Chapter 1: Introduction

Authors: Jackie Chandler, Julian PT Higgins, Jonathan J Deeks, Clare Davenport, Mike J Clarke.

This chapter should be cited as: Chandler J, Higgins JPT, Deeks JJ, Davenport C, Clarke MJ. Chapter 1: Introduction. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS (editors), *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0 (updated June 2017), Cochrane, 2017. Available from www.training.cochrane.org/handbook

This extract is from *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0. The previous version of this chapter (5.1.0, 2011) is available online at handbook.cochrane.org.

An earlier version (version 5.0.2, 2008) of the *Handbook* is also published by John Wiley & Sons, Ltd under “The Cochrane Book Series” Imprint, as Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on www.wiley.com.

Key Points

- Systematic reviews seek to collate all evidence that fits pre-specified eligibility criteria in order to address a specific research question.
- Systematic reviews aim to minimize bias by using explicit, systematic methods documented in advance with a protocol.
- Cochrane prepares, maintains and promotes systematic reviews to inform decisions about health and social care (Cochrane Reviews).
- Cochrane Reviews are published in the *Cochrane Database of Systematic Reviews* in the Cochrane Library*.*
- The *Cochrane Handbook for Systematic Reviews of Interventions* contains methodological guidance for the preparation and maintenance of Cochrane Intervention Reviews, Overviews of Reviews and Methodology Reviews.

- Methodological advice on Cochrane Diagnostic Test Accuracy Reviews can be found in the separate *Cochrane Handbook for Diagnostic Test Accuracy Reviews*.
- Cochrane has developed conduct and reporting standards.

1.1 Cochrane

1.1.1 What is Cochrane?

Trusted evidence. Informed decisions. Better health.

Cochrane is a global independent network of health practitioners, researchers, patient advocates and others, responding to the challenge of making the vast amounts of evidence generated through research useful for informing decisions about health (www.cochrane.org). Previously known as The Cochrane Collaboration, it is a not-for-profit organization where collaborators aim to produce credible, accessible health information that is free from commercial sponsorship and other conflicts of interest.

Cochrane's mission is to promote evidence-informed health decision-making by producing high quality, relevant, accessible systematic reviews and other synthesized research evidence. The work of Cochrane is underpinned by a set of 10 key principles, listed in Box 1.1.a

Box 1.1.a: The 10 principles of Cochrane

1	Collaboration	by fostering global co-operation, teamwork, and open and transparent communication and decision-making.
2	Building on the enthusiasm of individuals	by involving, supporting and training people of different skills and backgrounds.
3	Avoiding duplication of effort	by good management, co-ordination and effective internal communications to maximize economy of effort.
4	Minimizing bias	through a variety of approaches such as scientific rigour, ensuring broad participation, and avoiding conflicts of interest.
5	Keeping up-to-date	by a commitment to ensure that Cochrane Systematic Reviews are maintained through identification and incorporation of new evidence.
6	Striving for relevance	by promoting the assessment of health questions using outcomes that matter to people making choices in health and health care.
7	Promoting access	by wide dissemination of our outputs, taking advantage of strategic alliances, and by promoting appropriate access models and delivery solutions to meet the needs of users worldwide.
8	Ensuring quality	by applying advances in methodology, developing systems for quality improvement, and being open and responsive to criticism.
9	Continuity	by ensuring that responsibility for reviews, editorial processes and key functions is maintained and renewed.
10	Enabling wide participation	in our work by reducing barriers to contributing and by encouraging diversity.

1.1.2 A brief history of Cochrane

The Cochrane Collaboration was founded in 1993, a year after the establishment of the UK Cochrane Centre in Oxford, UK. The UK Cochrane Centre arose from a vision to extend a ground-breaking programme of work by Iain Chalmers and colleagues in the area of pregnancy and childbirth to the rest of health care. Inspired by Archie Cochrane’s claim that “It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials" (Cochrane 1979), Chalmers and colleagues developed the *Oxford Database of Perinatal Trials* and a series of systematic reviews published in *Effective Care in Pregnancy and Childbirth* (Chalmers 1989). The database became a regularly updated electronic publication in 1989, developed into *Cochrane Pregnancy and Childbirth Database* in early 1993, and formed the basis of the broader *Cochrane Database of Systematic Reviews* (*CDSR*), launched in 1995. Work on a handbook to support authors of Cochrane Reviews had begun in 1993, and the first version was published in May 1994. Over its first 20 years, Cochrane grew from an initial group of 77 people from nine countries who met at the first Cochrane Colloquium in Oxford in 1993 to over 31,000 contributors from more than 120 countries in 2015, making it the largest organization involved in this kind of work (Allen 2006, Allen 2007, Allen 2011). Cochrane is now an internationally renowned initiative (Clarke 2005, Green 2005).

1.1.3 Cochrane organization and structure

Cochrane currently involves over fifty Cochrane Review Groups (CRGs), responsible for supporting the production and publication of reviews within specific areas of health. The review authors working with these groups include researchers, health professionals and people using healthcare services (consumers), all of whom share a common enthusiasm for generating reliable, up-to-date evidence relevant to the prevention and treatment of specific health problems or groups of problems.

CRGs are supported in this work by Methods Groups, Centres, Fields and by the Cochrane Editorial Unit (CEU). Cochrane Methods Groups provide a forum for methodologists to discuss development, evaluation and application of methods used to conduct Cochrane Reviews. They play a major role in the production of the *Cochrane Handbook for Systematic Reviews of Interventions* and, where appropriate, chapters in this volume contain information about relevant Methods Groups. Members of these Methods Groups have made major contributions to systematic review methodology (Chandler 2013). Cochrane Centres are located in different countries. Collectively, they represent all regions of the world and provide training and support for review authors and CRGs in addition to advocacy and promotion of access to Cochrane Reviews. Cochrane Fields focus on broad dimensions of health, such as the setting of care (e.g. primary care), the type of consumer (e.g. children), or the type of intervention (e.g. vaccines). People associated with Fields help to ensure that priorities and perspectives in their sphere of interest reflect the work of CRGs. The CEU provides strategic support and direction, and leads initiatives to improve and assure the quality of review activity across Cochrane.

1.2 Systematic reviews

1.2.1 The need for systematic reviews

Healthcare providers, consumers, researchers, and policy makers are inundated with unmanageable amounts of information, including evidence from health research. It is unlikely that they will have the time, skills and resources to find, appraise and interpret all this evidence and to incorporate it into healthcare decisions. Cochrane Reviews respond to this challenge by identifying, appraising and synthesizing research-based evidence and presenting it in an accessible format (Mulrow 1994). The requirement for systematic reviews to appraise the ever-growing proliferation of individual research studies has, if anything, become more important in recent years (Mallett 2003, Bastian 2010).

1.2.2 What is a systematic review?

A systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimizing bias, thus providing more reliable findings from which conclusions can be drawn and decisions made (Antman 1992, Oxman 1993). The key characteristics of a systematic review are:

- a clearly stated set of objectives with pre-defined eligibility criteria for studies;
- an explicit, reproducible methodology;
- a systematic search that attempts to identify all studies that meet the eligibility criteria;
- an assessment of the validity of the findings of the included studies, for example through the assessment of risk of bias; and
- a systematic presentation, and synthesis, of the characteristics and findings of the included studies.

Many systematic reviews contain meta-analyses. Meta-analysis is the use of statistical methods to summarize the results of independent studies (Glass 1976). By combining information from all relevant studies, meta-analyses can provide more precise estimates of the effects of health care than those derived from the individual studies included within a review (see Chapter 9, Section 9.1.3). Meta-analyses facilitate investigations of the consistency of evidence across studies, and the exploration of differences across studies.

1.3 Cochrane Reviews

Cochrane has developed a rigorous approach to the preparation of systematic reviews, with a structured review model. Cochrane publishes four main types of systematic reviews, summarized in Sections 1.3.1 to 1.3.4 and has a programme to explore development of review methods for other types of research question.

1.3.1 Reviews of the effects of interventions

Most Cochrane Reviews consider evidence on the effects of health or healthcare interventions. These reviews focus primarily on randomized studies as the most robust research design for assessment of the effects of interventions. Where evidence is unlikely to be found in randomized studies, for example for many adverse effects of interventions, or for large-scale interventions such as in public health or organizational change, reviews include non-randomized studies. Intervention reviews may additionally address broader issues such as economic issues or patient experiences of the intervention.

Cochrane has recently developed quality standards for the conduct and reporting of reviews. These standards summarize attributes for the conduct, and reporting, of reviews of interventions as set out in this *Handbook* (see Chapter 2, Section 2.4).

1.3.2 Reviews of diagnostic test accuracy

Cochrane has published systematic reviews of diagnostic test accuracy (DTA) in *CDSR* since 2008 (Leeflang 2013). These reviews evaluate how correctly a test detects the presence or absence of a target condition. Cochrane DTA reviews cover target conditions across health, including both pathologically defined diseases and more loosely defined indications for which treatments may be available. All types of tests are eligible, including: signs and symptoms from the patient history and examination; questionnaire-based tools, scores and decision rules; laboratory tests including biochemical, immunological, genetic, genomic and other 'pan-omic' technologies; imaging tests; and physiological measurements. Evaluation of the accuracy of a test is one component of the assessment of whether test use could lead to improvement in patient outcomes. Direct evaluation of how a test (and consequent decision-making and interventions) actually affect patient outcomes is best assessed by randomized studies that incorporate the effects of interventions that follow the test result. Such studies fit within the structure of Cochrane Intervention Reviews. However, randomized studies of test use are rare (especially outside the context of screening; Ferrante di Ruffano 2012), whereas accuracy studies are relatively common and provide most of the available evidence to guide test use, which makes them worthy of detailed systematic review. Although the stages in a DTA review are the same as for reviews of interventions, specific methodological challenges are encountered at each step: from formulation of review questions, through searching for and locating studies, assessing study quality, meta-analysis and interpretation of findings. Full methodological details are described in a separate *Cochrane Handbook for Diagnostic Test Accuracy Reviews* (http://srdta.cochrane.org/handbook-dta-reviews).

1.3.3 Overviews of Reviews

Cochrane Overviews of Reviews (Overviews) compile evidence from multiple systematic reviews into a single accessible and usable document. They are intended primarily to synthesize multiple Cochrane Reviews addressing a set of related interventions, populations, outcomes, or conditions, although other published non-Cochrane reviews may also be included. Cochrane Overviews provide the reader with a quick and comprehensive guide to reviews relevant to a specific decision. Overviews are aimed at decision makers, such as clinicians, policy makers, or informed consumers**,** who are accessing the *CDSR* for evidence on a specific problem. Overviews of Reviews on the effects of interventions are addressed in detail in Chapter 22 (see Section 22.1).

An overview of systematic reviews of diagnostic test accuracy (DTA Overview) can be used to synthesize and compare findings from a related set of test accuracy reviews. For example, an overview might bring together and compare the findings of separate reviews of alternative tests used to diagnose the same condition at the same point in the patient pathway. DTA Overviews also have a role in evaluating the accuracy of tests for the detection of closely related target conditions (particularly when they form part of a set of differential diagnoses), and in evaluating the performance of the same test across different settings. DTA Overviews are best planned when commencing on a portfolio of related individual systematic reviews, and plans for incorporation in a DTA Overview should be mentioned in the protocols of the individual reviews.

1.3.4 Reviews of methodology

Cochrane Methodology Reviews seek to answer questions about various aspects of the methods for systematic reviews, randomized studies and other evaluations of health and social care. They provide an evidence base for the methods of these evaluations, as well as providing descriptive accounts of other relevant issues, for example, to show the scale of problems faced by researchers working on systematic reviews or making decisions about health and social care. Cochrane Methodology Reviews use the widest range of study designs of Cochrane Reviews, including:

- experimental studies such as randomized studies, for example to compare different strategies to increase response rates to surveys;
- comparative observational studies, for example to examine the relationship between the use of reporting guidelines and the quality of research reports; and
- descriptive observational studies, for example of the proportion of studies presented at conferences that are also published in full.

Cochrane Methodology Reviews have a particular structure, based on the structure of Cochrane Intervention Reviews but with changes to some of the headings and sub-headings. The Cochrane Methodology Review Group has editorial responsibility for all Methodology Reviews. Appendix A provides a guide to the contents of a Cochrane Methodology protocol and review.

1.4 Publication of Cochrane Reviews

1.4.1 The Cochrane Library

Cochrane Reviews are published in full online in the *CDSR*, which is a core component of the Cochrane Library (www.thecochranelibrary.com). The Cochrane Library was first published in 1996, and is now an online collection of six databases (listed in 1.4.a) published by Wiley-Blackwell. In addition to the *CDSR*, the Cochrane Library includes additional resources that are provided by the Centre for Reviews and Dissemination (CRD) in York, UK. It is available free at the point of use in some countries, thanks to national licences and free one-click access provided by Wiley-Blackwell and Cochrane in most low- and middle-income countries, in association with Evidence Aid. Elsewhere it is subscription based, or pay-per-view. Since February 2013, reviews that have been published in full, or updated in full for the first time, now become freely available to all 12 months after their initial publication under an open access model.

Box 1.4.a: Databases published in the Cochrane Library

Active databases

The *CDSR* contains the full text (including methods, results and conclusions) of Cochrane Reviews and protocols.

The Cochrane Central Register of Controlled Trials (CENTRAL) is a highly concentrated source of reports of randomized and quasi-randomized studies. The majority of CENTRAL records are taken from bibliographic databases (mainly MEDLINE and Embase), but records are also derived from other published and unpublished sources.

The Health Technology Assessment database contains details of completed and ongoing health technology assessments (studies of the medical, social, ethical, and economic implications of healthcare interventions). It is produced by CRD, using information obtained from members of International Network of Agencies for Health Technology Assessment (INAHTA) and other health technology assessment organizations.

Archived databases

The Database of Abstracts of Reviews of Effects (DARE), assembled and previously maintained by CRD, contains critical assessments and structured abstracts of other systematic reviews, conforming to explicit quality criteria. This database was archived in March 2015.

The Cochrane Methodology Register (CMR) contains bibliographic information on articles and books on the science of reviewing research, and a prospective register of methodological studies. This database was archived in July 2012.

NHS Economic Evaluation Database (EED) contains appraised economic evaluations highlighting their relative strengths and weaknesses. It was produced by CRD. This database was archived in March 2015.

1.5 Handbook structure

There are three parts to the *Handbook*. Part 1 provides general information on Cochrane, its principles and the specific structure of Cochrane Reviews, their preparation, reporting, publication and maintenance. Part 2 provides the requisite methods to conduct a review with the required minimum standards. Part 3 covers a range of special topics for consideration when undertaking a Cochrane Review.

1.6 Chapter information

Acknowledgements: We thank previous chapter authors Sally Green, Philip Alderson, Cynthia Mulrow and Andrew Oxman on whose text this version is based. We also thank Ruth Foxlee for her contribution to 1.4.a.

1.7 References

Allen 2006

Allen C, Clarke M. International activity in Cochrane Review Groups with particular reference to China. *Chinese Journal of Evidence-based Medicine* 2006; 6: 541-545.

Allen 2007

Allen C, Clarke M, Tharyan P. International activity in Cochrane Review Groups with particular reference to India. *National Medical Journal of India* 2007; 20: 250-255.

Allen 2011

Allen C, Richmond K. The Cochrane Collaboration: international activity within Cochrane Review Groups in the first decade of the twenty-first century. *Journal of Evidence Based Medicine* 2011; 4: 2-7.

Antman 1992

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

Bastian 2010

Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: How will we keep up? *PLoS Medicine* 2010; 7: e1000326.

Chalmers 1989

Chalmers I, Enkin M, Keirse MJNC. *Effective care in pregnancy and childbirth*. Oxford (UK): Oxford University Press, 1989.

Chandler 2013

Chandler J, Hopewell S. Cochrane methods - twenty years experience in developing systematic review methods. *Systematic Reviews* 2013; 2: 76.

Clarke 2005

Clarke M. Cochrane Collaboration. In: Armitage P, Colton T, editor(s). *Encyclopedia of Biostatistics*. 2nd edition. Chichester (UK): John Wiley & Sons, 2005.

Cochrane 1979

Cochrane AL. 1931-1971: a critical review, with particular reference to the medical profession. In: Teeling-Smith G, Wells N, editor(s). *Medicines for the year 2000*. London (UK): Office of Health Economics, 1979.

Ferrante di Ruffano 2012

Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *Journal of Clinical Epidemiology* 2012; 65: 282-287.

Glass 1976

Glass GV. Primary, secondary and meta-analysis of research. *Educational Researcher* 1976; 5: 3-8.

Green 2005

Green S, McDonald S. The Cochrane Collaboration: More than systematic reviews? *Internal Medicine Journal* 2005; 35: 4-5.

Leeflang 2013

Leeflang MMG, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. *Systematic Reviews* 2013; 2: 82.

Mallett 2003

Mallett S, Clarke M. How many Cochrane Reviews are needed to cover existing evidence on the effects of healthcare interventions? *Evidence Based Medicine* 2003; 8: 100-101.

Mulrow 1994

Mulrow CD. Rationale for systematic reviews. *BMJ* 1994; 309: 597-599.

Oxman 1993

Oxman AD, Guyatt GH. The science of reviewing research. *Annals of the New York Academy of Sciences* 1993; 703: 125-133.

Chapter 8: Assessing risk of bias in included studies

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

This chapter should be cited as: Higgins JPT, Altman DG, Sterne JAC (editors). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS (editors), *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0 (updated June 2017), Cochrane, 2017. Available from www.training.cochrane.org/handbook.

This extract is from *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0. The previous version of this chapter (5.1.0, 2011) is available online at handbook.cochrane.org.

An earlier version (version 5.0.2, 2008) of the *Handbook* is also published by John Wiley & Sons, Ltd under "The Cochrane Book Series" Imprint, as Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on www.wiley.com.

Key Points

- Problems with the design and execution of individual studies of healthcare interventions raise questions about the validity of their findings; empirical evidence provides support for this concern.
- An assessment of the validity of studies included in a Cochrane Review should emphasize the risk of bias in their results, i.e. the risk that they will overestimate or underestimate the true intervention effect.
- Numerous tools are available for assessing methodological quality of clinical trials. The use of scales that yield a summary score is emphatically discouraged.
- Cochrane recommends a specific tool for assessing risk of bias in each included study. This comprises a judgement and a support for the judgement for each entry in a 'Risk

of bias' table, where each entry addresses a specific feature of the study. The judgement for each entry involves assessing the risk of bias as 'low risk', 'high risk, or 'unclear risk', with the last category indicating either lack of information or uncertainty over the potential for bias.

- Plots of 'Risk of bias' assessments can be created in Review Manager (RevMan).
- In clinical trials, biases can be categorized broadly as selection bias, performance bias, detection bias, attrition bias, reporting bias and other biases that do not fit into these categories.
- For parallel group trials, the features of interest in a standard 'Risk of bias' table of a Cochrane Review are sequence generation (selection bias), allocation sequence concealment (selection bias), blinding of participants and personnel (performance bias), blinding of outcome assessment (detection bias), incomplete outcome data (attrition bias), selective outcome reporting (reporting bias) and other potential sources of bias.
- Detailed considerations for the assessment of these features are provided in this chapter.

8.1 Introduction

The extent to which a Cochrane Review can draw conclusions about the effects of an intervention depends on whether the data and results from 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 11 (Section 11.2.2) and Chapter 12 (Section 12.2).

The second dimension of a study's validity relates to whether it answers its research question 'correctly', that is, in a manner that is 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 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, and Chapter 14 includes further considerations for adverse effects. Assessments of internal validity are frequently referred to as 'assessments of methodological quality' or 'quality assessment'. However, we will avoid the term quality, for reasons that will be explained in Section 8.2.2. In the next section we define 'bias' and distinguish it from the 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: different biases 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 apparent finding may be entirely due to bias). Even a particular source of bias may 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). In fact, as the results of a study may 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 closer 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).

C52 Cochrane Reviews must assess the risk of bias in all studies included in the review. This must be done 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 8.6 to 8.8). Sections 8.9 to 8.15 provide background considerations to assist review authors in using the tool.

Bias should not be confused with **imprecision**. Bias refers to *systematic error*, meaning that multiple replications of the same study would reach the wrong answer on average. Imprecision 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. Imprecision 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 phrase '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:

- 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.
- A study may be performed to the highest possible standards yet still have an important risk of bias. For example, in many situations it is impractical or impossible to blind participants or study personnel regarding intervention group. It is inappropriately judgemental to describe all such studies as of 'low quality', but that does not mean they are free of bias resulting from knowledge of intervention status.
- 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 (Schulz 2010), are unlikely to have direct implications for risk of bias.
- 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 have 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 1995a). This study demonstrated that trials in which randomization was inadequately concealed or inadequately reported yielded exaggerated estimates of intervention effect compared with trials that reported adequate concealment, and found a similar (but smaller) association 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 intervention 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 nine checklists that had been used to assess the validity or 'quality' of randomized trials (Moher 1995, Moher 1996). These scales and checklists included between three 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 were mentioned in textbooks. Many instruments 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 did not relate directly to internal validity.

The Cochrane 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, and is described in Section 8.5. It was developed between 2005 and 2007 by a working group of methodologists, editors and review authors. Since 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 in Moher 1995 were liable to confuse these separate issues. 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 1995a). 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).

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 websites (including trials registries). Published protocols are a particularly valuable source of information when they are available. The extraction of information from such reports is discussed in Chapter 7. Data collection forms should include space to extract sufficient details to allow implementation of the Cochrane 'Risk of bias' tool (Section 8.5). When extracting this information, it is highly desirable to record the source of each piece of information (including the precise location within a document). It is helpful to test data collection 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 other studies suggested little benefit from blind assessments (Berlin 1997, Kjaergard 2001). Blinded assessments are very time consuming, they may not be possible when the studies are well known to the review authors, 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 assessment of certain information in study reports.

Review authors with different levels of methodological training and experience may identify different sources of evidence and reach different judgements about risk of bias. Although experts in content areas may have preformed opinions that can influence their assessments (Oxman 1993), nonetheless, they may give more consistent assessments of

C53

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 potential biases that are not at first apparent. 'Risk of bias' assessments in Cochrane Reviews must be made independently by at least two people, with the process for resolving disagreements defined in advance. It is desirable that review authors should include both content experts and methodologists and 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 intervention the next patient should get?” More focused 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 trial 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. A useful classification of biases is into selection bias, performance bias, attrition bias, detection bias and reporting bias. In this section we describe each of these biases and introduce seven corresponding domains that are assessed in the Cochrane 'Risk of bias' tool. These are summarized in Table 8.4.a. We describe the tool for assessing the seven domains in Section 8.5. We provide more detailed consideration of each issue in Sections 8.9 to 8.15.

8.4.1 Selection bias

Selection bias refers to systematic differences between baseline characteristics of the groups that are compared. The unique strength of randomization is that, if successfully accomplished, it prevents selection bias in allocation of 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 foreknowledge of the forthcoming allocations. This process is often termed **allocation concealment**, although could more accurately be described as allocation sequence concealment. Thus, one suitable method for assigning interventions would be to use a simple random (and therefore unpredictable) sequence, and to conceal the upcoming allocations from those involved in enrolment into the trial.

8.4.2 Performance bias

Performance bias refers to systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest. 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. Effective blinding can also ensure that the groups being compared receive a similar amount of attention, ancillary treatment and diagnostic investigations. Blinding is not always possible, however. For example, it is usually impossible to blind people to whether or not major surgery has been undertaken.

8.4.3 Detection bias

Detection bias refers to systematic differences between groups in how outcomes are determined. **Blinding** (or masking) **of outcome assessors** may reduce the risk that knowledge of which intervention was received, rather than the intervention itself, affects outcome measurement. Blinding of outcome assessors can be especially important for assessment of subjective outcomes, such as degree of postoperative pain.

8.4.4 Attrition bias

Attrition bias refers to systematic differences between groups in withdrawals from a study. Withdrawals from the study lead to **incomplete outcome data**. There are two reasons for withdrawals or incomplete outcome data in clinical trials. *Exclusions* refer to situations in which some participants are omitted from reports of analyses, despite outcome data being available to the trialists. *Attrition* refers to situations in which outcome data are not available.

8.4.5 Reporting bias

Reporting bias refers to systematic differences between reported and unreported findings. Within a published report those analyses with statistically significant differences between intervention 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).

8.4.6 Other biases

In addition there are **other sources of bias** that are relevant only in certain circumstances. These relate mainly to 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. contamination, whereby the experimental and control interventions get 'mixed', for example if participants pool their drugs); and there may be sources of bias that are only found in a particular clinical setting.

For all potential sources of bias, it is important to consider the likely magnitude and direction of the bias. For example, if all methodological limitations of studies were expected to bias the results towards a lack of effect, and the evidence indicates that the intervention is effective, then it may be concluded that the intervention is effective even in the presence of these potential biases.

Table 8.4.a: A common classification scheme for bias

Type of bias	Description	Relevant domains in the Cochrane 'Risk of bias' tool
Selection bias	Systematic differences between baseline characteristics of 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 and personnel • Other potential threats to validity
Detection bias	Systematic differences between groups in how outcomes are determined	• Blinding of outcome assessment • Other potential threats to validity
Attrition bias	Systematic differences between groups in withdrawals from a study	• Incomplete outcome data
Reporting bias	Systematic differences between reported and unreported findings	• Selective outcome reporting (see also Chapter 10)

8.5 The Cochrane tool for assessing risk of bias

8.5.1 Overview

C52 This section describes the approach that must be used for assessing risk of bias in randomized studies included in Cochrane Reviews. It is a two-part tool, addressing the seven specific domains discussed in Sections 8.9 to 8.15 (namely sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective outcome reporting and (optionally) ‘other issues’). The tool is summarized in Table 8.5.a. Note that the tool was revised in late 2010 after an evaluation project. Changes made at that point are summarized in Table 8.5.b.

Each domain in the tool includes one or more specific entries in a ‘Risk of bias’ table. Within each entry, the first part of the tool describes what was reported to have happened in the study, in sufficient detail to support a judgement about the risk of bias. The second part of the tool assigns a judgement relating to the risk of bias for that entry. This is achieved by assigning a judgement of ‘low risk’ of bias, ‘high risk’ of bias, or ‘unclear risk’ of bias.

C56 The domains of sequence generation, allocation concealment and selective outcome reporting should each be addressed in the tool by a single entry for each study. For blinding of participants and personnel, blinding of outcome assessment and for incomplete outcome data, two or more entries 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 try to limit the number of entries used by grouping outcomes, for example, as ‘subjective’ or ‘objective’ outcomes for the purposes of assessing blinding of outcome assessment; or as ‘patient-reported at six months’ or ‘patient-reported at 12 months’ for incomplete outcome data. The same groupings of outcomes will be applied to every study in the review. The final domain (‘other bias’) can be assessed as a single entry for studies as a whole (this is the default setting in RevMan). However, it is strongly recommended that prespecified entries be used to address specific other risks of bias. Such author-specified entries may be for studies as a whole or for individual (or grouped) outcomes within every study.

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

Domain	**Support for judgement**	**Review authors' judgement**
Selection bias		
Random 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.	Risk of selection bias (biased allocation to interventions) due to inadequate generation of a randomized sequence.
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, enrolment.	Risk of selection bias (biased allocation to interventions) due to inadequate concealment of allocations prior to assignment.
Performance bias		
Blinding of participants and personnel *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.	Risk of performance bias due to knowledge of the allocated interventions by participants and personnel during the study.
Detection bias		
Blinding of outcome assessment *Assessments should be made for each main*	Describe all measures used, if any, to blind outcome assessors from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.	Risk of detection bias due to knowledge of the allocated interventions by outcome assessors.

outcome (or class of outcomes).		
Attrition bias		
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 in each intervention group (compared with total randomized participants), reasons for attrition/exclusions (where reported), and any reinclusions in analyses performed by the review authors.	Risk of attrition bias due to amount, nature or handling of incomplete outcome data.
Reporting bias		
Selective reporting	State how the possibility of selective outcome reporting was examined by the review authors, and what was found.	Risk of reporting bias due to selective outcome reporting.
Other bias		
Other sources of bias	State any important concerns about bias that are not addressed in the other domains of the tool. If particular questions/entries were prespecified in the review's protocol, responses should be provided for each question/entry.	Risk of bias due to problems not covered elsewhere in the table.

Table 8.5.b: Differences between the ‘Risk of bias’ tool described in *Handbook* versions 5.0.1/5.0.2 and the revised ‘Risk of bias’ tool described in *Handbook* version 5.1/5.2 (this version)

Separation of blinding	In the earlier version of the tool, biases related to blinding of participants, personnel and outcome assessors were all assessed within a single domain (although they may have been assessed separately for different outcomes). In the revised tool, bias related to blinding of participants and personnel is assessed in a separate domain from bias related to blinding of outcome assessment.
Nature of the judgement	The judgements are now expressed simply as ‘low risk’, ‘high risk’ or ‘unclear risk’ of bias. The domains are no longer expressed as questions, and the responses ‘Yes’ indicating low risk of bias and ‘No’ indicating high risk of bias have been removed.
Minor rewording	The items have been renamed in RevMan with the removal of question-based judgements: ‘*Adequate sequence generation?*’ became ‘*Random sequence generation*’. ‘*Allocation concealment?*’ became ‘*Allocation concealment*’. ‘*Blinding?*’ became ‘*Blinding of participants and personnel*’ and ‘*Blinding of outcome assessment*’. ‘*Incomplete outcome data addressed?*’ became ‘*Incomplete outcome data*’. ‘*Free of selective reporting?*’ became ‘*Selective reporting*’. ‘*Free of other bias?*’ became ‘*Other bias*’.
Insertion of categories of bias	The revised tool clarifies the category of bias within which each domain falls: selection bias (random sequence generation and allocation concealment), performance bias (blinding of participants and personnel), detection bias (blinding of outcome assessment), attrition bias (incomplete outcome data), reporting bias (selective reporting) and other bias.
Reconsideration of eligible issues for other	The guidance for the other bias domain has been edited to strengthen the guidance that additional items should be used only exceptionally, and that these items should relate

bias, including early stopping of a trial	to issues that may lead directly to bias. In particular, the mention of early stopping of a trial has been removed, because: 1) simulation evidence suggests that inclusion of trials that stopped early in meta-analyses will not lead to substantial bias, and 2) exclusion of trials that stopped early has the potential to bias meta-analyses towards the null (as well as leading to loss of precision).

8.5.2 The support for judgement

C54 C55 All judgements of risk of bias in the 'Risk of bias' tool must be supported by a succinct summary of the evidence or rationale underlying the judgement. This aims to ensure transparency in how these judgements are reached. The source of information in the supporting statement should be made clear. For a specific study, information for the support for a judgement 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 support for judgement 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 judgements 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 judgement, this should be stated explicitly. Examples of proposed formatting for the description are provided in Table 8.5.c.

Table 8.5.c: Examples of supports for judgement for sequence generation entry (fictional)

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 judgement

Review authors’ judgements should be categorized as ‘low risk’ of bias, ‘high risk’ of bias or ‘unclear risk’ of bias. The assessments should consider the risk of *material* bias rather than *any* bias. We define ‘material bias’ as bias of sufficient magnitude to have a notable impact on the results or conclusions of the trial, recognizing that subjectivity is involved in any such judgement.

Table 8.5.d provides criteria for making judgements about risk of bias from each of the seven domains in the tool. If insufficient detail about what happened in the study is reported, the judgement will usually be ‘unclear risk’ of bias. An ‘unclear’ judgement should also be made if what happened in the study is known, but the risk of bias is unknown; or if an entry is not relevant to the study at hand (particularly for assessing blinding and incomplete outcome data, when the outcome being assessed by the entry has not been measured in the study).

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

Random sequence generation

Selection bias (biased allocation to interventions) due to inadequate generation of a randomized sequence

Criteria for a judgement of '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.* *Minimization may be implemented without a random element, and this is considered to be equivalent to being random.
Criteria for the judgement of '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 rule based on date (or day) of admission; • sequence generated by some rule based on hospital or clinic record number. Other non-random approaches happen much less frequently than the systematic approaches mentioned here and tend to be obvious. They usually involve judgement or some method of non-random categorization of participants, for example: • allocation by judgement 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 judgement of 'unclear risk' of bias	Insufficient information about the sequence generation process available to permit a judgement of 'low risk' or 'high risk'.

Allocation concealment

Selection bias (biased allocation to interventions) due to inadequate concealment of allocations prior to assignment

Criteria for a judgement of '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: • central allocation (including telephone, web-based and pharmacy-controlled randomization); • sequentially numbered drug containers of identical appearance;

	• sequentially numbered, opaque, sealed envelopes.
Criteria for the judgement of 'high risk' of bias	Participants or investigators enrolling participants could possibly foresee assignments, and thus introduce selection bias, due to allocation based on: • use of an open random allocation schedule (e.g. a list of random numbers); • use of assignment envelopes without appropriate safeguards (e.g. 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 judgement of 'unclear risk' of bias	Insufficient information available to permit a judgement of 'low risk' or 'high risk'. This is usually the case if the method of concealment is not described or not described in sufficient detail to allow a definite judgement – for example if the use of assignment envelopes was described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.

Blinding of participants and personnel

Performance bias due to knowledge of the allocated interventions by participants and personnel during the study

Criteria for a judgement of 'low risk' of bias	Either of the following:

	• no blinding or incomplete blinding, but the review authors judge that the outcome was 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.
Criteria for the judgement of 'high risk' of bias	Either of the following: • no blinding or incomplete blinding, and the outcome was 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, and the outcome was likely to be influenced by lack of blinding.
Criteria for the judgement of 'unclear risk' of bias	Either of the following: • insufficient information available to permit a judgement of 'low risk' or 'high risk'; • the study did not address this outcome.

Blinding of outcome assessment

Detection bias due to knowledge of the allocated interventions by outcome assessors

Criteria for a judgement of 'low risk' of bias	Either of the following: • no blinding of outcome assessment, but the review authors judge that the outcome measurement was not likely to be influenced by lack of blinding; • blinding of outcome assessment ensured, and unlikely that the blinding could have been broken.

