underlying sequence is not random. We acknowledge that a randomized sequence is not always completely unpredictable, even if mechanisms for allocation concealment are in place. This may sometimes be the case, for example, if blocked randomization is used, and all allocations are known after enrolment. Nevertheless, we do not consider this special situation under either sequence generation or allocation concealment, and address it as a separate consideration in Section 8.14.1.6.

Methodological studies have assessed the importance of sequence generation. At least three of those studies have avoided confounding by disease or intervention, which is critical to the assessment (Schulz 1995b, Moher 1998, Kjaergard 2001). The inadequate generation of allocation sequences was not consistently associated with biased treatment effects across the studies.

In one study that restricted the analysis to 79 trials that had reported adequately concealed allocation, trials with inadequate sequence generation yielded more beneficial estimates of treatment effects, on average, than trials with adequate sequence generation (relative odds ratio of 0.75; 95% CI of 0.55 to 1.02; p=0.07). This suggests that if assignments are non-random, some deciphering of the sequence can occur, even with apparently adequate allocation concealment (Schulz 1995b).

8.9.2 Assessing risk of bias in relation to adequate or inadequate sequence generation

Sequence generation is often improperly addressed in the design and implementation phases of RCTs, and is often neglected in published reports, which causes major problems in assessing the risk of bias. The following considerations may help review authors assess whether sequence generation is suitable to protect against bias using the Collaboration's tool (Section 8.5).

8.9.2.1 Adequate methods of sequence generation

The use of a random component should be sufficient for adequate sequence generation.

Randomization with no constraints is called **simple randomization** or **unrestricted randomization**. This can be achieved by allocating interventions entirely by chance, using methods such as repeated coin-tossing, throwing dice or dealing previously shuffled cards (Schulz 2002c, Schulz 2006). More usually it is achieved by referring to a published list of random numbers, or to a list of random assignments generated by a computer. In trials using large samples (large usually meaning at least 100 in each randomized group (Schulz 2002d, Schulz 2002c, Schulz 2006), simple randomization generates unbiased comparison groups of relatively similar sizes. In trials using small samples, simple

randomization will sometimes result in groups that differ, by chance, quite substantially in size or in the occurrence of prognostic factors (i.e. 'case-mix' variation) (Altman 1999).

Example (of low risk of bias): We generated the two comparison groups using simple randomization, with an equal allocation ratio, by referring to a table of random numbers.

Sometimes **restricted randomization** is used to ensure balance between intervention groups. Blocked randomization (random permuted blocks) is a common form of restricted randomization (Schulz 2002c, Schulz 2006). Blocking ensures that the numbers of participants to be assigned to each of the comparison groups will be balanced within blocks of, for example, five in one group and five in the other for every 10 consecutively entered participants. The block size may be randomly varied to reduce the likelihood of foreknowledge of treatment assignment.

Example (of low risk of bias): We used blocked randomization to form the allocation list for the two comparison groups. We used a computer random number generator to select random permuted blocks with a block size of 8 and an equal allocation ratio.

Also common is stratified randomization, in which restricted randomization is performed separately within strata. This generates separate randomization schedules for subsets of participants defined by potentially important prognostic factors, such as disease severity and study centres. If simple (rather than restricted) randomization was used in each stratum, then stratification would be ineffective but the randomization would still be valid. Risk of bias may be judged in the same way whether or not a trial claims to have stratified.

Another approach that incorporates both the general concepts of stratification and restricted randomization is minimization, which can be used to make small groups closely similar with respect to several characteristics. The use of minimization should not be considered to put a study at risk of bias. However, some methodologists remain cautious about the acceptability of minimization, particularly when it is used without any random component, while others consider it to be very attractive (Brown 2005).

Other adequate types of randomization that are sometimes used are biased coin or urn randomization, replacement randomization, mixed randomization, and maximal randomization (Schulz 2002c, Schulz 2002d, Berger 2003). If these or other approaches are encountered, consultation with a statistician may be necessary.

8.9.2.2 Inadequate methods of sequence generation

Systematic methods, such as alternation, assignment based on date of birth, case record number, and date of presentation are sometimes referred to as 'quasi-random'. Alternation (or rotation, for more than two intervention groups) might in principle result in similar groups, but many other systematic methods of sequence generation may not. For example, the day on which a patient is admitted to hospital is not solely a matter of chance, and allocation made using case record number could be seriously problematic if this number always ends in 1 for men and 2 for women.

An important weakness with all systematic methods is that concealing the allocation schedule is usually impossible, which allows foreknowledge of treatment assignment among those recruiting participants to the study, and biased allocations (see Section 8.10).

Example (of high risk of bias): We allocated patients to the intervention group based on the week of the month.

Example (of high risk of bias): Patients born on even days were assigned to Treatment A and patients born on odd days were assigned to Treatment B.

8.9.2.3 Methods of sequence generation with unclear risk of bias

A simple statement such as 'we randomly allocated' or 'using a randomized design' is often insufficient to be confident that the allocation sequence was genuinely randomized. It is not uncommon for authors to use the term 'randomized' even when it is not justified – many trials with declared systematic allocation are described by the authors as randomized. If there is doubt, then the adequacy of sequence generation should be considered to be unclear.

Sometimes trial authors provide some information, but they incompletely define their approach and do not confirm some random component in the process. For example, authors may state that blocked randomization was used, but the process of selecting the blocks, such as a random number table or a computer random number generator, was not specified. The adequacy of sequence generation should then be classified as unclear.

8.10 Allocation concealment

8.10.1 Rationale for concern about bias

Randomized sequence generation is a necessary but not a sufficient safeguard against bias in treatment allocation. Efforts made to generate unpredictable and unbiased sequences are likely to be ineffective if those sequences are not protected by adequate allocation concealment from those involved in the enrolment and assignment of participants.

Knowledge of the next assignment – for example, from a table of random numbers openly posted on a bulletin board – can cause selective enrolment of participants on the basis of prognostic factors. Participants who would have been assigned to an intervention deemed to be 'inappropriate' may be rejected. Other participants may be deliberately directed to the 'appropriate' intervention, which can often be accomplished by delaying a participant's entry into the trial until the next appropriate allocation appears. Deciphering of allocation schedules may occur even if concealment was attempted. For example, unsealed allocation envelopes may be opened, while translucent envelopes may be held against a bright light to reveal the contents (Schulz 1995b, Schulz 1995a, Jüni 2001). Personal accounts suggest that many allocation schemes have been deciphered by investigators simply because the methods of concealment were inadequate (Schulz 1995a).

Avoidance of such selection biases depends on preventing foreknowledge of intervention assignment. Decisions on participants' eligibility and their decision whether to give informed consent should be made in ignorance of the upcoming assignment. Adequate **concealment of allocation** shields those who admit participants to a study from knowing the upcoming assignments.

Several methodological studies have looked at whether concealment of allocation is associated with magnitude of effect estimates in controlled clinical trials while avoiding confounding by disease or intervention (Schulz 1995b, Moher 1998, Kjaergard 2001, Balk 2002, Egger 2003, Pildal 2007). A pooled analysis of seven methodological studies found that effect estimates from trials with inadequate concealment of allocation or unclear reporting of the technique used for concealment of allocation were on average 18% more beneficial than effect estimates from trials with adequate concealment of

allocation (95% confidence interval 5 to 29%) (Pildal 2007). A recent detailed analysis of three of these data sets combined (1346 trials from 146 meta-analyses) sheds some light on the heterogeneity of these studies. Intervention effect estimates were exaggerated when there was inadequate allocation concealment in trials where a subjective outcome was analyzed, but there was little evidence of bias in trials with objective outcomes (Wood 2006).

8.10.2 Assessing risk of bias in relation to adequate or inadequate allocation concealment

The following considerations may help review authors assess whether concealment of allocation was sufficient to protect against bias using the Collaboration's tool (Section 8.5).

Proper allocation concealment secures strict implementation of an allocation sequence without foreknowledge of intervention assignments. Methods for allocation concealment refer to techniques used to implement the sequence, **not** to generate it (Schulz 1995b). However, most allocation *sequences* that are deemed inadequate, such as allocation based on day of admission or case record number, cannot be adequately concealed, and so fail on both counts. It is theoretically possible, yet unlikely, that an inadequate sequence is adequately concealed (the person responsible for recruitment and assigned interventions would have to be unaware that the sequence being implemented was inappropriate). However, it is not uncommon for an adequate (i.e. randomized) allocation sequence to be inadequately concealed, for example if the sequence is posted on the staff room wall.

Some review authors confuse allocation concealment with blinding of allocated treatments. Allocation concealment seeks to prevent selection bias by protecting the allocation sequence **before and until** assignment, and can always be successfully implemented regardless of the study topic (Schulz 1995b, Jüni 2001). In contrast, blinding seeks to prevent performance and detection bias by protecting the sequence **after** assignment (Jüni 2001, Schulz 2002a), and cannot always be implemented – for example, in trials comparing surgical with medical treatments. Thus, allocation concealment up to the point of assignment of the intervention and blinding after that point address different sources of bias and differ in their feasibility.

The importance of allocation concealment may depend on whether strong beliefs exist among investigators and participants regarding the benefits or harms of assigned interventions, or whether equipoise of interventions is accepted by all people involved (Schulz 1995a). Among the different methods used to conceal allocation, those using envelopes are more susceptible to manipulation than other approaches (Schulz 1995b). If investigators use envelopes, they should develop and monitor the allocation process to preserve concealment. In addition to use of sequentially numbered, opaque, sealed envelopes, they should ensure that the envelopes are opened sequentially, and only after the envelope has been irreversibly assigned to the participant.

8.10.2.1 Adequate methods of allocation concealment

Table 8.10.a provides minimal criteria for a judgment of adequate concealment of allocation (left) and extended criteria, which provide additional assurance that concealment of allocation was indeed adequate (right).

Examples (of low risk of bias) [published descriptions of concealment procedures judged to be adequate, as compiled by Schulz and Grimes (Schulz 2002b)]:

" . . . that combined coded numbers with drug allocation. Each block of ten numbers was transmitted from the central office to a person who acted as the randomization authority in each centre. This individual (a pharmacist or a nurse not involved in care of the trial patients and independent of the

site investigator) was responsible for allocation, preparation, and accounting of trial infusion. The trial infusion was prepared at a separate site, then taken to the bedside nurse every 24 h. The nurse infused it into the patient at the appropriate rate. The randomization schedule was thus concealed from all care providers, ward physicians, and other research personnel." (Bellomo 2000).

". . . concealed in sequentially numbered, sealed, opaque envelopes, and kept by the hospital pharmacist of the two centres." (Smilde 2001).

"Treatments were centrally assigned on telephone verification of the correctness of inclusion criteria . . ." (de Gaetano 2001).

"Glenfield Hospital Pharmacy Department did the randomization, distributed the study agents, and held the trial codes, which were disclosed after the study." (Brightling 2000).

Table 8.10.a: Minimal and extended criteria for judging concealment of allocation to be adequate (low risk of bias)

Minimal criteria for a judgment of adequate concealment	Extended criteria providing additional assurance
Sequentially numbered, opaque, sealed envelopes	Envelopes were sequentially numbered and opened sequentially only after participant details were written on the envelope. Pressure sensitive or carbon paper inside the envelope transferred the participant's details to the assignment card. Cardboard or aluminium foil inside the envelope rendered the envelope impermeable to intense light. Envelopes were sealed using tamper proof security tape.
Sequentially numbered drug containers	Drug containers prepared by an independent pharmacy were sequentially numbered and opened sequentially. Containers were of identical appearance, tamper-proof and equal in weight.
Central randomization	The central randomization office was remote from patient recruitment centres. Participant details were provided, for example, by phone, fax or email and the allocation sequence was concealed to individuals staffing the randomization office until a participant was irreversibly registered.

8.11 Blinding of participants, personnel and outcome assessors

8.11.1 Rationale for concern about bias

Blinding (sometimes called masking) refers to the process by which study participants and personnel, including people assessing outcomes, are kept unaware of intervention assignments after inclusion of

participants into the study. Blinding may reduce the risk that knowledge of which intervention was received, rather than the intervention itself, affects outcomes and assessments of outcomes.

Different types of participants and personnel can be blinded in a clinical trial (Gøtzsche 1996, Haahr 2006):

1. Participants (e.g. patients or healthy people);
2. Health care providers (e.g. the doctors or nurses responsible for care);
3. Outcome assessors, including primary data collectors (e.g. interview staff responsible for measurement or collection of outcome data) and any secondary assessors (e.g. external data endpoint committees);
4. Data analysts (e.g. statisticians);
5. Manuscript writers.

Lack of blinding of participants or health care providers could bias the results by affecting the *actual* outcomes of the participants in the trial. This may be due lack of expectations in a control group, or due to differential behaviours across intervention groups (for example, differential drop-out, differential cross-over to an alternative intervention, or differential administration of co-interventions). Lack of blinding of any of the persons in the first three categories could lead to bias in *assessments* of outcome, depending on who measures the outcomes. Lack of blinding of the last two categories may lead to reporting biases. In assessing blinding in Cochrane reviews, the emphasis should be placed on the first three in the list. Given the overlapping considerations when participants or healthcare providers are also those assessing outcomes, we consider them together.

In empirical studies, lack of blinding in randomized trials has been shown to be associated with exaggerated estimated intervention effects of 9%, on average, measured as odds ratio (Pildal 2007). These studies have dealt with a variety of outcomes, some of which are objective and would not be expected to be much influenced by lack of blinding. The estimated effect could therefore be expected to be more biased, on average, in trials with more subjective outcomes (Wood 2006). Lack of blinding might also lead to bias caused by additional investigations or co-interventions regardless of the type of outcomes.

Almost all outcome assessments can be influenced by lack of blinding, although more subjective outcomes (e.g. pain or number of days with a common cold) are particularly vulnerable. It is therefore important to consider how subjective or objective an outcome is when considering blinding. The importance of blinding and whether blinding is possible may differ across outcomes within a study. Seemingly objective assessments, e.g. doctors assessing the degree of psychological or physical impairment, can also be somewhat subjective (Noseworthy 1994).

Blinding can be impossible for at least some people (e.g. for many types of surgery). However, such studies can take other measures to reduce the risk of bias, such as treating patients according to a strict protocol to reduce the risk of differential behaviours by patients and health care providers.

An attempt to blind participants and personnel does not ensure successful blinding in practice. Blinding can be compromised for most interventions. For many double blind drug trials, the side effects of the drugs allows the possible detection of which intervention is being received for some participants, unless the study compares two rather similar interventions, e.g. drugs with similar side effects, or uses an active placebo (Boutron 2006).

8.11.2 Assessing risk of bias in relation to adequate or inadequate blinding

Study reports often describe blinding in broad terms, e.g. 'double blind', that makes it impossible to know who was blinded (Schulz 2002a). Such terms are also used very inconsistently (Devereaux 2001, Boutron 2005, Haahr 2006), and the frequency of explicit reporting of the blinding status of study participants and personnel remains low even in trials published in top journals (Montori 2002), despite recommendations to the contrary in the CONSORT Statement (Moher 2001). A review of methods used for blinding highlights the variety of methods used in practice (Boutron 2006). The following considerations may help review authors assess whether any blinding used in a study was sufficient to protect against bias using the Collaboration's tool (Section 8.5).

When considering the risk of bias from lack of blinding it is important to consider specifically:

1. Who was and was not blinded;
2. Risk of bias in actual outcomes due to lack of blinding during the study (e.g. due to co-intervention or differential behaviour);
3. Risk of bias in outcome assessments (considering how subjective or objective an outcome is);

Assessors of some outcomes may be blinded while assessors of other outcomes are not. For example, in a surgical trial in which patients are aware of their own intervention, patient-reported outcomes (e.g. quality of life) would be collected in knowledge of intervention assignment, whereas other outcomes measured by an independent clinician (e.g. physical ability) might be blinded. Thus, assessments of who was blinded may need to be made separately for different outcomes.