Criteria for the judgement of 'high risk' of bias	Either of the following: • no blinding of outcome assessment, and the outcome measurement was likely to be influenced by lack of blinding; • blinding of outcome assessment, but likely that the blinding could have been broken, and the outcome measurement was likely to be influenced by lack of blinding.
Criteria for the judgement of 'unclear risk' of bias	Either of the following: • insufficient information available to permit a judgement of 'low risk' or 'high risk'; • the study did not address this outcome.

Incomplete outcome data

Attrition bias due to amount, nature or handling of incomplete outcome data

Criteria for a judgement of 'low risk' of bias	Any one of the following: • 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 the observed event risk is not enough to have had a clinically relevant impact on the intervention effect estimate; • for continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not enough to have had a clinically relevant impact on the observed effect size; • missing data have been imputed using appropriate methods.
Criteria for the judgement of ‘high risk’ of bias	Any one of the following: • reason for missing outcome data is 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 the observed event risk is enough to have induced clinically relevant bias in the intervention effect estimate; • for continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is enough to have induced clinically relevant bias in the observed effect size; • ‘as-treated’ analysis done with substantial departure of the intervention received from that assigned at randomization; • potentially inappropriate application of simple imputation.
Criteria for the judgement of ‘unclear risk’ of bias	Either of the following: • insufficient reporting of attrition/exclusions to permit a judgement of ‘low risk’ or ‘high risk’ (e.g. number randomized not stated, no reasons for missing data provided);

	• the study did not address this outcome.

Selective reporting

Reporting bias due to selective outcome reporting

Criteria for a judgement of 'low risk' of bias	Either of the following: • the study protocol is available and all of the study's prespecified (primary and secondary) outcomes that are of interest in the review have been reported in the prespecified way; • the study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were prespecified (convincing text of this nature may be uncommon).
Criteria for the judgement of 'high risk' of bias	Any one of the following: • not all of the study's prespecified primary outcomes have been reported; • one or more primary outcomes have been reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not prespecified; • one or more reported primary outcomes were not prespecified (unless clear justification for their reporting is provided, such as an unexpected adverse effect); • one or more outcomes of interest in the review have been reported incompletely so that they cannot be entered in a meta-analysis;

	• the study report failed to include results for a key outcome that would be expected to have been reported for such a study.
Criteria for the judgement of 'unclear risk' of bias	Insufficient information available to permit a judgement of 'low risk' or 'high risk'. It is likely that the majority of studies will fall into this category.

Other bias

Bias due to problems not covered elsewhere in the table

Criteria for a judgement of 'low risk' of bias	The study appears to be free of other sources of bias.
Criteria for the judgement of '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; • has been claimed to have been fraudulent; • had some other problem.
Criteria for the judgement of 'unclear' 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; • 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 RevMan for inclusion in a Cochrane Review as part of the 'Characteristics of included studies' table. For each entry, the judgement ('low risk' of bias; 'high risk' of bias, or 'unclear risk' of bias) is followed by a text box for a description of the design, conduct or observations that underlie the judgement. Figure 8.6.a provides an example of how it might look. If the text box is left empty, and the judgement is left as 'unclear risk', then the entry will be omitted from the 'Risk of bias' table for the study on publication in the *Cochrane Database of Systematic Reviews* (*CDSR*).

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

Three types of figures may be generated using RevMan to present 'Risk of bias' assessments in a published review. Firstly, a 'Risk of bias' graph illustrates the proportion of studies with each of the judgements ('low risk', 'high risk', 'unclear risk' of bias) for each entry in the tool (see Figure 8.6.b). Secondly, a 'Risk of bias' summary figure presents all of the judgements in a cross-tabulation of study by entry (see Figure 8.6.c). Thirdly (in RevMan 5.3 onwards), a standard forest plot can present the judgements as they appear in the 'Risk of bias' summary figure, alongside the results for each study. Where different judgements have been recorded for different outcome groups (i.e. for performance bias, detection bias, attrition bias and any user-defined domains assigned to assessment at the outcome level, as indicated in Section 8.5.1), the outcome illustrated in the forest plot must be linked to the correct outcome-level 'Risk of bias' assessments within RevMan.

An alternative, and perhaps preferable, version of the first figure (the 'Risk of bias' graph) would be to restrict attention to studies in a particularly important meta-analysis, and to represent the proportion of information (rather than the proportion of studies) at low risk, unclear risk and high risk of bias. The proportion of information may be measured by the sums of weights awarded to the studies in the meta-analysis. Currently, however, such plots cannot be produced within RevMan.

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

Entry	Judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Quote: "patients were randomly allocated." Comment: Probably done, since earlier reports from the same investigators clearly describe use of random sequences (Cartwright 1980).
Allocation concealment (selection bias)	High risk	Quote: ". . . using a table of random numbers." Comment: probably not done

Blinding of participants and personnel (performance bias)	Low risk	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 of outcome assessment (detection bias; patient-reported outcomes)	Low risk	Quote: “double blind” Comment: probably done
Blinding of outcome assessment (detection bias; all-cause mortality)	Low risk	Obtained from medical records; review authors do not believe this will introduce bias.
Incomplete outcome data addressed (attrition bias; short-term (2-6 weeks))	High risk	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 (attrition bias; long-term (> 6 weeks))	High risk	12 weeks: 31/110 missing from intervention group; 18/113 missing from control group. Reasons differed across groups.
Selective reporting (reporting bias)	High risk	Three rating scales for cognition listed in Methods, but only one (with statistically significant results) was reported.

Figure 8.6.b: Example of a ‘Risk of bias’ graph

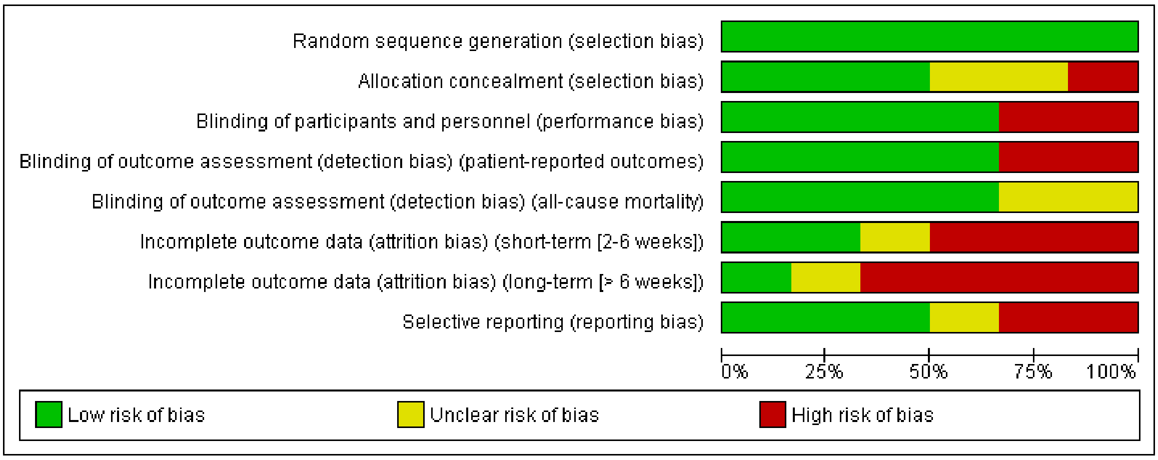
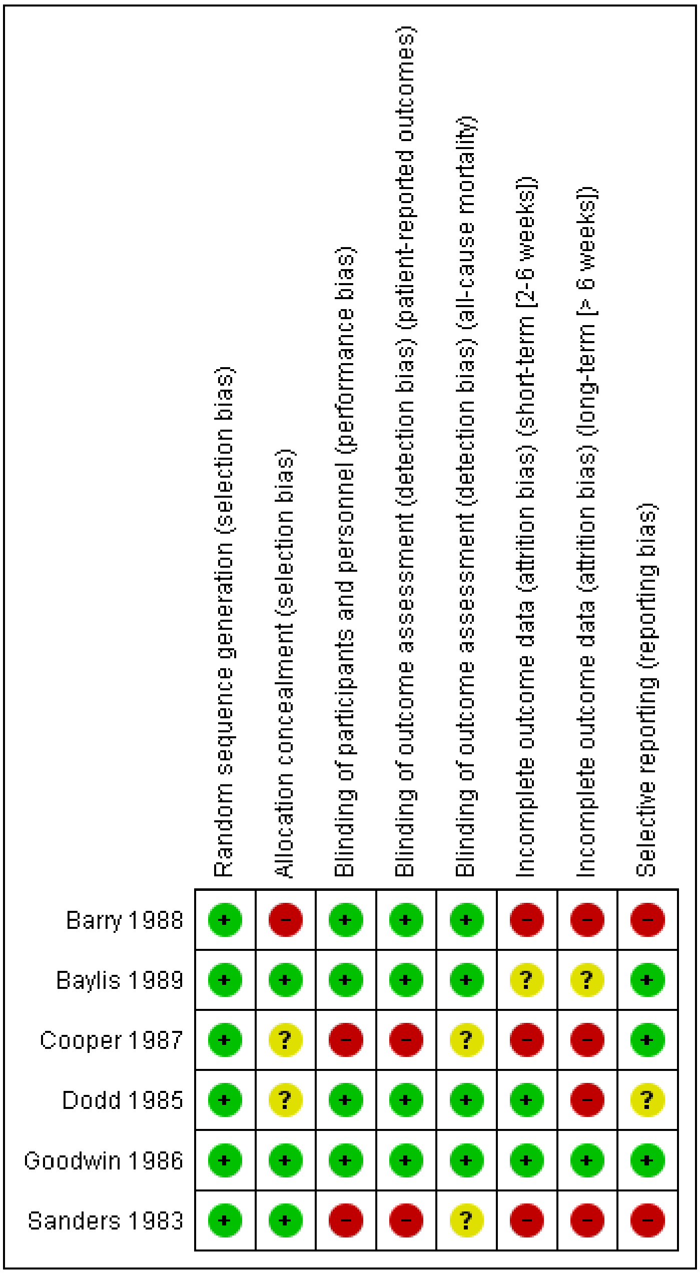


Figure 8.6.c: Example of a 'Risk of bias' summary figure



	Random sequence generation (selection bias)	Allocation concealment (selection bias)	Blinding of participants and personnel (performance bias)	Blinding of outcome assessment (detection bias) (patient-reported outcomes)	Blinding of outcome assessment (detection bias) (all-cause mortality)	Incomplete outcome data (attrition bias) (short-term [2-6 weeks])	Incomplete outcome data (attrition bias) (long-term [> 6 weeks])	Selective reporting (reporting bias)
Barry 1988	+	−	+	+	+	−	−	−
Baylis 1989	+	+	+	+	+	?	?	+
Cooper 1987	+	?	−	−	?	−	−	+
Dodd 1985	+	?	+	+	+	+	−	?
Goodwin 1986	+	+	+	+	+	+	+	+
Sanders 1983	+	+	−	−	?	−	−	−

8.7 Summary assessments of risk of bias

Cochrane's recommended tool for assessing risk of bias in included studies involves the assessment and presentation of individual domains, 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 domains. A review author will have to make judgements about which domains 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. How such judgements are reached should be made explicit and they should be informed by:

- **Empirical evidence of bias**: Sections 8.5 to 8.15 summarize empirical evidence of the association between domains 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 criteria, such as adequate allocation concealment, is associated with overestimates of effect. If the likely direction of bias for a domain is such that effects will be underestimated (biased towards the null), then, providing the review demonstrates an important effect of the intervention, such a domain may be of less concern.

- **Likely magnitude of bias**: The likely magnitude of bias associated with any domain 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 about 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 sequence concealment and a small estimate of effect might substantially reduce confidence in the estimate, whereas minor inadequacies in how incomplete outcome data were addressed might not reduce confidence in a large estimate of effect substantially.

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

- **Summarizing risk of bias for a study across outcomes**: Some domains affect the risk of bias across outcomes in a study: e.g. sequence generation and allocation sequence concealment. Other domains, such as blinding and incomplete outcome data, may have different risks of bias for different outcomes within a study. Thus, review authors should not assume 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 domains)**: 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. Indeed, it is highly recommended that risk of bias is summarized at this level. A summary assessment of the risk of bias for an outcome should include all of the entries relevant to that outcome: i.e. both study-level entries, such as allocation sequence concealment, and outcome specific entries, such as blinding. C58

- **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 11.2). As explained in Section 8.8, including study results at high risk of bias in a meta-analysis may lead to the quality of evidence being lower than if such trials were excluded.

- **Summarizing risk of bias for a review as a whole (across studies and outcomes)**: Summarizing the overall risk of bias in a review should be avoided for two reasons. Firstly, this requires value judgements about which outcomes are critical to a decision. 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 outcomes that are critical to such an assessment. Secondly, judgements about which outcomes are critical to a decision may vary from setting to setting, because of differences in both societal values and other factors, such as baseline risk. 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 domains ('key domains') 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 domains) 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 domains	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 domains	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 domains	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 some 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).

C59 C60 When performing and presenting meta-analyses, review authors must address risk of bias in the results of included studies, and when randomized studies are involved, this must be based on the Cochrane 'Risk of bias' tool. 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, and the lower will be the grading of the quality of the evidence.

8.8.2 Exploring the impact of risk of bias

8.8.2.1 Graphing results according to risk of bias

The discussion that follows applies both individual bias domains and 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 judgements on each 'Risk of bias' entry are available in RevMan 5. Such plots give a visual impression 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 is usually sensible to restrict such plots to key bias domains (see Section 8.7).

8.8.2.2 Studies assessed as at unclear risk of bias

Studies are assessed as being at an 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 entry is not relevant to a study (for example because the study did not address any of the outcomes in the group of outcomes to which the entry applies). 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 a 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 Schulz 1995a study found that intervention odds ratios were exaggerated by 41% for trials with inadequate concealment (high risk of bias) and by 30% for trials with unclear concealment (unclear risk of bias). However, most empirical studies 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, Section 9.6.4). 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 a 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 exaggerated in studies at high or unclear risk of bias compared with studies at a low risk of bias.

When a meta-analysis includes many studies, meta-regression analyses can include more than one domain (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. As meta-analyses usually contain a small number of studies, usually the ratio of odds ratios is estimated imprecisely. 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 bias and a substantial effect of bias.

A test for differences across subgroups provides an alternative to meta-regression for examination of a single entry (e.g. comparing studies with adequate versus 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 in any case the P values will be too small in the presence of heterogeneity 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 a 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 most frequently used approach to incorporating 'Risk of bias' assessments in Cochrane Reviews is to **restrict** meta-analyses to studies at a low (or lower) risk of bias, or to **stratify** studies according to 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, when deciding which result to present in the Abstract). The intended strategy should be described in the protocol for the review.

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

The first 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 eligibility criteria, or using reasoned argument (which may draw on empirical evidence of bias from meta-epidemiological studies). In rare cases, within-meta-analysis comparisons of studies at high and low risk of bias may produce evidence of differences between intervention effect estimates and justify restricting analyses to studies at a low risk of bias (see Section 8.8.2.3). If the primary analysis includes studies at an unclear risk of bias, review authors should justify 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, in theory, studies may 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 number. Review authors who restrict their primary analysis in this way are encouraged to perform **sensitivity analyses** to show how conclusions might be affected if studies at a high risk of bias were included.

2. **Present multiple (stratified) analyses**

Stratifying according to the summary risk of bias may produce at least three estimates of the intervention effect: from studies at high and low risks of bias and from all studies. Two or more such estimates might be presented with equal prominence, for example, one including all studies and one including only those at a 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. Furthermore, usually ‘Summary of findings’ tables will present only a single result for each outcome. On the other hand, a stratified forest plot presents all the information transparently.

The choice between strategies 1 and 2 should be based on the context of the particular review and the balance between the potential for bias and the loss of precision when studies at a high or unclear risk of bias are excluded. As explained in Section 8.8.2.3, lack of a statistically significant difference between studies at a high and low risk of bias should not be interpreted as implying an absence of bias, because meta-regression analyses typically have low power.

3. 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 intervention 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 a high risk, all are at an 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. Firstly, 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 ‘Authors’ conclusions’, ‘Abstract’ and ‘Summary of findings’ table, so that the final interpretation ignores the risk of bias and decisions continue to be based, at least in part, on flawed evidence. Secondly, such an analysis fails to down-weight studies at a high risk of bias and so will lead to an overall intervention that is too precise, as well as being potentially biased.

When the primary analysis is based on all studies, summary assessments of risk of bias must be incorporated into explicit measures of the quality of evidence for each important outcome, for example using the GRADE system (Guyatt 2008). This can help to ensure that judgements about the risk of bias, as well as other factors affecting the quality of evidence, such as imprecision, heterogeneity and publication bias, are taken into consideration appropriately in interpreting the results of the review (See Chapter 11, Section 11.2).

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 method for combining results of multiple studies is to 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 additionally 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, Section 16.8; Turner 2009). 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; currently they are not sufficiently well developed for widespread adoption.

8.9 Random sequence generation

8.9.1 Rationale for concern about bias

Under the domain of random sequence generation in the Cochrane tool for assessing risk of bias, we address whether or not the study used a randomized sequence of assignments. This is the first of two domains in the Cochrane tool that addresses the allocation process, the second being concealment of the allocation sequence (allocation concealment). We start by explaining the distinction between these domains.

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. Firstly, an allocation sequence must be used that, if perfectly implemented, would balance prognostic factors, on average, evenly across intervention groups. Randomization plays a fundamental role here. It can be argued that other assignment rules, such as alternation (alternating between two interventions) or rotation (cycling through more than two interventions), can achieve the same thing (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 and non-enrolment of participants into a study in the light of the upcoming intervention assignment.

Future assignments may be anticipated for several reasons. These include: 1) knowledge of a deterministic assignment rule, such as by alternation, date of birth or day of admission; 2) knowledge of the sequence of assignments, whether randomized or not (e.g. if a sequence of random assignments is posted on the wall); 3) 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 practical aspect is concealment of the allocation sequence, that is, the use of mechanisms to prevent foreknowledge of the next assignment. Historically this has been assessed in Cochrane Reviews, with empirical justification. We address allocation sequence concealment as a separate domain in the tool (see Section 8.10).

Randomization allows for the sequence to be unpredictable. An unpredictable sequence, combined with allocation sequence 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 sequence 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. We do not consider this special situation under either sequence generation or allocation concealment, but address it as a separate consideration in Section 8.15.1.3.

Methodological studies have assessed the importance of sequence generation, including several that have avoided confounding by disease or intervention, which is critical to the assessment (Schulz 1995a, Moher 1998, Kjaergard 2001, Siersma 2007). The BRANDO (Bias in Randomized and Observational Studies) project, which combined data from all available meta-epidemiologic studies, included a reanalysis of 112 meta-analyses from multiple methodological studies that indicated an average exaggeration of 11% in studies with inadequate or unclear sequence generation (relative odds ratio 0.8; 95% confidence interval (CI) 0.82 to 0.96; (Savovic 2012a). In one study, which restricted the analysis to 79 trials that had reported an adequately concealed **allocation sequence**, trials with inadequate **sequence generation** yielded exaggerated estimates of intervention effects, on average, when compared against trials with adequate sequence generation (relative odds ratio of 0.75; 95% CI 0.55 to 1.02; P = 0.07). These results suggest that, if assignments are non-random, some deciphering of the sequence can occur, even with apparently adequate concealment of the allocation sequence (Schulz 1995a).

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

Sequence generation is often improperly addressed in the design and implementation phases of randomized controlled trials, and is often neglected in published reports, which causes major problems when assessing the risk of bias. The following considerations may help review authors assess whether sequence generation is suitable to protect against bias, when using the Cochrane 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.

When randomization is used, without constraints, to generate an allocation sequence it is called **simple randomization** or **unrestricted randomization**. In principle, this could be achieved by allocating interventions using methods such as repeated coin-tossing, throwing dice or dealing previously shuffled cards (Schulz 2002a, 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 (usually meaning at least 100 in each randomized group (Schulz 2002a, Schulz 2002b, Schulz 2006), simple randomization generates comparison groups of relatively similar sizes. In trials that use small samples, simple randomization will sometimes result in an allocation sequence that leads to 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 generate a sequence to ensure particular allocation ratios to the intervention groups (e.g. 1:1). Blocked randomization (random permuted blocks) is a common form of restricted randomization (Schulz 2002a, 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 intervention 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 eight and an equal allocation ratio.*

Stratified randomization is also common; in this, 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 were used in each stratum, then stratification would have no effect, 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 used stratification.

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 for several characteristics. Use of minimization should not automatically be considered as putting 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 include biased coin or urn randomization, replacement randomization, mixed randomization, and maximal randomization (Schulz 2002a, Schulz 2002b, 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.

An important weakness with all systematic methods is that concealment of the allocation schedule is usually impossible; this allows foreknowledge of intervention 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 Intervention A and patients born on odd days were assigned to Intervention 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 define their approach incompletely and do not confirm some random component in the process. For example, authors may state that blocked randomization was used, but the process for selecting the blocks, such as a random number table or a computer random number generator, may not be specified. The adequacy of sequence generation should then be classified as unclear.

8.10 Allocation sequence concealment

8.10.1 Rationale for concern about bias

Randomized sequence generation is a necessary, but not a sufficient, safeguard against bias in intervention allocation. Efforts made to generate unpredictable and unbiased sequences are likely to be ineffective if those sequences are not protected by adequate concealment of the allocation sequence 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 1995a, Schulz 1995b, Jüni 2001). Personal accounts suggest that many allocation schemes have been deciphered by investigators because the methods of concealment were inadequate (Schulz 1995b).

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 sequence** shields those who admit participants to a study from knowing the upcoming assignments.

Several methodological studies have looked at whether concealment of allocation sequence is associated with magnitude of effect estimates in controlled clinical trials while avoiding confounding by disease or intervention. 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% CI 5% to 29%; (Pildal 2007). The BRANDO project, which combined data from all available meta-epidemiologic studies, included a reanalysis of 146 meta-analyses and observed an exaggeration in intervention effect by an average of 7% (relative odds ratio 0.93; 95% CI 0.87 to 0.99; (Savovic 2012b). There was evidence of a larger impact among meta-analyses with subjectively assessed outcomes (relative odds ratio 0.85), but less impact on objectively assessed outcomes (relative odds ratio 0.97), such as all-cause mortality (relative odds ratio 0.98).

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

The following considerations may help review authors assess whether concealment of allocation is sufficient to protect against bias, when using the Cochrane tool (Section 8.5).

Proper concealment of the allocation sequence 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 1995a). 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 staffroom wall.

Some review authors confuse allocation concealment with blinding of allocated interventions. Allocation concealment seeks to prevent selection bias in intervention assignment by protecting the allocation sequence *before and until* assignment, and can always be successfully implemented regardless of the study topic (Schulz 1995a, Jüni 2001). In contrast, blinding seeks to prevent performance and detection bias by protecting the sequence *after* assignment (Jüni 2001, Schulz 2002c), and cannot always be implemented – for example, in trials comparing surgical with medical interventions. 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 the extent to which potential participants in the study have different prognoses, whether strong beliefs exist among investigators and participants regarding the benefits or harms of assigned interventions, and whether uncertainty about the interventions is accepted by all people involved (Schulz 1995b). Among the different methods used to conceal allocation, central randomization by a third party is perhaps the most desirable. Methods that use envelopes are more susceptible to manipulation than other approaches (Schulz 1995a). 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 sequence concealment

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

Examples (of low risk of bias; published descriptions of concealment procedures judged to be adequate, as compiled (Schulz 2002d)):

" *. . . 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 [the] 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 sequence to be adequate (low risk of bias)

Minimal criteria for a judgement of adequate concealment of the allocation sequence	Extended criteria to provide additional assurance

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.
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.
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.

8.11 Blinding of participants and personnel

8.11.1 Rationale for concern about bias

Several types of people can be blinded in a clinical trial: see Box 8.11.a. The first of the two domains in the tool that specifically address blinding focuses on participants and personnel (healthcare providers). It is highly desirable for blinding of participants and personnel to be separated from blinding of outcome assessors, which is covered in the second blinding-related domain (see Section 8.12). Lack of blinding of participants or healthcare providers could bias the results by affecting the *actual* outcomes of the participants in the trial. This may be due to a 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 cointerventions).

Empirical evidence of bias due to lack of blinding of participants and personnel is not currently available. However, there is evidence for studies described as 'blind' or 'double-blind', which usually includes blinding of one or both of these groups of people. In empirical studies, lack of blinding in randomized trials has been shown to be associated with more exaggerated estimated intervention effects – by 13% on average – measured as odds ratios (Savovic 2012b). These studies have dealt with a variety of outcomes, some of which were objective. The estimated effect has been observed to be more biased, on average, in trials with more subjective outcomes (Wood 2008). Lack of blinding might also lead to bias caused by additional investigations or cointerventions regardless of the type of outcomes, if these occur differentially across intervention groups.

Blinding can be impossible for at least some people (e.g. most patients receiving 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 healthcare providers. An attempt to blind participants and personnel does not ensure successful blinding in practice. Blinding can be compromised for most interventions. For many blinded drug trials, the side effects of the drugs allow 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).

In blinded studies, especially placebo-controlled trials, there may be concern about whether the participants were truly blinded (and sometimes also whether those caring for the participants were). Several groups have suggested that it would be sensible to ask trial participants to guess which intervention they have been receiving at the end of the trial (Fergusson 2004, Rees 2005), and some reviews of such reports have been published (Fergusson 2004, Hróbjartsson 2007). Evidence of correct guesses exceeding 50% would seem to suggest that blinding may have been broken, but in fact can simply reflect the patients' experiences in the trial: a good outcome, or a marked side effect, will tend to be more often attributed to an active intervention, and a poor outcome to a placebo (Sackett 2007). It follows that we would expect to see some successful 'guessing' when there is a difference in either efficacy or adverse effects, but none when the interventions have very similar effects, even when the blinding has been preserved. As a consequence, review authors should consider carefully whether to take any notice of the findings of such an exercise.

Box 8.11.a: A note on blinding in clinical trials

In general, blinding (sometimes called masking) refers to the process by which study participants, health providers and investigators, including people assessing outcomes, are kept unaware of intervention allocations after inclusion of participants in the study. Blinding may reduce the risk that knowledge of which intervention was received – rather than the intervention itself – will affect outcomes and assessments of outcomes.

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

- participants (e.g. patients or healthy people);

- healthcare providers/personnel (e.g. the doctors or nurses responsible for care);
- 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 outcome adjudication committees);
- data analysts (e.g. statisticians); and
- manuscript writers.

The first two types of people are addressed in the tool under the item 'Blinding of participants and personnel'. The third is addressed by the item 'Blinding of outcome assessment'. The last two are not explicitly covered by the tool.

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

Study reports often describe blinding in broad terms, such as 'double blind'. This term makes it impossible to know who was blinded (Schulz 2002c). 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 in the CONSORT Statement to be explicit (Schulz 2010). A review of methods used for blinding highlighted the variety of methods used in practice (Boutron 2006). The following considerations may help review authors assess whether any blinding of participants and personnel in a study was likely to be sufficient to protect against bias, when using the Cochrane tool (Section 8.5).

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

- who was and was not blinded; and
- risk of bias in actual outcomes due to lack of blinding during the study (e.g. due to cointervention or differential behaviour).

Risk of bias may be high for some outcomes and low for others, even if the same people were unblinded in the study. For example, knowledge of the assigned intervention may impact on behavioural outcomes (such as number of clinic visits), while not impacting on physiological outcomes or mortality. Thus, it is highly desirable for assessments of risk of bias resulting from lack of blinding 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. C56

8.12 Blinding of outcome assessment

8.12.1 Rationale for concern about bias

Several types of people can be blinded in a clinical trial: see Box 8.11.a. The second of the two domains in the tool that specifically addresses blinding focuses on blinding of outcome assessors. If people who determine outcome measurements are aware of intervention assignments, bias could be introduced into assessments. Outcome assessments may be made by the participants themselves, by their healthcare providers, or by independent assessors.

Empirical studies have shown that lack of blinding in randomized trials is associated with more exaggerated estimated intervention effects – by 13% on average – measured as odds ratios (Savovic 2012b). These studies have dealt with a variety of outcomes, some of which are objective. The estimated effect has been observed to be more biased, on average, in trials with more subjective outcomes (Wood 2008, Savovic 2012b). Recently, a systematic review of trials with both blinded and non-blinded assessment of the same outcome showed biased effect estimates in unblinded assessment, which, for subjective outcomes, exaggerated the odds ratios by 36% (Hróbjartsson 2012).

All outcome assessments can be influenced by lack of blinding, although there are particular risks of bias with more subjective outcomes (e.g. pain or number of days with a common cold). 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.

Blinding of outcome assessment can be impossible (e.g. when patients have received major surgery). However, this does not mean that potential biases can be ignored, and review authors should still assess the risk of bias due to lack of blinding of outcome assessment for all studies in their review.

8.12.2 Assessing risk of bias in relation to adequate or inadequate blinding of outcome assessment

Study reports often describe blinding in broad terms, such as 'double blind'. This term makes it impossible to know who was blinded (Schulz 2002c). 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 in the CONSORT Statement to be explicit (Moher 2001). A review of methods used for blinding highlighted the variety of methods used in practice (Boutron 2006). The following considerations may help review authors assess whether any blinding of outcome assessment used in a study was likely to be sufficient to protect against bias, when using the Cochrane tool (Section 8.5).

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

- who is assessing the outcome; and

- the risk of bias in the outcome assessment (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 obviously be collected with knowledge of the intervention received, whereas other outcomes, measured by an independent clinician (e.g. physical ability), might be blinded. Furthermore, risk of bias may be high for some outcomes and low for others, even if the same people were unblinded in the study. For example, knowledge of the assigned intervention may impact on patient-reported outcomes (such as level of pain), while not impacting on other outcomes such as mortality. In many circumstances the assessment of total mortality might be considered to be unbiased, even if outcome assessors were aware of intervention assignments. Thus, it is highly desirable for assessments of risk of bias resulting from lack of blinding to be made separately for different outcomes. C56

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.

8.13 Incomplete outcome data

8.13.1 Rationale for concern about bias

Missing outcome data, due to attrition (drop out of participants) 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 for the following reasons.

- 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.

In addition, some participants may be excluded from analysis for the following reasons.

- Some participants are enrolled in the study, but 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 8.13.2).
- The study analysis excluded some participants for other reasons.

Some exclusions of participants may be 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, and preferably on the basis of decisions made while blinded to assignment. The intention to exclude such participants should be specified before the outcome data are seen.

An intention-to-treat (ITT) analysis is often recommended as the least biased way to estimate intervention effects in randomized trials (Newell 1992): see Chapter 16 (Section 16.2). The principles of ITT analyses are to:

- keep participants in the intervention groups to which they were randomized, regardless of the intervention they actually received;
- measure outcome data on all participants; and
- include all randomized participants in the analysis.

The first principle can always be applied. However, the second is often impossible due to attrition beyond the control of the trialists. Consequently, the third principle of conducting an analysis that includes all participants can only be followed by making assumptions about the missing values. Thus very few trials can perform a true ITT analysis without making imputations (see Section 8.13.2.3), especially when there is extended follow-up. 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 outlined above, and should interpret any studies that use the term without clarification with care.

Review 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 intervention. 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 (**reinclusions**), if the review author considers the reasons for exclusions to be inappropriate and the data are available. 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 1995a, Kjaergard 2001, Balk 2002, Siersma 2007). Tierney 2005 observed a tendency for analyses, conducted after trial authors had excluded participants, to favour the experimental intervention compared with analyses that included all participants. There are notable examples of biased 'per-protocol' analyses (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 in the pre-CONSORT era before 1996 (Moher 2001). For example, Schulz 1996observed 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).

8.13.2 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 of incomplete outcome data across intervention groups, the reasons for outcomes being missing, the likely difference in outcome between participants with and without data, what the 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 a 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, when using the Cochrane tool (Section 8.5).

It is often assumed that a high proportion of missing outcomes, or a large difference in these proportions between intervention groups, is the main cause for concern over bias. However, these characteristics on their own are insufficient to introduce bias. Here we elaborate on situations in which an analysis can be judged to be at a low or high risk of bias. It is essential to consider the reasons for outcomes being missing as well as the numbers missing.

Risk of bias may be high for some outcomes (or time points) and low for others. For example, there may be fewer dropouts at one-month follow-up than at two-year follow-up. Thus, it is highly desirable for assessments of risk of bias resulting from incomplete outcome data to be made separately for different outcomes (or time points). 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 short-term outcomes that is different from a common assessment of blinding for all long-term outcomes. C57

8.13.2.1 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 in Section 8.13.1, 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 considered to be 'censored' on the date of their last follow-up (we do not know whether the outcome event occurred after follow-up ended, see Chapter 9, Section 9.2.6). The important consideration for this type of analysis is whether such censoring can be assumed to be unbiased, i.e. that the intervention 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 data on **dichotomous outcomes** depends on the frequency (or risk) of the outcome. For example, if 10% of participants have missing outcomes, then their potential impact on the results is much greater if the risk of the event is 10% than if it is 50%. Table 8.13.a illustrates the potential impact of observed risks. A and B represent two hypothetical trials of 1000 participants in which 90% 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%) and those in the control group have a much lower risk (20%). The only difference between trials A and B is the risk among the observed participants. In trial A the risk is 50%, and the impact of the missing data, had they been observed, would be low. In trial B the risk is 10%, and the impact of the same missing data, had they been observed, would be 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 Study B it was 100/90 (1.1).

The potential impact of missing data on **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. Table 8.13.b 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% of participants are observed and 10% 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.

Table 8.13.a: Potential impact of missing data: dichotomous outcomes

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

Table 8.13.b: Potential impact of missing data: continuous outcomes

	Number randomized	Number observed	Observed mean	Number missing	Hypothetical extreme mean among missing participants	Overall mean (weighted average)	Mean difference based on all participants
Study A							

Intervention	500	**450**	10	**50**	15	10.5	**1**
Control	500	**450**	10	**50**	5	9.5	

Study B

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

8.13.2.2 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, if participants with poorer clinical outcomes are more likely to drop out due to adverse effects, and this happens mainly in the experimental group, then 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 nine 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 (i.e. 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 (Schulz 2010).