Furthermore, risk of bias may be high for some outcomes and low for others, even if the same people were unblinded in the trial. For example, knowledge of intervention assignment may impact on behavioural outcomes (such as number of clinic visits), while not impacting on physiological outcomes. In many circumstances assessment of total mortality might be considered to be unbiased, even if outcome assessors were aware of intervention assignments. Thus, assessments of risk of bias resulting from lack of blinding may need to be made separately for different outcomes.

Rather than assessing risk of bias for each outcome separately, it is often convenient to group outcomes with similar risks of bias (see Section 8.5). For example, there may be a common assessment of risk of bias for all subjective outcomes that is different from a common assessment of blinding for all objective outcomes.

8.12 Incomplete outcome data

8.12.1 Rationale for concern about bias

Missing outcome data, due to attrition (drop-out) during the study or exclusions from the analysis, raise the possibility that the observed effect estimate is biased. We shall use the term '**incomplete outcome data**' to refer to both attrition and exclusions. When an individual participant's outcome is not available we shall refer to it as '**missing**'.

Attrition may occur because:

- Participants withdraw, or are withdrawn, from the study;
- Participants do not attend an appointment at which outcomes should have been measured;
- Participants attend an appointment but do not provide relevant data;

- Participants fail to complete diaries or questionnaires;
- Participants cannot be located (lost to follow-up);
- The study investigators decide, usually inappropriately, to cease follow-up;
- Data or records are lost, or are unavailable for other reasons.

Exclusions from analysis may occur because:

- There was attrition from the study so that some participants could not be assessed;
- Enrolled participants were later found to be ineligible;
- An 'as-treated' (or per-protocol) analysis is performed (in which participants are included only if they received the intended intervention in accordance with the protocol; see Section 8.12.2);
- The study analysis excluded participants for other reasons.

Some exclusions of participants are justifiable, in which case they need not be considered as leading to missing outcome data (Fergusson 2002). For example, participants who are randomized but are subsequently found not to have been eligible for the trial may be excluded, as long as the discovery of ineligibility could not have been affected by the randomized intervention.

An intention-to-treat (ITT) analysis is often recommended as the least biased way to estimate intervention effects in randomized controlled trials. The principles of ITT analyses are

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.

The first principle can always be applied. However, the second is often impossible due to attrition beyond trialists' control. Consequently, the third principle of conducting an analysis that includes all participants can only be followed by making assumptions about the missing values (see below). In practice, study authors may describe an analysis as ITT even when some outcome data are missing. The term 'ITT' does not have a clear and consistent definition, and it is used inconsistently in study reports (Hollis 1999). Review authors should use the term only to imply all three of the principles above, and should interpret with care any studies that use the term without clarification.

Authors may also encounter analyses described as "modified intention-to-treat", which usually means that participants were excluded if they did not receive a specified minimum amount of the intended treatment. This term is also used in a variety of ways so review authors should always seek information about precisely who was included.

Note that it might be possible to conduct analyses that include participants who were excluded by the study authors ('**re-inclusions**'), if the reasons for exclusions are considered inappropriate and the data are available to the review author. Review authors are encouraged to do this when possible and appropriate.

Concerns over bias resulting from incomplete outcome data are driven mainly by theoretical considerations. Several empirical studies have looked at whether various aspects of missing data are associated with the magnitude of effect estimates. Most found no clear evidence of bias (Schulz 1995b, Kjaergard 2001, Balk 2002). Tierney et al. observed a tendency for analyses conducted after trial authors excluded participants to favour the experimental treatment compared with analyses

including all participants (Tierney 2005). There are notable examples of biased 'per-protocol' analyses, however, (Melander 2003), and a review has found more exaggerated effect estimates from 'per-protocol' analyses compared with 'ITT' analyses of the same trials (Porta 2007). Interpretation of empirical studies is difficult because exclusions are poorly reported, particularly before 1996 in the pre-CONSORT era (Moher 2001). For example, Schulz observed that the *apparent* lack of exclusions was associated with more beneficial effect sizes as well as with less likelihood of adequate allocation concealment. Hence, failure to report exclusions in trials in Schulz's study may have been a marker of poor trial conduct rather than true absence of any exclusions.

Empirical research has also investigated the adequacy with which incomplete outcome data are addressed in reports of trials. One study, of 71 trial reports from four general medical journals, concluded that missing data are common and often inadequately handled in the statistical analysis (Wood 2004).

8.12.2 Assessing risk of bias from incomplete outcome data

The risk of bias arising from incomplete outcome data depends on several factors, including the amount and distribution across intervention groups, the reasons for outcomes being missing, the likely difference between participants with and without data, what study authors have done to address the problem in their reported analyses, and the clinical context. Therefore it is not possible to formulate a simple rule for judging a study to be at low or high risk of bias. The following considerations may help review authors assess whether incomplete outcome data could be addressed in a way that protects against bias using the Collaboration's tool (Section 8.5).

It is often assumed that a high proportion of missing outcomes, or a large difference in proportions between intervention groups, is the main cause for concern over bias. However, these characteristics are on their own not sufficient to introduce bias. Here we elaborate on situations in which an analysis can be judged to be at low or high risk of bias. It is essential to consider the reasons for outcomes being missing as well as the numbers missing.

8.12.2.1 Low risk of bias due to incomplete outcome data

To conclude that there are no missing outcome data, review authors should be confident that the participants included in the analysis are exactly those who were randomized into the trial. If the numbers randomized into each intervention group are not clearly reported, the risk of bias is unclear. As noted above, participants randomized but subsequently found not to be eligible need not always be considered as having missing outcome data.

Example (of low risk of bias): "All patients completed the study and there were no losses to follow up, no treatment withdrawals, no trial group changes and no major adverse events".

Acceptable reasons for missing data

A healthy person's decision to move house away from the geographical location of a clinical trial is unlikely to be connected with their subsequent outcome. For studies with a long duration of follow-up, some withdrawals for such reasons are inevitable.

For studies reporting time-to-event data, all participants who did not experience the event of interest are, by definition, considered to be censored on the date of their last follow-up (as we do not know whether the outcome event occurred after follow-up ended). The important consideration for this type of analysis is whether censoring can be assumed to be unbiased, i.e. that the treatment effect (e.g.

assessed by a hazard ratio) in individuals who were censored before the scheduled end of follow up is the same as the hazard ratio in other individuals. In other words, there is no bias if censoring is unrelated to prognosis.

If outcome data are missing in both intervention groups, but reasons for these are both reported and balanced across groups, then important bias would not be expected unless the reasons have different implications in the compared groups. For example, 'refusal to participate' may mean unwillingness to exercise in an exercise group, whereas refusal might imply dissatisfaction with the advice not to exercise in the other group. In practice, incomplete reporting of reasons for missing outcomes may prevent review authors from making this assessment.

Potential impact of missing data on effect estimates

The potential impact of missing **dichotomous outcomes** depends on the frequency (or risk) of the outcome. For example, if 10 percent of participants have missing outcomes, then their potential impact on the results is much greater if the risk of the event is 10 percent than if it is 50 percent. The following table illustrates the potential impact of observed risks. A and B represent two hypothetical trials of 1000 participants in which 90 percent of the individuals are observed, and the risk ratio among these 900 observed participants is 1. Furthermore, in both trials we suppose that missing participants in the intervention group have a high risk of event (80 percent) and those in the control group have a much lower risk (20 percent), The only difference between trials A and B is the risk among the observed participants. In trial A the risk is 50 percent, and the impact of the missing data, had they been observed, is low. In trial B the risk is 10 percent, and the impact of the same missing data, had they been observed, is large. Generally, the higher the ratio of participants with missing data to participants with events, the greater potential there is for bias. In trial A this ratio was 100/450 (0.2), whereas in Trial B it was 100/90 (1.1).

	Number randomized	Risk among observed	Observed data	Hypothetical extreme risks among missing	Missing data	Complete data	Risk ratio based on all participants
Trial A							
Intervention	500	**50%**	225/450	80%	40/50	265/500	**1.13**
Control	500	**50%**	225/450	20%	10/50	235/500	
Trial B							
Intervention	500	**10%**	45/450	80%	40/50	85/500	**1.55**
Control	500	**10%**	45/450	20%	10/50	55/500	

The potential impact of missing **continuous outcomes** increases with the proportion of participants with missing data. It is also necessary to consider the plausible intervention effect among participants with missing outcomes. The following table illustrates the impact of different proportions of missing outcomes. A and B represent two hypothetical trials of 1000 participants in which the difference in mean response between intervention and control among the observed participants is 0. Furthermore, in both trials we suppose that missing participants in the intervention arm have a higher mean and those in the control arm have a lower mean. The only difference between trials A and B is the number of missing participants. In trial A, 90 percent of participants are observed and 10 percent missing, and the impact of the missing data on the observed mean difference is low. In trial B, half of the participants are missing, and the impact of the same missing data on the observed mean difference is large.

Trial A

	Number randomized	Number observed	Observed mean	Number missing	Hypothetical extreme mean among missing	Overall mean (weighted average)	Mean difference based on all participants
Intervention	500	**450**	10	**50**	15	10.5	**1**
Control	500	**450**	10	**50**	5	9.5	

Trial B

Intervention	500	**250**	10	**250**	15	12.5	**5**
Control	500	**250**	10	**250**	5	7.5	

8.12.2.2 High risk of bias due to incomplete outcome data

Unacceptable reasons for missing data

A difference in the proportion of incomplete outcome data across groups is of concern if the availability of outcome data is determined by the participants' true outcomes. For example, participants with poorer clinical outcomes are more likely to drop out due to adverse effects. If such attrition occurs mainly in the experimental group, the effect estimate will be biased in favour of the experimental intervention. Exclusion of participants due to 'inefficacy', or 'failure to improve' will introduce bias if the numbers excluded are not balanced across intervention groups. Note that a non-significant result of a statistical test for differential missingness does not confirm the absence of bias, especially in small studies.

Example (of high risk of bias): "In a trial of sibutramine versus placebo to treat obesity, 13/35 were withdrawn from the sibutramine group, 7 of these due to lack of efficacy. 25/34 were withdrawn from the placebo group, 17 due to lack of efficacy. An 'intention-to-treat' analysis included only those remaining" (Cuellar 2000) (i.e. only 9 of 34 in the placebo group) .

Even if incomplete outcome data are balanced in numbers across groups, bias can be introduced if the reasons for missing outcomes differ. For example, in a trial of an experimental intervention aimed at smoking cessation it is feasible that a proportion of the control intervention participants could leave the study due to a lack of enthusiasm at receiving nothing novel (and continue to smoke), and that a similar proportion of the experimental intervention group could leave the study due to successful cessation of smoking.

The common approach to dealing with missing outcome data in smoking cessation studies (to assume that everyone who leaves the study continues to smoke) may therefore not always be free from bias. The example highlights the importance of considering *reasons* for incomplete outcome data when assessing risk of bias. In practice, knowledge of why most participants drop out is often unavailable, although an empirical study has observed that 38 out of 63 trials with missing data provided information on reasons (Wood 2004), and this is likely to improve through the use of the CONSORT Statement (Moher 2001).

'As-treated' (per-protocol) analyses

Eligible participants should be analyzed in the groups to which they were randomized, regardless of the intervention that they actually received. Thus, in a study comparing surgery with radiotherapy for treatment of localized prostate cancer, patients who refused surgery and chose radiotherapy subsequent to randomization should be included in the surgery group for analysis. This is because participants' propensity to change groups may be related to prognosis, in which case switching intervention groups introduces selection bias. Although this is strictly speaking an issue of inappropriate analysis rather than incomplete outcome data, studies in which 'as treated' analyses are reported should be rated as at high risk of bias due to incomplete outcome data, unless the number of switches is too small to make any important difference to the estimated intervention effect.

A similarly inappropriate approach to analysis of a study is to focus only on participants who complied with the protocol. A striking example is provided by a trial of the lipid lowering drug, clofibrate (Coronary Drug Project Research Group 1980). The five-year mortality in 1103 men treated with clofibrate was 20.0 per cent, as compared with 20.9 per cent in 2789 men given placebo (P=0.55). Good adherers in the clofibrate group (patients who took >80% of the protocol prescription) had a substantially lower five-year mortality than did poor adherers to clofibrate (15.0% vs. 24.6%). However, similar findings were noted in the placebo group (15.1% for good adherers and 28.3% for poor adherers). Thus, adherence was a marker of prognosis rather than modifying the effect of clofibrate. These findings show the serious difficulty of evaluating treatment efficacy in subgroups determined by patient responses to the treatments. Because non-receipt of intervention can be more informative than non-availability of outcome data, there is a high risk of bias in analyses restricted to compliers, even with low rates of incomplete data.

Attempts to address missing data in reports: imputation

A common, but dangerous, approach to dealing with missing outcome data is to **impute** outcomes and treat them as if they were real measurements. For example, individuals with missing outcome data might be assigned the mean outcome for their intervention group, or be assigned a treatment success or failure. Such procedures can lead both to serious bias and to confidence intervals that are too narrow. A variant of this whose validity is more difficult to assess is the use of 'last observation carried forward' (LOCF). Here, the most recently observed outcome measure is assumed to hold for all subsequent outcome assessment times (Lachin 2000, Unnebrink 2001). LOCF procedures can also lead to serious bias. For example, in a trial of a drug for a degenerative condition, such as Alzheimer's disease, attrition may be related to side effects of the drug. Because outcomes tend to deteriorate with time, using LOCF will bias the effect estimate in favour of the drug. On the other hand, use of LOCF might be appropriate if most people for whom outcomes are carried forward had a genuine measurement relatively recently.

There is a substantial literature on statistical methods that deal with missing data in a valid manner. Such methods include 'multiple imputation', 'weighted estimation' and 'full likelihood-based estimation'. There are relatively few practical applications of these methods in clinical trial reports (Wood 2004). Statistical advice is recommended if review authors encounter their use. A good starting point for learning about them is www.missingdata.org.uk.

8.13 Selective outcome reporting

8.13.1 Rationale for concern about bias

Selective outcome reporting has been defined as the selection on the basis of the results of a subset of the original variables recorded for inclusion in publication of trials (Hutton 2000). The particular concern is that non-significant results might be selectively withheld from publication. Until recently, published evidence of selective outcome reporting was limited. There were initially a few case studies.

Then a small study of a complete cohort of applications approved by a single Local Research Ethics Committee found that the primary outcome was stated in only six of the protocols for the 15 publications obtained. Eight protocols made some reference to an intended analysis, but seven of the publications did not follow this analysis plan (Hahn 2002). Within-study selective reporting was evident or suspected in several trials included in a review of a cohort of five meta-analyses on the Cochrane Library (Williamson 2005a).

Convincing direct empirical evidence for the existence of within-study selective reporting bias comes from 3 recent studies. In the first study (Chan 2004a), 102 trials with 122 publications and 3736 outcomes were identified. Overall, (a median of) 38% of efficacy and 50% of safety outcomes per parallel group trial were incompletely reported, i.e. with insufficient information to be included in a meta-analysis. Statistically significant outcomes had a higher odds of being fully reported when compared with non-significant outcomes for both efficacy (pooled odds ratio 2.4; 95% confidence interval 1.4 to 4.0) and harms (4.7, 1.8 to 12) data. Further, when comparing publications with protocols, 62% of trials had at least one primary outcome that was changed, introduced or omitted. A second study of 48 trials funded by the Canadian Institutes of Health Research found closely similar results (Chan 2004b). A third study, involving a retrospective review of 519 trial publications and a follow-up survey of authors, compared the presented results with the outcomes mentioned in the methods section of the same article (Chan 2005). On average, over 20% of the outcomes measured in parallel group trials were incompletely reported. Within trials, such outcomes had a higher odds of being statistically non-significant compared with fully reported outcomes (odds ratio 2.0; 1.6 to 2.7 for efficacy outcomes; 1.9 (1.1 to 3.5) for harm outcomes). These three studies suggest an odds ratio of about 2.4 associated with selective outcome reporting which corresponds, for example, to about 50% of non-significant outcomes being published compared to 72% of significant ones.