'As-treated' (per-protocol) analyses

Eligible participants should be analysed 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 being at a 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 rate in the 1103 men assigned to clofibrate was 20.0%, and was 20.9% in the 2789 men assigned to placebo (P = 0.55). Those who adhered well to the protocol in the clofibrate group had lower five-year mortality rate (15.0%) than those who did not (24.6%). However, a similar difference between 'good adherers' and 'poor adherers' was observed in the placebo group (15.1% versus 28.3%). Thus, adherence was a marker of prognosis rather than modifying the effect of clofibrate. These findings show the serious difficulty of evaluating intervention efficacy in subgroups determined by patient responses to the interventions. As 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.

8.13.2.3 Attempts to address missing data in reports: imputation

A common, but potentially dangerous, approach to dealing with missing outcome data is to **impute** outcomes and treat them as if they were real measurements (see also Chapter 16, Section 16.2). 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, the validity of which 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. Since 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: see Chapter 16 (Section 16.1). 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.

8.14 Selective outcome reporting

8.14.1 Rationale for concern about bias

Selective outcome reporting has been defined as the selection of a subset of the original variables recorded, on the basis of the results, for inclusion in publication of trials (Hutton 2000); see also Chapter 10 (Section 10.2.2.5). The particular concern about selective outcome reporting is that statistically 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 in the *CDSR* (Williamson 2005a).

Convincing direct empirical evidence for the existence of within-study selective reporting bias comes from several studies that compared protocols to publications (Dwan 2013). In one early 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, that is, 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, both for efficacy (pooled odds ratio 2.4; 95% CI 1.4 to 4.0) and for harms data (pooled odds ratio 4.7; 95% CI 1.8 to 12). Furthermore, when comparing publications with protocols, 62% of trials had at least one primary outcome that was changed, introduced or omitted. A subsequent study of 48 trials funded by the Canadian Institutes of Health Research found very 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, 95% CI 1.6 to 2.7 for efficacy outcomes; odds ratio 1.9, 95% CI 1.1 to 3.5 for harm outcomes). These three studies suggest an odds ratio of about 2.4 associated with selective outcome reporting that corresponds, for example, to about 50% of non-significant outcomes being published compared to 72% of significant ones.

In all three of these 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. Furthermore, authors commonly failed to mention the existence of unpublished outcomes even when those outcomes had been mentioned in the protocol or publication.

Other studies have found similar results (Ghersi 2006, von Elm 2006). In a different type of study, the effect in meta-analyses was larger 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. Kirkham and colleagues showed that outcome reporting bias affects the conclusions in a substantial proportion of Cochrane Reviews (Kirkham 2010): the median amount of review outcome data missing for any reason was 10%, whereas 50% or more of the potential data were missing in 70 (25%) reviews. A survey of trialists showed that in almost all trials in which prespecified outcomes had been analysed but not reported, this under-reporting resulted in bias (Smyth 2011). Other researchers have highlighted the value of clinical trials registries to identify selective reporting of outcomes (Mathieu 2009).

Bias associated with selective reporting of different measures of the same characteristic seems likely. In trials of treatments for schizophrenia, an intervention 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 statistically significant or that, following analysis, unfavourable items may have been dropped to create an apparent beneficial 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 is 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 data for an outcome were 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 to 5 below):

1. **Selective omission of outcomes from reports:** Only some of the analysed 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 analysed. For example, continuous outcomes such as blood pressure reduction might be analysed 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 comparison of final values to a comparison of changes from baseline because of an observed baseline imbalance actually introduces bias rather than removes 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 subscales of a full measurement scale or a subset of events. For example, fungal infections may be identified at baseline or within a couple of days after randomization or may be referred to as 'break-through' fungal infections that are detected some days after randomization, and selection of a subset of these infections may lead to reporting bias (Jørgensen 2007, Jørgensen 2014).

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'.

Other forms of selective reporting are not addressed here. These 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 2004a, Chan 2004b). This issue alone should not generally be of concern to review authors (who do not take note of which outcomes are labelled as such in each study), provided it does not influence which results are published.

8.14.2 Assessing risk of bias from selective reporting of outcomes

Although the possibility of *between-study* publication bias can be examined only by considering a complete set of studies (see Chapter 10), 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 Cochrane 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 for which results are reported. If non-significant results are mentioned but not reported adequately, bias is likely to occur in a meta-analysis. 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 reported in Chan 2004a and Chan 2004b did so.

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 for 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, for example with rows as studies and columns as outcomes. Complete and incomplete reporting can also be indicated. This matrix will allow review authors to see which studies did not report outcomes reported by most other studies.

PubMed, other major reference databases and the internet should be searched for a study protocol; in rare cases the web address may be given in the study report. Alternatively, and more often in the future as mandatory registration 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.

Missing information that seems sure to have been recorded is of particular interest. For example, some measurements are expected to appear together, such as systolic and diastolic blood pressure, so if only one is reported we should wonder why. 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 standard deviation of the changes. Williamson 2005a gives several examples, including a Cochrane Review in which nine trials reported the outcome of treatment failure but only five reported mortality. Yet since mortality was part of the definition of treatment failure, those data must have been collected in the four trials that did not contribute to 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.

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

It is not generally recommended to try to 'adjust for' reporting bias in the main meta-analysis. Sensitivity analysis is a better 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 'support for judgement' 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 judgement provides an assessment of the overall susceptibility of the study to selective reporting bias.

8.15 Other potential threats to validity

8.15.1 Rationale for concern about bias

The preceding domains (sequence generation, allocation concealment, blinding, incomplete outcome data and selective outcome reporting) relate to important potential sources of bias in clinical studies across all healthcare areas. Beyond these specific domains, however, review authors should be alert for further issues that may raise concerns about the possibility of bias. This seventh domain in the 'Risk of bias' assessment tool is a 'catch-all' for other such sources of bias. For reviews in some topic areas, there may be additional questions that should be asked of all studies. In particular, some study designs warrant special consideration when they are encountered. If particular study designs are anticipated (e.g. cross-over trials, or types of non-randomized study), additional questions relating to the risk of bias in these types of studies may be posed. Assessing risk of bias in non-randomized studies is addressed in Chapter 13, and risk of bias for cluster-randomized trials, cross-over trials and trials with multiple intervention groups is addressed in Chapter 16. Furthermore, some major, unanticipated, problems with specific studies may be identified during the course of the systematic review or meta-analysis. For example, a trial may have substantial imbalance of participant characteristics at baseline. Several examples are discussed in the sections that follow.

8.15.1.1 Design-specific risks of bias

The principal concern over risk of bias in non-randomized studies is selection bias in the form of differences in types of participants between experimental and control intervention groups. Review authors should refer to the full discussion in Chapter 13 (Section 13.5). The main concerns over risk of bias in cluster-randomized trials are: 1) recruitment bias (differential participant recruitment in clusters for different interventions); 2) baseline imbalance; 3) loss of clusters; 4) incorrect analysis; and 5) comparability with individually randomized trials. The main concerns over risk of bias in cross-over trials are: 1) whether the cross-over design is suitable; 2) whether there is a carry-over effect; 3) whether only first-period data are available; 4) incorrect analysis; and 5) comparability of results with those from parallel-group trials. These are discussed in detail in Chapter 16 (Sections 16.3

and 16.4). Risk of bias in studies with more than two intervention groups is also discussed in Chapter 16 (Section 16.5).

8.15.1.2 Baseline imbalance

Baseline imbalance in factors that are strongly related to outcome measures can cause bias in the intervention 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 entries 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 intervention allocation.

Example (of high risk of bias): *A trial of captopril versus a 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).*

8.15.1.3 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, for example because of characteristic side effects. When blocked randomization is used, and when the assignments are revealed after a person has been recruited 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.d (Berger 2005).

8.15.1.4 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 if the experimental intervention is a nurse visiting a patient at home, and the control intervention is no visit.

8.15.1.5 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).
- Prior to randomization, there is administration of an intervention that could enhance or diminish the effect of a subsequent, randomized, intervention.
- There is inappropriate administration of an intervention (or cointervention).
- There is contamination (e.g. participants pooling drugs).
- There is occurrence of 'null bias' due to interventions being insufficiently well delivered or overly wide eligibility 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).
- There is selective reporting of subgroups.
- Fraud is identified or suspected.

8.15.1.6 Other issues

In this section we comment on some further issues that have been raised in relation to risk of bias, but for which we are unable to provide definitive guidance at present.

Influence of funders

Inappropriate influence of funders (or, more generally, of people with a vested interest in the results) is often regarded as an important risk of bias. For example, in one empirical study, more than half of the protocols for industry-initiated trials stated that the sponsor either owned the data or needed to approve the manuscript, or both; none of these constraints were stated in any of the trial publications (Gøtzsche 2006). It is important that information about vested interests is collected and presented when relevant. However, review authors should provide this information in the 'Characteristics of included studies' table (see Section 11.2.2). The 'Risk of bias' table should be used to assess specific aspects of methodology that might be been influenced by vested interests and which may lead directly to a risk of bias. Note that some decisions that may be influenced by those with a vested interest, such as choice of a particularly low dose of a comparator drug, should be addressed as a source of heterogeneity rather than through the 'Risk of bias' tool, since they do not impact directly on the internal validity of the findings.

Early stopping

There is a debate related to the risk of bias of trials that stop early because of benefit. A systematic review and a meta-epidemiologic study showed that truncated randomized trials were associated with greater effect sizes than trials not stopped early, particularly for trials with small sample size (Montori 2005, Bassler 2010). These results were widely discussed (Goodman 2010), and recommendations relating to this item will be provided in future. Currently, review authors should record systematically whether the trial was stopped early for benefit and report this information in the 'Characteristics of included studies' table.

Single-centre versus multi-centre studies

Recent meta-epidemiologic studies of binary and continuous outcomes showed that intervention effect estimates in single-centre randomized trials were significantly larger than in multi-centre trials even after controlling for sample size (Dechartres 2011, Bafeta 2012). The BRANDO project, which combined data from all available meta-epidemiologic studies (Savovic 2012b), found consistent results for subjective outcomes (relative odds ratio 0.86; 95% CI 0.68 to 1.05). Several reasons may explain these results: small study effect, reporting bias, higher risk of bias in single centre studies, or factors related to the selection of the participants, intervention administration, care providers' expertise, etc. Further studies are needed to explore the role and effect of these different mechanisms. However, information related to the number of centres should be systematically collected and reported in the 'Characteristics of included studies' table.

8.15.2 Assessing risk of bias from other sources

Some general guidelines for determining suitable topics for assessment as 'other sources of bias' are provided here. In particular, suitable topics should constitute potential sources of bias and not sources of imprecision, sources of diversity (heterogeneity) or measures of research quality that are unrelated to bias. The topics covered in this domain of the tool include primarily the examples provided in Section 8.15.1. 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 judgements 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 Cochrane tool (Section 8.5).

Wherever possible, a review protocol should prespecify any questions to be addressed that would lead to separate entries 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.

As the tool addresses internal biases only, 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 2014). Alternatively, participants may be chosen selectively 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 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 (see Chapter 11).

Many judgements 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). Such factors may be important, and should be presented in the table of 'Characteristics of included studies' or in 'Additional tables' (see Chapter 11).

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 domains in the tool. For example, in Alzheimer's disease, patients deteriorate significantly over time during the trial. Generally, the effects of interventions are small but have appreciable toxicity. Dealing satisfactorily with participant losses is very difficult. Those on the experimental intervention are likely to drop out earlier due to adverse effects or death, 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 in the 'Incomplete outcome data' section.

8.16 Methodological standards for the conduct of new Cochrane Intervention Reviews

No.	Status	Name	Standard	Rationale & elaboration	Handbook sections

C52	Mandatory	Assessing risk of bias	Assess the risk of bias for each included study. For randomized trials, the Cochrane 'Risk of bias' tool should be used, involving judgements and supports for those judgements across a series of domains of bias, as described in Chapter 8 of the *Handbook* (version 5 or later).	The risk of bias of every included study in a Cochrane Review must be explicitly considered to determine the extent to which its findings can be believed, noting that risks of bias might vary by outcome. Recommendations for assessing bias in randomized studies included in Cochrane Reviews are now well-established. The new tool – as described in the *Handbook* – must be used for all randomized trials in new reviews and all newly included randomized trials in updated reviews. This does not prevent other tools being used. The discussions in Chapters 8 and 13 of the *Handbook* should be used to inform the selection of an appropriate tool for non-randomized studies.	8.2.1 8.5.1 8.9 8.10 8.11 8.12 8.13 8.14 8.15
C53	Mandatory	Assessing risk of bias in duplicate	Use (at least) two people working independently to apply the 'Risk of bias' tool to each included study, and define in advance the process for resolving disagreements.	Duplicating the 'Risk of bias' assessment reduces both the risk of making mistakes and the possibility that assessments are influenced by a single person's biases.	8.3.4
C54	Mandatory	Supporting judgemen	Justify judgements of risk of bias (high, low and	Providing support for the judgement makes the process transparent. Items	8.5.2

		ts of risk of bias	unclear) and provide this information in the ‘Risk of bias’ tables (as ‘Support for judgement’).	which are judged to be at an unclear risk of bias but without accompanying information supporting the judgment appear as empty cells in the graphical plots based on the ‘Risk of bias' tool in the published review.	
C55	Highly desirable	Providing sources of information for ‘Risk of bias’ assessments	Collect the source of information for each ‘Risk of bias’ judgement (e.g. quotation, summary of information from a trial report, correspondence with investigator etc.). Where judgements are based on assumptions made on the basis of information provided outside publicly available documents, this should be stated.	Readers, editors and referees should have the opportunity to see for themselves where supports for judgments have been obtained.	8.5.2
C56	Highly desirable	Assessing risk of bias due to lack of blinding for different outcomes	Consider blinding separately for different key outcomes.	The risk of bias due to lack of blinding may be different for different outcomes (e.g. for unblinded outcome assessment, risk of bias for all-cause mortality may be very different from that for a patient-reported pain scale). When there are multiple outcomes, they should be grouped (e.g. objective versus subjective).	8.5.1 8.11.2 8.12.2

C57	Highly desirable	Assessing completeness of data for different outcomes	Consider the impact of missing data separately for different key outcomes to which an included study contributes data.	Often, considering risk of bias due to incomplete (missing) outcome data, this often cannot reliably be done for the study as a whole. The risk of bias due to missing outcome data may be different for different outcomes. For example, there may be less drop-out for a three-month outcome than for a six-year outcome. When there are multiple outcomes, they should be grouped (e.g. short term versus long term). Judgements should be attempted about which outcomes are thought to be at high or low risk of bias.	8.5.1 8.13.2
C58	Highly desirable	Summarizing risk of bias assessments	Summarize the risk of bias for each key outcome for each study.	This reinforces the link between the characteristics of the study design and their possible impact on the results of the study, and is an important pre-requisite for the GRADE approach to assessing the quality of the body of evidence.	8.7
C59	Highly desirable	Addressing risk of bias in the synthesis	Address risk of bias in the synthesis (whether quantitative or non-quantitative). For example, present analyses stratified according to summary risk of bias, or restricted to studies at low risk of bias.	Review authors should consider how study biases affect conclusions. This is useful in determining the strength of conclusions and how future research should be designed and conducted.	8.8.1

C60	Mandatory	Incorporating assessments of risk of bias	*If randomized trials have been assessed using one or more tools in addition to the Cochrane 'Risk of bias' tool,* use the Cochrane tool as the primary assessment of bias for interpreting results, choosing the primary analysis, and drawing conclusions.	For consistency of approach across Cochrane Reviews, the Cochrane 'Risk of bias' tool should take precedence when two or more tools are used. The Cochrane tool also feeds directly into the GRADE approach for assessing the quality of the body of evidence.	8.8.1

8.17 Chapter information

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

Contributing authors: Douglas Altman, Gerd Antes, Isabelle Boutron, Peter Gøtzsche, Julian Higgins, Peter Jüni, Steff Lewis, David Moher, Andrew Oxman, Ken Schulz, Jonathan Sterne and Simon Thompson.

Acknowledgements: Thanks to Hilda Bastian, Rachelle Buchbinder, Iain Chalmers, 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. For details of the Cochrane Statistical Methods Group, see Chapter 9 (Box 9.8.a), and for the Cochrane Bias Methods Group, see Chapter 10 (Box 10.5.a).

8.18 References

Altman 1999

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

Bafeta 2012

Bafeta A, Dechartres A, Trinquart L, Yavchitz A, Boutron I, Ravaud P. Impact of single centre status on estimates of intervention effects in trials with continuous outcomes: meta-epidemiological study. *BMJ* 2012; 344: e813.

Balk 2002

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

Bassler 2010

Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA* 2010; 303: 1180-1187.

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 2003

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

Berger 2005

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

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 2005

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

Boutron 2006

Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, Hróbjartsson A, et al. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. *PLoS Medicine* 2006; 3: 1931-1939.

Brightling 2000

Brightling CE, Monteiro W, Ward R, Parker D, Morgan MD, Wardlaw AJ, et al. 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. *Statistics in Medicine* 2005; 24: 3715-3727.

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. *Canadian Medical Association Journal* 2004; 171: 735-740.

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.

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 England Journal of Medicine* 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. *Obesity Research* 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.

Dechartres 2011

Dechartres A, Boutron I, Trinquart L, Charles P, Ravaud P. Single-center trials show larger treatment effects than multicenter trials: evidence from a meta-epidemiologic study. *Annals of Internal Medicine* 2011; 155: 39-51.

Detsky 1992

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology* 1992; 45: 255-265.

Devereaux 2001

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

Dwan 2013

Dwan K, Gamble C, Williamson PR, Kirkham JJ, Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLOS ONE* 2013; 8: e66844.

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. *Controlled Clinical Trials* 1990; 11: 339-352.

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.

Fergusson 2004

Fergusson D, Glass KC, Waring D, Shapiro S. Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. *BMJ* 2004; 328: 432.

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*; 2006 Oct 23-26; Dublin, Ireland.

Goodman 2010

Goodman S, Berry D, Wittes J. Bias and trials stopped early for benefit. *JAMA* 2010; 304: 157; author reply 158-159.

Gøtzsche 1996

Gøtzsche PC. Blinding during data analysis and writing of manuscripts. *Controlled Clinical 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.

Guyatt 2008

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336: 924-926.

Haahr 2006

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

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. *Journal of Evaluation in Clinical Practice* 2002; 8: 353-359.

Hansson 1999

Hansson L, Lindholm LH, Niskanen L, Lanke J, Hedner T, Niklason A, et al. 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. *Controlled Clinical 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.

Hróbjartsson 2007

Hróbjartsson A, Forfang E, Haahr MT, ls-Nielsen B, Brorson S. Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. *International Journal of Epidemiology* 2007; 36: 654-663.

Hróbjartsson 2012

Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ* 2012; 344: e1119.

Hutton 2000

Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2000; 49: 359-370.

Jadad 1996

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials* 1996; 17: 1-12.

Johansen 2014

Johansen HK, Gøtzsche PC. Amphotericin B lipid soluble formulations versus amphotericin B in cancer patients with neutropenia. *Cochrane Database of Systematic Reviews* Issue 9. CD000969. DOI: 10.1002/14651858.CD000969.pub2.

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.

Jørgensen 2014

Jørgensen KJ, Gøtzsche PC, Dalboge CS, Johansen HK. Voriconazole versus amphotericin B or fluconazole in cancer patients with neutropenia. *Cochrane Database of Systematic Reviews* Issue 2. CD004707. DOI: 10.1002/14651858.CD004707.pub3.

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.

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.

Kirkham 2010

Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010; 340: c365.

Kjaergard 2001

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

Lachin 2000

Lachin JM. Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials* 2000; 21: 167-189.

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: 249-252.

Mathieu 2009

Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA* 2009; 302: 977-984.

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.

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. *Controlled Clinical 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. *International Journal of Technology Assessment in Health Care* 1996; 12: 195-208.

Moher 1998

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. 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. *Journal of Clinical Epidemiology* 2002; 55: 787-790.

Montori 2005

Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. 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.

Newell 1992

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

Oxman 1993

Oxman AD, Guyatt GH. The science of reviewing research. *Annals of the New York Academy of Sciences* 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 randomized trials. *International Journal of Epidemiology* 2007; 36: 847-857.

Porta 2007

Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. *Journal of Clinical Epidemiology* 2007; 60: 663-669.

Rees 2005

Rees JR, Wade TJ, Levy DA, Colford JM, Jr., Hilton JF. Changes in beliefs identify unblinding in randomized controlled trials: a method to meet CONSORT guidelines. *Contemporary Clinical Trials* 2005; 26: 25-37.

Sackett 2007

Sackett DL. Commentary: Measuring the success of blinding in RCTs: don't, must, can't or needn't? *International Journal of Epidemiology* 2007; 36: 664-665.

Safer 2002

Safer DJ. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. *Journal of Nervous and Mental Disease* 2002; 190: 583-592.

Savovic 2012a

Savovic J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Annals of Internal Medicine* 2012; 157: 429-438.

Savovic 2012b

Savovic J, Jones H, Altman D, Harris R, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technology Assessment* 2012; 16: 1-82.

Schulz 1995a

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 1995b

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

Schulz 1996

Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ* 1996; 312: 742-744.

Schulz 2002a

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

Schulz 2002b

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

Schulz 2002c

Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. *Annals of Internal Medicine* 2002; 136: 254-259.

Schulz 2002d

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

Schulz 2006

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

Schulz 2010

Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340: c332.

Senn 1991

Senn S. Baseline comparisons in randomized clinical trials. *Statistics in Medicine* 1991; 10: 1157-1159.

Siersma 2007

Siersma V, ls-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. *Statistics in Medicine* 2007; 26: 2745-2758.

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.

Smyth 2011

Smyth RM, Kirkham JJ, Jacoby A, Altman DG, Gamble C, Williamson PR. Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. *BMJ* 2011s; 342: c7153.

Spiegelhalter 2003

Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* 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. *Statistics in Medicine* 2002; 21: 1513-1524.

Tierney 2005

Tierney JF, Stewart LA. Investigating patient exclusion bias in meta-analysis. *International Journal of Epidemiology* 2005; 34: 79-87.

Turner 2009

Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008; 172: 21-47.

Unnebrink 2001

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

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 published articles. *14th Cochrane Colloquium*; 2006 Oct 23-26; Dublin, Ireland.

Williamson 2005a

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

Williamson 2005b

Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research* 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. *Clinical Trials* 2004; 1: 368-376.

Wood 2008

Wood L, Egger M, Gluud LL, Schulz K, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008; 336: 601-605.

Woods 1995

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

Chapter 9: Analysing data and undertaking meta-analyses

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

This chapter should be cited as: Deeks JJ, Higgins JPT, Altman DG (editors) on behalf of the Cochrane Statistical Methods Group. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS (editors), *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0 (updated June 2017), Cochrane, 2017. Available from www.training.cochrane.org/handbook.

This extract is from *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0. The previous version of this chapter (5.1.0, 2011) is available online at handbook.cochrane.org.

An earlier version (version 5.0.2, 2008) of the *Handbook* is also published by John Wiley & Sons, Ltd under "The Cochrane Book Series" Imprint, as Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on www.wiley.com.

Key Points

- Meta-analysis is the statistical combination of results from two or more separate studies.
- Potential advantages of meta-analyses include an increase in power, an improvement in precision, the ability to answer questions not posed by individual studies, and the opportunity to settle controversies arising from conflicting claims. However, they also have the potential to mislead seriously, particularly if specific study designs, within-study biases, variation across studies, and reporting biases are not carefully considered.
- It is important to be familiar with the type of data (e.g. dichotomous, continuous) that result from measurement of an outcome in an individual study, and to choose suitable effect measures for comparing intervention groups.

- Most meta-analysis methods are variations on a weighted average of the effect estimates from the different studies.
- Variation across studies (heterogeneity) must be considered, although most Cochrane Reviews do not have enough studies to allow the reliable investigation of the reasons for it. Random-effects meta-analyses allow for heterogeneity by assuming that underlying effects follow a normal distribution.
- Many judgements are required in the process of preparing a Cochrane Review or meta-analysis. Sensitivity analyses should be used to examine whether overall findings are robust to potentially influential decisions.

9.1 Introduction

9.1.1 Do not start here!

It can be tempting to jump prematurely into a statistical analysis when undertaking a systematic review. The production of a diamond at the bottom of a plot is an exciting moment for many authors, but results of meta-analyses can be very misleading if suitable attention has not been given to formulating the review question; specifying eligibility criteria; identifying, selecting and critically appraising studies; collecting appropriate data; and deciding what would be meaningful to analyse. Review authors should consult the chapters that precede this one before a meta-analysis is undertaken.

9.1.2 Planning the analysis

While in primary studies the investigators select and collect data from individual patients, in systematic reviews the investigators select and collect data from primary studies. While primary studies include analyses of their participants, Cochrane Reviews contain analyses of the primary studies. Analyses may be narrative, such as a structured summary and discussion of the studies' characteristics and findings, or quantitative, that is involving statistical analysis. **Meta-analysis** – the statistical combination of results from two or more separate studies – is the most commonly used statistical technique. The Cochrane Review writing software (RevMan) can perform a variety of meta-analyses, but it must be stressed that meta-analysis is not appropriate in all Cochrane Reviews. Issues to consider when deciding whether a meta-analysis is appropriate in a review are discussed in this section and in Section 9.1.4.

Studies comparing healthcare interventions, notably randomized trials, use the outcomes of participants to compare the effects of different interventions. Meta-analyses focus on pair-wise comparisons of interventions, such as an experimental intervention versus a control intervention, or the comparison of two experimental interventions. The terminology used here (experimental versus control interventions) implies the former, although the methods apply equally to the latter.

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

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

Meta-analysis provides a statistical method for questions 1 to 3. Assessment of question 4 relies additionally on judgements based on assessments of study design and risk of bias, as well as statistical measures of uncertainty.

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

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

1. The most straightforward Cochrane Review assembles studies that make one particular comparison between two intervention options, for example, comparing kava extract versus placebo for treating anxiety (Pittler 2003). Meta-analysis and related techniques can be used if there is a consistent outcome measure to:

   i. establish whether there is evidence of an effect;

   ii. estimate the size of the effect and the uncertainty surrounding that size; and

   iii. investigate whether the effect is consistent across studies.

2. Some reviews may have a broader focus than a single comparison. The first is where the intention is to identify and collate studies of numerous interventions for the same disease or condition. An example of such a review is that of topical treatments for fungal infections of the skin and nails of the foot, which included studies of any topical intervention (Crawford 2007). The second, related aim is that of identifying a ‘best’ intervention. A review of interventions for emergency contraception sought that which was most effective (while also considering potential adverse effects). Such reviews may include multiple comparisons and meta-analyses between all possible pairs of interventions, and require care when it comes to planning analyses (see Section 9.1.6 and Chapter 16, Section 16.6).

3. Occasionally review comparisons have particularly wide scopes that make the use of meta-analysis problematic. For example, a review of workplace interventions for smoking cessation covered diverse types of interventions (Moher 2005). When reviews contain very diverse studies a meta-analysis might be useful to answer the overall question of whether there is evidence that, for example, work-based interventions can work (but see Section 9.1.4), but use of meta-analysis to describe the size of effect may

not be meaningful if the implementations are so diverse that an effect estimate cannot be interpreted in any specific context.

4. An aim of some reviews is to investigate the relationship between the size of an effect and some characteristic(s) of the studies. This is uncommon as a primary aim in Cochrane Reviews, but may be a secondary aim. For example, in a review of beclomethasone versus placebo for chronic asthma, there was interest in whether the administered dose of beclomethasone affected its efficacy (Adams 2005). Such investigations of heterogeneity need to be undertaken with care (see Section 9.6).

9.1.2.1 Checking data before synthesis

Before embarking on a synthesis, it is important to be confident that the findings from the individual studies have been collated correctly. Therefore, review authors must compare the magnitude and direction of effects reported by studies with how they are to be presented in the review. This is a reasonably straightforward way for authors to check a number of potential problems, including typographical errors in studies' reports, accuracy of data collection and manipulation, and data entry into RevMan. For example, the direction of a standardized mean difference may accidentally be wrong in the review. A basic check is to ensure the same qualitative findings (e.g. direction of effect and statistical significance) between the data as presented in the review and the data as available from the original study. C51

Results in forest plots should agree with data in the original report (point estimate and confidence interval) if the same effect measure and statistical model is used. There are legitimate reasons for differences, however, including: using a different measure of intervention effect; making different choices between change-from-baseline measures, postintervention measures alone or postintervention measures adjusted for baseline values; grouping similar intervention groups; or making adjustments for unit-of-analysis errors in the reports of the primary studies.

9.1.3 Why perform a meta-analysis in a review?

The value a meta-analysis can add to a review depends on the context in which it is used, as described in Section 9.1.2. The following are reasons for considering including a meta-analysis in a review.

1. To increase power. Power is the chance of detecting a real effect as statistically significant if it exists. Many individual studies are too small to detect small effects, but when several are combined there is a higher chance of detecting an effect.

2. To improve precision. The estimation of an intervention effect can be improved when it is based on more information.

3. To answer questions not posed by the individual studies. Primary studies often involve a specific type of patient and explicitly defined interventions. A selection of studies in which these characteristics differ can allow investigation of the consistency of effect and, if relevant, allow reasons for differences in effect estimates to be investigated.

4. To settle controversies arising from apparently conflicting studies or to generate new hypotheses. Statistical analysis of findings allows the degree of conflict to be assessed formally, and reasons for different results to be explored and quantified.

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

9.1.4 When not to use meta-analysis in a review

C62

If used appropriately, meta-analysis is a powerful tool for deriving meaningful conclusions from data and can help prevent errors in interpretation. However, it must be used only if participants, interventions, comparisons and outcomes are judged to be sufficiently similar to ensure an answer that is clinically meaningful. There are situations in which a meta-analysis can be more of a hindrance than a help.

1. A common criticism of meta-analyses is that they 'combine apples with oranges'. If studies are clinically diverse then a meta-analysis may be meaningless, and genuine differences in effects may be obscured. A particularly important type of diversity is in the comparisons being made by the primary studies. Often it is nonsensical to combine all included studies in a single meta-analysis: sometimes there is a mix of comparisons of different interventions with different comparators, each combination of which may need to be considered separately. Furthermore, it is important not to combine outcomes that are too diverse. Decisions concerning what should and should not be combined are inevitably subjective, and are not amenable to statistical solutions but require discussion and clinical judgement. In some cases consensus may be hard to reach.

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

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

9.1.5 What does a meta-analysis entail?

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

1. Which comparisons should be made?

2. Which study results should be used in each comparison?

3. What is the best summary of effect for each comparison?

4. Are the results of studies similar within each comparison?

5. How reliable are those summaries?

The first step in addressing these questions is to decide which comparisons to make (see Section 9.1.6) and what sorts of data are appropriate for the outcomes of interest (see Section 9.2). The next step is to prepare tabular summaries of the characteristics and results of the studies that are included in each comparison (extraction of data and conversion to the desired format is discussed in Chapter 7, Section 7.7). It is then possible to derive estimates of effect across studies in a systematic way (Section 9.4), to measure and investigate differences among studies (Sections 9.5 and 9.6) and to interpret the findings and conclude how much confidence should be placed in them (see Chapter 11 and Chapter 12).

9.1.6 Which comparisons should be made?

The first and most important step in planning the analysis is to specify the pair-wise comparisons that will be made. The comparisons addressed in the review should relate clearly and directly to the questions or hypotheses that are posed when the review is formulated (see Chapter 5). It should be possible to specify in the protocol of a review the main comparisons that will be made. However, it will often be necessary to modify comparisons and add new ones in light of the data that are collected. For example, important variations in the intervention may only be discovered after data are collected.

Decisions about which studies are similar enough for their results to be grouped together require an understanding of the problem that the review addresses, and judgement by the review author and, subsequently, the user. The formulation of the questions that a review addresses is discussed in Chapter 5. Essentially the same considerations apply to deciding which comparisons to make, which outcomes to combine and which key characteristics (of study design, participants, interventions and outcomes) to consider when investigating variation in effects (heterogeneity). These considerations must be addressed when setting up the ‘Data and analyses’ tables in RevMan and in deciding what information to put in the ‘Characteristics of included studies’ table.

9.1.7 Writing the analysis section of the protocol

The analysis section of a Cochrane Review protocol may be more susceptible to change than other protocol sections (such as criteria for including studies and how methodological quality will be assessed). It is rarely possible to anticipate all the statistical issues that may arise, for example, finding outcomes that are similar but not the same as each other; outcomes measured at multiple or varying time points; and use of concomitant interventions.

However the protocol should provide a strong indication of how the review author will approach the statistical evaluation of studies’ findings. At least one member of the review team should be familiar with the majority of the contents of this chapter when the protocol is written. As a guideline we recommend that the following be addressed.

1. Ensure that the analysis strategy firmly addresses the stated objectives of the review (see Section 9.1.2).

2. Consider which types of study design would be appropriate for the review. Parallel group trials are the norm, but other randomized designs may be appropriate to the topic (e.g. cross-over trials, cluster-randomized trials, factorial trials). Decide how such studies will be addressed in the analysis (see Section 9.3).

3. Decide whether a meta-analysis is intended and consider how the decision about whether a meta-analysis is appropriate will be made (see Sections 9.1.3 and 9.1.4).

4. Determine the probable nature of outcome data (e.g. dichotomous, continuous, etc.; see Section 9.2).

5. Consider whether it is possible to specify in advance what intervention effect measures will be used (e.g. risk ratio, odds ratio or risk difference for dichotomous outcomes, mean difference or standardized mean difference for continuous outcomes; see Sections 9.4.4.4 and 9.4.5.1).

6. Decide how statistical heterogeneity will be identified or quantified (see Section 9.5.2).

7. Decide whether random-effects meta-analyses, fixed-effect meta-analyses or both methods will be used for each planned meta-analysis (see Section 9.5.4).

8. Consider how clinical and methodological diversity (heterogeneity) will be assessed and whether (and how) these will be incorporated into the analysis strategy (see Sections 9.5 and 9.6).

9. Decide how the risk of bias in included studies will be assessed and addressed in the analysis (see Chapter 8).

10. Prespecify characteristics of the studies that may be examined as potential causes of heterogeneity (see Section 9.6.5).

11. Consider how missing data will be handled (e.g. imputing data for intention-to-treat analyses; see Chapter 16, Sections 16.1 and 16.2).

12. Decide whether (and how) evidence of possible publication and/or reporting biases will be sought (see Chapter 10).

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

9.2 Types of data and effect measures

9.2.1 Types of data

The starting point of all meta-analyses of studies of effectiveness involves the identification of the data type for the outcome measurements. Throughout this chapter we consider outcome data to be of five different types:

1. dichotomous (or binary) data, where each individual’s outcome is one of only two possible categorical responses;

2. continuous data, where each individual's outcome is a measurement of a numerical quantity;

3. ordinal data (including measurement scales), where the outcome is one of several ordered categories, or generated by scoring and summing categorical responses;

4. counts and rates calculated from counting the number of events that each individual experiences; and

5. time-to-event (typically survival) data that analyse the time until an event occurs, but where not all individuals in the study experience the event (censored data).

The ways in which the effect of an intervention can be measured depend on the nature of the data being collected. In this section we briefly examine the types of outcome data that might be encountered in systematic reviews of clinical trials, and review definitions, properties and interpretation of standard measures of intervention effect. In Sections 9.4.4.4 and 9.4.5.1 we discuss issues in the selection of one of these measures for a particular meta-analysis.

9.2.2 Effect measures for dichotomous outcomes

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

1. the risk ratio (RR; also called the relative risk);

2. the odds ratio (OR);

3. the risk difference (RD; also called the absolute risk reduction); and

4. the number needed to treat for an additional beneficial or harmful outcome (NNT).

Details of the calculations of the first three of these measures are given in Box 9.2.a. Numbers needed to treat are discussed in detail in Chapter 12 (Section 12.4).

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

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

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

	Event ('Success')	No event ('Fail')	Total
Experimental intervention	S_E	F_E	N_E
Control intervention	S_C	F_C	N_C

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

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

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

$$\begin{aligned}\mathrm{RD} &= \text{risk of event in experimental group} - \text{risk of event in control group}\\ &= \frac{S_E}{N_E} - \frac{S_C}{N_C}\end{aligned}$$

9.2.2.1 Risk and odds

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

Risk is the concept more familiar to patients and health professionals. Risk describes the probability with which a health outcome (usually an adverse event) will occur. In research, risk is commonly expressed as a decimal number between 0 and 1, although it is occasionally converted into a percentage. In 'Summary of findings' tables in Cochrane Reviews, it is often expressed as a number of individuals per 1000 (see Chapter 11, Section 11.5). It is simple to grasp the relationship between a risk and the likely occurrence of events: in a sample of 100 people the number of events observed will on average be the risk multiplied by 100. For example, when the risk is 0.1, about 10 people out of every 100 will have the event; when the risk is 0.5, about 50 people out of every 100 will have the event. In a sample of 1000 people, these numbers are 100 and 500 respectively.

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

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

The interpretation of odds is more complicated than for a risk. The simplest way to ensure that the interpretation is correct is first to convert the odds into a risk. For example, when the odds are 1:10, or 0.1, one person will have the event for every 10 who do not, and, using the formula, the risk of the event is 0.1/(1+0.1) = 0.091. In a sample of 100, about 9 individuals will have the event and 91 will not. When the odds are equal to 1, one person will have the event for every person who does not, so in a sample of 100, 100 × 1/(1+1) = 50 will have the event and 50 will not.

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

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

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

Measures of relative effect express the outcome in one group relative to that in the other. The **risk ratio** (or relative risk) is the ratio of the risk of an event in the two groups, whereas the **odds ratio** is the ratio of the odds of an event (see Box 9.2.a). For both measures a value of 1 indicates that the estimated effects are the same for both interventions.

Neither the risk ratio nor the odds ratio can be calculated for a study if there are no events in the control group. This is because, as can be seen from the formulae in Box 9.2.a, we would be trying to divide by zero. The odds ratio also cannot be calculated if everybody in the intervention group experiences an event. In these situations, and others where standard errors cannot be computed, it is customary to add ½ to each cell of the 2 × 2 table (RevMan automatically makes this correction when necessary). In the case where no events (or all events) are observed in both groups the study provides no information about relative probability of the event and is automatically omitted from the meta-analysis. This is entirely appropriate. Zeros arise particularly when the event of interest is rare – such events are often unintended adverse outcomes. For further discussion of choice of effect measures for such sparse data (often with lots of zeros) see Chapter 16 (Section 16.9).

Risk ratios describe the multiplication of the risk that occurs with use of the experimental intervention. For example, a risk ratio of 3 for an intervention implies that events with intervention are three times more likely than events without intervention. Alternatively we can say that intervention increases the risk of events by 100 × (RR – 1)% = 200%. Similarly a risk ratio of 0.25 is interpreted as the probability of an event with intervention being one-quarter of that without intervention. This may be expressed alternatively by saying that intervention decreases the risk of events by 100 × (1 –RR)% = 75%. This is known as the relative risk reduction (see also Chapter 12, Section 12.4.1). The interpretation of the clinical importance of a given risk ratio cannot be made without knowledge of the typical risk of events without intervention: a risk ratio of 0.75 could correspond to a clinically important reduction in events from 80% to 60%, or a small, less clinically important reduction from 4% to 3%.

The numerical value of the observed risk ratio must always be between 0 and 1/CGR, where CGR (abbreviation of ‘control group risk’, sometimes referred to as the control event rate) is the observed risk of the event in the control group (expressed as a number between 0 and 1). This means that for common events large values of risk ratio are impossible. For example, when the observed risk of events in the control group is 0.66 (or 66%) then the observed risk ratio cannot exceed 1.5. This problem applies only for increases in risk, and causes problems only when the results are extrapolated to risks above those observed in the study.

Odds ratios, like odds, are more difficult to interpret (Sinclair 1994, Sackett 1996). Odds ratios describe the multiplication of the odds of the outcome that occur with use of the intervention. To understand what an odds ratio means in terms of changes in numbers of events it is simplest to first convert it into a risk ratio, and then interpret the risk ratio in the context of a typical control group risk, as outlined above. The formula for converting an odds ratio to a risk ratio is provided in Chapter 12 (Section 12.4.4.4). Sometimes it may be sensible to calculate the RR for more than one assumed control group risk.

9.2.2.3 Warning: OR and RR are not the same

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

9.2.2.4 Measure of absolute effect: the risk difference

The **risk difference** is the difference between the observed risks (proportions of individuals with the outcome of interest) in the two groups (see Box 9.2.a). The risk difference can be calculated for any study, even when there are no events in either group. The risk difference is straightforward to interpret: it describes the actual difference in the observed risk of events between experimental and control interventions; for an individual it describes the estimated difference in the probability of experiencing the event. However, the clinical importance of a risk difference may depend on the underlying risk of events. For example, a risk difference of 0.02 (or 2%) may represent a small, clinically insignificant change from a risk of 58% to 60% or a proportionally much larger and potentially important change from 1% to 3%. Although the risk difference provides more directly relevant information than relative measures (Laupacis 1988, Sackett 1997), it is still important to be aware of the underlying risk of events, and consequences of the events, when interpreting a risk difference. Absolute measures, such as the risk difference, are particularly useful when considering trade-offs between likely benefits and likely harms of an intervention.

The risk difference is naturally constrained (like the risk ratio), which may create difficulties when applying results to other patient groups and settings. For example, if a study or meta-analysis estimates a risk difference of –0.1 (or –10%), then for a group with an initial risk of, say, 7% the outcome will have an impossible estimated negative probability of –3%. Similar scenarios for increases in risk occur at the other end of the scale. Such problems can arise only when the results are applied to patients with different risks from those observed in the studies.

The number needed to treat is obtained from the risk difference. Although it is often used to summarize results of clinical trials, NNTs cannot be combined in a meta-analysis (see Section 9.4.4.4). However, odds ratios, risk ratios and risk differences may be usefully converted to NNTs and used when interpreting the results of a meta-analysis as discussed in Chapter 12 (Section 12.4).

9.2.2.5 What is the event?

In the context of dichotomous outcomes, healthcare interventions are intended either to reduce the risk of occurrence of an adverse outcome or increase the chance of a good outcome. All of the effect measures described in Section 9.2.2 apply equally to both scenarios.

In many situations it is natural to talk about one of the outcome states as being an event. For example, when participants have particular symptoms at the start of the study the event of interest is usually recovery or cure. If participants are well or, alternatively, at risk of some adverse outcome at the beginning of the study, then the event is the onset of disease or occurrence of the adverse outcome. Since the focus is usually on the experimental intervention group, a study in which the experimental intervention reduces the occurrence of an adverse outcome will have an odds ratio and risk ratio less than 1, and a negative risk difference. A study in which the experimental intervention increases the occurrence of a good outcome will have an odds ratio and risk ratio greater than 1, and a positive risk difference (see Box 9.2.a).

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

By contrast, switching the outcome can make a substantial difference for risk ratios, affecting the effect estimate, its significance, and the consistency of intervention effects across studies. This is because the precision of a risk ratio estimate differs markedly between those situations where risks are low and those where risks are high. In a meta-analysis the effect of this reversal cannot be predicted easily. The identification, before data analysis, of which risk ratio is more likely to be the most relevant summary statistic is therefore important and discussed further in Section 9.4.4.4.

9.2.3 Effect measures for continuous outcomes

The term ‘continuous’ in statistics conventionally refers to data that can take any value in a specified range. When dealing with numerical data, this means that any number may be measured and reported to an arbitrary number of decimal places. Examples of truly continuous data are weight, area and volume. In practice, in Cochrane Reviews we can use the same statistical methods for other types of data, most commonly measurement scales and counts of large numbers of events (see Section 9.2.4).

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

9.2.3.1 The mean difference (or difference in means)

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

Aside: Analyses based on this effect measure have historically been termed weighted mean difference (WMD) analyses in the *Cochrane Database of Systematic Reviews* (*CDSR*). This name is potentially confusing: although the meta-analysis computes a weighted average of these differences in means, no weighting is involved in calculation of a statistical summary of a single study. Furthermore, all meta-analyses involve a weighted combination of estimates, yet we do not use the word ‘weighted’ when referring to other methods.

9.2.3.2 The standardized mean difference

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

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

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

However, the method assumes that the differences in standard deviations among studies reflect differences in measurement scales and not real differences in variability among study populations. This assumption may be problematic in some circumstances where real differences in variability between the participants in different studies are expected. For example, where pragmatic and explanatory trials are combined in the same review, pragmatic trials may include a wider range of participants and may consequently have higher standard deviations. The overall intervention effect can also be difficult to interpret as it is reported in units of standard deviation rather than in units of any of the measurement scales used in the review, but in some circumstances it is possible to transform the effect back to the units used in a specific study (see Chapter 12, Section 12.5).

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

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

C61

9.2.4 Effect measures for ordinal outcomes and measurement scales

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

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

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

Specialist methods are available for analysing ordinal outcome data that describe effects in terms of **proportional odds ratios**, but they are not available in RevMan, and become unwieldy (and unnecessary) when the number of categories is large. In practice, longer ordinal scales are often analysed in meta-analyses as continuous data, whilst shorter ordinal scales are often made into dichotomous data by combining adjacent categories together. The latter is especially appropriate if an established, defensible cut-point is available. Inappropriate choice of a cut-point can induce bias, particularly if it is chosen to maximize the difference between two intervention arms in a clinical trial.

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

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

9.2.5 Effect measures for counts and rates

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

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

Counts of more common events, such as counts of decayed, missing or filled teeth, may often be treated in the same way as continuous outcome data. The intervention effect used will be the mean difference which will compare the difference in the mean number of events (possibly standardized to a unit time period) experienced by participants in the intervention group compared with participants in the control group.

9.2.5.1 Warning: counting events or counting participants?

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

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

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

Time-to-event data can sometimes be analysed as dichotomous data. This requires the status of all patients in a study to be known at a fixed time point. For example, if all patients have been followed for at least 12 months, and the proportion who have incurred the event before 12 months is known for both groups, then a 2 × 2 table can be constructed (see Box 9.2.a) and intervention effects expressed as risk ratios, odds ratios or risk differences.

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

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

9.2.7 Expressing intervention effects on log scales

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

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

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

9.3 Study designs and identifying the unit of analysis

9.3.1 Unit-of-analysis issues

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

1. groups of individuals were randomized together to the same intervention (i.e. cluster-randomized trials);

2. individuals undergo more than one intervention (e.g. in a cross-over trial, or simultaneous treatment of multiple sites on each individual); and

3. there are multiple observations for the same outcome (e.g. repeated measurements, recurring events, measurements on different body parts).

C70 Review authors must consider the impact on the analysis of any such clustering, matching or other non-standard design features of the included studies. A more detailed list of situations in which unit-of-analysis issues commonly arise follows, together with directions to relevant discussions elsewhere in this *Handbook*.

9.3.2 Cluster-randomized trials

In a cluster-randomized trial, groups of participants are randomized to different interventions. For example, the groups may be schools, villages, medical practices, patients of a single doctor or families. See Chapter 16 (Section 16.3).

9.3.3 Cross-over trials

In a cross-over trial, all participants receive all interventions in sequence: they are randomized to an ordering of interventions, and participants act as their own control. See Chapter 16 (Section 16.4).

9.3.4 Repeated observations on participants

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

1. Obtain individual patient data and perform an analysis (such as time-to-event analysis) that uses the whole follow-up for each participant. Alternatively, compute an effect measure for each individual participant that incorporates all time points, such as total number of events, an overall mean, or a trend over time. Occasionally, such analyses are available in published reports.

2. Define several different outcomes, based on different periods of follow-up, and perform separate analyses. For example, time frames might be defined to reflect short-term, medium-term and long-term follow-up.

3. Select a single time point and analyse only data at this time for studies in which it is presented. Ideally this should be a clinically important time point. Sometimes it might be chosen to maximize the data available, although authors should be aware of the possibility of reporting biases.

4. Select the longest follow-up from each study. This may induce a lack of consistency across studies, giving rise to heterogeneity.

9.3.5 Events that may re-occur

If the outcome of interest is an event that can occur more than once, then care must be taken to avoid a unit-of-analysis error. Count data should not be treated as if they are dichotomous data. See Section 9.2.5.

9.3.6 Multiple treatment attempts

Similarly, multiple treatment attempts per participant can cause a unit-of-analysis error. Care must be taken to ensure that the number of participants randomized, and not the number of treatment attempts, is used to calculate confidence intervals. For example, in subfertility studies, women may undergo multiple cycles, and authors might erroneously use cycles as the denominator rather than women. This is similar to the situation in cluster-randomized trials, except that each participant is the 'cluster'. See methods described in Chapter 16 (Section 16.3).

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

In some studies, people are randomized, but multiple parts (or sites) of the body receive the same intervention, a separate outcome judgement being made for each body part, and the number of body parts is used as the denominator in the analysis. For example, eyes may be mistakenly used as the denominator without adjustment for the non independence between eyes. This is similar to the situation in cluster-randomized studies, except that participants are the 'clusters'. See methods described in Chapter 16 (Section 16.3).

9.3.8 Multiple body parts II: body parts receive different interventions

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

9.3.9 Multiple intervention groups

Studies that compare more than two intervention groups need to be treated with care. Such studies are often included in meta-analysis by making multiple pair-wise comparisons between all possible pairs of intervention groups. A serious unit-of-analysis problem arises if the same group of participants is included twice in the same meta-analysis (for example, if 'Dose 1 vs Placebo' and 'Dose 2 vs Placebo' are both included in the same meta-analysis, with the same placebo patients in both comparisons). Review authors must analyse multiple intervention groups in an appropriate way that avoids arbitrary omission of relevant groups and double-counting of participants. See Chapter 16 (Section 16.5). C66

9.4 Summarizing effects across studies

9.4.1 Meta-analysis

An important step in a systematic review is the thoughtful consideration of whether it is appropriate to combine the numerical results of all, or perhaps some, of the studies. Such a **meta-analysis** yields an overall statistic (together with its confidence interval) that summarizes the effectiveness of the experimental intervention compared with a control intervention (see Section 9.1.2). This section describes the principles and methods used to carry out a meta-analysis for the main types of data encountered.

Formulae for all the methods described are provided in a supplementary document Statistical algorithms in Review Manager 5 (available at cochrane.org/handbook), and a longer discussion of the issues discussed in this section appear in Deeks 2001.

9.4.2 Principles of meta-analysis

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

1. Meta-analysis is typically a two-stage process. In the first stage, a summary statistic is calculated for each study, to describe the observed intervention effect. For example, the summary statistic may be a risk ratio if the data are dichotomous, or a difference between means if the data are continuous.

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

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

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

3. The combination of intervention effect estimates across studies may optionally incorporate an assumption that the studies are not all estimating the same intervention effect, but estimate intervention effects that follow a distribution across studies. This is the basis of a **random-effects meta-analysis** (see Section 9.5.4). Alternatively, if it is assumed that each study is estimating exactly the same quantity a **fixed-effect meta-analysis** is performed.

4. The standard error of the summary (pooled) intervention effect can be used to derive a confidence interval, which communicates the precision (or uncertainty) of the

summary estimate, and to derive a P value, which communicates the strength of the evidence against the null hypothesis of no intervention effect.

5. As well as yielding a summary quantification of the pooled effect, all methods of meta-analysis can incorporate an assessment of whether the variation among the results of the separate studies is compatible with random variation, or whether it is large enough to indicate inconsistency of intervention effects across studies (see Section 9.5).

6. The problem of missing data is one of the numerous practical considerations that must be thought through when undertaking a meta-analysis. In particular, Review authors should consider the implications of missing outcome data from individual participants (due to losses to follow-up or exclusions from analysis). Missing data is addressed in more detail in Chapter 16, Section 16.1. C64

9.4.3 A generic inverse-variance approach to meta-analysis

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

The inverse variance method is so named because the weight given to each study is chosen to be the inverse of the variance of the effect estimate (i.e. 1 over the square of its standard error). Thus larger studies, which have smaller standard errors, are given more weight than smaller studies, which have larger standard errors. This choice of weight minimizes the imprecision (uncertainty) of the pooled effect estimate.

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

$$\text{generic inverse-variance weighted average} = \frac{\sum Y_i \left(1/\text{SE}_i^2\right)}{\sum \left(1/\text{SE}_i^2\right)}$$

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

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

A variation on the inverse-variance method is to incorporate an assumption that the different studies are estimating different, yet related, intervention effects. This produces a random-effects meta-analysis, and the simplest version is known as the DerSimonian and Laird method (DerSimonian 1986). Random-effects meta-analysis is discussed in Section 9.5.4. To undertake a random-effects meta-analysis, the standard errors of the study-specific estimates (SE_i in Section 9.4.3) are adjusted to incorporate a measure of the extent of variation, or heterogeneity, among the intervention effects observed in different studies (this variation is often referred to as tau-squared, τ^2, or Tau^2). The amount of variation, and hence the adjustment, can be estimated from the intervention effects and standard errors of the studies included in the meta-analysis.

9.4.3.2 The generic inverse variance outcome type in RevMan

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

Note that the ability to enter estimates and standard errors directly into RevMan creates a high degree of flexibility in meta-analysis. For example, it facilitates the analysis of properly analysed cross-over trials, cluster-randomized trials and non randomized trials, as well as outcome data that are ordinal, time-to-event or rates. However, in most situations for analyses of continuous and dichotomous outcome data it is preferable to enter more detailed data into RevMan (i.e. specifically as simple summaries of dichotomous or continuous data for each group). This avoids the need for the author to calculate effect estimates, and allows the use of methods targeted specifically at different types of data (see Sections 9.4.4 and 9.4.5). Also, it is helpful for the readers of the review to see the summary statistics for each intervention group in each study.

9.4.4 Meta-analysis of dichotomous outcomes

There are four widely used methods of meta-analysis for dichotomous outcomes, three fixed-effect methods (Mantel-Haenszel, Peto and inverse variance) and one random-effects method (DerSimonian and Laird). All of these methods are available as analysis options in RevMan. The Peto method can only pool odds ratios, whilst the other three methods can pool odds ratios, risk ratios and risk differences. Formulae for all of the meta-analysis methods are given in Deeks 2001.

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

9.4.4.1 Mantel-Haenszel methods

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

9.4.4.2 Peto odds ratio method

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

The approximation used in the computation of the log odds ratio works well when intervention effects are small (odds ratios are close to 1), events are not particularly common and the studies have similar numbers in experimental and control groups. In other situations it has been shown to give biased answers. As these criteria are not always fulfilled, Peto's method is not recommended as a default approach for meta-analysis.

Corrections for zero cell counts are not necessary when using Peto's method. Perhaps for this reason, this method performs well when events are very rare (Bradburn 2007; see Chapter 16, Section 16.9). Also, Peto's method can be used to combine studies with dichotomous outcome data with studies using time-to-event analyses where log-rank tests have been used (see Section 9.4.9).

9.4.4.3 Random-effects method

The random-effects method incorporates an assumption that the different studies are estimating different, yet related, intervention effects (DerSimonian 1986). As described in Section 9.4.3.1, the method is based on the inverse-variance approach, making an adjustment to the study weights according to the extent of variation, or heterogeneity, among the varying intervention effects. The random-effects method and the fixed-effect method will give identical results when there is no heterogeneity among the studies. Where there is heterogeneity, confidence intervals for the average intervention effect will be wider if the random-effects method is used rather than a fixed-effect method, and corresponding claims of statistical significance will be more conservative. It is also possible that the central estimate of the intervention effect will change if there are relationships between observed intervention effects and sample sizes. See Section 9.5.4 for further discussion of these issues.

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

9.4.4.4 Which measure for dichotomous outcomes?

Summary statistics for dichotomous data are described in Section 9.2.2. The effect of an intervention can be expressed as either a relative or an absolute effect. The risk ratio (relative risk) and odds ratio are relative measures, while the risk difference and number needed to treat for an additional beneficial effect are absolute measures. A further complication is that there are, in fact, two risk ratios. We can calculate the risk ratio of an event occurring or the risk ratio of no event occurring. These give different pooled results in a meta-analysis, sometimes dramatically so.

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

Consistency: Empirical evidence suggests that relative effect measures are, on average, more consistent than absolute measures (Engels 2000, Deeks 2002, Rücker 2009). For this reason it is wise to avoid performing meta-analyses of risk differences, unless there is a clear reason to suspect that risk differences will be consistent in a particular clinical situation. On average there is little difference between the odds ratio and risk ratio in terms of consistency (Deeks 2002). When the study aims to reduce the incidence of an adverse outcome (see Section 9.2.2.5), there is empirical evidence that risk ratios of the adverse outcome are more consistent than risk ratios of the non-event (Deeks 2002). Selecting an effect measure on the basis of what is the most consistent in a *particular* situation is not a generally recommended strategy, since it may lead to a selection that spuriously maximizes the precision of a meta-analysis estimate.

Mathematical properties: The most important mathematical criterion is the availability of a reliable variance estimate. The number needed to treat for an additional beneficial outcome does not have a simple variance estimator and cannot easily be used directly in meta-analysis, although it can be computed from the other summary statistics (see Chapter 12, Section 12.4). There is no consensus regarding the importance of two other often-cited mathematical properties: the fact that the behaviour of the odds ratio and the risk difference do not rely on which of the two outcome states is coded as the event, and the odds ratio being the only statistic which is unbounded (see Section 9.2.2).

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

It seems important to avoid using summary statistics for which there is empirical evidence that they are unlikely to give consistent estimates of intervention effects (the risk difference), and it is impossible to use statistics for which meta-analysis cannot be performed (the number needed to treat for an additional beneficial outcome). Thus it is generally recommended that analysis proceeds using risk ratios (taking care to make a sensible choice over which category of outcome is classified as the event) or odds ratios. It may be wise to plan to undertake a sensitivity analysis to investigate whether choice of summary statistic (and selection of the event category) is critical to the conclusions of the meta-analysis (see Section 9.7).

It is often sensible to use one statistic for meta-analysis and to re-express the results using a second, more easily interpretable statistic. For example, often meta-analysis may be best performed using relative effect measures (risk ratios or odds ratios) and the results re-expressed using absolute effect measures (risk differences or numbers needed to treat for an additional beneficial outcome – see Chapter 12, Section 12.3). This is one of the key motivations for 'Summary of findings' tables in Cochrane Reviews: see Chapter 11 (Section 11.1). If odds ratios are used for meta-analysis they can also be re-expressed as risk ratios (see Chapter 12, Section 12.4.4.4). In all cases the same formulae can be used to convert upper and lower confidence limits. However, it is important to note that all of these transformations require specification of a value of baseline risk that indicates the likely risk of the outcome in the 'control' population to which the experimental intervention will be applied. Where the chosen value for this assumed control risk is close to the typical observed control group risks across the studies, similar estimates of absolute effect will be obtained regardless of whether odds ratios or risk ratios are used for meta-analysis. Where the assumed control risk differs from the typical observed control group risk, the predictions of absolute benefit will differ according to which summary statistic was used for meta-analysis.

9.4.5 Meta-analysis of continuous outcomes

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

Authors should be aware that one assumption underlying methods for meta-analysis of continuous data is that the outcomes have a normal distribution in each intervention arm in each study. This assumption may not always be met, although it is unimportant in very large studies. It is useful to consider the possibility of skewed data (see Section 9.4.5.3).

9.4.5.1 Which measure for continuous outcomes?

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

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

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

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

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

9.4.5.2 Meta-analysis of change scores

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

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

In practice an author is likely to discover that the studies included in a review may include a mixture of change-from-baseline and final value scores. However, mixing of outcomes is not a problem when it comes to meta-analysis of mean differences. There is no statistical reason why studies with change-from-baseline outcomes should not be combined in a meta-analysis with studies with final measurement outcomes when using the (unstandardized) mean difference method in RevMan. In a randomized study, mean differences based on changes from baseline can usually be assumed to be addressing exactly the same underlying intervention effects as analyses based on final measurements. That is to say, the difference in mean final values will on average be the same as the difference in mean change scores. If the use of change scores does increase precision, appropriately, the studies presenting change scores will be given higher weights in the analysis than they would have received if final values had been used, as they will have smaller standard deviations.

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

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

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

9.4.5.3 Meta-analysis of skewed data

C65 Analyses based on means are appropriate for data that are at least approximately normally distributed, and for data from very large trials. If the true distribution of outcomes is asymmetrical, then the data are said to be skewed. Review authors should consider the possibility and implications of skewed data when analysing continuous outcomes. Skew can sometimes be diagnosed from the means and standard deviations of the outcomes. A rough check is available, but it is only valid if a lowest or highest possible value for an outcome is known to exist. Thus the check may be used for outcomes such as weight, volume and blood concentrations, which have lowest possible values of 0, or for scale outcomes with minimum or maximum scores, but it may not be appropriate for change from baseline measures. The check involves calculating the observed mean minus the lowest possible value (or the highest possible value minus the observed mean), and dividing this by the standard deviation. A ratio less than 2 suggests skew (Altman 1996). If the ratio is less than 1, there is strong evidence of a skewed distribution.

Transformation of the original outcome data may reduce skew substantially. Reports of trials may present results on a transformed scale, usually a log scale. Collection of appropriate data summaries from the trialists, or acquisition of individual patient data, is currently the approach of choice. Appropriate data summaries and analysis strategies for the individual patient data will depend on the situation. Consultation with a knowledgeable statistician is advised.

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

9.4.6 Combining dichotomous and continuous outcomes

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

There are several options for handling combinations of dichotomous and continuous data. Generally, it is useful to summarize results from all the relevant, valid studies in a similar way, but this is not always possible. It may be possible to collect missing data from investigators so that this can be done. If not, it may be useful to summarize the data in three ways: by entering the means and standard deviations as continuous outcomes, by entering the counts as dichotomous outcomes and by entering all of the data in text form as 'Other data' outcomes.

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

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

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

9.4.7 Meta-analysis of ordinal outcomes and measurement scales

Ordinal and measurement scale outcomes are most commonly meta-analysed as dichotomous data (if so, see Section 9.4.4) or continuous data (if so, see Section 9.4.5) depending on the way that the study authors performed the original analyses.

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

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

Estimates of log odds ratios and their standard errors from a proportional odds model may be meta-analysed using the generic inverse-variance method in RevMan (see Section 9.4.3.2). Both fixed-effect and random-effects methods of analysis are available. If the same ordinal scale has been used in all studies, but in some reports has been presented as a dichotomous outcome, it may still be possible to include all studies in the meta-analysis. In the context of the three-category model, this might mean that for some studies category 1 constitutes a success, while for others both categories 1 and 2 constitute a success. Methods are available for dealing with this, and for combining data from scales that are related but have different definitions for their categories (Whitehead 1994).

9.4.8 Meta-analysis of counts and rates

Results may be expressed as **count data** when each participant may experience an event, and may experience it more than once (see Section 9.2.5). For example, ‘number of strokes’, or ‘number of hospital visits’ are counts. These events may not happen at all, but if they do happen there is no theoretical maximum number of occurrences for an individual.

As described in Chapter 7 (Section 7.7.5), count data may be analysed using methods for dichotomous (see Section 9.4.4), continuous (see Section 9.4.5) and time-to-event data (see Section 9.4.9), as well as being analysed as rate data.

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

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

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

The results of a study may be expressed as a **rate ratio**, that is the ratio of the rate in the experimental intervention group to the rate in the control group. Suppose E_E events occurred during T_E participant-years of follow-up in the experimental intervention group, and E_C events during T_C participant-years in the control intervention group. The rate ratio is:

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

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

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

A correction of 0.5 may be added to each count in the case of zero events. Note that the choice of time unit (i.e. patient-months, women-years, etc.) is irrelevant since it is cancelled out of the rate ratio and does not figure in the standard error. However, the units should still be displayed when presenting the study results. An alternative means of estimating the rate ratio is through the approach of (Whitehead 1991).

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

It is possible also to focus attention on the rate difference:

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

An approximate standard error for the rate difference is:

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

The analysis again requires use of the generic inverse-variance method in RevMan. One of the only discussions of meta-analysis of rates, which is still rather short, is that by Hasselblad and McCrory (Hasselblad 1995).

9.4.9 Meta-analysis of time-to-event outcomes

Two approaches to meta-analysis of time-to-event outcomes are available in RevMan. The choice of which to use will depend on the type of data that have been extracted from the primary studies, or obtained from reanalysis of individual patient data.

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

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

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

9.4.10 A summary of meta-analysis methods available in RevMan

Table 9.4.a lists the options for statistical analysis that are available in RevMan. RevMan requires the author to select one preferred method for each outcome. If these are not specified then the software defaults to the fixed-effect Mantel-Haenszel odds ratio for dichotomous outcomes, the fixed-effect mean difference for continuous outcomes and the fixed-effect model for generic inverse-variance outcomes. It is important that authors make it clear which method they are using when results are presented in the text of a review, since it cannot be guaranteed that the meta-analysis displayed to the user will coincide with the selected preferred method.

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

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

9.4.11 Use of vote counting for meta-analysis

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

9.5 Heterogeneity

9.5.1 What is heterogeneity?

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

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

Differences between studies in terms of methodological factors, such as use of blinding and concealment of allocation, or if there are differences between studies in the way the outcomes are defined and measured, may be expected to lead to differences in the observed intervention effects. Significant statistical heterogeneity arising from methodological diversity or differences in outcome assessments suggests that the studies are not all estimating the same quantity, but does not necessarily suggest that the true intervention effect varies. In particular, heterogeneity associated solely with methodological diversity would indicate that the studies suffer from different degrees of bias. Empirical evidence suggests that some aspects of design can affect the result of clinical trials, although this is not always the case. Further discussion appears in Chapter 8.

The scope of a review will largely determine the extent to which studies included in a review are diverse. Sometimes a review will include studies addressing a variety of questions, for example when several different interventions for the same condition are of interest (see also Chapter 5, Section 5.6). Studies of each intervention should be analysed and presented separately. Meta-analysis should only be considered when a group of studies is sufficiently homogeneous in terms of participants, interventions and outcomes to provide a meaningful summary. It is often appropriate to take a broader perspective in a meta-analysis than in a single clinical trial. A common analogy is that systematic reviews bring together apples and oranges, and that combining these can yield a meaningless result. This is true if apples and oranges are of intrinsic interest on their own, but may not be if they are used to contribute to a wider question about fruit. For example, a meta-analysis may reasonably evaluate the average effect of a class of drugs by combining results from trials where each evaluates the effect of a different drug from the class.

There may be specific interest in a review in investigating how clinical and methodological aspects of studies relate to their results. Where possible these investigations should be specified a priori, i.e. in the protocol for the systematic review. It is legitimate for a systematic review to focus on examining the relationship between some clinical characteristic(s) of the studies and the size of intervention effect, rather than on obtaining a summary effect estimate across a series of studies (see Section 9.6). Meta-regression may best be used for this purpose, although it is not implemented in RevMan (see Section 9.6.4).

9.5.2 Identifying and measuring heterogeneity

It is essential to consider to what extent the results of studies are consistent with each other. If confidence intervals for the results of individual studies (generally depicted graphically using horizontal lines) have poor overlap, this generally indicates the presence of statistical heterogeneity. More formally, a statistical test for heterogeneity is available. This chi-squared (χ^2, or Chi^2) test is included in the forest plots in Cochrane Reviews. It assesses whether observed differences in results are compatible with chance alone. A low P value (or a large Chi^2 statistic relative to its degree of freedom) provides evidence of heterogeneity of intervention effects (variation in effect estimates beyond chance). C63

Care must be taken in the interpretation of the Chi^2 test, since it has low power in the (common) situation of a meta-analysis when studies have small sample size or are few in number. This means that while a statistically significant result may indicate a problem with heterogeneity, a non-significant result must not be taken as evidence of no heterogeneity. This is also why a P value of 0.10, rather than the conventional level of 0.05, is sometimes used to determine statistical significance. A further problem with the test, which seldom occurs in Cochrane Reviews, is that when there are many studies in a meta-analysis, the test has high power to detect a small amount of heterogeneity that may be clinically unimportant.