For all the three studies, authors were asked whether there were unpublished outcomes, whether those showed significant differences and why those outcomes had not been published. The most common reasons for non-publication of results were ‘lack of clinical importance’ or lack of statistical significance. Therefore, meta-analyses excluding unpublished outcomes are likely to overestimate intervention effects. Further, authors commonly failed to mention the existence of unpublished outcomes even when those outcomes had been mentioned in the protocol or publication.

Recent studies have found similar results (Ghersi 2006, von Elm 2006). In a different type of study, the effect in meta-analyses was smaller when fewer of the available trials contributed data to that meta-analysis (Furukawa 2007). This finding also suggests that results may have been selectively withheld by trialists on the basis of the magnitude of effect.

Bias associated with selective reporting of different measures of the same characteristic seems likely. In trials of treatments for schizophrenia, a treatment effect has been observed to be more likely when unpublished, rather than published, rating scales were used (Marshall 2000). The authors hypothesized that data from unpublished scales may be less likely to be published when they are not significant or that, following analysis, unfavourable items may have been dropped to create an apparent effect.

In many systematic reviews, only a few eligible studies can be included in a meta-analysis for a specific outcome because the necessary information was not reported by the other studies. While that outcome may not have been assessed in some studies, there is almost always a risk of biased reporting for some studies. Review authors need to consider whether an outcome was collected but not reported or simply not collected.

Selective reporting of outcomes may arise in several ways, some affecting the study as a whole (point 1 below) and others relating to specific outcomes (points 2-6 below):

1. Selective omission of outcomes from reports: Only some of the analyzed outcomes may be included in the published report. If that choice is made based on the results, in particular the statistical significance, the corresponding meta-analytic estimates are likely to be biased.
2. Selective choice of data for an outcome: For a specific outcome there may be different time points at which the outcome has been measured, or there may have been different instruments used to measure the outcome at the same time point (e.g. different scales, or different assessors). For example, in a report of a trial in osteoporosis, there were 12 different data sets to choose from for estimating bone mineral content. The standardized mean difference for these 12 possibilities varied between −0.02 and 1.42 (Gøtzsche 2007).If study authors make choices in relation to such results, then the meta-analytic estimate will be biased.
3. Selective reporting of analyses using the same data: There are often several different ways in which an outcome can be analyzed. For example, continuous outcomes such as blood pressure reduction might be analyzed as a continuous or dichotomous variable, with the further possibility of selecting from multiple cut-points. Another common analysis choice is between endpoint scores versus changes from baseline (Williamson 2005b). Switching from an intended analysis of final values to change from baseline as a result of an observed baseline imbalance actually introduces bias rather than removing it (as the study authors may suppose) (Senn 1991, Vickers 2001).
4. Selective reporting of subsets of the data: Selective reporting may occur if outcome data can be subdivided, for example selecting sub-scales of a full measurement scale or a subset of events. For example, fungal infections at baseline or within a couple of days after randomization versus so-called 'break-through' fungal infections that are detected some days after randomization (Jørgensen 2006, Jørgensen 2007). [This is not a sentence.]
5. Selective under-reporting of data: Some outcomes may be reported but with inadequate detail for the data to be included in a meta-analysis. Sometimes this is explicitly related to the result, for example reported only as "not significant" or "P>0.05".

Yet other forms of selective reporting are not addressed here; they include selected reporting of subgroup analyses or adjusted analyses, and presentation of the first period results in cross-over trials (Williamson 2005a). Also, descriptions of outcomes as 'primary', 'secondary' etc may sometimes be altered retrospectively in the light of the findings (Chan 2004b, Chan 2004a). This issue alone will not generally be of concern to review authors (who do not take note of which outcomes are so labelled in each study), provided it does not influence which results are published.

8.13.2 Assessing risk of bias from selective reporting of outcomes

Although the possibility of *between-study* publication bias can be examined only by considering a set of studies, the possibility of *within-study* selective outcome reporting can be examined for each study included in a systematic review. The following considerations may help review authors assess whether outcome reporting is sufficiently complete and transparent to protect against bias using the Collaboration's tool (Section 8.5).

Statistical methods to detect within-study selective reporting are, as yet, not well developed. There are, however, other ways of detecting such bias although a thorough assessment is likely to be labour intensive. If the protocol is available, then outcomes in the protocol and published report can be compared. If not, then outcomes listed in the methods section of an article can be compared with those whose results are reported. If non-significant results are mentioned but not reported adequately, bias in a meta-analysis is likely to occur. Further information can also be sought from authors of the study reports, although it should be realized that such information may be unreliable (Chan 2004a).

Some differences between protocol and publication may be explained by legitimate changes to the protocol. Although such changes should be reported in publications, none of the 150 studies in the two samples of Chan et al. did so (Chan 2004b, Chan 2004a).

Review authors should look hard for evidence of collection by study investigators of a small number of key outcomes that are routinely measured in the area in question, and report which studies report data on these and which do not. Review authors should consider the *reasons* why data might be missing from a meta-analysis (Williamson 2005b). Methods for seeking such evidence are not well-established, but we describe some possible strategies.

A useful first step is to construct a matrix indicating which outcomes were recorded in which studies, e.g. with rows as studies and columns as outcomes. Complete and incomplete reporting can also be indicated. This matrix will show to the review authors which studies did not report outcomes reported by most other studies.

PubMed and the internet should be searched for a study protocol; in rare cases the web address will be given in the study report. Alternatively, and more often in the future as mandatory registrations of trials becomes more common, a detailed description of the study may be available in a trial registry. Abstracts of presentations relating to the study may contain information about outcomes not subsequently mentioned in publications. In addition, review authors should examine carefully the methods section of published articles for details of outcomes that were assessed.

Of particular interest is missing information that seems sure to have been recorded. For example, some measurements are expected to appear together, such as systolic and diastolic blood pressure, so we should wonder why only one is reported. An alternative example is a study reporting the proportion of participants whose change in a continuous variable exceeded some threshold; the investigators must have had access to the raw data and so could have shown the results as mean and SD of the changes. Williamson et al give several examples, including a Cochrane review in which 9 trials reported the outcome treatment failure but only five reported mortality. Yet mortality was part of the definition of treatment failure so those data must have been collected in the four trials missing from the analysis of mortality. Bias was suggested by the marked difference in results for treatment failure for trials with or without separate reporting of mortality (Williamson 2005a).

When there is suspicion of or direct evidence for selective outcome reporting it is desirable to contact the study authors asking for additional information. For example, authors could be asked to supply the study protocol and full information for outcomes reported inadequately. In addition, for outcomes mentioned in article or protocol but not reported they could be asked to clarify whether those outcome measures were in fact analyzed, and if so to supply the data.

It is not generally recommended to try to make a formal allowance for reporting bias in the main meta-analysis. Sensitivity analysis is a more promising approach to investigate the possible impact of selective outcome reporting (Hutton 2000, Williamson 2005a).

The assessment of risk of bias due to selective reporting of outcomes should be made for the study as a whole, rather than for each outcome. Although it may be clear for a particular study that some specific outcomes are subject to selective reporting while others are not, we recommend the study-level approach because it is not practical to list all fully reported outcomes in the 'Risk of bias' table. The Description part of the tool (see Section 8.5.2) should be used to describe the outcomes for which there is particular evidence of selective (or incomplete) reporting. The study-level judgment provides an assessment of the overall susceptibility of the study to selective reporting bias.