Some argue that, since clinical and methodological diversity always occur in a meta-analysis, statistical heterogeneity is inevitable (Higgins 2003). Thus the test for heterogeneity is irrelevant to the choice of analysis; heterogeneity will always exist whether or not we happen to be able to detect it using a statistical test. Methods have been developed for quantifying inconsistency across studies that move the focus away from testing whether heterogeneity is present to assessing its impact on the meta-analysis. A useful statistic for quantifying inconsistency is:

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

In this equation, Q is the Chi^2 statistic and df is its degrees of freedom (Higgins 2002, Higgins 2003). This describes the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance).

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

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

*The importance of the observed value of I^2 depends on 1) magnitude and direction of effects, and 2) strength of evidence for heterogeneity (e.g. P value from the Chi^2 test, or a confidence interval for I^2: uncertainty in the value of I^2 is substantial when the number of studies is small).

9.5.3 Strategies for addressing heterogeneity

Review authors must take into account any statistical heterogeneity when interpreting results, particularly when there is variation in the direction of effect. A number of options are available if heterogeneity is identified among a group of studies that would otherwise be considered suitable for a meta-analysis. C69

1. Check again that the data are correct

Severe heterogeneity can indicate that data have been incorrectly extracted or entered into RevMan. For example, if standard errors have mistakenly been entered as standard deviations for continuous outcomes, this could manifest itself in overly narrow confidence intervals with poor overlap and hence substantial heterogeneity. Unit-of-analysis errors may also be causes of heterogeneity (see Section 9.3).

2. Do not do a meta-analysis

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

3. Explore heterogeneity

It is clearly of interest to determine the causes of heterogeneity among results of studies. This process is problematic since there are often many characteristics that vary across studies from which one may choose. Heterogeneity may be explored by conducting subgroup analyses (see Section 9.6.3) or meta-regression (see Section 9.6.4), though this latter method is not implemented in RevMan. Ideally, investigations of characteristics of studies that may be associated with heterogeneity should be prespecified in the protocol of a review (see Section 9.1.7). Reliable conclusions can only be drawn from analyses that are truly prespecified before inspecting the studies' results, and even these conclusions should be interpreted with caution. In practice, authors will often be familiar with some study results when writing the protocol, so true prespecification is not possible. Explorations of heterogeneity that are devised after heterogeneity is identified can at best lead to the generation of hypotheses. They should be interpreted with even more caution and should generally not be listed among the conclusions of a review. Also, investigations of heterogeneity when there are very few studies are of questionable value.

4. Ignore heterogeneity

Fixed-effect meta-analyses ignore heterogeneity. The pooled effect estimate from a fixed-effect meta-analysis is normally interpreted as being the best estimate of the intervention effect. However, the existence of heterogeneity suggests that there may not be a single intervention effect but a distribution of intervention effects. Thus the pooled fixed-effect estimate may be an intervention effect that does not actually exist in any population, and therefore have a confidence interval that is meaningless as well as being too narrow, (see Section 9.5.4). The P value obtained from a fixed-effect meta-analysis does however provide a meaningful test of the null hypothesis that there is no effect in any of the studies.

5. Perform a random-effects meta-analysis

A random-effects meta-analysis may be used to incorporate heterogeneity among studies. This is not a substitute for a thorough investigation of heterogeneity. It is intended primarily for heterogeneity that cannot be explained. An extended discussion of this option appears in Section 9.5.4.

6. Change the effect measure

Heterogeneity may be an artificial consequence of an inappropriate choice of effect measure. For example, when studies collect continuous outcome data using different scales or different units, extreme heterogeneity may be apparent when using the mean difference but not when the more appropriate standardized mean difference is used. Furthermore, choice of effect measure for dichotomous outcomes (odds ratio, risk ratio, or risk difference) may affect the degree of heterogeneity among results. In particular, when control group risks vary, homogeneous odds ratios or risk ratios will necessarily lead to heterogeneous risk differences, and vice versa. However, it remains unclear whether homogeneity of intervention effect in a particular meta-analysis is a suitable criterion for choosing between these measures (see also Section 9.4.4.4).

7. Exclude studies

Heterogeneity may be due to the presence of one or two outlying studies with results that conflict with the rest of the studies. In general it is unwise to exclude studies from a meta-analysis on the basis of their results as this may introduce bias. However, if an obvious reason for the outlying result is apparent, the study might be removed with more confidence. Since usually at least one characteristic can be found for any study in any meta-analysis which makes it different from the others, this criterion is unreliable because it is all too easy to fulfil. It is advisable to perform analyses both with and without outlying studies as part of a sensitivity analysis (see Section 9.7). Whenever possible, potential sources of clinical diversity that might lead to such situations should be specified in the protocol.

9.5.4 Incorporating heterogeneity into random-effects models

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

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

Note that a random-effects model does not ‘take account’ of the heterogeneity, in the sense that it is no longer an issue. It is always advisable to explore possible causes of heterogeneity, although there may be too few studies to do this adequately (see Section 9.6).

For random-effects analyses in RevMan, the pooled estimate and confidence interval refer to the centre of the distribution of intervention effects, but do not describe the width of the distribution. Often the pooled estimate and its confidence interval are quoted in isolation as an alternative estimate of the quantity evaluated in a fixed-effect meta-analysis, which is inappropriate. The confidence interval from a random-effects meta-analysis describes uncertainty in the location of the mean of systematically different effects in the different studies. It does not describe the degree of heterogeneity among studies, as may be commonly believed. For example, when there are many studies in a meta-analysis, one may obtain a tight confidence interval around the random-effects estimate of the mean effect even when there is a large amount of heterogeneity.

In common with other meta-analysis software, RevMan presents an estimate of the between-study variance in a random-effects meta-analysis (known as tau-squared, τ^2 or Tau^2). The square root of this number (i.e. tau) is the estimated standard deviation of underlying effects across studies. For absolute measures of effect (e.g. risk difference, mean difference, standardized mean difference), an approximate 95% range of underlying effects can be obtained by creating an interval from 2 × tau below the random-effects pooled estimate, to 2 × tau above it. For relative measures (e.g. odds ratio, risk ratio), the interval needs to be centred on the natural logarithm of the pooled estimate, and the limits anti-logged (exponentiated) to obtain an interval on the ratio scale. Alternative intervals, for the predicted effect in a new study, have been proposed (Higgins 2009). The range of the intervention effects observed in the studies may be thought to give a rough idea of the spread of the distribution of true intervention effects, but in fact it will be slightly too wide as it also describes the random error in the observed effect estimates.

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

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

For any particular set of studies in which heterogeneity is present, a confidence interval around the random-effects pooled estimate is wider than a confidence interval around a fixed-effect pooled estimate. This will happen if the I^2 statistic is greater than zero, even if the heterogeneity is not detected by the Chi^2 test for heterogeneity (Higgins 2003; see Section 9.5.2). The choice between a fixed-effect and a random-effects meta-analysis should never be made on the basis of a statistical test for heterogeneity.

In a heterogeneous set of studies, a random-effects meta-analysis will award relatively more weight to smaller studies than such studies would receive in a fixed-effect meta-analysis. This is because small studies are more informative for learning about the distribution of effects across studies than for learning about an assumed common intervention effect. Care must be taken that random-effects analyses are applied only when the idea of a 'random' distribution of intervention effects can be justified. In particular, if results of smaller studies are systematically different from results of larger ones, which can happen as a result of publication bias or within-study bias in smaller studies (Egger 1997, Poole 1999, Kjaergard 2001), then a random-effects meta-analysis will exacerbate the effects of the bias (see also Chapter 10, Section 10.4.4.1). A fixed-effect analysis will be affected less, although strictly it will also be inappropriate. In this situation it may be wise to present neither type of meta-analysis, or to perform a sensitivity analysis in which small studies are excluded.

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

RevMan implements a version of random-effects meta-analysis that is described by DerSimonian and Laird (DerSimonian 1986). The attraction of this method is that the calculations are straightforward, but it has a theoretical disadvantage in that the confidence intervals are slightly too narrow to encompass full uncertainty resulting from having estimated the degree of heterogeneity. Alternative methods exist that encompass full uncertainty, but they require more advanced statistical software (see also Chapter 16, Section 16.8). In practice, the difference in the results is likely to be small unless there are few studies. For dichotomous data, RevMan implements two versions of the DerSimonian and Laird random-effects model (see Section 9.4.4.3).

9.6 Investigating heterogeneity

9.6.1 Interaction and effect modification

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

9.6.2 What are subgroup analyses?

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

Subgroup analyses of subsets of participants within studies are uncommon in systematic reviews of the literature because sufficient details to extract data about separate participant types are seldom published in reports. By contrast, such subsets of participants are easily analysed when individual patient data have been collected (see Chapter 18). The methods we describe in Section 9.6.3 are for subgroups of trials.

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

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

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

9.6.3 Undertaking subgroup analyses

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

9.6.3.1 Is the effect different in different subgroups?

Valid investigations of whether an intervention works differently in different subgroups involve comparing the subgroups with each other. It is a mistake to compare within-subgroup inferences such as P values. If one subgroup analysis is statistically significant and another is not, then the latter may simply reflect a lack of information rather than a smaller (or absent) effect. When there are only two subgroups, non overlap of the confidence intervals indicates statistical significance, but note that the confidence intervals can overlap to a small degree and the difference still be statistically significant.

A formal statistical approach must be used to examine differences among subgroups. A simple significance test to investigate differences between two or more subgroups is described by Borenstein 2008. This method is implemented from RevMan version 5.1 C67 onwards for all types of meta-analysis. This procedure consists of undertaking a standard test for heterogeneity across subgroup results rather than across individual study results. When the meta-analysis uses a fixed-effect inverse-variance weighted average approach, the method is exactly equivalent to the test described by Deeks 2001. An I^2 statistic is also computed for subgroup differences. This describes the percentage of the variability in effect estimates from the different subgroups that is due to genuine subgroup differences rather than sampling error (chance). Note that these methods for examining subgroup differences should be used only when the data in the subgroups are independent (i.e. they should not be used if the same study participants contribute to more than one of the subgroups in the forest plot).

If fixed-effect models are used for the analysis within each subgroup, then these statistics relate to differences in typical effects across different subgroups. If random-effects models are used for the analysis within each subgroup, then the statistics relate to variation in the mean effects in the different subgroups. An alternative method for testing for differences between subgroups is to use meta-regression techniques, in which case a random-effects model is generally preferred (see Section 9.6.4). Tests for subgroup differences based on random-effects models may be regarded as preferable to those based on fixed-effect models, due to the high risk of false-positive results when a fixed-effect model is used to compare subgroups (Higgins 2004).

9.6.4 Meta-regression

If studies are divided into subgroups (see Section 9.6.2), this may be viewed as an investigation of how a categorical study characteristic is associated with the intervention effects in the meta-analysis. For example, studies in which allocation sequence concealment was adequate may yield different results from those in which it was inadequate. Here, allocation sequence concealment, being either adequate or inadequate, is a categorical characteristic at the study level. Meta-regression is an extension to subgroup analyses that allows the effect of continuous, as well as categorical, characteristics to be investigated, and in principle allows the effects of multiple factors to be investigated simultaneously (although this is rarely possible due to inadequate numbers of studies; Thompson 2002). Meta-regression should generally not be considered when there are fewer than ten studies in a meta-analysis.

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

The regression coefficient obtained from a meta-regression analysis will describe how the outcome variable (the intervention effect) changes with a unit increase in the explanatory variable (the potential effect modifier). The statistical significance of the regression coefficient is a test of whether there is a linear relationship between intervention effect and the explanatory variable. If the intervention effect is a ratio measure, the log-transformed value of the intervention effect should always be used in the regression model (see Section 9.2.7), and the exponential of the regression coefficient will give an estimate of the relative change in intervention effect with a unit increase in the explanatory variable.

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

Meta-regression may be performed using the ‘metareg’ macro available for the Stata statistical package, or using the ‘metafor’ package for R, as well as other packages.

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

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

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

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

9.6.5.2 Specify characteristics in advance

Authors should, whenever possible, prespecify characteristics in the protocol that later will be subject to subgroup analyses or meta-regression. The plan specified in the protocol must then be followed (data permitting), without undue emphasis on any particular findings. Prespecifying characteristics reduces the likelihood of spurious findings, firstly by limiting the number of subgroups investigated and secondly by preventing knowledge of the studies’ results influencing which subgroups are analysed. True prespecification is difficult in systematic reviews, because the results of some of the relevant studies are often known when the protocol is drafted. If a characteristic was overlooked in the protocol, but is clearly of major importance and justified by external evidence, then C68 authors should not be reluctant to explore it. However, such post hoc analyses should be identified as such.

9.6.5.3 Select a small number of characteristics

The likelihood of a false positive result among subgroup analyses and meta-regression increases with the number of characteristics investigated. It is difficult to suggest a maximum number of characteristics to look at, especially since the number of available studies is unknown in advance. If more than one or two characteristics are investigated it may be sensible to adjust the level of significance to account for making multiple comparisons. The help of a statistician is recommended (see Chapter 16, Section 16.7).

9.6.5.4 Ensure there is scientific rationale for investigating each characteristic

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

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

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

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

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

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

9.6.6 Interpretation of subgroup analyses and meta-regressions

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

1. Subgroup comparisons are observational

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

2. Was the analysis prespecified or post hoc?

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

3. Is there indirect evidence in support of the findings?

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

4. Is the magnitude of the difference practically important?

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

5. Is there a statistically significant difference between subgroups?

To establish whether there is a different effect of an intervention in different situations, the magnitudes of effects in different subgroups should be compared directly with each other. In particular, statistical significance of the results within separate subgroup analyses should not be compared (see Section 9.6.3.1).

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

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

9.6.7 Investigating the effect of underlying risk

One potentially important source of heterogeneity among a series of studies is when the underlying average risk of the outcome event varies between the studies. The underlying risk of a particular event may be viewed as an aggregate measure of case-mix factors such as age or disease severity. It is generally measured as the observed risk of the event in the control group of each study (the control group risk, or CGR). The notion is controversial in its relevance to clinical practice since underlying risk represents a summary of both known and unknown risk factors. Problems also arise because control group risk will depend on the length of follow-up, which often varies across studies. However, underlying risk has received particular attention in meta-analysis because the information is readily available once dichotomous data have been prepared for use in meta-analyses. Sharp provides a full discussion of the topic (Sharp 2001).

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

Use of different summary statistics (risk ratio, odds ratio and risk difference) will demonstrate different relationships with underlying risk. Summary statistics that show close to no relationship with underlying risk are generally preferred for use in meta-analysis (see Section 9.4.4.4).

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

9.6.8 Dose-response analyses

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

9.7 Sensitivity analyses

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

It is highly desirable to prove that the findings from a systematic review are not dependent on such arbitrary or unclear decisions by using sensitivity analysis. A sensitivity analysis is a repeat of the primary analysis or meta-analysis, in which alternative decisions or ranges of values are substituted for decisions that were arbitrary or unclear. For example, if the eligibility of some studies in the meta-analysis is dubious because they do not contain full details, sensitivity analysis may involve undertaking the meta-analysis twice: the first time including all studies and, secondly, including only those that are definitely known to be eligible. A sensitivity analysis asks the question, 'Are the findings robust to the decisions made in the process of obtaining them?' C71

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

Searching for studies:

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

Eligibility criteria:

1. Characteristics of participants: where a majority but not all people in a study meet an age range, should the study be included?

2. Characteristics of the intervention: what range of doses should be included in the meta-analysis?

3. Characteristics of the comparator: what criteria are required to define usual care to be used as a comparator group?

4. Characteristics of the outcome: what time point or range of time points are eligible for inclusion?

5. Study design: should blinded and unblinded outcome assessment be included, or should study inclusion be restricted by other aspects of methodological criteria?

What data should be analysed?

1. Time-to-event data: what assumptions of the distribution of censored data should be made?

2. Continuous data: where standard deviations are missing, when and how should they be imputed? Should analyses be based on change scores or on final values?

3. Ordinal scales: what cut-point should be used to dichotomize short ordinal scales into two groups?

4. Cluster-randomized trials: what values of the intraclass correlation coefficient should be used when trial analyses have not been adjusted for clustering?

5. Cross-over trials: what values of the within-subject correlation coefficient should be used when this is not available in primary reports?

6. All analyses: what assumptions should be made about missing outcomes to facilitate intention-to-treat analyses? Should adjusted or unadjusted estimates of intervention effects be used?

Analysis methods:

1. Should fixed-effect or random-effects methods be used for the analysis?

2. For dichotomous outcomes, should odds ratios, risk ratios or risk differences be used?

3. For continuous outcomes, where several scales have assessed the same dimension, should results be analysed as a standardized mean difference across all scales or as mean differences individually for each scale?

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

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

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

9.8 Methodological standards for the conduct of new Cochrane Intervention Reviews

No.	Status	Name	Standard	Rationale & elaboration	Handbook sections
C51	Mandatory	Checking accuracy of numeric data in the review	Compare magnitude and direction of effects reported by studies with how they are presented in the review, taking account of legitimate differences.	This is a reasonably straightforward way for authors to check a number of potential problems, including typographical errors in studies' reports, accuracy of data collection and manipulation, and data entry into RevMan. For example, the direction of a standardized mean difference may accidentally be wrong in the review. A basic check is to ensure the same qualitative findings (e.g. direction of effect and statistical significance) between the data as presented in the review and the data as available from the original study. Results in forest plots should agree with data in the original report (point estimate and confidence interval) if the same effect measure and statistical model are used.	9.1.2.1
C61	Mandatory	Combining different scales	*If studies are combined with different scales, ensure that higher scores for continuous outcomes all have the same meaning for any particular outcome*; explain the direction of	Sometimes scales have higher scores that reflect a 'better' outcome and sometimes lower scores reflect 'better' outcome. Meaningless (and misleading) results arise	9.2.3.2

			interpretation; and report when directions are reversed.	when effect estimates with opposite clinical meanings are combined.	
C62	Mandatory	Ensuring meta-analyses are meaningful	Undertake (or display) a meta-analysis only if participants, interventions, comparisons and outcomes are judged to be sufficiently similar to ensure an answer that is clinically meaningful.	Meta-analyses of very diverse studies can be misleading, for example of studies using different forms of control. Clinical diversity does not indicate necessarily that a meta-analysis should not be performed. However, authors must be clear about the underlying question that all studies are addressing.	9.1.4
C63	Mandatory	Assessing statistical heterogeneity	Assess the presence and extent of between-study variation when undertaking a meta-analysis.	The presence of heterogeneity affects the extent to which generalizable conclusions can be formed. It is important to identify heterogeneity in case there is sufficient information to explain it and offer new insights. Authors should recognize that there is much uncertainty in measures such as the I^2 and Tau^2 when there are few studies. Thus, use of simple thresholds to diagnose heterogeneity should be avoided.	9.5.2
C64	Highly desirable	Addressing missing outcome data	Consider the implications of missing outcome data from individual participants (due to losses to follow-up or exclusions from analysis).	Incomplete outcome data can introduce bias. In most circumstances, authors should follow the principles of intention-to-treat analyses as far as possible (this may not be appropriate for adverse effects or if trying to demonstrate equivalence). Risk of bias due to incomplete outcome data is addressed in	9.4.2

				the Cochrane 'Risk of bias' tool. However, statistical analyses and careful interpretation of results are additional ways in which the issue can be addressed by review authors. Imputation methods can be considered (accompanied by, or in the form of, sensitivity analyses).	
C65	Highly desirable	Addressing skewed data	Consider the possibility and implications of skewed data when analysing continuous outcomes.	Skewed data are sometimes not summarized usefully by means and standard deviations. While statistical methods are approximately valid for large sample sizes, skewed outcome data can lead to misleading results when studies are small.	9.4.5.3
C66	Mandatory	Addressing studies with more than two groups	*If multi-arm studies are included*, analyse multiple intervention groups in an appropriate way that avoids arbitrary omission of relevant groups and double-counting of participants.	Excluding relevant groups decreases precision and double-counting increases precision spuriously; both are inappropriate and unnecessary. Alternative strategies include combining intervention groups, separating comparisons into different forest plots and using multiple treatments meta-analysis.	9.3.9
C67	Mandatory	Comparing subgroups	*If subgroup analyses are to be compared, and there are judged to*	Concluding that there is a difference in effect in different subgroups on the basis of	9.6.3.1

			be sufficient studies to do this meaningfully, use a formal statistical test to compare them.	differences in the level of statistical significance within subgroups can be very misleading.	
C68	Mandatory	Interpreting subgroup analyses	*If subgroup analyses are conducted*, follow the subgroup analysis plan specified in the protocol without undue emphasis on particular findings.	Selective reporting, or over-interpretation, of particular subgroups or particular subgroup analyses should be avoided. This is a problem especially when multiple subgroup analyses are performed. This does not preclude the use of sensible and honest post hoc subgroup analyses.	9.6.5.2
C69	Mandatory	Considering statistical heterogeneity when interpreting the results	Take into account any statistical heterogeneity when interpreting the results, particularly when there is variation in the direction of effect.	The presence of heterogeneity affects the extent to which generalizable conclusions can be formed. If a fixed-effect analysis is used, the confidence intervals ignore the extent of heterogeneity. If a random-effects analysis is used, the result pertains to the mean effect across studies. In both cases, the implications of notable heterogeneity should be addressed. It may be possible to understand the reasons for the heterogeneity, if there are sufficient studies.	9.5.4
C70	Mandatory	Addressing non-standard designs	Consider the impact on the analysis of clustering, matching or other non-standard design features of the included studies.	Cluster-randomized studies, cross-over studies, studies involving measurements on multiple body parts, and other designs need to be addressed specifically, since a naive analysis might underestimate or	9.3.1

				overestimate the precision of the study. Failure to account for clustering is likely to overestimate the precision of the study, that is, to give it confidence intervals that are too narrow and a weight that is too large. Failure to account for correlation is likely to underestimate the precision of the study, that is, to give it confidence intervals that are too wide and a weight that is too small.	
C71	Highly desirable	Sensitivity analysis	Use sensitivity analyses to assess the robustness of results, such as the impact of notable assumptions, imputed data, borderline decisions and studies at high risk of bias.	It is important to be aware when results are robust, since the strength of the conclusion may be strengthened or weakened.	9.7

9.9 Chapter information

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

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

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

Box 9.9.a: The Cochrane Statistical Methods Group

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

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

Specifically, the aims of the group are:

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

Website: smg.cochrane.org

9.10 References

Adams 2005

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

Agresti 1996

Agresti A. *An Introduction to Categorical Data Analysis*. New York (NY): John Wiley & Sons, 1996.

Altman 1996

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

Antman 1992

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

Berlin 1993

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

Berlin 1994

Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trials data. *Online Journal of Current Clinical Trials* 1994; Doc No 134.

Berlin 2002

Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman KA, Anti-Lymphocyte Antibody Induction Therapy Study Group. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine* 2002; 21: 371-387.

Borenstein 2008

Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to Meta-analysis*. Chichester (UK): John Wiley & Sons, 2008.

Bradburn 2007

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

Chinn 2000

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

Cooper 1980

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

Crawford 2007

Crawford F, Hollis S. Topical treatments for fungal infections of the skin and nails of the feet. *Cochrane Database of Systematic Reviews* 2007, Issue 3. CD001434. DOI: 10.1002/14651858.CD001434.pub2.

Deeks 1998

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

Deeks 2001

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

Deeks 2002

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

DerSimonian 1986

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

Egger 1997

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

Engels 2000

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

Greenland 1985

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

Greenland 1987

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

Greenland 1992

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

Hasselblad 1995

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

Higgins 2002

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

Higgins 2003

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

Higgins 2004

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

Higgins 2008

Higgins JP, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. *Statistics in Medicine* 2008; 27: 6072-6092.

Higgins 2009

Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008; 172: 137-159.

Kjaergard 2001

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

Laupacis 1988

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

Mantel 1959

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

McIntosh 1996

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

Moher 2005

Moher M, Hey K, Lancaster T. Workplace interventions for smoking cessation. *Cochrane Database of Systematic Reviews* 2005, Issue 2. CD003440. DOI: 10.1002/14651858.CD003440.pub2.

Morgenstern 1982

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

O'Rourke 1989

O'Rourke K, Detsky AS. Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. *Journal of Clinical Epidemiology* 1989; 42: 1021-1024.

Oxman 1992

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

Pittler 2003

Pittler MH, Ernst E. Kava extract versus placebo for treating anxiety. *Cochrane Database of Systematic Reviews* 2003, Issue 1. CD003383. DOI: 10.1002/14651858.CD003383.

Poole 1999

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

Rücker 2009

Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* 2009; 28: 721-738.

Sackett 1996

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

Sackett 1997

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

Sharp 2001

Sharp SJ. Analysing the relationship between treatment benefit and underlying risk: precautions and practical recommendations. In: Egger M, Davey Smith G, Altman DG, editor(s). *Systematic Reviews in Health Care: Meta-analysis in Context*. 2nd edition. London (UK): BMJ Publication Group, 2001.

Sinclair 1994

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

Thompson 1997

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

Thompson 1999

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

Thompson 2002

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

Whitehead 1991

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

Whitehead 1994

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

Yusuf 1985

Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Progress in Cardiovascular Diseases* 1985; 27: 335-371.

Yusuf 1991

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

Chapter 10: Addressing reporting biases

Editors: Jonathan AC Sterne, Matthias Egger, David Moher and Isabelle Boutron on behalf of the Cochrane Bias Methods Group.

This chapter should be cited as: Sterne JAC, Egger M, Moher D, Boutron I (editors). Chapter 10: Addressing reporting biases. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS (editors), *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0 (updated June 2017), Cochrane, 2017. Available from www.training.cochrane.org/handbook.

This extract is from *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0. The previous version of this chapter (5.1.0, 2011) is available online at handbook.cochrane.org.

An earlier version (version 5.0.2, 2008) of the *Handbook* is also published by John Wiley & Sons, Ltd under "The Cochrane Book Series" Imprint, as Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on www.wiley.com.

Key Points

- Only a proportion of research projects will be published in sources easily identifiable by authors of systematic reviews. Reporting biases arise when the dissemination of research findings is influenced by the nature and direction of results.
- The contribution made to the totality of the evidence in systematic reviews by studies with statistically non-significant results is as important as that from studies with statistically significant results.
- The convincing evidence for the presence of several types of reporting biases (outlined in this chapter) demonstrates the need to search comprehensively for studies that meet the eligibility criteria for a Cochrane Review.

- Prospective trial registration, now a requirement for publication in many journals, has the potential to reduce the effects of publication bias substantially.
- Funnel plots can be used for reviews with sufficient numbers of included studies, but an asymmetrical funnel plot should not be equated with publication bias.
- Several methods are available to test for asymmetry in a funnel plot and recommendations are included in the chapter for selecting an appropriate test.

10.1 Introduction

The dissemination of research findings should not be considered as being divided into published or unpublished work, but as a continuum that ranges from the sharing of draft papers among colleagues, through presentations at meetings and published abstracts, to papers in journals that are indexed in the major bibliographic databases (Smith 1999). It has long been recognized that only a proportion of research projects ultimately reach publication in an indexed journal, and thus become easily identifiable for systematic reviews.

Reporting biases arise when the dissemination of research findings is influenced by the nature and direction of results. Statistically significant, 'positive' results that indicate that an intervention works are more likely to be published, published rapidly, published in English, published more than once, published in high impact journals and, with respect to the last point, more likely to be cited by others. The contribution made to the totality of the evidence in systematic reviews by studies with non-significant results is as important as that from studies with statistically significant results. It is highly desirable to consider the potential impact of reporting biases on the results of the review or the meta-analyses it contains. C73

Table 10.1.a summarizes some different types of reporting biases. These are considered in more detail in Section 10.2, highlighting in particular the evidence supporting the presence of each bias. Approaches for avoiding reporting biases in Cochrane Reviews are discussed in Section 10.3, and funnel plots and statistical methods for detecting potential biases are addressed in Section 10.4. Although for the purpose of discussing these biases statistically significant ($P < 0.05$) results will sometimes be denoted as 'positive' results and statistically non-significant or null results as 'negative' results, Cochrane review authors should not use such labels.

Table 10.1.a: Definitions of some types of reporting biases

Type of reporting bias	Definition
Publication bias	The *publication* or *non-publication* of research findings, depending on the nature and direction of the results
Time lag bias	The *rapid* or *delayed* publication of research findings, depending on the nature and direction of the results
Multiple (duplicate) publication bias	The *multiple* or *singular* publication of research findings, depending on the nature and direction of the results
Location bias	The publication of research findings in journals with different *ease of access* or *levels of indexing* in standard databases, depending on the nature and direction of results
Citation bias	The *citation* or *non-citation* of research findings, depending on the nature and direction of the results
Language bias	The publication of research findings *in a particular language*, depending on the nature and direction of the results
Outcome reporting bias	The *selective reporting* of some outcomes but not others, depending on the nature and direction of the results

10.2 Types of reporting biases and the supporting evidence

10.2.1 Publication bias

In a 1979 article (Rosenthal 1979), “The ‘file drawer problem’ and tolerance for null results”, Rosenthal described a gloomy scenario where “the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g. P > 0.05) results”. The file drawer problem has long been suspected in the social sciences: a review of psychology journals found that 97.3% of 294 studies published in the 1950s rejected the null hypothesis at the 5% level (P < 0.05; Sterling 1959). This study was updated and complemented with three other journals (*New England Journal of Medicine*, *American Journal of Epidemiology,* *American Journal of Public Health*; Sterling 1995). Little had changed in the psychology journals (95.6% reported significant results) and high proportions of statistically significant results (85.4%) were also found in the general medical and public health journals. Similar results have been reported in many different areas such as emergency medicine (Moscati 1994), alternative and complementary medicine (Vickers 1998, Pittler 2000), and acute stroke trials (Liebeskind 2006). A recent study of 758 articles across health research in general observed 78% of first-reported results to be statistically significant, and found two noticeable discontinuities of the distribution of P values at P = 0.01 and P = 0.05 (Albarqouni 2017).

It is possible that studies that suggest a beneficial intervention effect or a larger effect size are published, while a similar amount of data that points in the other direction remains unpublished. In this situation, a systematic review of the published studies could identify a spurious beneficial intervention effect, or miss an important adverse effect of an intervention. In cardiovascular medicine, investigators who, in 1980, found an increased death rate among patients with acute myocardial infarction treated with a class 1 anti-arrhythmic drug dismissed it as a chance finding and did not publish their trial at the time (Cowley 1993). Their findings would have contributed to a more timely detection of the increased mortality that has since become known to be associated with the use of class I anti-arrhythmic agents (Teo 1993, CLASP Collaborative Group 1994).

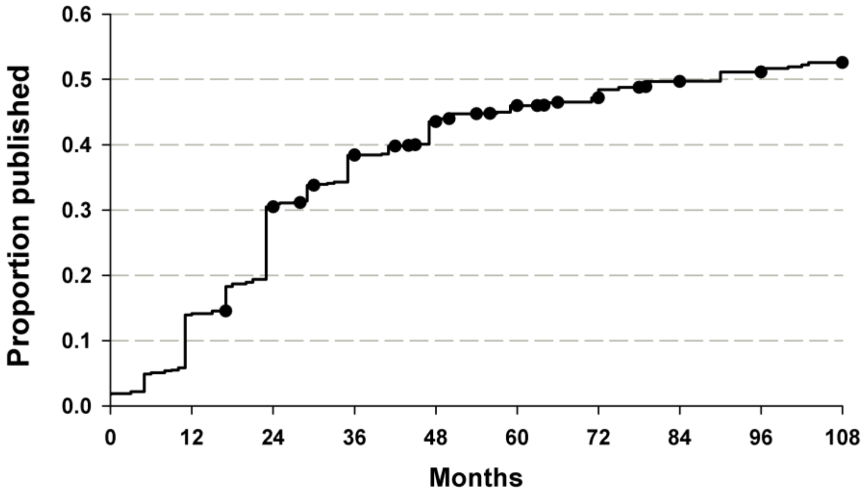
Studies that examine the existence of publication bias empirically can be viewed in two categories: namely, indirect and direct evidence. Surveys of published results, such as some of those already described (Sterling 1995, Albarqouni 2017), can provide only indirect evidence of publication bias, as the proportion of all hypotheses tested for which the null hypothesis is truly false is unknown. There is also substantial direct evidence of publication bias. Roberta Scherer and colleagues updated a systematic review that summarized 79 studies which described subsequent full publication of research initially presented in abstract or short report form (Scherer 2007). The data from 45 of these studies that included data on time to publication are summarized in Figure 10.2.a. Only about half of the abstracts presented at conferences were later published in full (63% for randomized trials), and subsequent publication was associated with positive results (Scherer 2007).

Additional direct evidence is available from a number of cohort studies of proposals submitted to ethics committees and institutional review boards (Easterbrook 1991, Dickersin 1992, Stern 1997, Decullier 2005, Decullier 2007), trials submitted to licensing authorities (Bardy 1998), analyses of trials registries (Simes 1987), or from cohorts of trials funded by specific funding agencies (Dickersin 1993). Several years later researchers contacted the principal investigators for each cohort of research proposals to determine the publication status of each completed study. In all these studies publication was more likely if the intervention effects were large and statistically significant.

Hopewell and colleagues completed a methodology review of such studies, restricting their attention to studies of clinical trials (Hopewell 2009). Five studies were included in the review, and the percentages of trials that resulted in full publication as journal articles ranged from 36% to 94% across these five studies (Table 10.2.a). Positive results were consistently more likely to have been published than negative results; the odds of publication were approximately four times greater if results were statistically significant (odds ratio (OR) 3.90, 95% confidence interval (CI) 2.68 to 5.68) as shown in Figure 10.2.b. Other factors such as the study size, funding source, and academic rank and sex of primary investigator were not consistently associated with the probability of publication, or were not possible to assess separately for clinical trials (Hopewell 2009).

Recently, the US Food and Drug Administration (FDA) database has been used in several cohort studies to explore reporting bias. Turner and colleagues compared reviews from the FDA and matched publications for 74 studies of antidepressant agents (Turner 2008). They found that 31% of studies were not published. Within the published literature, 94% of the trials were positive, compared with 51% of trials known to the FDA. Meta-analysis from published data showed an increase in effect size that ranged from 11% to 69% compared with FDA reviews. Other work using the FDA database has shown similar results, although the magnitude of publication bias varies by drugs and outcomes (Rising 2008, Hart 2012, Turner 2012). These trials also highlight that FDA reports, which are freely available on the FDA website, can be a useful resource when searching systematically for unpublished trials.

Figure 10.2.a: Cumulative full publication of results initially presented as abstracts from 45 studies reporting time to publication that followed up research presented at meetings and conferences



month	0	12	24	36	48	60	72	84	96	108
# published	362	2,460	3,348	1,519	800	280	282	84	27	10
# remaining	20,227	19,091	16,313	10,758	9,032	6,518	4,030	1,803	1,352	246

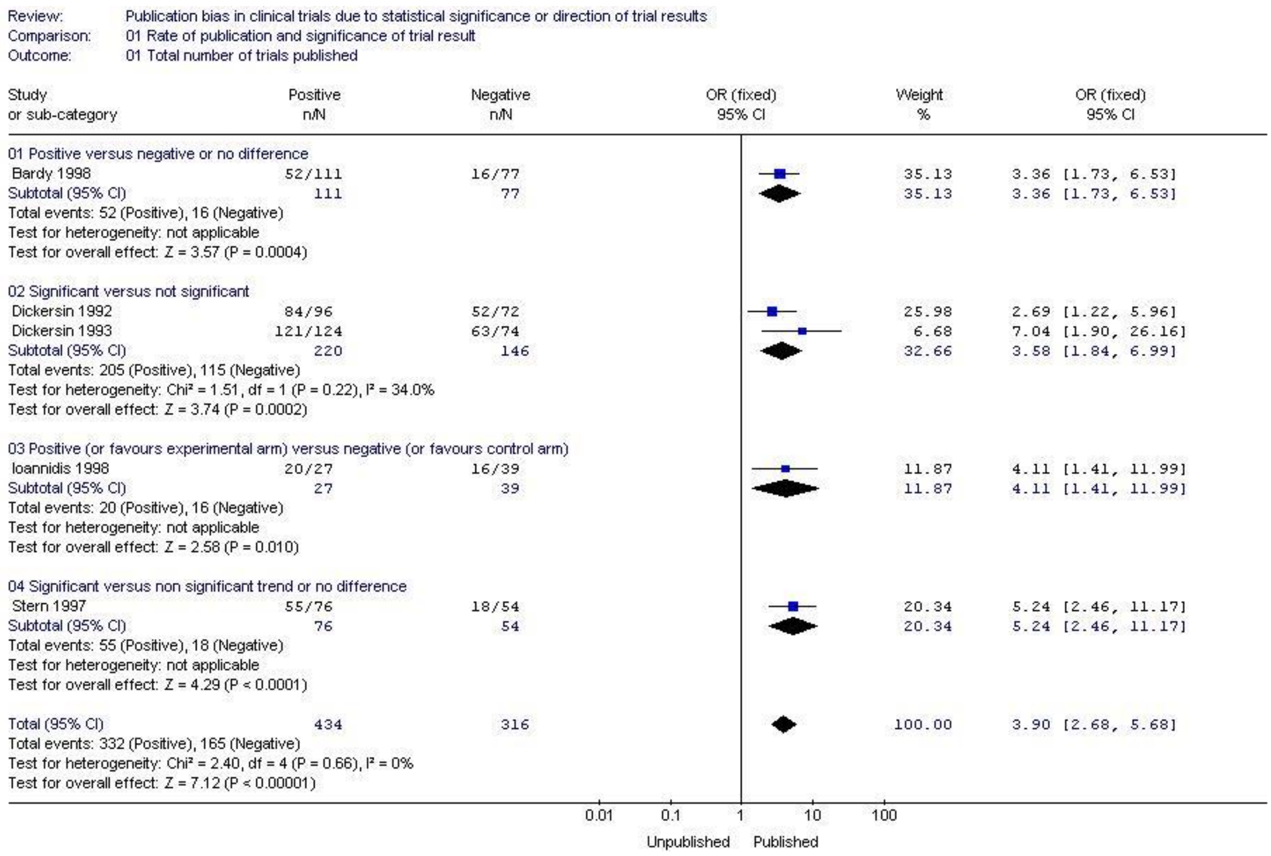
N = 20,227 abstracts
Circles show points where data censored because reports stopped follow-up.

Table 10.2.a: Publication status of five cohorts of research projects approved by ethics committees or institutional review boards that had been completed and analysed at the time of follow-up (adapted from Hopewell 2009)

	Johns Hopkins University, Baltimore	National Institutes of Health, USA	Royal Prince Alfred Hospital, Sydney	National Agency for Medicine, Finland	National Institutes of Health, USA, Multi-centre trials in HIV/AIDS
Reference	Dickersin 1992	Dickersin 1993	Stern 1997	Bardy 1998	Ioannidis 1998
Period of approval	1980	1979	1979-88	1987	1986-1996
Year of follow-up	1988	1988	1992	1995	1996
Number approved	168	198	130	188	66
Published	136 (81%)	184 (93%)	73 (56%)	68 (36%)	36 (54%)
Positive*	84/96 (87%)	121/124 (98%)	55/76 (72%)	52/111 (47%)	20/27 (75%)
Negative*	52/72 (72%)	63/74 (85%)	3/15 (20%)	5/44 (11%)	16/39 (41%)
Inconclusive/ null (if assessed separately)	Not assessed	Not assessed	15/39 (38%)	11/33 (33%)	Not assessed

*-Definitions differed by study.

Figure 10.2.b: Publication bias in clinical trials due to statistical significance or direction of trial results (adapted from Hopewell 2009)



Review: Publication bias in clinical trials due to statistical significance or direction of trial results
Comparison: 01 Rate of publication and significance of trial result
Outcome: 01 Total number of trials published

Study or sub-category	Positive n/N	Negative n/N	OR (fixed) 95% CI	Weight %	OR (fixed) 95% CI
01 Positive versus negative or no difference					
Bardy 1998	52/111	16/77		35.13	3.36 [1.73, 6.53]
Subtotal (95% CI)	111	77		35.13	3.36 [1.73, 6.53]
Total events: 52 (Positive), 16 (Negative)					
Test for heterogeneity: not applicable					
Test for overall effect: Z = 3.57 (P = 0.0004)					
02 Significant versus not significant					
Dickersin 1992	84/96	52/72		25.98	2.69 [1.22, 5.96]
Dickersin 1993	121/124	63/74		6.68	7.04 [1.90, 26.16]
Subtotal (95% CI)	220	146		32.66	3.58 [1.84, 6.99]
Total events: 205 (Positive), 115 (Negative)					
Test for heterogeneity: Chi² = 1.51, df = 1 (P = 0.22), I² = 34.0%					
Test for overall effect: Z = 3.74 (P = 0.0002)					
03 Positive (or favours experimental arm) versus negative (or favours control arm)					
Ioannidis 1998	20/27	16/39		11.87	4.11 [1.41, 11.99]
Subtotal (95% CI)	27	39		11.87	4.11 [1.41, 11.99]
Total events: 20 (Positive), 16 (Negative)					
Test for heterogeneity: not applicable					
Test for overall effect: Z = 2.58 (P = 0.010)					
04 Significant versus non significant trend or no difference					
Stern 1997	55/76	18/54		20.34	5.24 [2.46, 11.17]
Subtotal (95% CI)	76	54		20.34	5.24 [2.46, 11.17]
Total events: 55 (Positive), 18 (Negative)					
Test for heterogeneity: not applicable					
Test for overall effect: Z = 4.29 (P < 0.0001)					
Total (95% CI)	434	316		100.00	3.90 [2.68, 5.68]
Total events: 332 (Positive), 165 (Negative)					
Test for heterogeneity: Chi² = 2.40, df = 4 (P = 0.66), I² = 0%					
Test for overall effect: Z = 7.12 (P < 0.00001)					

10.2.1.1 Time lag bias

Studies continue to appear in print many years after approval by ethics committees. Hopewell and colleagues reviewed studies examining time to publication for results of clinical trials (Hopewell 2007a). The two studies included in this review, Ioannidis 1998 and Stern 1997, found that about half of all trials were published and that those with positive results were published, on average, approximately two to three years earlier than trials with null or negative results.

Among proposals submitted to the Royal Prince Alfred Hospital Ethics Committee in Sydney, Australia, an estimated 85% of studies with significant results had been published after 10 years compared to 65% of studies with null results (Stern 1997). The median time to publication was 4.7 years for studies with significant results and 8.0 years for studies with negative/null results. Similarly, trials conducted by multi-centre trial groups in the field of HIV infection in the USA appeared on average 4.3 years after the start of patient enrolment if results were statistically significant, but took 6.5 years to be published if the results were negative (Ioannidis 1998). Since then another study has found similar results (Decullier 2005). The fact that a substantial proportion of studies remain unpublished even a decade after the study had been completed and analysed is troubling, as potentially important information remains hidden from systematic review authors and consumers.

Ioannidis 1998 also found that trials with positive and negative results differed little in the time they took to complete follow-up. Rather, the time lag was attributable to differences in the time from completion to publication. These findings indicate that time lag bias may be introduced in systematic reviews even in situations when most or all studies will eventually be published. Studies with positive results will dominate the literature and introduce bias for several years until the negative, but equally important, results finally appear. Furthermore, rare adverse events are likely to be found later in the research process than short-term beneficial effects.

10.2.1.2 Who is responsible for publication bias?

Studies with negative results could remain unpublished because authors fail to write manuscripts and submit them to journals, as such studies are peer reviewed less favourably, or because editors simply do not want to publish negative results. The peer review process is sometimes unreliable and susceptible to subjectivity, bias and conflict of interest (Peters 1982, Godlee 1999). Experimental studies in which test manuscripts were submitted to peer reviewers or journals showed that peer reviewers were more likely to referee favourably if results were in accordance with their own views (Mahoney 1977, Epstein 1990, Ernst 1994). For example, when a selected group of authors was asked to peer review a fictitious paper on transcutaneous electrical nerve stimulation (TENS) they were influenced by their own findings and preconceptions. Other studies have shown no association between publication of submitted manuscripts and study outcomes (Abbot 1998, Olson 2002), suggesting that although peer reviewers may hold strong beliefs that will influence their assessments, there is no general bias for or against positive findings. Recently, a group of journal editors explored the impact of positive findings during the peer review process (Emerson 2010). They found that peer reviewers were more likely to recommend the positive version of a fabricated manuscript for publication than the no-difference version of the same manuscript (97.3% versus 80.0%; P < 0.001).

A number of studies have asked authors directly why they had not published their findings. The most frequent answer was that the findings were not interesting enough to merit publication (e.g. journals would be unlikely to accept the manuscripts; Easterbrook 1991, Dickersin 1992, Stern 1997, Weber 1998, Decullier 2005), or the investigators did not have enough time to prepare a manuscript (Weber 1998, Hartling 2004). Rejection of a manuscript by a journal was rarely mentioned as a reason for not publishing. In addition, Dickersin and colleagues examined the time from manuscript submission (to the journal *JAMA*) to full publication and found no association between this time and any study characteristics examined, including statistical significance of the study results (Dickersin 2002). Thus, time-lag bias may also result from delayed submission of manuscripts for publication by authors rather than by delayed publication by journals.

10.2.1.3 The influence of external funding and commercial interests

External funding has been found to be associated with publication independently of the statistical significance of the results (Dickersin 1997). Funding by government agencies was significantly associated with publication in three cohorts of proposals submitted to ethics committees (Easterbrook 1991, Dickersin 1992, Stern 1997), whereas studies sponsored by the pharmaceutical industry were less likely to be published (Easterbrook

1991, Dickersin 1992). Indeed, a large proportion of clinical trials submitted by drug companies to licensing authorities remain unpublished (Hemminki 1980, Bardy 1998).

In a systematic review, Lexchin and colleagues identified 30 studies published between 1966 and 2002 that examined whether funding of drug studies by the pharmaceutical industry was associated with outcomes that were favourable to the funder. They found that research funded by drug companies was less likely to be published than research funded from other sources, and that studies sponsored by pharmaceutical companies were more likely to have outcomes that favoured the sponsor than studies with other sponsors (Lexchin 2003). Other studies have since examined these associations and have found similar results (Bhandari 2004, Heres 2006). A study of head-to-head comparisons of antipsychotics found that the overall outcome of the trials favoured the drug manufactured by the industry sponsor in 90% of studies considered, and further found that some of the studies that were apparently similar in conduct reported opposing conclusions, each supporting the product of the study sponsor (Heres 2006).

The implication is that the pharmaceutical industry tends to discourage the publication of negative studies that it has funded. For example, a manuscript reporting on a trial that compared the bioequivalence of generic and brand levothyroxine products, which had failed to produce the results desired by the sponsor of the study, Boots Pharmaceuticals, was withdrawn because Boots took legal action against the university and the investigators. The actions of Boots, recounted in detail by one of the editors of *JAMA*, Drummond Rennie (Rennie 1997), meant that publication of the paper, Dong 1997, was delayed by about seven years. In a national survey of life-science faculty members in the USA, 20% reported that they had experienced delays of more than six months in publication of their work and reasons for not publishing included “to delay the dissemination of undesired results” (Blumenthal 1997). Delays in publication were associated with involvement in commercialization and academic-industry research relationship, as well as with male sex and higher academic rank of the investigator (Blumenthal 1997).

Industry documents made available after legal challenges have provided more insight into the different strategies of reporting bias used by the pharmaceutical industry (Vedula 2009). For example, the documents released from litigation brought by consumers against Pfizer for fraudulent sales practices in the marketing of gabapentin showed the implementation of different strategies to delay publication allowing a delay of seven years before full reporting (Vedula 2012).

10.2.2 Other reporting biases

While publication bias has long been recognized and much discussed, other factors can contribute to biased inclusion of studies in meta-analyses. Indeed, among published studies, the probability of identifying relevant studies for meta-analysis is also influenced by their results. These biases have received much less consideration than publication bias, but their consequences could be of equal importance.

10.2.2.1 Duplicate (multiple) publication bias

In 1989, Gøtzsche found that 44 (18%) out of 244 reports of trials comparing non-steroidal anti-inflammatory drugs in rheumatoid arthritis were redundant, multiple publications, which overlapped substantially with a previously published article. Twenty trials were published twice, ten trials three times and one trial four times (Gøtzsche 1989). The production of multiple publications from single studies can lead to bias in a number of ways (Huston 1996). Most importantly, studies with significant results are more likely to lead to multiple publications and presentations (Easterbrook 1991), which makes it more likely that they will be located and included in a meta-analysis. It is not always obvious that multiple publications come from a single study, and one set of study participants may be included in an analysis twice. The inclusion of duplicated data may therefore lead to overestimation of intervention effects, as was demonstrated for trials of the efficacy of ondansetron to prevent postoperative nausea and vomiting (Tramèr 1997).

Other authors have described the difficulties and frustration caused by redundancy and the ‘disaggregation’ of medical research when results from a multi-centre trial are presented in several publications (Huston 1996, Johansen 1999). Redundant publications often fail to cross-reference each other (Bailey 2002, Barden 2003), and there are examples where two articles reporting the same trial do not share a single common author (Gøtzsche 1989, Tramèr 1997). Thus, without contacting the authors, it may be difficult or impossible for review authors to determine whether two papers represent duplicate publications of one study or two separate studies, which may result in biasing a meta-analysis of these data.

10.2.2.2 Location bias

Research suggests that various factors related to the accessibility of study results are associated with effect sizes in trials. For example, in a series of trials in the field of complementary and alternative medicine, Pittler and colleagues examined the relationship between trial outcome, methodological quality and sample size with characteristics of the journals of publication of these trials (Pittler 2000). They found that trials published in low- or non-impact factor journals were more likely to report significant results than those published in high-impact mainstream medical journals and that the quality of the studies was also associated with the journal of publication. More recently, Siontis and colleagues conducted a meta-epidemiological trial that showed that small studies of experimental interventions published in prestigious journals (namely the *New England Journal of Medicine*, *JAMA* and the *Lancet*) showed more favourable results than trials in other journals, particularly for trials that were published early (Siontis 2011). Similarly, some trials suggest that trials published in English language journals are more likely to show strong significant effects than those published in non-English language journals (Egger 1997a), however this has not been shown consistently (Moher 2000, Jüni 2002, Pham 2005); see Section 10.2.2.4.

‘Location bias’ is also used to refer to the accessibility of studies based on variable indexing in electronic databases. Depending on the clinical question, choices regarding which databases to search may bias the effect estimate in a meta-analysis. For example, one study found that trials published in journals that were not indexed in MEDLINE might show a more beneficial effect than trials published in MEDLINE-indexed journals (Egger

2003). Another study of 61 meta-analyses found that, in general, trials published in journals indexed in Embase but not in MEDLINE reported smaller estimates of effect than those indexed in MEDLINE, but that the risk of bias may be minor, given the lower prevalence of the Embase unique trials (Sampson 2003). As for previous examples, these findings may vary substantially with the clinical topic being examined.

A final form of location bias is regional or developed country bias. Research supporting the evidence of this bias suggests that studies published in certain countries may be more likely than others to produce research showing significant effects of interventions. Vickers and colleagues demonstrated the potential existence of this bias (Vickers 1998).

10.2.2.3 Citation bias

The perusal of the reference lists of articles is widely used to identify additional articles that may be relevant, although there is little evidence to support this methodology. The problem with this approach is that the act of citing previous work is far from objective, and retrieving literature by scanning reference lists may thus produce a biased sample of studies. There are many possible motivations for citing an article. Brooks interviewed academic authors from various faculties at the University of Iowa and asked for the reasons for citing each reference in one of the authors' articles (Brooks 1985). Persuasiveness, that is the desire to convince peers and substantiate their own point of view, emerged as the most important reason for citing articles. Brooks concluded that authors advocate their own opinions and use the literature to justify their point of view: "Authors can be pictured as intellectual partisans of their own opinions, scouring the literature for justification" (Brooks 1985).

In Gøtzsche's analysis of trials of non-steroidal anti-inflammatory drugs in rheumatoid arthritis, trials that demonstrated a superior effect of a new drug were more likely to be cited than trials with negative results (Gøtzsche 1987). Similar results were shown in an analysis of randomized trials of hepato-biliary diseases (Kjaergard 2002). Similarly, trials of cholesterol-lowering to prevent coronary heart disease were cited almost six times more often if they were supportive of cholesterol-lowering (Ravnskov 1992). Over-citation of unsupportive studies can also occur. Hutchison and colleagues examined reviews of the effectiveness of pneumococcal vaccines and found that unsupportive studies were more likely to be cited than studies showing that vaccines worked (Hutchison 1995).

Citation bias may affect the 'secondary' literature. For example, the *ACP Journal Club* aims to summarize original and review articles so that physicians can keep abreast of the latest evidence. However, Carter and colleagues found that, after controlling for other reasons for selection, trials with a positive outcome were more likely to be summarized (Carter 2006). If positive studies are more likely to be cited, they may be more likely to be located and, thus, more likely to be included in a systematic review, thus biasing the findings of the review.

10.2.2.4 Language bias

Reviews have often been exclusively based on studies published in English. For example, among 36 meta-analyses reported in leading English-language general medicine journals from 1991 to 1993, 26 (72%) had restricted their search to studies reported in English

(Grégoire 1995). This trend may be changing, as a review of 300 systematic reviews found approximately 16% of them were limited to trials published in English, while systematic reviews published in paper-based journals were more likely than Cochrane Reviews to report having limited their search to trials published in English (Moher 2007). In addition, for reviews with a therapeutic focus, Cochrane Reviews were more likely than non-Cochrane reviews to report the absence of language restrictions (62% versus 26%; Moher 2007).

Investigators working in a non-English speaking country will publish some of their work in local journals (Dickersin 1994). It is conceivable that authors are more likely to report in an international, English-language journal if results are positive, but publish negative findings in a local journal. This has been demonstrated for the German-language literature (Egger 1997a).

Bias could thus be introduced in reviews exclusively based on English-language reports (Grégoire 1995, Moher 1996). However, the results of research examining this issue conflict. In a study of 50 reviews that employed comprehensive literature searches and included both English and non-English-language trials, Jüni and colleagues reported that non-English trials were more likely to produce significant results at $P < 0.05$, and that estimates of intervention effects were, on average, 16% (95% CI 3% to 26%) more beneficial in non-English-language trials than in English-language trials (Jüni 2002). Conversely, Moher and colleagues examined the effect of inclusion or exclusion of English language trials in two studies of meta-analyses and found, overall, that the exclusion of trials reported in a language other than English did not significantly affect the results of the meta-analyses (Moher 2003). These results were similar when the analysis was limited to meta-analyses of trials of conventional medicines. When the analyses were conducted separately for meta-analyses of trials of complementary and alternative medicines, however, the effect size of meta-analyses significantly decreased by excluding reports in languages other than English (Moher 2003).

The extent and effects of language bias may have diminished recently because of the shift towards publication of studies in English. In 2006, Galandi and colleagues reported a dramatic decline in the number of randomized trials published in German-language healthcare journals: with fewer than two randomized trials published per journal per year after 1999 (Galandi 2006). While the potential impact of studies published in languages other than English in a meta-analysis may be minimal, it is difficult to predict the cases in which this exclusion may bias a systematic review. Review authors may want to search without language restrictions and decisions about including reports from languages other than English may need to be taken on a case-by-case basis.

10.2.2.5 Outcome reporting bias

In many studies, a range of outcome measures is recorded, but not all are reported (Pocock 1987, Tannock 1996). The choice of outcomes that are reported can be influenced by the results, potentially making published results misleading. For example, two separate analyses of a double-blind placebo-controlled trial that assessed the efficacy of amoxicillin in children with non-suppurative otitis media reached opposite conclusions mainly because different 'weight' was given to the various outcome measures that were assessed in the study (Mandel 1987, Cantekin 1991). This disagreement was conducted in the public arena, since it was accompanied by accusations of impropriety against the team producing the findings favourable to amoxicillin. The leader of this team had received substantial fiscal support, both in research grants and as personal honoraria, from the manufacturers of amoxicillin (Rennie 1991). It is a good example of how reliance upon the data chosen to be presented by the investigators can lead to distortion (Anonymous 1991). Such 'outcome reporting bias' may be particularly important for adverse effects. Hemminki examined reports of clinical trials submitted by drug companies to licensing authorities in Finland and Sweden and found that unpublished trials gave information on adverse effects more often than published trials (Hemminki 1980). Since then several other studies have shown that the reporting of adverse events and safety outcomes in clinical trials is often inadequate and selective (Ioannidis 2001, Melander 2003, Heres 2006). A group from Canada, Denmark and the UK pioneered empirical research into the selective reporting of study outcomes (Chan 2004a, Chan 2004b, Chan 2005). These studies are described in Chapter 8 (Section 8.14), along with a more detailed discussion of outcome reporting bias.

10.3 Avoiding reporting biases

10.3.1 Implications of the evidence concerning reporting biases

The convincing evidence for the presence of reporting biases, described in Section 10.2, demonstrates the need to search comprehensively for studies that meet the eligibility criteria for a Cochrane Review. Review authors should ensure that multiple sources are searched; for example, a search of MEDLINE alone would not be considered sufficient. Sources and methods for searching are described in detail in Chapter 6. Comprehensive searches do not necessarily remove bias, however, and review authors should bear in mind, for example, that study reports may present results selectively; that reference lists may cite sources selectively; and that duplicate publication of results can be difficult to spot. Furthermore, the availability of study information may be subject to time-lag bias, particularly in fast-moving research areas. Two further means of reducing, or potentially avoiding, reporting biases will now be discussed: the inclusion of unpublished studies, and the use of trial registries.

10.3.2 Including unpublished studies in systematic reviews

Publication bias clearly is a major threat to the validity of any type of review, but particularly of unsystematic, narrative reviews. Obtaining and including data from unpublished trials appears to be one obvious way of avoiding this problem. Hopewell and colleagues conducted a review of studies comparing the effect of the inclusion or exclusion of 'grey' literature (defined here as reports that are produced by all levels of government, academics, business and industry in print and electronic formats but that are not controlled by commercial publishers) in meta-analyses of randomized trials (Hopewell 2007b). They included five studies (Fergusson 2000, McAuley 2000, Burdett 2003, Egger 2003, Hopewell 2004), all of which showed that published trials had an overall greater intervention effect than grey trials. A meta-analysis of three of these studies suggested that, on average, published trials showed a 9% larger intervention effect than grey trials (Hopewell 2007b).

The inclusion of data from unpublished studies can itself introduce bias. The studies that can be located may be an unrepresentative sample of all unpublished studies. Unpublished studies may be of lower methodological quality than published studies: a study of 60 meta-analyses that included published and unpublished trials found that unpublished trials were less likely to conceal intervention allocation adequately and to blind outcome assessments (Egger 2003). In contrast, Hopewell and colleagues found no difference in the quality of reporting of this information (Hopewell 2004).

A further problem relates to the willingness of investigators of any unpublished studies located to provide data. This may depend upon the findings of the study, more favourable results being provided more readily. Again, this could bias the findings of a systematic review. Interestingly, when Hetherington and colleagues, in a massive effort to obtain information about unpublished trials in perinatal medicine, approached 42,000 obstetricians and paediatricians in 18 countries they identified only 18 unpublished trials that had been completed for more than two years (Hetherington 1989).

A questionnaire assessing the attitudes toward inclusion of unpublished data was sent to the authors of 150 meta-analyses and to the editors of the journals that published them (Cook 1993). Researchers and editors differed in their views about including unpublished data in meta-analyses. Support for the use of unpublished material was evident among a clear majority (78%) of meta-analysts while journal editors were less convinced (47%; Cook 1993). This study was repeated in 2006, with a focus on the inclusion of grey literature in systematic reviews, and it was found that acceptance of inclusion of grey literature had increased, and, although differences between the two groups remained (systematic review authors: 86%, editors: 69%), these may have decreased since the Cook 1993 paper was published (Tetzlaff 2006).

Reasons for reluctance to include grey literature include the absence of peer-review for unpublished literature. It should be kept in mind, however, that the refereeing process has not always been a successful way of ensuring that published results are valid (Godlee 1999). Teams involved in preparing Cochrane Reviews should have at least a similar level of expertise for appraising unpublished studies as peer reviewers for a journal. On the other hand, meta-analyses of unpublished data from interested sources are clearly a cause for concern.

To minimize reporting bias, it is highly desirable to seek key unpublished information in a systematic way. These include data from studies that have been completed but not published, as well as data available to the researcher but missing from reports of included studies. There are several potential sources of unpublished information on trials methods and results (Chan 2012). These include trial registries such as the World Health Organization’s International Clinical Trials Registry Platform Search Portal (www.who.int/trialsearch/), as well as the ClinicalTrials.gov results database, and pharmaceutical companies’ voluntary trial registers and results databases for drugs that have received regulatory approval. Other sources concern regulatory agencies (the FDA and the European Medicines Agency) and contacting trialists and sponsors.

10.3.3 Trial registries and publication bias

In September 2004 a number of major medical journals belonging to the International Committee of Medical Journal Editors (ICMJE) announced they would no longer publish trials that were not registered at inception (Abbasi 2004). All trials that began enrolment of participants after September 2005 had to be registered in a public trials registry at or before the onset of enrolment in order to be considered for publication in those journals. The ICMJE described 'acceptable' registers; these were to be electronically searchable, freely accessible to the public, open to all registrants, and managed by a non-profit organization. Similarly, the ICMJE asked clinical trialists to adhere to a minimum dataset proposed by the World Health Organization.

In September 2007, the Food and Drug Administration Amendments Act (FDAAA) expanded the registration requirement for the ClinicalTrials.gov registry to mandate investigators to submit basic summary results within one year after study completion (Zarin 2008, Zarin 2011). This requirement concerns most trials of drugs, devices or biologics regulated by the FDA having at least one site in the USA. The ClinicalTrials.gov results database should improve transparency. If this initiative is successful, it has the potential to reduce the effects of publication bias substantially. However, this would depend on review authors identifying all relevant trials by searching online trial registries, and also on the results of unpublished trials identified via registries being made available to them. While there is emerging evidence to suggest that some of the data fields requested in the registries are incomplete (Zarin 2005, Prayle 2012), this is likely to improve over time. The extent to which trial registration will facilitate the work of Cochrane review authors is unclear at present. For advice on searching trial registries, see Chapter 6 (Section 6.2.3).

10.4 Detecting reporting biases

10.4.1 Funnel plots

A funnel plot is a simple scatter plot of the intervention effect estimates from individual studies against some measure of each study's size or precision. In common with forest plots, it is most common to plot the effect estimates on the horizontal scale, and thus the measure of study size on the vertical axis. This is the opposite of conventional graphical displays for scatter plots, in which the outcome (e.g. intervention effect) is plotted on the vertical axis and the covariate (e.g. study size) is plotted on the horizontal axis.

The name 'funnel plot' arises from the fact that precision of the estimated intervention effect increases as the size of the study increases. Effect estimates from small studies will therefore scatter more widely at the bottom of the graph, with the spread narrowing among larger studies. In the absence of bias the plot should approximately resemble a symmetrical (inverted) funnel. This is illustrated in Panel A of Figure 10.4.a, in which the effect estimates in the larger studies are close to the true intervention odds ratio of 0.4.

If there is bias, for example because smaller studies without statistically significant effects (shown as open circles in Figure 10.4.a, Panel A) remain unpublished, this will lead to an asymmetrical appearance of the funnel plot with a gap in a bottom corner of the graph (Panel B). In this situation the effect calculated in a meta-analysis will tend to overestimate the intervention effect (Egger 1997b, Villar 1997). The more pronounced the asymmetry, the more likely it is that the amount of bias will be substantial.

Funnel plots were first used in educational research and psychology, with effect estimates plotted against total sample size (Light 1984). It is now usually recommended that the standard error of the intervention effect estimate be plotted, rather than the total sample size, on the vertical axis (Sterne 2001). This is because the statistical power of a trial is determined by factors in addition to sample size, such as the number of participants experiencing the event for dichotomous outcomes, and the standard deviation of responses for continuous outcomes. For example, a study with 100,000 participants and 10 events is less likely to show a statistically significant intervention effect than a study with 1000 participants and 100 events. The standard error summarizes these other factors. Plotting standard errors on a reversed scale places the larger, or most powerful, studies towards the top of the plot. Another potential advantage of using standard errors is that a simple triangular region can be plotted, within which 95% of studies would be expected to lie in the absence of both biases and heterogeneity. These regions are included in Figure 10.4.a. Funnel plots of effect estimates against their standard errors (on a reversed scale) can be created using RevMan. A triangular 95% confidence region based on a fixed-effect meta-analysis can be included in the plot, and different plotting symbols allow studies in different subgroups to be identified.

Publication bias need not lead to asymmetry in funnel plots. In the absence of any intervention effect, selective publication based on the P value alone will lead to a symmetrical funnel plot in which studies on the extreme left or right are more likely to be published than those in the middle. This could bias the estimated between-study heterogeneity variance.

Ratio measures of intervention effect (such as odds ratios and risk ratios) should be plotted on a logarithmic scale. This ensures that effects of the same magnitude but opposite directions (for example odds ratios of 0.5 and 2) are equidistant from 1.0. For outcomes measured on a continuous (numerical) scale (e.g. blood pressure, depression score) intervention effects are measured as mean differences or standardized mean differences, which should therefore be used as the horizontal axis in funnel plots. As far as we are aware, no empirical investigations have examined choice of axes for funnel plots for continuous outcomes. For mean differences, the standard error is approximately proportional to the inverse of the square root of the number of participants, and therefore seems an uncontroversial choice for the vertical axis.

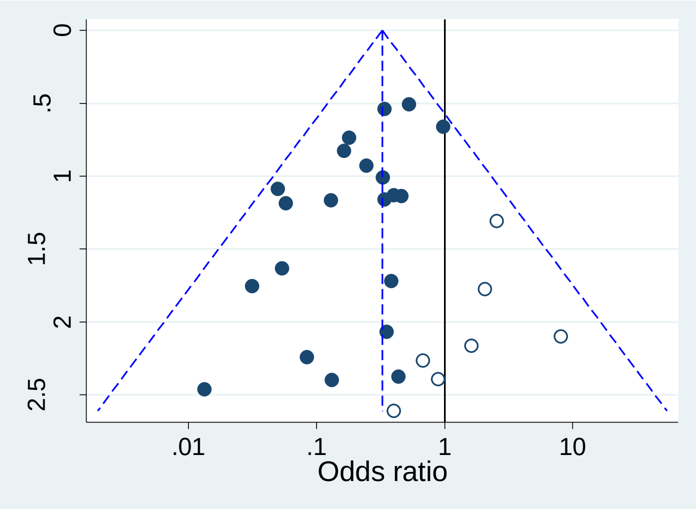
Some authors have argued that visual interpretation of funnel plots is too subjective to be useful. In particular, Terrin and colleagues found that researchers had only a limited ability to identify funnel plots from meta-analyses subject to publication bias correctly (Terrin 2005).

A further, important, problem with funnel plots is that some effect estimates (e.g. odds ratios and standardized mean differences) are naturally correlated with their standard errors, and can produce spurious asymmetry in a funnel plot. This problem is discussed in more detail in Section 10.4.3.