8.14 Topic-specific, design-specific or other potential threats to validity

8.14.1 Rationale for concern about bias

The preceding topics relate to important potential sources of bias across all types of studies in all healthcare areas. In some topic areas, there may be additional questions that should be asked of all studies; some study designs warrant special consideration when they are encountered; and some major, unanticipated, problems may be identified during the course of the systematic review or meta-analysis. Several examples are given here.

8.14.1.1 Cross-over trials

Randomized cross-over trials are discussed in Chapter 16. Many cross-over trials use a 'two period, two treatment' design in which each participant receives two study treatments in sequence in random order. The main concerns over risk of bias 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 predicated on the expectation that there will not be any 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. That 'two stage analysis' is now 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 authors 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 analyzed 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 versus parallel trials (Khan 1996), they had looked at treatments for infertility, an area notorious for the inappropriateness of the cross-over design, and a careful reanalysis did not support the original findings (te Velde 1998).

Other issues

- 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
- 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 9 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?

8.14.1.2 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 RCTs, and found possible recruitment bias in 14 (39%) (Puffer 2003).

(ii) Cluster randomized trials often randomize all clusters at once, so lack of allocation concealment should not usually be an issue if the allocation is done in an independent and secure fashion. 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 (see Section 8.12), this may lead to bias. In addition, missing outcomes for individuals within clusters also causes a risk of bias in cluster RCTs.

(iv) Many cluster RCTs are analyzed by incorrect statistical methods, not taking the clustering into account. For example, Eldridge et al reviewed 152 cluster RCTs 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 Chapter 16. 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, contamination would lead to under-estimates 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.

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).

8.14.1.3 Biases in trials with multiple treatment groups

Two other types of trial design are considered here briefly: multi-arm parallel group trials, and factorial trials.

Multi-arm parallel group trials

A substantial minority of parallel group randomized trials have more than two treatment arms. Methods for incorporating such trials into a review are discussed in Chapter 16. For example, issues related to splitting control group data when there are two active treatment arms to avoid double-counting of the control group are discussed there and are not relevant to concerns about risk of bias.

Bias may be introduced in a multi-arm trial if the decisions regarding data analysis are made after seeing the data. For example, groups receiving different doses of same treatment 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 multi-arm RCTs, of which over a third had at least 4 treatment arms (Juszczak 2007). 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.

Some suggested questions for assessing risk of bias in multi-arm trials 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 treatment arms for some outcomes?

Factorial trials

In a factorial trial two (or more) treatment comparisons are carried out simultaneously. Thus for example, participants may be randomized to receive aspirin or placebo and also to receive an educational intervention or standard advice. Most factorial trials are like the example just cited, with two 'factors', 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. 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 get 'two trials for the price of one', and the assumption is made that the effects of the different active treatments 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 but multi-arm trials, as discussed above.

The 2×2 factorial design can be displayed as a 2×2 table, with the rows indicating one comparison (say A vs control) and the columns the other (B vs control). Sometimes only one of these comparisons will be of relevance to any particular review. A 2×2 factorial trial can be seen as two trials answering 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 A vs control, including all participants regardless of whether they had intervention B or the control for B, and likewise for B. 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 control 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 reanalysis, 2/44 trials (6%) had $P<0.05$, which is a rate that 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 2 (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.

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

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

8.14.1.4 Early stopping

Studies that were stopped early (whether or not as a result of a formal stopping rule) are more likely to show extreme treatment effects than those that continue to the end, particularly if they have very few events (Montori 2005). This is especially the case when a study stops because early results show a large, statistically significant, treatment effect, although may also be the case if a study stops early because of harm. If a study does not describe having a pre-specified sample size, or any formal stopping rules, or the attained sample size is much less than the intended size but no explanation is given, then the study may have stopped at a point chosen because of the observed results and the available results may therefore be biased. Early stopping may be more common than is reported. For example, in a study of 44 industry-initiated trials, the trial protocols showed that the sponsor of had access to accumulating data in 16 (e.g. through interim analyses and participation in data and safety monitoring committees), but such access was disclosed in only one corresponding trial report. An additional 16 protocols noted that the sponsor had the right to stop the trial at any time, for any reason; this was not noted in any of the trial publications (Gøtzsche 2006). Even when trials are known to have stopped early, systematic reviews frequently fail to note this (Bassler 2007).

Bias-adjusted analyses are available for studies that stop early due to a formal stopping rule, but such analyses are seldom implemented.

Studies that fail to attain a pre-specified sample size for reasons unrelated to the observed treatment effect (e.g. a lower than expected recruitment rate; insufficient funds; no supply of drug, etc) are not more likely to show extreme results, and should not generally be considered to be prone to bias due to early stopping.

Example (of high risk of bias): The data and safety monitoring board recommended stopping the trial because the test statistic for the primary outcome measure exceeded the stopping boundary for benefit.

8.14.1.5 Baseline imbalance

Baseline imbalance in factors that are strongly related to the outcome measure can cause bias in the treatment effect estimate. This can happen through chance alone, but imbalance may also arise through non-randomized (unconcealed) allocation of interventions. Sometimes trial authors may exclude some randomized individuals, causing imbalance in participant characteristics in the different intervention groups. Sequence generation, lack of allocation concealment or exclusion of participants should each be addressed using the specific items for these in the tool. If further inexplicable baseline imbalance is observed that is sufficient to lead to important exaggeration of effect estimates, then it should be noted. Tests of baseline imbalance have no value in truly randomized trials, but very small P values could suggest bias in the randomization process.

Example: A trial of captopril vs conventional anti-hypertensive had small but highly significant imbalances in height, weight, systolic and diastolic BP: P=10^{-4} to 10^{-18} (Hansson 1999). Such an

imbalance suggests failure of randomization (which was by sealed envelopes) at some centres (Peto 1999).

8.14.1.6 Blocked randomization in unblinded trials

Some combinations of methods for sequence generation, allocation concealment and blinding act together to create a risk of selection bias in the allocation of interventions. One particular combination is the use of blocked randomization in an unblinded trial, or in a blinded trial where the blinding is broken, e.g. because of characteristic side effects. When blocked randomization is used, and when the assignments are revealed subsequent to the person recruiting into the trial, then it is sometimes possible to predict future assignments. This is particularly the case when blocks are of a fixed size and are not divided across multiple recruitment centres. This ability to predict future assignments can happen even when allocation concealment is adequate according to the criteria suggested in Table 8.5.c (Berger 2005).

8.14.1.7 Differential diagnostic activity

Outcome assessments can be biased despite effective blinding. In particular, increased diagnostic activity could lead to increased diagnosis of true but harmless cases of disease. For example, many stomach ulcers give no symptoms and have no clinical relevance, but such cases could be detected more frequently on gastroscopy in patients who receive a drug that causes unspecific stomach discomfort and therefore leads to more gastroscopies. Similarly, if a drug causes diarrhoea, this could lead to more digital, rectal examinations, and, therefore, also to the detection of more harmless cases of prostatic cancer. Obviously, assessment of beneficial effects can also become biased through such a mechanism. Interventions may also lead to different diagnostic activity. For example, randomizing whether or not participants will have a nurse visit them at home may create differential diagnostic activity between intervention groups.

8.14.1.8 Further examples of potential biases

The following list of other potential sources of bias in a clinical study may aid detection of further problems.

- The conduct of the study is affected by interim results (e.g. recruiting additional participants from a subgroup showing more benefit);
- There is deviation from the study protocol in a way that does not reflect clinical practice (e.g. post-hoc stepping-up of doses to exaggerated levels);
- There is pre-randomization administration of an intervention that could enhance or diminish the effect of a subsequent, randomized, intervention;
- Administration of an intervention (or co-intervention) is inappropriate;
- Contamination (e.g. participants pooling drugs);
- Ocurrence of ‘null bias’ due to interventions being insufficiently delivered or overly wide inclusion criteria for participants (Woods 1995);
- An insensitive instrument is used to measure outcomes (which can lead to under-estimation of both beneficial and harmful effects);
- Selective reporting of subgroups;
- Fraud;
- Inappropriate influence of funders (e.g. in one empirical study, more than half of the protocols for industry-initiated trials stated that the sponsor either owns the data or needs to approve the manuscript, or both; none of these constraints were stated in any of the trial publications (Gøtzsche 2006)).