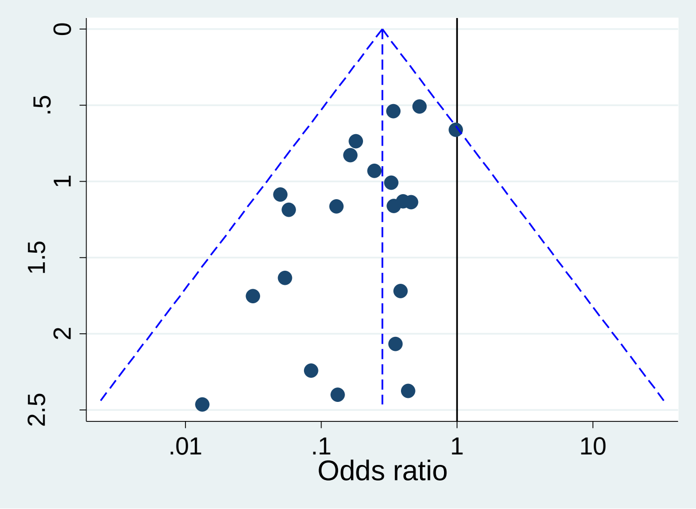
Figure 10.4.a: Hypothetical funnel plots

Panel A: symmetrical plot in the absence of bias. Panel B: asymmetrical plot in the presence of reporting bias. Panel C: asymmetrical plot in the presence of bias because some smaller studies (open circles) are of lower methodological quality and therefore produce exaggerated intervention effect estimates.

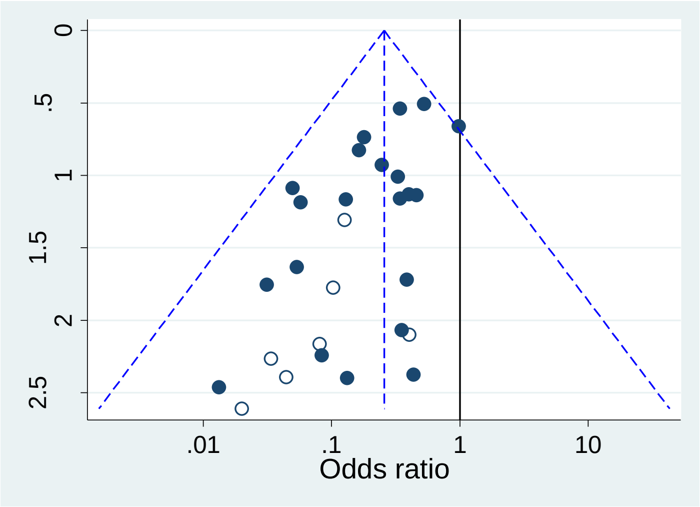
Panel A



Panel B



Panel C



10.4.2 Different reasons for funnel plot asymmetry

Although funnel plot asymmetry has long been equated with publication bias (Light 1984, Begg 1988), the funnel plot should be seen as a generic means of displaying *small-study effects* – a tendency for the intervention effects estimated in smaller studies to differ from those estimated in larger studies (Sterne 2000, Sterne 2011). Small-study effects may be due to reasons other than publication bias (Egger 1997b, Sterne 2000, Sterne 2011). Some of these are shown in Table 10.4.a.

Differences in methodological quality are an important potential source of funnel plot asymmetry. Smaller studies tend to be conducted and analysed with less methodological rigour than larger studies (Egger 2003). Trials of lower quality also tend to show larger intervention effects (Schulz 1995). Therefore, trials that would have been 'negative', if conducted and analysed properly, may become 'positive' (Figure 10.4.a, Panel C).

True heterogeneity in intervention effects may also lead to funnel plot asymmetry (Sterne 2011). For example, substantial benefit may be seen only in patients at high risk for the outcome which is affected by the intervention, and usually these high risk patients are more likely to be included in small, early studies (Davey Smith 1994, Glasziou 1995). In addition, small trials are generally conducted before larger trials are established, and, in the intervening years standard interventions may improve (resulting in smaller intervention effects in the larger trials). Furthermore, some interventions may have been implemented less thoroughly in larger trials and may, therefore, have resulted in smaller estimates of the intervention effect (Stuck 1998). Finally, it is of course possible that an asymmetrical funnel plot arises merely by the play of chance. Terrin and colleagues have suggested that the funnel plot is inappropriate for heterogeneous meta-analyses, and drew attention to the premise that the studies come from a single underlying population given by the originators of the funnel plot (Light 1984, Terrin 2003).

A proposed enhancement to the funnel plot is to include contour lines corresponding to perceived 'milestones' of statistical significance (P = 0.01, 0.05, 0.1 etc.; Peters 2008). This allows the statistical significance of study estimates, and areas in which studies are perceived to be missing, to be considered. Such ‘contour-enhanced’ funnel plots may help review authors to differentiate asymmetry that is due to publication bias from that due to other factors. For example if studies appear to be missing in areas of statistical non-significance (see Figure 10.4.b, Panel A for example) then this adds credence to the possibility that the asymmetry is due to publication bias. Conversely, if the supposed missing studies are in areas of higher statistical significance (see Figure 10.4.b, Panel B for example), this would suggest the cause of the asymmetry may be more likely to be due to factors other than publication bias (see Table 10.4.a). If there are no statistically significant studies then publication bias may not be a plausible explanation for funnel plot asymmetry (Ioannidis 2007a).

Therefore, when interpreting funnel plots, systematic review authors need to distinguish the different possible reasons for funnel plot asymmetry listed in Table 10.4.a. Knowledge of the particular intervention, and the circumstances in which it was implemented in different studies, can help identify true heterogeneity as a cause of funnel plot asymmetry, but a concern remains that visual interpretation of funnel plots is inherently subjective. Therefore, statistical tests for funnel plot asymmetry, and the extent to which they may assist in the objective interpretation of funnel plots will now be discussed. When review authors are concerned that small study effects are influencing the results of a meta-analysis, they may want to conduct sensitivity analyses in order to explore the robustness of the meta-analysis’ conclusions to different assumptions about the causes of funnel plot asymmetry: these are discussed in Section 10.4.4.

Table 10.4.a: Possible sources of asymmetry in funnel plots

Adapted from Egger 1997b.

1. Selection biases

- Publication bias:
    - delayed publication (also known as ‘time-lag’ or ‘pipeline’) bias;
    - location biases:
        - language bias;
        - citation bias;
        - multiple publication bias.
- Selective outcome reporting.

2. Poor methodological quality leading to spuriously inflated effects in smaller studies

- Poor methodological design.
- Inadequate analysis.
- Fraud.

3. True heterogeneity

- Size of effect differs according to study size (for example, due to differences in the intensity of interventions or differences in underlying risk between studies of different sizes).

4. Artefactual

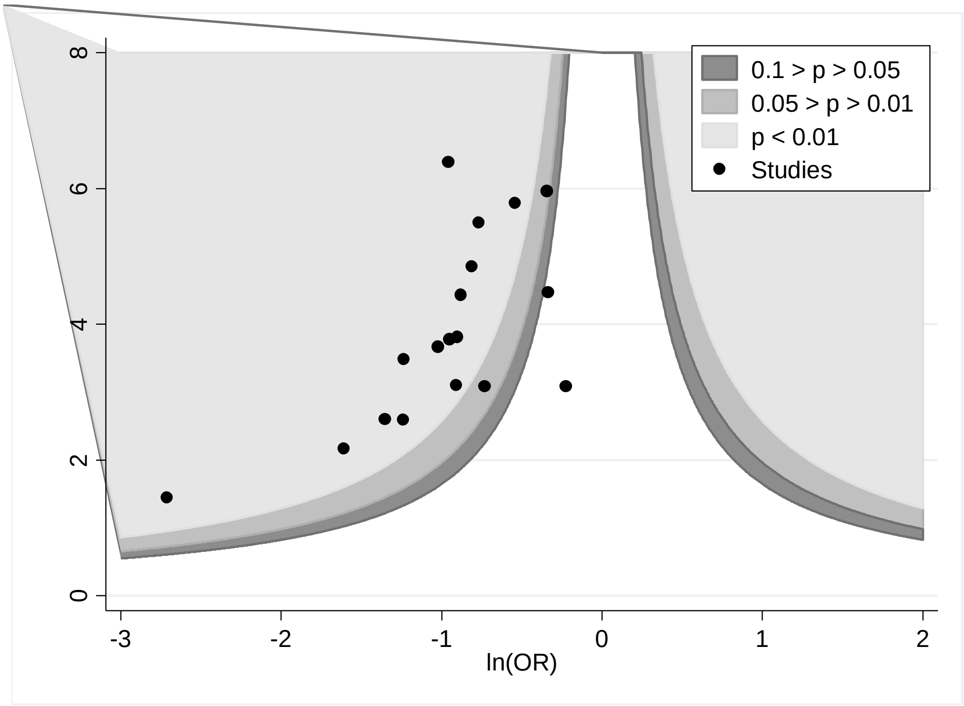
- In some circumstances (see Section 10.4.3), sampling variation can lead to an association between the intervention effect and its standard error.

5. Chance

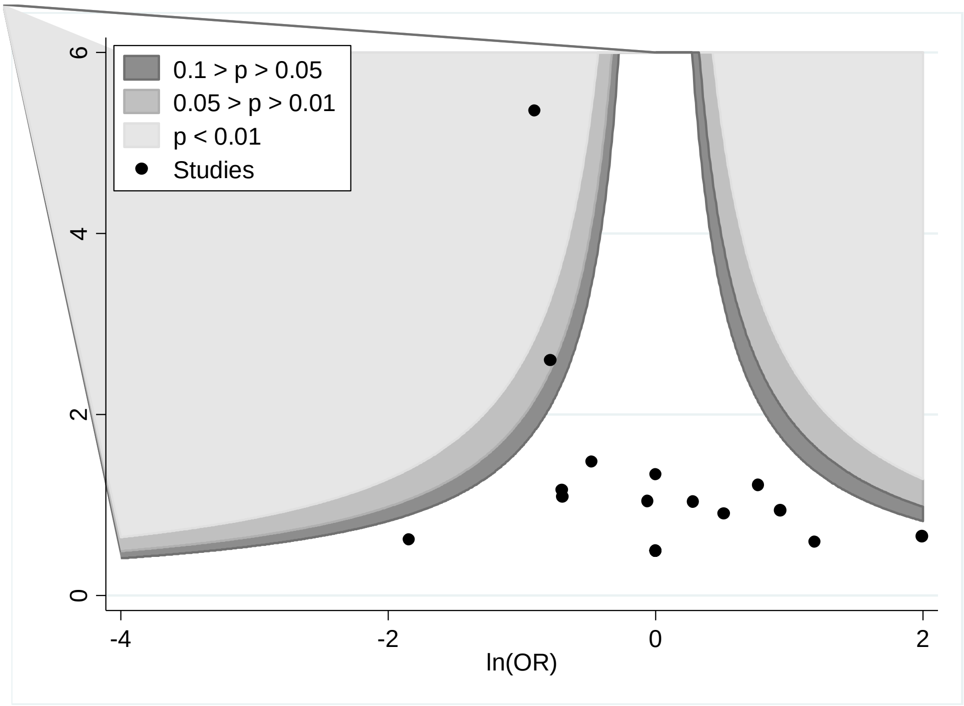
Figure 10.4.b: Contour-enhanced funnel plots

Panel A: there is a suggestion of missing studies on the right-hand side of the plot, broadly in the area of non-significance (i.e. the white area where P > 0.1), for which publication bias is a plausible explanation. Panel B: there is a suggestion of missing studies on the bottom left-hand side of the plot. Since most of this area contains regions of high statistical significance (i.e. indicated by darker shading), this reduces the plausibility that publication bias is the underlying cause of this funnel asymmetry.

Panel A



Panel B



10.4.3 Tests for funnel plot asymmetry

A test for funnel plot asymmetry (small study effects) formally examines whether the association between estimated intervention effects and a measure of study size (such as the standard error of the intervention effect) is greater than might be expected to occur by chance. For outcomes measured on a continuous (numerical) scale this is reasonably straightforward. Using an approach proposed by Egger 1997b, it is possible to perform a linear regression of the intervention effect estimates on their standard errors, weighting by 1/(variance of the intervention effect estimate). This looks for a straight-line relationship between intervention effect and its standard error. Under the null hypothesis of no small study effects (e.g. Panel A in Figure 10.4.a) such a line would be vertical. The greater the association between intervention effect and standard error (e.g. as in Panel B in Figure 10.4.a), the more the slope would move away from the vertical. Note that the weighting is important to ensure the regression estimates are not dominated by the smaller studies.

When outcomes are dichotomous, and intervention effects are expressed as odds ratios, the approach proposed by Egger 1997b corresponds to a linear regression of the log odds ratio on its standard error, weighted by the inverse of the variance of the log odds ratio (Sterne 2000). This is the most widely used and cited approach to testing for funnel plot asymmetry. Unfortunately, there are statistical problems with this approach, because the standard error of the log odds ratio is mathematically linked to the size of the odds ratio, even in the absence of small study effects (Irwig 1998; see Deeks 2005 for an algebraic explanation of this phenomenon). This can cause funnel plots plotted using log odds ratios (or odds ratios on a log scale) to appear asymmetrical and can mean that P values from the test of Egger and colleagues are too small, leading to false-positive test results. These problems are especially prone to occur when the intervention has a large effect, there is substantial between-study heterogeneity, there are few events per study, or when all studies are of similar sizes.

Therefore, a number of authors have proposed alternative tests for funnel plot asymmetry: these are summarized in Table 10.4.b. Because it is impossible to know the precise mechanism for publication bias, simulation studies (in which the tests are evaluated on a large number of computer-generated datasets) are required to evaluate the characteristics of the tests under a range of assumptions about the mechanism for publication bias (Sterne 2000, Macaskill 2001, Harbord 2006, Peters 2006, Schwarzer 2007). The most comprehensive study (in terms of scenarios examined, simulations carried out and the range of tests compared) was reported by Rücker 2008. Results of this and the other published simulation studies inform the recommendations on testing for funnel plot asymmetry in Section 10.4.3.1 (Sterne 2011). Although simulation studies provide useful insights, they inevitably evaluate circumstances that differ from a particular meta-analysis of interest, so their results must be interpreted carefully.

Most of this methodological work has focused on intervention effects measured as odds ratios. While it seems plausible to expect that corresponding problems will arise for intervention effects measured as risk ratios or standardized mean differences, further investigations of these situations are required.

There is ongoing debate over the representativeness of the parameter values used in the simulation studies, and the mechanisms used to simulate publication bias and small study effects, which are often chosen with little explicit justification. Some potentially useful variations on the different tests remain unexamined. Therefore, it is not possible to make definitive recommendations on choice of tests for funnel plot asymmetry. Nevertheless, we can identify three tests that should be considered by review authors wishing to test for funnel plot asymmetry.

None of the tests described here is implemented in RevMan, and consultation with a statistician is recommended for their implementation.

Table 10.4.b: Proposed tests for funnel plot asymmetry

N_{tot} is the total sample size, N_E and N_C are the sizes of the experimental and control intervention groups, S is the total number of events across both groups and $F = N_{tot} – S$. Note that only the first three of these tests, Begg 1994, Egger 1997b and Tang 2000, can be used for continuous outcomes.

Reference	Basis of test
Begg 1994	Rank correlation between standardized intervention effect and its standard error
Egger 1997b	Linear regression of intervention effect estimate against its standard error, weighted by the inverse of the variance of the intervention effect estimate
Tang 2000	Linear regression of intervention effect estimate on $1/\sqrt{N_{tot}}$, with weights N_{tot}
Macaskill 2001*	Linear regression of intervention effect estimate on N_{tot}, with weights $S \times F/N_{tot}$
Deeks 2005*	Linear regression of log odds ratio on $1/\sqrt{ESS}$ with weights ESS, where effective sample size $ESS = 4N_E \times N_C / N_{tot}$
Harbord 2006*	Modified version of the test proposed by Egger and colleagues, based on the 'score' (O–E) and 'score variance' (V) of the log odds ratio
Peters 2006*	Linear regression of intervention effect estimate on $1/N_{tot}$, with weights $S \times F/N_{tot}$
Schwarzer 2007*	Rank correlation test, using mean and variance of the non-central hypergeometric distribution
Rücker 2008	Test based on arcsine transformation of observed risks, with explicit modelling of between-study heterogeneity

* Test formulated in terms of odds ratios, but may be applicable to other measures of intervention effect.

10.4.3.1 Recommendations on testing for funnel plot asymmetry

For **all types of outcome**:

- As a rule of thumb, tests for funnel plot asymmetry should be used only when there are at least 10 studies included in the meta-analysis, because when there are fewer studies the power of the tests is too low to distinguish chance from real asymmetry. In some situations, the minimum numbers of studies may be substantially more than 10.

- Tests for funnel plot asymmetry should not be used if all studies are of similar sizes (similar standard errors of intervention effect estimates). However, we are not aware of evidence from simulation studies that provides specific guidance on when study sizes should be considered 'too similar'.

- Results of tests for funnel plot asymmetry should be interpreted in the light of visual inspection of the funnel plot. For example, do small studies tend to lead to more or less beneficial intervention effect estimates? Are there studies with markedly different intervention effect estimates (outliers), or studies that are highly influential in the meta-analysis? Is a small P value caused by one study alone? Examining a contour-enhanced funnel plot, as outlined in Section 10.4.1, may further help interpretation of a test result.

- When there is evidence of small-study effects, publication bias should be considered as only one of a number of possible explanations (see Table 10.4.a). Although funnel plots, and tests for funnel plot asymmetry, may alert review authors to a problem that needs to be considered, they do not provide a solution to this problem. Finally, review authors should remember that, because the tests typically have relatively low power, even when a test does not provide evidence of funnel plot asymmetry, bias (including publication bias) cannot be excluded.

For **continuous outcomes with intervention effects measured as mean differences**:

- The test proposed in Egger 1997b may be used to test for funnel plot asymmetry. There is currently no reason to prefer any of the more recently proposed tests in this situation, although their relative advantages and disadvantages have not been formally examined. While we know of no research specifically on the power of the approach in the continuous case, general considerations suggest that the power will be greater than for dichotomous outcomes, and that use of the method with fewer than 10 studies would be unwise.

For **dichotomous outcomes with intervention effects measured as odds ratios:**

- The tests proposed in Harbord 2006 and Peters 2006 avoid the mathematical association between the log odds ratio and its standard error (and hence false-positive test results) that occurs for the test proposed by Egger 1997b when there is a substantial intervention effect, while retaining power compared with alternative tests. However, false-positive results may still occur in the presence of substantial between-study heterogeneity.

- The test proposed in Rücker 2008 avoids false-positive results both when there is a substantial intervention effect, and in the presence of substantial between-study heterogeneity. As a rule of thumb, when the estimated between-study heterogeneity variance of log odds ratios, tau-squared (also known as τ^2, or Tau^2), is more than 0.1, only the version of the arcsine test including random-effects (referred to as 'AS+RE' in Rücker 2008) has been shown to work reasonably well. However, it is slightly conservative in the absence of heterogeneity, and its interpretation is less familiar because it is based on an arcsine transformation. (Note that although this recommendation is based on the magnitude of Tau^2, other factors – including the sizes of the different studies and their distribution – influence a test's performance. We are not currently able to incorporate these other factors in our recommendations.)

- When the heterogeneity variance Tau^2 is less than 0.1, one of the tests proposed by Harbord 2006, Peters 2006 or Rücker 2008 can be used. (Test performance generally deteriorates as Tau^2 increases.)

- As far as possible, review authors should specify their testing strategy in advance (noting that test choice may be dependent on the degree of heterogeneity observed). They should apply only one test, appropriate to the context of the particular meta-analysis, from the list recommended in Table 10.4.b and report only the result from their chosen test. Application of two or more tests is undesirable, since interpretation of the most extreme (largest or smallest) P value from a set of tests is not well-characterized.

For **dichotomous outcomes with intervention effects measured as risk ratios or risk differences, and continuous outcomes with intervention effects measured as standardized mean differences:**

- Potential problems in funnel plots have been less extensively studied for these effect measures than for odds ratios, and firm guidance is not yet available.

- Meta-analyses of risk differences are generally considered less appropriate than meta-analyses using a ratio measure of effect (see Chapter 9, Section 9.4.4.4). For similar reasons, funnel plots using risk differences should seldom be of interest. If the risk ratio (or odds ratio) is constant across studies, then a funnel plot using risk differences will be asymmetrical if smaller studies have higher (or lower) baseline risk.

Based on a survey of meta-analyses published in the *Cochrane Database of Systematic Reviews*, these criteria imply that tests for funnel plot asymmetry should be used in only a minority of meta-analyses (Ioannidis 2007a).

Tests for which there is insufficient evidence to recommend use

The following comments apply to all intervention measures. The test proposed in Begg 1994 has the same statistical problems but lower power than the test in Egger 1997b, and is therefore not recommended. The test proposed in Tang 2000 has not been evaluated in simulation studies, while the test proposed in Macaskill 2001 has lower power than more recently proposed alternatives. The test proposed in Schwarzer 2007 avoids the mathematical association between the log odds ratio and its standard error, but has low power relative to the tests discussed in Table 10.4.b.

In the context of meta-analyses of intervention studies considered in this chapter, the test proposed in Deeks 2005 is likely to have lower power than more recently proposed alternatives. This test was not designed as a test for publication bias in systematic reviews of randomized trials: rather it is aimed at meta-analyses of diagnostic test accuracy studies, where very large odds ratios and very imbalanced studies cause problems for other tests.

10.4.4 Sensitivity analyses

When review authors find evidence of small-study effects, they should consider sensitivity analyses to examine how the results of the meta-analysis change under different assumptions relating to the reasons for these effects. We stress the exploratory nature of such analysis, due to the inherent difficulty in adjusting for publication bias and a lack of research into the performance of such methods applied conditionally based on the results of tests for publication bias considered in Section 10.4.3. This area is relatively underdeveloped; the following approaches have been suggested.

10.4.4.1 Comparing fixed-effect and random-effects estimates

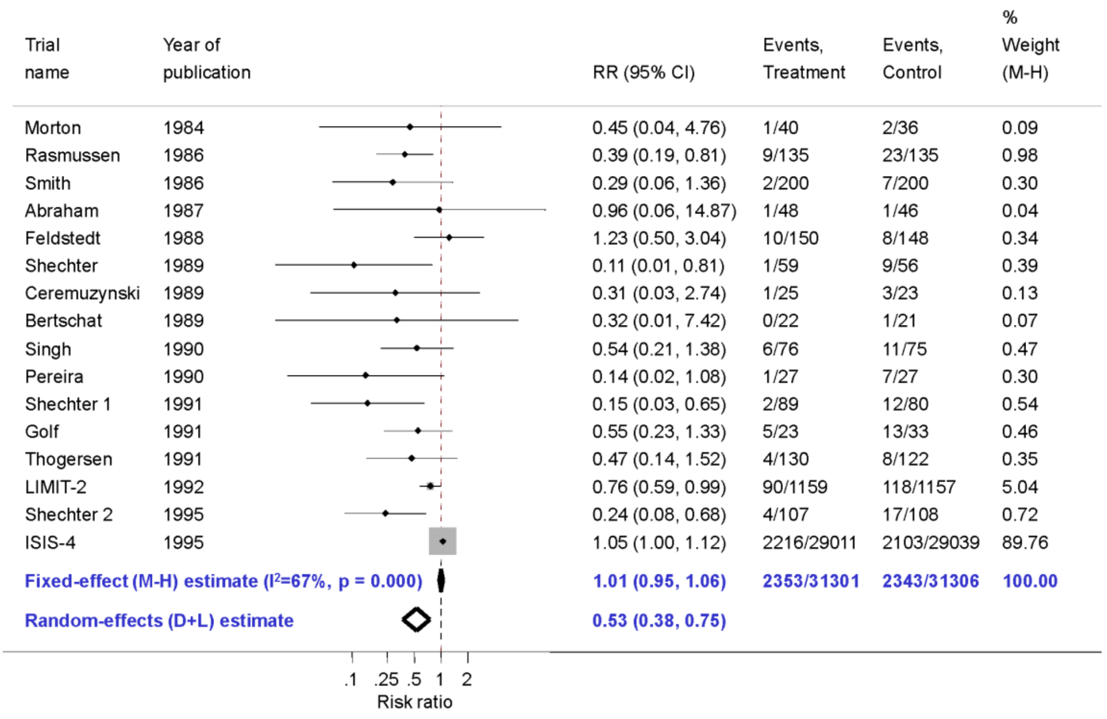
In the presence of heterogeneity, a random-effects meta-analysis weights the studies relatively more equally than a fixed-effect analysis. It follows that in the presence of small-study effects such as those displayed in Figure 10.2.a, in which the intervention effect is more beneficial in the smaller studies, the random-effects estimate of the intervention effect will be more beneficial than the fixed-effect estimate. Poole and Greenland summarized this by noting that “random-effects meta-analyses are not always conservative” (Poole 1999). This issue is also discussed in Chapter 9 (Section 9.5.4).

An extreme example of the differences between fixed-effect and random-effects analyses that can arise in the presence of small-study effects is shown in Figure 10.4.c, which displays both fixed-effect and random-effects estimates of the effect of intravenous magnesium on mortality following myocardial infarction. This is a well-known example in which beneficial effects of intervention were found in a meta-analysis of small studies, but were subsequently contradicted when the very large ISIS-4 study found no evidence that magnesium affected mortality.

Because there is substantial between-trial heterogeneity, the studies are weighted much more equally in the random-effects analysis than in the fixed-effect analysis. In the fixed-effect analysis the ISIS-4 trial gets 90% of the weight and so there is no evidence of a beneficial intervention effect. In the random-effects analysis the small studies dominate, and there appears to be clear evidence of a beneficial effect of intervention. To interpret the accumulated evidence, it is necessary to make a judgement about the likely validity of the combined evidence from the smaller studies, compared with that from the ISIS-4 trial.

We recommend that when review authors are concerned about the influence of small-study effects on the results of a meta-analysis in which there is evidence of between-study heterogeneity ($I^2 > 0$), they compare the fixed-effect and random-effects estimates of the intervention effect. If the estimates are similar, then any small-study effects have little effect on the intervention effect estimate. If the random-effects estimate is more beneficial, review authors should consider whether it is reasonable to conclude that the intervention was more effective in the smaller studies. If the larger studies tend to be those conducted with more methodological rigour, or conducted in circumstances more typical of the use of the intervention in practice, then review authors should consider reporting the results of meta-analyses restricted to the larger, more rigorous studies. Formal evaluation of such strategies in simulation studies would be desirable. Note that formal statistical comparisons of the fixed-effect and random-effects estimates of intervention effect are not possible, and that it is still possible for small-study effects to bias the results of a meta-analysis in which there is no evidence of heterogeneity, even though the fixed-effect and random-effects estimates of intervention effect will be identical in this situation.

Figure 10.4.c: Comparison of fixed-effect and random-effects meta-analytic estimates of the effect of intravenous magnesium on mortality following myocardial infarction



Trial name	Year of publication	RR (95% CI)	Events, Treatment	Events, Control	% Weight (M-H)
Morton	1984	0.45 (0.04, 4.76)	1/40	2/36	0.09
Rasmussen	1986	0.39 (0.19, 0.81)	9/135	23/135	0.98
Smith	1986	0.29 (0.06, 1.36)	2/200	7/200	0.30
Abraham	1987	0.96 (0.06, 14.87)	1/48	1/46	0.04
Feldstedt	1988	1.23 (0.50, 3.04)	10/150	8/148	0.34
Shechter	1989	0.11 (0.01, 0.81)	1/59	9/56	0.39
Ceremuzynski	1989	0.31 (0.03, 2.74)	1/25	3/23	0.13
Bertschat	1989	0.32 (0.01, 7.42)	0/22	1/21	0.07
Singh	1990	0.54 (0.21, 1.38)	6/76	11/75	0.47
Pereira	1990	0.14 (0.02, 1.08)	1/27	7/27	0.30
Shechter 1	1991	0.15 (0.03, 0.65)	2/89	12/80	0.54
Golf	1991	0.55 (0.23, 1.33)	5/23	13/33	0.46
Thogersen	1991	0.47 (0.14, 1.52)	4/130	8/122	0.35
LIMIT-2	1992	0.76 (0.59, 0.99)	90/1159	118/1157	5.04
Shechter 2	1995	0.24 (0.08, 0.68)	4/107	17/108	0.72
ISIS-4	1995	1.05 (1.00, 1.12)	2216/29011	2103/29039	89.76
Fixed-effect (M-H) estimate (I^2=67%, p = 0.000)		**1.01 (0.95, 1.06)**	**2353/31301**	**2343/31306**	**100.00**
Random-effects (D+L) estimate		**0.53 (0.38, 0.75)**			

10.4.4.2 Trim and fill

The 'trim and fill' method aims both to identify and correct for funnel plot asymmetry arising from publication bias (Taylor 1998, Duval 2000). The basis of the method is to: 1) 'trim' (remove) the smaller studies causing funnel plot asymmetry; 2) use the trimmed funnel plot to estimate the true 'centre' of the funnel; then 3) replace the omitted studies and their missing 'counterparts' around the centre (filling). As well as providing an estimate of the number of missing studies, an adjusted intervention effect is derived by performing a meta-analysis that includes the filled studies.

The trim and fill method requires no assumptions about the mechanism leading to publication bias, provides an estimate of the number of missing studies, and also provides an estimated intervention effect that is 'adjusted' for the publication bias (based on the filled studies). However, it is built on the strong assumption that there should be a symmetrical funnel plot, and there is no guarantee that the adjusted intervention effect matches what would have been observed in the absence of publication bias, since one cannot know the true mechanism for publication bias. Equally importantly, the trim and fill method does not take into account reasons for funnel plot asymmetry other than publication bias. Therefore, 'corrected' intervention effect estimates from this method should be interpreted with great caution. The method is known to perform poorly in the presence of substantial between-study heterogeneity (Terrin 2003, Peters 2007). Additionally, estimation and inferences are based on a dataset that contains imputed intervention effect estimates. Such estimates, it can be argued, inappropriately contribute information that reduces the uncertainty in the summary intervention effect.

10.4.4.3 Fail-safe N

Rosenthal suggested assessing the potential for publication bias to have influenced the results of a meta-analysis by calculating the 'fail-safe N', that is, the number of additional 'negative' studies (studies in which the intervention effect was zero) that would be needed to increase the P value for the meta-analysis to above 0.05 (Rosenthal 1979). However the estimate of fail-safe N is highly dependent on the mean intervention effect that is assumed for the unpublished studies (Iyengar 1988), and available methods lead to widely varying estimates of the number of additional studies (Becker 2005). The method also runs against the principle that in medical research in general, and systematic reviews in particular, one should concentrate on the size of the estimated intervention effect and the associated confidence intervals, rather than on whether the P value reaches a particular, arbitrary threshold, although related methods for effect sizes have also been proposed (Orwin 1983). Therefore this, and related methods, are not recommended for use in Cochrane Reviews.

10.4.4.4 Other selection models

Other authors have proposed more sophisticated methods that avoid strong assumptions about the association between study P value and publication probability (Dear 1992, Hedges 1992). These methods can be extended to estimate intervention effects, corrected for the estimated publication bias (Vevea 1995). However, they require a large number of studies so that a sufficient range of study P values is included. A Bayesian approach in which the number and outcomes of unobserved studies are simulated has also been proposed as a means of correcting intervention effect estimates for publication bias

(Givens 1997). Some work has examined the possibility of assessing robustness over a range of weight functions, thus avoiding the need for large numbers of studies (Vevea 2005). The complexity of the statistical methods, and the large number of studies needed, probably explain why selection models have not been widely used in practice.

10.4.4.5 Sensitivity analyses based on selection models

Copas developed a model in which the probability that a study is included in a meta-analysis depends on its standard error. Since it is not possible to estimate all model parameters precisely, he advocates sensitivity analyses in which the value of the estimated intervention effect is computed under a range of assumptions about the severity of the selection bias (Copas 1999). Rather than a single intervention effect estimated 'corrected' for publication bias, the reader can see how the estimated effect (and confidence interval) varies as the assumed amount of selection bias increases. Application of the method to epidemiological studies of environmental tobacco smoke and lung cancer suggests that publication bias may explain some of the association observed in meta-analyses of these studies (Copas 2000).

10.4.4.6 Testing for excess of studies with significant results

Ioannidis and Trikalinos have proposed a simple test that aims to evaluate whether there is an excess of studies that have formally statistically significant results (Ioannidis 2007b). The test compares the number of studies that have formally statistically significant results with the number of statistically significant results expected under different assumptions about the magnitude of the effect size. The simplest assumption is that the effect size is equal to the observed summary effect in the meta-analysis (but this may introduce an element of circularity). Other values for the underlying effect size, and different thresholds of significance, may be used. Hence, like the contour funnel plots described in Section 10.4.1, but unlike the regression tests, this method considers the distribution of the significance of study results. However, unlike either the regression tests or contour funnel plots, the test does not make any assumption about small-study effects. An excess of significant results can reflect either suppression of whole studies or related selective/manipulative analysis and reporting practices that would cause similar excess.

The test has limited power, as do most other tests, when there are few studies and when there are few studies with significant results. As the test has not been rigorously evaluated through simulation in comparison with alternative tests and under different scenarios, currently we do not recommend it as an alternative to those described in Section 10.4.3.

A novel feature of the test is that it can be applied across a large number of meta-analyses on the same research field to examine the extent of publication and selective reporting biases across a whole domain of clinical research. Again, further evaluation of this approach would be welcome.

10.4.4.7 Regression based methods

A further approach to dealing with potential reporting bias is a regression approach based on the tests used for examining funnel plot asymmetry (Stanley 2008, Moreno 2009a). This approach fits a regression line to the funnel plot, and extrapolates the line to a study with infinite precision (or infinite size). The effect size at this 'ideal' point is regarded as an estimate of effect size, after adjusting for small-study effects. Numerous options are available for the choice of explanatory variable in the regression, including the options listed in Table 10.4.b (Moreno 2009b).