8.14.2 Assessing risk of bias from other sources

This sixth domain in the 'Risk of bias' assessment tool described below is a 'catch-all' for sources of bias other than those due to sequence generation, allocation concealment, blinding, incomplete outcome data and selective outcome reporting. Some general guidelines for determining suitable topics for assessment are provided below. The topics covered in this domain of the tool include primarily the examples provided in Section 8.14.1. Beyond these specific issues, however, review authors should be alert for study-specific issues that may raise concerns about the possibility of bias, and should formulate judgments about them under this domain of the tool. The following considerations may help review authors assess whether a study is free of risk of bias from other sources using the Collaboration's tool (Section 8.5).

Wherever possible, a review protocol should pre-specify any questions to be addressed, which would lead to separate items in the 'Risk of bias' table. For example, if cross-over trials are the usual study design for the question being addressed by the review, then specific questions related to bias in cross-over trials should be formulated in advance.

Issues covered by the risk of bias tool must be a potential source of bias, and not just a cause of *imprecision* (see section 8.2), and this applies to aspects that are assessed under this 'other sources of bias' domain. A potential source of bias must be able to change the magnitude of the effect estimate, whereas sources of imprecision affect only the uncertainty in the estimate (i.e. its confidence interval). Potential factors affecting precision of an estimate include technological variability (e.g. measurement error), and observer variability.

Because the tool addresses only internal biases, any issue covered by this domain should be a potential source of internal bias, and not a source of *diversity*. Possible causes of diversity include differences in dose of drug, length of follow up, and characteristics of participants (e.g. age, stage of disease). Studies may select doses that favour the experimental drug over the control drug. For example, old drugs are often overdosed (Safer 2002) or may be given under clearly suboptimal circumstances that do not reflect clinical practice (Jørgensen 2007, Johansen 2000). Alternatively, participants may be selectively chosen for inclusion in a study on the basis of previously demonstrated 'response' to the experimental intervention. It is important that such biased choices are addressed in Cochrane reviews. Although they may not be covered by the 'Risk of bias' tool described in the current chapter, they may be sometimes be addressed in the analysis (e.g. by subgroup analysis and meta-regression) and should be considered in the grading and interpretation of evidence in a 'Summary of findings' table (Chapter 11 and Chapter 12).

Many judgments can be made about the design and conduct of a clinical trial, but not all of them may be associated with bias. Measures of 'quality' alone are often strongly associated with aspects that could introduce bias. However, review authors should focus on the mechanisms that lead to bias rather than descriptors of studies that reflect only 'quality'. Some examples of 'quality' indicators that should not be assessed within this domain include criteria related to applicability, 'generalizability' or 'external validity (including those noted above), criteria related to precision (e.g. sample size or use of a sample size (or power) calculation), reporting standards, and ethical criteria (e.g. whether the study had ethical approval or participants gave informed consent).

Finally, to avoid double-counting, potential sources of bias should not be included as 'bias from other sources' if they are more appropriately covered by earlier items in the tool. For example, in Alzheimer's disease, patients deteriorate significantly over time during the trial. Generally, the effects of treatments are small and treatments have appreciable toxicity. Dealing satisfactorily with

participant losses is very difficult. Those on treatment are likely to drop out earlier due to adverse effects and hence the measurements on these people, tending to be earlier in the study, will favour the intervention. It is often difficult to get continued monitoring of these participants in order to carry out an analysis of all randomized participants. This issue, although it might at first seem to be a topic-specific cause of bias would be more appropriately covered under Incomplete Outcome Data.

8.15 Contributions to this chapter

Editors: Julian PT Higgins and Douglas G Altman.

Contributing authors: Doug Altman, Gerd Antes, Peter C Gøtzsche, Julian Higgins, Peter Jüni, Steff Lewis, David Moher, Andy Oxman, Ken Schulz, Jonathan Sterne, Simon Thompson.

Working group: Doug Altman (co-lead), Gerd Antes, Chris Cates, Mike Clarke, Jon Deeks, Peter C Gøtzsche, Julian Higgins (co-lead), Sally Hopewell, Peter Jüni (core group), Steff Lewis, Philippa Middleton, David Moher (core group), Andy Oxman, Ken Schulz (core group), Nandi Siegfried, Jonathan Sterne, Simon Thompson.

Acknowledgements: We thank Hilda Bastian, Rachelle Buchbinder, Miranda Cumpston, Sally Green, Peter Herbison, Victor Montori, Hannah Rothstein, Georgia Salanti, Guido Schwarzer, Ian Shrier, Jayne Tierney, Ian White and Paula Williamson for helpful comments.

8.16 References

Altman 1999

Altman DG, Bland JM. How to randomize. BMJ 1999; 319: 703-704.

Balk 2002

Balk EM, Bonis PAL, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, Lau J. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002; 287: 2973-2982.

Bassler 2007

Bassler D, Ferreira-Gonzalez I, Briel M, Cook DJ, Devereaux PJ, Heels-Ansdell D, Kirpalani H, Meade MO, Montori VM, Rozenberg A, Schunemann HJ, Guyatt GH. Systematic reviewers neglect bias that results from trials stopped early for benefit. J Clin Epidemiol 2007; 60: 869-873.

Bellomo 2000

Bellomo R, Chapman M, Finfer S, Hickling K, Myburgh J. Low-dose dopamine in patients with early renal dysfunction: a placebo-controlled randomised trial. Australian and New Zealand Intensive Care Society (ANZICS) Clinical Trials Group. Lancet 2000; 356: 2139-2143.

Berger 2005

Berger VW. Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. Biom J 2005; 47: 119-127.

Berger 2003

Berger VW, Ivanova A, Knoll MD. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. Stat Med 2003; 22: 3017-3028.

Berlin 1997

Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. Lancet 1997; 350: 185-186.

Boutron 2006

Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, Hróbjartsson A, Ravaud P. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. PLOS Med 2006; 3: 1931-1939.

Boutron 2005

Boutron I, Estellat C, Ravaud P. A review of blinding in randomized controlled trials found results inconsistent and questionable. J Clin Epidemiol 2005; 58: 1220-1226.

Brightling 2000

Brightling CE, Monteiro W, Ward R, Parker D, Morgan MD, Wardlaw AJ, Pavord ID. Sputum eosinophilia and short-term response to prednisolone in chronic obstructive pulmonary disease: a randomised controlled trial. Lancet 2000; 356: 1480-1485.

Brown 2005

Brown S, Thorpe H, Hawkins K, Brown J. Minimization: reducing predictability for multi-centre trials whilst retaining balance within centre. Stat Med 2005; 24: 3715-3727.

Chan 2005

Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 2005; 330: 753.

Chan 2004a

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004; 291: 2457-2465.

Chan 2004b

Chan AW, Krleža-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. CMAJ 2004; 171: 735-740.

Coronary Drug Project Research Group 1980

Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. New Eng J Med 1980; 303: 1038-1041.

Cuellar 2000

Cuellar GEM, Ruiz AM, Monsalve MCR, Berber A. Six-month treatment of obesity with sibutramine 15 mg; a double-blind, placebo-controlled monocenter clinical trial in a Hispanic population. Obes Res 2000; 8: 71-82.

de Gaetano 2001

de Gaetano G. Low-dose aspirin and vitamin E in people at cardiovascular risk: a randomised trial in general practice. Collaborative Group of the Primary Prevention Project. Lancet 2001; 357: 89-95.

Detsky 1992

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol 1992; 45: 255-265.

Devereaux 2001

Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, Bhandari M, Guyatt GH. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 2001; 285: 2000-2003.

Egger 2003

Egger M, Jüni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003; 7: 1-76.

Elbourne 2002

Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vaillancourt JM. Meta-analyses involving cross-over trials: methodological issues. Int J Epidemiol 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. Clin Trials 2004; 1: 80-90.

Emerson 1990

Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Control Clin Trials 1990; 11: 339-352.

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. Clin Trials 2005; 2: 119-124.

Fergusson 2002

Fergusson D, Aaron SD, Guyatt G, Hébert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ 2002; 325: 652-654.

Freeman 1989

Freeman PR. The performance of the two-stage analysis of two-treatment, two-period cross-over trials. Stat Med 1989; 8: 1421-1432.

Furukawa 2007

Furukawa TA, Watanabe N, Omori IM, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. JAMA 2007; 297: 468-470.

Ghersi 2006

Ghersi D, Clarke M, Simes J. Selective reporting of the primary outcomes of clinical trials: a follow-up study. 14th Cochrane Colloquium, Dublin, 2006.

Gøtzsche 1996

Gøtzsche PC. Blinding during data analysis and writing of manuscripts. Control Clin Trials 1996; 17: 285-290.

Gøtzsche 2006

Gøtzsche PC, Hróbjartsson A, Johansen HK, Haahr MT, Altman DG, Chan AW. Constraints on publication rights in industry-initiated clinical trials. JAMA 2006; 295: 1645-1646.

Gøtzsche 2007

Gøtzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. JAMA 2007; 298: 430-437.

Greenland 2001

Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2001; 2: 463-471.

Haahr 2006

Haahr MT, Hróbjartsson A. Who is blinded in randomised clinical trials? A study of 200 trials and a survey of authors. Clin Trials 2006; 3: 360-365.

Hahn 2005

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

Hahn 2002

Hahn S, Williamson PR, Hutton JL. Investigation of within-study selective reporting in clinical research: follow-up of applications submitted to a local research ethics committee. J Eval Clin Pract 2002; 8: 353-359.

Hansson 1999

Hansson L, Lindholm LH, Niskanen L, Lanke J, Hedner T, Niklason A, Luomanmaki K, Dahlof B, de Faire U, Morlin C, Karlberg BE, Wester PO, Bjorck JE. Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial. Lancet 1999; 353: 611-616.

Hill 1990

Hill AB. Memories of the British streptomycin trial in tuberculosis: the first randomized clinical trial. Control Clin Trials 1990; 11: 77-79.

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.

Hutton 2000

Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. J R Statist Soc C 2000; 49: 359-370.

Jadad 1996

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay H. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996; 17: 1-12.

Johansen 2000

Johansen HK, Gøtzsche PC. Amphotericin B lipid soluble formulations versus amphotericin B in cancer patients with neutropenia. Cochrane Database Syst Rev 2000, Issue 3. Art No: CD000969.

Jørgensen 2006

Jørgensen KJ, Johansen HK, Gøtzsche PC. Voriconazole versus amphotericin B in cancer patients with neutropenia. Cochrane Database Syst Rev 2006, Issue 1. Art No: CD004707.

Jørgensen 2007

Jørgensen KJ, Johansen HK, Gøtzsche PC. Flaws in design, analysis and interpretation of Pfizer's antifungal trials of voriconazole and uncritical subsequent quotations. Trials 2007; 7: 3.

Jüni 2001

Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 2001; 323: 42-46.

Jüni 1999

Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999; 282: 1054-1060.

Juszczak 2007

Juszczak E, Chan AW, Altman DG. A review of the methodology and reporting of multi-arm parallel group randomised trials. 2007.

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. Fertil Steril 1996; 65: 939-945.

Kjaergard 2001

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

Lachin 2000

Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials 2000; 21: 167-189.

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) (PhD thesis). University of London, 2005.

Marshall 2000

Marshall M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. British Journal of Psychiatry 2000; 176:249-52.: 249-252.

McAlister 2003

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

Melander 2003

Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine - selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003; 326: 1171-1173.

Mills 2005

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

Moher 1995

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Control Clin Trials 1995; 16: 62-73.

Moher 1996

Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: Current issues and future directions. Int J Technol Assess Health Care 1996; 12: 195-208.

Moher 1998

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998; 352: 609-613.

Moher 2001

Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001; 357: 1191-1194.

Montori 2002

Montori VM, Bhandari M, Devereaux PJ, Manns BJ, Ghali WA, Guyatt GH. In the dark: the reporting of blinding status in randomized controlled trials. J Clin Epidemiol 2002; 55: 787-790.

Montori 2005

Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, Lacchetti C, Leung TW, Darling E, Bryant DM, Bucher HC, Schünemann HJ, Meade MO, Cook DJ, Erwin PJ, Sood A, Sood R, Lo B, Thompson CA, Zhou Q, Mills E, Guyatt GH. Randomized trials stopped early for benefit: a systematic review. JAMA 2005; 294: 2203-2209.

Naylor 1997

Naylor CD. Meta-analysis and the meta-epidemiology of clinical research. BMJ 1997; 315: 617-619.

Noseworthy 1994

Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. Neurology 1994; 44: 16-20.

Oxman 1993

Oxman AD, Guyatt GH. The science of reviewing research. Ann N Y Acad Sci 1993; 703: 125-133.

Peto 1999

Peto R. Failure of randomisation by "sealed" envelope. Lancet 1999; 354: 73.

Pildal 2007

Pildal J, Hróbjartsson A, Jørgensen KJ, Hilden J, Altman DG, Gøtzsche PC. Impact of allocation concealment on conclusions drawn from meta-analyses of randomised trials. 2007.

Porta 2007

Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. J Clin Epidemiol 2007; 60: 663-669.

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.

Safer 2002

Safer DJ. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. J Nerv Ment Dis 2002; 190: 583-592.

Schulz 1995a

Schulz KF. Subverting randomization in controlled trials. JAMA 1995; 274: 1456-1458.

Schulz 2002a

Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. Ann Intern Med 2002; 136: 254-259.

Schulz 1995b

Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273: 408-412.

Schulz 2002b

Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet 2002; 359: 614-618.

Schulz 2002c

Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet 2002; 359: 515-519.

Schulz 2002d

Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. Lancet 2002; 359: 966-970.

Schulz 2006

Schulz KF, Grimes DA. The Lancet Handbook of Essential Concepts in Clinical Research. Edinburgh: Elsevier, 2006.

Senn 1991

Senn S. Baseline comparisons in randomized clinical trials. Stat Med 1991; 10: 1157-1159.

Smilde 2001

Smilde TJ, van Wissen S, Wollersheim H, Trip MD, Kastelein JJ, Stalenhoef AF. Effect of aggressive versus conventional lipid lowering on atherosclerosis progression in familial hypercholesterolaemia (ASAP): a prospective, randomised, double-blind trial. Lancet 2001; 357: 577-581.

Spiegelhalter 2003

Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. Stat Med 2003; 22: 3687-3709.

Sterne 2002

Sterne JA, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med 2002; 21: 1513-1524.

te Velde 1998

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

Tierney 2005

Tierney JF, Stewart LA. Investigating patient exclusion bias in meta-analysis. Int J Epidemiol 2005; 34: 79-87.

Unnebrink 2001

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

Vickers 2001

Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Medical Research Methodology 2001; 1: 6.

von Elm 2006

von Elm E, Röllin A, Blümle A, Senessie C, Low N, Egger M. Selective reporting of outcomes of drug trials. Comparison of study protocols and pulbished articles. 14th Cochrane Colloquium, Dublin, 2006.

Williamson 2005a

Williamson PR, Gamble C. Identification and impact of outcome selection bias in meta-analysis. Stat Med 2005; 24: 1547-1561.

Williamson 2005b

Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in meta-analysis. Stat Methods Med Res 2005; 14: 515-524.

Wood 2004

Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials 2004; 1: 368-376.

Wood 2006

Wood L. The epidemiology of bias in randomised (clinical) controlled trials: a meta-epidemiological study (PhD thesis). The University of Bristol, 2006.

Woods 1995

Woods KL. Mega-trials and management of acute myocardial infarction. Lancet 1995; 346: 611-614.