Moreno 2012 addresses in detail a particular model that is not included in this list, in which effect size is regressed on within-study variance, and in which heterogeneity is incorporated as a multiplicative rather than an additive component. Moreno 2012 shows that more weight is given to the larger studies than in either a standard fixed-effect or random-effects meta-analysis, so the adjusted estimate will, as intended, lie closer to the effects observed in the larger studies. Rücker and colleagues used a similar approach and combined it with a shrinkage procedure (Rücker 2011b, Rücker 2011a). The underlying model is an extended random-effects model, with an additional parameter representing the bias introduced by small-study effects.

In common with tests for funnel plot asymmetry, the methods should be used only when there are sufficient studies (at least 10) to allow appropriate estimation of the regression line. When all the studies are small, extrapolation to an infinitely sized study may produce effect estimates that are more extreme than any of the existing studies, and if the approach is used in such a situation it might be more appropriate to extrapolate only as far as the largest observed study.

10.4.5 Summary

Although there is clear evidence that publication and other reporting biases lead to over-optimistic estimates of intervention effects, overcoming, detecting and correcting for reporting bias is problematic. Comprehensive searches are important, particularly to identify research as well defined as randomized trials. However, these methods are not sufficient to prevent some substantial potential biases. Publication bias should be seen as one of a number of possible causes of ‘small-study effects’ – a tendency for estimates of the intervention effect to be more beneficial in smaller studies. Funnel plots allow review authors to make a visual assessment of whether small-study effects may be present in a meta-analysis. For continuous (numerical) outcomes with intervention effects measured as mean differences, funnel plots and statistical tests for funnel plot asymmetry are valid. However, for dichotomous outcomes with intervention effects expressed as odds ratios, the standard error of the log odds ratio is mathematically linked to the size of the odds ratio, even in the absence of small-study effects. This can cause funnel plots plotted using log odds ratios (or odds ratios on a log scale) to appear asymmetrical and can mean that P values from the test of Egger and colleagues are too small. For other effect measures, firm guidance is not yet offered. Three statistical tests for small-study effects are recommended for use in Cochrane Reviews, provided that there are at least 10 studies. However, none is implemented in RevMan and statistical support is usually required. Only one test has been shown to work when the between-study heterogeneity variance exceeds 0.1. Results from tests for funnel plot asymmetry should be interpreted cautiously. When there is evidence of small-study effects, publication bias should be considered as only one of a number of possible explanations. In these circumstances, review authors should attempt to understand the source of the small-study effects, and consider their implications in sensitivity analyses.

10.5 Methodological standards for the conduct of new Cochrane Intervention Reviews

No.	Status	Name	Standard	Rationale & elaboration	Handbook sections
C73	Highly desirable	Investigating reporting biases	Consider the potential impact of reporting biases on the results of the review or the meta-analyses it contains.	There is overwhelming evidence of reporting biases of various types. These can be addressed at various points in the review. A thorough search, and attempts to obtain unpublished results, might minimize the risk. Analyses of the results of included studies, for example using funnel plots, can sometimes help determine the possible extent if the problem, as can attempts to identify study protocols, which should be a more routine feature of a review.	10.1 10.2

10.6 Chapter information

Editors: Jonathan AC Sterne, Matthias Egger, David Moher and Isabelle Boutron on behalf of the Cochrane Bias Methods Group.

Contributing authors: Isabelle Boutron, James Carpenter, Matthias Egger, Roger Harbord, Julian Higgins, David Jones, David Moher, Jonathan Sterne, Alex Sutton, Jennifer Tetzlaff, Lucy Turner.

Acknowledgements: We thank Doug Altman, Jon Deeks, John Ioannidis, Jaime Peters and Gerta Rücker for helpful comments.

Declarations of interest: James Carpenter, Jon Deeks, Matthias Egger, Roger Harbord, David Jones, Jaime Peters, Gerta Rücker, Jonathan Sterne and Alex Sutton are all authors on papers proposing tests for funnel plot asymmetry.

Box 10.6.a: The Cochrane Bias Methods Group

The Bias Methods Group (BMG), previously the Reporting Bias Methods Group, was formally registered as a Methods Group in 2000. The BMG addresses a range of different forms of bias, such as publication bias, language bias, selective outcome reporting bias and biases arising from study design and conduct. A major initiative of the group, in collaboration with the Statistical Methods Group, was the development of the new guidance for assessing risk of bias of included studies in Cochrane Reviews.

Activities of BMG members include:

- undertaking empirical research to examine whether, and in which circumstances, various biases may have a substantial impact on systematic reviews, including the preparation of Cochrane Methodology Reviews;
- undertaking methodological research on how to identify and address potential biases in systematic reviews and meta-analyses;
- helping to complete and co-ordinate Methods systematic reviews pertinent to the Group’s remit;
- providing advice to Cochrane entities; and
- offering training to both Cochrane and non-Cochrane systematic review authors via formal and informal opportunities.

The BMG membership emailing list is used as a forum for discussion and dissemination of information. The annual *Cochrane Methods* publication, Cochrane Connect (Cochrane’s official international newsletter) and Cochrane Community (internal newsletter), are also used for dissemination of group activities.

Website: bmg.cochrane.org

10.7 References

Abbasi 2004

Abbasi K. Compulsory registration of clinical trials. *BMJ* 2004; 329: 637-638.

Abbot 1998

Abbot NC, Ernst E. Publication bias: direction of outcome is less important than scientific quality. *Perfusion* 1998; 11: 182-184.

Albarqouni 2017

Albarqouni LN, Lopez-Lopez JA, Higgins JP. Indirect evidence of reporting biases was found in a survey of medical research studies. *Journal of Clinical Epidemiology* 2017 Jan 11; [Epub ahead of print]. DOI: 10.1016/j.jclinepi.2016.11.013.

Anonymous 1991

Anonymous. Subjectivity in data analysis. *Lancet* 1991; 337: 401-402.

Bailey 2002

Bailey BJ. Duplicate publication in the field of otolaryngology-head and neck surgery. *Otolaryngology and Head and Neck Surgery* 2002; 126: 211-216.

Barden 2003

Barden J, Edwards JE, McQuay HJ, Moore RA. Oral valdecoxib and injected parecoxib for acute postoperative pain: a quantitative systematic review. *BMC Anesthesiology* 2003; 3: 1.

Bardy 1998

Bardy AH. Bias in reporting clinical trials. *British Journal of Clinical Pharmacology* 1998; 46: 147-150.

Becker 2005

Becker BJ. Failsafe *N* or file-drawer number. In: Rothstein HR, Sutton AJ, Borenstein M, editor(s). *Publication Bias in Meta-Analysis*. Chichester (UK): John Wiley & Sons, 2005.

Begg 1988

Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 1988; 151: 419-463.

Begg 1994

Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994; 50: 1088-1101.

Bhandari 2004

Bhandari M, Busse JW, Jackowski D, Montori VM, Schünemann H, Sprague S, et al. Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. *Canadian Medical Association Journal* 2004; 170: 477-480.

Blumenthal 1997

Blumenthal D, Campbell EG, Anderson MS, Causino N, Louis KS. Withholding research results in academic life science. Evidence from a national survey of faculty. *JAMA* 1997; 277: 1224-1228.

Brooks 1985

Brooks TA. Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science* 1985; 36: 223-229.

Burdett 2003

Burdett S, Stewart LA, Tierney JF. Publication bias and meta-analyses: a practical example. *International Journal of Technology Assessment in Health Care* 2003; 19: 129-134.

Cantekin 1991

Cantekin EI, McGuire TW, Griffith TL. Antimicrobial therapy for otitits media with effusion ('secretory' otitits media). *JAMA* 1991; 266: 3309-3317.

Carter 2006

Carter AO, Griffin GH, Carter TP. A survey identified publication bias in the secondary literature. *Journal of Clinical Epidemiology* 2006; 59: 241-245.

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. *Canadian Medical Association Journal* 2004; 171: 735-740.

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 2012

Chan AW. Out of sight but not out of mind: how to search for unpublished clinical trial evidence. *BMJ* 2012; 344: d8013.

CLASP Collaborative Group 1994

CLASP Collaborative Group. CLASP: a randomized trial of low-dose aspirin for the prevention and treatment of pre-eclampsia among 9364 pregnant women. *Lancet* 1994; 343: 619-629.

Cook 1993

Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, et al. Should unpublished data be included in meta-analyses? Current convictions and controversies. *JAMA* 1993; 269: 2749-2753.

Copas 1999

Copas J. What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 1999; 162: 95-109.

Copas 2000

Copas JB, Shi JQ. Reanalysis of epidemiological evidence on lung cancer and passive smoking. *BMJ* 2000; 320: 417-418.

Cowley 1993

Cowley AJ, Skene A, Stainer K, Hampton JR. The effect of lorcainide on arrhythmias and survival in patients with acute myocardial infarction: an example of publication bias. *International Journal of Cardiology* 1993; 40: 161-166.

Davey Smith 1994

Davey Smith G, Egger M. Who benefits from medical interventions? *BMJ* 1994; 308: 72-74.

Dear 1992

Dear KB, Begg CB. An approach to assessing publication bias prior to performing a meta-analysis. *Statistical Science* 1992; 7: 237-245.

Decullier 2005

Decullier E, Lheritier V, Chapuis F. Fate of biomedical research protocols and publication bias in France: retrospective cohort study. *BMJ* 2005; 331: 19.

Decullier 2007

Decullier E, Chapuis F. Oral presentation bias: a retrospective cohort study. *Journal of Epidemiology and Community Health* 2007; 61: 190-193.

Deeks 2005

Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology* 2005; 58: 882-893.

Dickersin 1992

Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA* 1992; 267: 374-378.

Dickersin 1993

Dickersin K, Min YI. NIH clinical trials and publication bias. *Online Journal of Current Clinical Trials* 1993; Doc No 50.

Dickersin 1994

Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994; 309: 1286-1291.

Dickersin 1997

Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Education and Prevention* 1997; 9: 15-21.

Dickersin 2002

Dickersin K, Olson CM, Rennie D, Cook D, Flanagin A, Zhu Q, et al. Association between time interval to publication and statistical significance. *JAMA* 2002; 287: 2829-2831.

Dong 1997

Dong BJ, Hauck WW, Gambertoglio JG, Gee L, White JR, Bubp JL, et al. Bioequivalence of generic and brand-name levothyroxine products in the treatment of hypothyroidism. *JAMA* 1997; 277: 1205-1213.

Duval 2000

Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; 56: 455-463.

Easterbrook 1991

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991; 337: 867-872.

Egger 1997a

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

Egger 1997b

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

Egger 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 Technology Assessment* 2003; 7: 1-76.

Emerson 2010

Emerson GB, Warme WJ, Wolf FM, Heckman JD, Brand RA, Leopold SS. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *Archives of Internal Medicine* 2010; 170: 1934-1939.

Epstein 1990

Epstein WM. Confirmational response bias among social work journals. *Science, Technology, & Human Values* 1990; 15: 9-38.

Ernst 1994

Ernst E, Resch KL. Reviewer bias: A blinded experimental study. *The Journal of Laboratory and Clinical Medicine* 1994; 124: 178-182.

Fergusson 2000

Fergusson D, Laupacis A, Salmi LR, McAlister FA, Huet C. What should be included in meta-analyses? An exploration of methodological issues using the ISPOT meta-analyses. *International Journal of Technology Assessment in Health Care* 2000; 16: 1109-1119.

Galandi 2006

Galandi D, Schwarzer G, Antes G. The demise of the randomised controlled trial: bibliometric study of the German-language health care literature, 1948 to 2004. *BMC Medical Research Methodology* 2006; 6: 30.

Givens 1997

Givens GH, Smith DD, Tweedie RL. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* 1997; 12: 221-250.

Glasziou 1995

Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *BMJ* 1995; 311: 1356-1359.

Godlee 1999

Godlee F, Dickersin K. Bias, subjectivity, chance, and conflict of interest in editorial decisions. In: Godlee F, Jefferson T, editor(s). *Peer Review in Health Sciences*. London (UK): BMJ Books, 1999.

Gøtzsche 1987

Gøtzsche PC. Reference bias in reports of drug trials. *British Medical Journal (Clinical Research Edition)* 1987; 295: 654-656.

Gøtzsche 1989

Gøtzsche PC. Multiple publication of reports of drug trials. *European Journal of Clinical Pharmacology* 1989; 36: 429-432.

Grégoire 1995

Grégoire G, Derderian F, LeLorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? *Journal of Clinical Epidemiology* 1995; 48: 159-163.

Harbord 2006

Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine* 2006; 25: 3443-3457.

Hart 2012

Hart B, Lundh A, Bero L. Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses. *BMJ* 2012; 344: d7202.

Hartling 2004

Hartling L, Craig WR, Russell K, Stevens K, Klassen TP. Factors influencing the publication of randomized controlled trials in child health research. *Archives of Pediatrics and Adolescent Medicine* 2004; 158: 983-987.

Hedges 1992

Hedges LV. Modeling publication selection effects in meta-analysis. *Statistical Science* 1992; 7: 246-255.

Hemminki 1980

Hemminki E. Study of information submitted by drug companies to licensing authorities. *British Medical Journal* 1980; 280: 833-836.

Heres 2006

Heres S, Davis J, Maino K, Jetzinger E, Kissling W, Leucht S. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. *American Journal of Psychiatry* 2006; 163: 185-194.

Hetherington 1989

Hetherington J, Dickersin K, Chalmers I, Meinert CL. Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. *Pediatrics* 1989; 84: 374-380.

Hopewell 2004

Hopewell S. Impact of grey literature on systematic reviews of randomized trials (PhD Thesis). University of Oxford, 2004.

Hopewell 2007a

Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials. *Cochrane Database of Systematic Reviews* 2007, Issue 2. MR000011. DOI: 10.1002/14651858.MR000011.pub2.

Hopewell 2007b

Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database of Systematic Reviews* 2007, Issue 2. MR000010. DOI: 10.1002/14651858.MR000010.pub3.

Hopewell 2009

Hopewell S, Louden K, Clarke M, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews* 2009, Issue 1. MR000006. DOI: 10.1002/14651858.MR000006.pub3.

Huston 1996

Huston P, Moher D. Redundancy, disaggregation, and the integrity of medical research. *Lancet* 1996; 347: 1024-1026.

Hutchison 1995

Hutchison BG, Oxman AD, Lloyd S. Comprehensiveness and bias in reporting clinical trials. Study of reviews of pneumococcal vaccine effectiveness. *Canadian Family Physician* 1995; 41: 1356-1360.

Ioannidis 1998

Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* 1998; 279: 281-286.

Ioannidis 2001

Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA* 2001; 285: 437-443.

Ioannidis 2007a

Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Canadian Medical Association Journal* 2007; 176: 1091-1096.

Ioannidis 2007b

Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clinical Trials* 2007; 4: 245-253.

Irwig 1998

Irwig L, Macaskill P, Berry G, Glasziou P. Bias in meta-analysis detected by a simple, graphical test. Graphical test is itself biased. *BMJ* 1998; 316: 470-471.

Iyengar 1988

Iyengar S, Greenhouse JB. Selection problems and the file drawer problem. *Statistical Science* 1988; 3: 109-135.

Johansen 1999

Johansen HK, Gøtzsche PC. Problems in the design and reporting of trials of antifungal agents encountered during meta-analysis. *JAMA* 1999; 282: 1752-1759.

Jüni 2002

Jüni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *International Journal of Epidemiology* 2002; 31: 115-123.

Kjaergard 2002

Kjaergard LL, Gluud C. Citation bias of hepato-biliary randomized clinical trials. *Journal of Clinical Epidemiology* 2002; 55: 407-410.

Lexchin 2003

Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 2003; 326: 1167-1170.

Liebeskind 2006

Liebeskind DS, Kidwell CS, Sayre JW, Saver JL. Evidence of publication bias in reporting acute stroke clinical trials. *Neurology* 2006; 67: 973-979.

Light 1984

Light RJ, Pillemer DB. *Summing Up: The Science of Reviewing Research*. Cambridge (MA): Harvard University Press, 1984.

Macaskill 2001

Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine* 2001; 20: 641-654.

Mahoney 1977

Mahoney MJ. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research* 1977; 1: 161-175.

Mandel 1987

Mandel EH, Rockette HE, Bluestone CD, Paradise JL, Nozza RJ. Efficacy of amoxicillin with and without decongestant-antihistamine for otitis media with effusion in children. *New England Journal of Medicine* 1987; 316: 432-437.

McAuley 2000

McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet* 2000; 356: 1228-1231.

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.

Moher 1996

Moher D, Fortin P, Jadad AR, Jüni P, Klassen T, Le Lorier J, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 1996; 347: 363-366.

Moher 2000

Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR, et al. What contributions do languages other than English make on the results of meta-analyses? *Journal of Clinical Epidemiology* 2000; 53: 964-972.

Moher 2003

Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health Technology Assessment* 2003; 7: 1-90.

Moher 2007

Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine* 2007; 4: e78.

Moreno 2009a

Moreno SG, Sutton AJ, Turner EH, Abrams KR, Cooper NJ, Palmer TM, et al. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ* 2009; 339: b2981.

Moreno 2009b

Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology* 2009; 9: 2.

Moreno 2012

Moreno SG, Sutton AJ, Thompson JR, Ades AE, Abrams KR, Cooper NJ. A generalized weighting regression-derived meta-analysis estimator robust to small-study effects and heterogeneity. *Statistics in Medicine* 2012; 31: 1407-1417.

Moscati 1994

Moscati R, Jehle D, Ellis D, Fiorello A, Landi M. Positive-outcome bias: comparison of emergency medicine and general medicine literatures. *Academic Emergency Medicine* 1994; 1: 267-271.

Olson 2002

Olson CM, Rennie D, Cook D, Dickersin K, Flanagin A, Hogan JW, et al. Publication bias in editorial decision making. *JAMA* 2002; 287: 2825-2828.

Orwin 1983

Orwin RG. A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics* 1983; 8: 157-159.

Peters 1982

Peters DP, Ceci SJ. Peer review practices of psychology journals: The fate of published articles, submitted again. *The Behavioral and Brain Sciences* 1982; 5: 187-255.

Peters 2006

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006; 295: 676-680.

Peters 2007

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine* 2007; 26: 4544-4562.

Peters 2008

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. The contour enhanced funnel plot: an aid to interpreting funnel asymmetry. *Journal of Clinical Epidemiology* 2008; 61: 991-996.

Pham 2005

Pham B, Klassen TP, Lawson ML, Moher D. Language of publication restrictions in systematic reviews gave different results depending on whether the intervention was conventional or complementary. *Journal of Clinical Epidemiology* 2005; 58: 769-776.

Pittler 2000

Pittler MH, Abbot NC, Harkness EF, Ernst E. Location bias in controlled clinical trials of complementary/alternative therapies. *Journal of Clinical Epidemiology* 2000; 53: 485-489.

Pocock 1987

Pocock S, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *New England Journal of Medicine* 1987; 317: 426-432.

Poole 1999

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

Prayle 2012

Prayle AP, Hurley MN, Smyth AR. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. *BMJ* 2012; 344: d7373.

Ravnskov 1992

Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *BMJ* 1992; 305: 15-19.

Rennie 1991

Rennie D. The Cantekin affair. *JAMA* 1991; 266: 3333-3337.

Rennie 1997

Rennie D. Thyroid Storm. *JAMA* 1997; 277: 1238-1243.

Rising 2008

Rising K, Bacchetti P, Bero L. Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation. *PLoS Medicine* 2008; 5: e217.

Rosenthal 1979

Rosenthal R. The 'file drawer problem' and tolerance for null results. *Psychological Bulletin* 1979; 86: 638-641.

Rücker 2008

Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* 2008; 27: 746-763.

Rücker 2011a

Rücker G, Schwarzer G, Carpenter JR, Binder H, Schumacher M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics* 2011; 12: 122-142.

Rücker 2011b

Rücker G, Carpenter JR, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal* 2011; 53: 351-368.

Sampson 2003

Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, et al. Should meta-analysts search Embase in addition to Medline? *Journal of Clinical Epidemiology* 2003; 56: 943-955.

Scherer 2007

Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews* 2007, Issue 2. MR000005. DOI: 10.1002/14651858.MR000005.pub3.

Schulz 1995

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.

Schwarzer 2007

Schwarzer G, Antes G, Schumacher M. A test for publication bias in meta-analysis with sparse binary data. *Statistics in Medicine* 2007; 26: 721-733.

Simes 1987

Simes RJ. Confronting publication bias: a cohort design for meta-analysis. *Statistics in Medicine* 1987; 6: 11-29.

Siontis 2011

Siontis KC, Evangelou E, Ioannidis JP. Magnitude of effects in clinical trials published in high-impact general medical journals. *International Journal of Epidemiology* 2011; 40: 1280-1291.

Smith 1999

Smith R. What is publication? A continuum. *BMJ* 1999; 318: 142.

Stanley 2008

Stanley TD. Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection. *Oxford Bulletin of Economics and Statistics* 2008; 70: 103-127.

Sterling 1959

Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association* 1959; 54: 30-34.

Sterling 1995

Sterling TD, Rosenbaum WL, Weinkam JJ. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician* 1995; 49: 108-112.

Stern 1997

Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997; 315: 640-645.

Sterne 2000

Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* 2000; 53: 1119-1129.

Sterne 2001

Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology* 2001; 54: 1046-1055.

Sterne 2011

Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011; 343: d4002.

Stuck 1998

Stuck AE, Rubenstein LZ, Wieland D. Bias in meta-analysis detected by a simple, graphical test. Asymmetry detected in funnel plot was probably due to true heterogeneity. *BMJ* 1998; 316: 469-471.

Tang 2000

Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology* 2000; 53: 477-484.

Tannock 1996

Tannock IF. False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. *Journal of the National Cancer Institute* 1996; 88: 206-207.

Taylor 1998

Taylor SJ, Tweedie RL. Practical estimates of the effect of publication bias in meta-analysis. *Australian Epidemiologist* 1998; 5: 14-17.

Teo 1993

Teo KK, Yusuf S, Furberg CD. Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction. An overview of results from randomized controlled trials. *JAMA* 1993; 270: 1589-1595.

Terrin 2003

Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine* 2003; 22: 2113-2126.

Terrin 2005

Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology* 2005; 58: 894-901.

Tetzlaff 2006

Tetzlaff J, Moher D, Pham B, Altman D. Survey of views on including grey literature in systematic reviews. *14th Cochrane Colloquium*; 2006 Oct 23-26; Dublin, Ireland.

Tramèr 1997

Tramèr MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997; 315: 635-640.

Turner 2008

Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine* 2008; 358: 252-260.

Turner 2012

Turner EH, Knoepflmacher D, Shapley L. Publication bias in antipsychotic trials: an analysis of efficacy comparing the published literature to the US Food and Drug Administration database. *PLoS Medicine* 2012; 9: e1001189.

Vedula 2009

Vedula SS, Bero L, Scherer RW, Dickersin K. Outcome reporting in industry-sponsored trials of gabapentin for off-label use. *New England Journal of Medicine* 2009; 361: 1963-1971.

Vedula 2012

Vedula SS, Goldman PS, Rona IJ, Greene TM, Dickersin K. Implementation of a publication strategy in the context of reporting biases. A case study based on new documents from Neurontin litigation. *Trials* 2012; 13: 136.

Vevea 1995

Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* 1995; 60: 419-435.

Vevea 2005

Vevea JL, Woods CM. Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological Methods* 2005; 10: 428-443.

Vickers 1998

Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. *Controlled Clinical Trials* 1998; 19: 159-166.

Villar 1997

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

Weber 1998

Weber EJ, Callaham ML, Wears RL, Barton C, Young G. Unpublished research from a medical specialty meeting: why investigators fail to publish. *JAMA* 1998; 280: 257-259.

Zarin 2005

Zarin DA, Tse T, Ide NC. Trial Registration at ClinicalTrials.gov between May and October 2005. *New England Journal of Medicine* 2005; 353: 2779-2787.

Zarin 2008

Zarin DA, Tse T. Medicine. Moving toward transparency of clinical trials. *Science* 2008; 319: 1340-1342.

Zarin 2011

Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database--update and key issues. *New England Journal of Medicine* 2011; 364: 852-860.

Chapter 11: Completing ‘Summary of findings’ tables and grading the confidence in or quality of the evidence

Authors: Holger J Schünemann, Andrew D Oxman, Julian PT Higgins, Gunn E Vist, Paul Glasziou, Elie Akl and Gordon H Guyatt on behalf of the Cochrane GRADEing Methods Group (formerly Applicability and Recommendations Methods Group) and the Cochrane Statistical Methods Group.

This chapter should be cited as: Schünemann HJ, Oxman AD, Higgins JPT, Vist GE, Glasziou P, Akl E, Guyatt GH on behalf of the Cochrane GRADEing Methods Group and the Cochrane Statistical Methods Group. Chapter 11: Completing ‘Summary of findings’ tables and grading the confidence in or quality of the evidence. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS (editors), *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0 (updated June 2017). Cochrane, 2017. Available from www.training.cochrane.org/handbook.

This extract is from *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0. The previous version of this chapter (5.1.0, 2011) is available online at handbook.cochrane.org.

An earlier version (version 5.0.2, 2008) of the *Handbook* is also published by John Wiley & Sons, Ltd under “The Cochrane Book Series” Imprint, as Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on www.wiley.com.

Key Points

- A ‘Summary of findings’ table provides key information concerning the quality of evidence, the magnitude of effect of the interventions examined, and the sum of available data on all important outcomes for a given comparison.

- The GRADE approach (Grading of Recommendations Assessment, Development and Evaluation), adopted by Cochrane, specifies four levels of quality for a body of evidence (high, moderate, low and very low). Review authors can downgrade the body of evidence depending on the presence of five factors and upgrade the quality of evidence of observational studies depending on three factors.
- Quality ratings according to GRADE are made separately for each outcome and express the confidence or certainty in an effect.

11.1 'Summary of findings' tables

11.1.1 Introduction to 'Summary of findings' tables

'Summary of findings' tables present the main findings of a review in a transparent and simple tabular format. In particular, they provide key information concerning the quality of evidence (i.e. the confidence or certainty in an effect estimate), the magnitude of effect of the interventions examined, and the sum of available data on the main outcomes. Cochrane Reviews should incorporate 'Summary of findings' tables during planning and publication (see Chapter 4, section 4.4.4.3), and should have at least one key 'Summary of findings' table representing the most important comparison. Some reviews may include more than one, for example if the review addresses more than one major comparison, or substantially different populations. In the *Cochrane Database of Systematic Reviews (CDSR),* the principal 'Summary of findings' table of a review will appear at the beginning, before the Background section. Other 'Summary of findings' tables will appear between the Results and Discussion sections.

The planning for the 'Summary of findings' table starts early in the systematic review, with the selection of the outcomes to be included in 1) the review, and 2) the 'Summary of findings' table. Since this is a crucial step, and one that review authors need to address carefully, we will review the issues in selecting outcomes here.

11.1.2 Selecting outcomes for 'Summary of findings' tables

Cochrane Reviews begin by developing a review question and by listing all main outcomes that are important to patients and other decision makers (see Chapter 5, Section 5.4) to ensure production of optimally useful information. Consultation and feedback on the review protocol can enhance this process.

Important outcomes are likely to include widely familiar events such as mortality and major morbidity (such as strokes and myocardial infarction). However, they may also represent frequent minor and rare major side effects, symptoms, quality of life, burdens associated with treatment, and resource issues (costs). Burdens include the demands of adhering to an intervention that patients or caregivers (e.g. family) may dislike, such as having to undergo more frequent tests, or the restrictions on lifestyle that certain interventions require.

Frequently, when formulating questions that include all patient-important outcomes for decision making, review authors will confront the fact that reports of randomized trials have not included all these outcomes. This is particularly true for adverse outcomes. For instance, randomized trials might contribute data on intended effects, and on frequent, relatively minor side effects, but not address the relative risk of rare adverse outcomes such as suicide attempts. Chapter 14 discusses strategies for addressing adverse effects adequately. To obtain data for all important outcomes it may be necessary to examine the results of observational (i.e. non-randomized) studies: see Chapter 13. Cochrane, in collaboration with others, has developed guidance for review authors to support their decision about when to look for and include observational studies (Schünemann 2013).

If a review focuses only on randomized trials, addressing all important outcomes may not be possible within the constraints of the review. Review authors should acknowledge these limitations, and make them transparent to readers.

Review authors who take on the challenge of compiling and summarizing the best evidence for all relevant outcomes may face a number of specific challenges. These include the fact that the analysis of harm may be carried out in (possibly observational) studies where participants differ from those included in the (typically randomized) studies used in the analysis of benefit. Thus, review authors will need to consider how much, if at all, the participants in observational studies differ from those in the randomized trials. This can influence the quality of evidence because of concerns about indirectness (see Chapter 12, Section 12.2). When review authors do not include information on these important outcomes in the review they should say so. Further discussion of these issues appears also in Chapter 13.

11.1.3 General template for ‘Summary of findings’ tables

While there may be good reasons for modifying the format of a ‘Summary of findings’ table for some reviews, a standard format for them has been developed with the aims of ensuring consistency and ease of use across reviews, inclusion of the most important information needed by decision makers, and optimal presentation of this information. Research on alternative formats of ‘Summary of findings’ tables has been conducted to improve understanding of the information they intend to convey.

In addition to describing the population, intervention and the comparison intervention, standard Cochrane ‘Summary of findings’ tables include the following seven elements using one of two fixed formats (see Figure 11.1.a and Figure 11.1.b).

1. A list of all important outcomes, both desirable and undesirable, limited to seven or fewer outcomes.

2. A measure of the typical burden of these outcomes (e.g. illustrative risk, or illustrative mean, on control intervention).

3. Absolute and relative magnitude of effect (if both are appropriate).

4. Numbers of participants and studies addressing these outcomes.

5. A grade of the overall quality of the body of evidence for each outcome (which may vary by outcome).

6. Space for comments.

7. Footnotes or explanations. These are detailed judgments informing the content of the ‘Summary of findings’, such as the overall GRADE assessment. The footnotes should explain the rationale for important aspects of the content. Further details of the issues should be described in the results and discussion section of the review if they cannot be sufficiently described in footnotes.

As a measure of the magnitude of effect, for dichotomous outcomes the table will usually provide both a relative measure (e.g. risk ratio or odds ratio) and measures of absolute risk. For other types of data, either an absolute measure alone (such as difference in means for continuous data) or a relative measure alone (e.g. hazard ratio for time-to-event data) might be provided. Where possible, however, both relative and absolute measures of effect should be provided. Reviews with more than one main comparison should have separate ‘Summary of findings’ tables for each comparison. Figure 11.1.a provides an example of a ‘Summary of findings’ table. Figure 11.1.b provides an alternative format that further facilitates users’ understanding and interpretation of the review’s findings (Johnston 2011, Carrasco-Labra 2016)

A detailed description of the contents of a ‘Summary of findings’ table appears in Section 11.1.6.

Summary of findings:

Compression stockings compared with no compression stockings for people taking long flights

Patients or population: anyone taking a long flight (lasting more than 6 hours)

Settings: international air travel

Intervention: compression stockings[1]

Comparison: without stockings

Outcomes	**Illustrative comparative risks*** (95% CI)		Relative effect (95% CI)	Number of participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk	Corresponding risk				
	Without stockings	**With stockings**				

Symptomatic deep vein thrombosis (DVT)	See comment	See comment	Not estimable	2821 (9 studies)	See comment	0 participants developed symptomatic DVT in these studies.
Symptomless DVT	**Low risk population**[2]		**RR 0.10** (0.04 to 0.26)	2637 (9 studies)	⊕⊕⊕⊕ **High**	
	10 per 1000	**1 per 1000** (0 to 3)				
	High risk population[2]					
	30 per 1000	**3 per 1000** (1 to 8)				
Superficial vein thrombosis	**13 per 1000**	**6 per 1000** (2 to 15)	**RR 0.45** (0.18 to 1.13)	1804 (8 studies)	⊕⊕⊕○ **Moderate**[3]	
Oedema Post-flight values measured on a scale from 0, no oedema, to 10, maximum oedema	The mean oedema score ranged across control groups from **6 to 9**	The mean oedema score in the intervention groups was on average **4.7 lower** (95% CI –4.9 to –4.5)		1246 (6 studies)	⊕⊕○○ **Low**[4]	
Pulmonary embolus	See comment	See comment	Not estimable	2821 (9 studies)	See comment	0 participants developed pulmonary embolus in these studies[5]
Death	See comment	See comment	Not estimable	2821 (9 studies)	See comment	0 participants died in these studies
Adverse effects	See comment	See comment	Not estimable	1182 (4 studies)	See comment	The tolerability of the stockings was described as very good with no complaints of side effects in 4 studies[6]

*The basis for the **assumed risk** is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the intervention group and the **relative effect** of the intervention (and its 95% CI).

CI: Confidence interval; RR: Risk ratio GRADE: GRADE Working Group grades of evidence (see explanations)

[1] All the stockings in the nine studies included in this review were below-knee compression stockings. In four studies the compression strength was 20 mmHg to 30 mmHg at the ankle. It was 10 mmHg to 20 mmHg in the other four studies. Stockings come in different sizes. If a stocking is too tight around the knee it can prevent essential venous return causing the blood to pool around the knee. Compression stockings should be fitted properly. A stocking that is too tight could cut into the skin on a long flight and potentially cause ulceration and increased risk of DVT. Some stockings can be slightly thicker than normal leg covering and can be potentially restrictive with tight foot wear. It is a good idea to wear stockings around the house prior to travel to ensure a good, comfortable fit. Participants put their stockings on two to three hours before the flight in most of the studies. The availability and cost of stockings can vary.

[2] Two studies recruited high risk participants defined as those with previous episodes of DVT, coagulation disorders, severe obesity, limited mobility due to bone or joint problems, neoplastic disease within the previous two years, large varicose veins or, in one of the studies, participants taller than 190 cm and heavier than 90 kg. The incidence for the seven studies that excluded high risk participants was 1.45% and the incidence for the two studies that recruited high-risk participants (with at least one risk factor) was 2.43%. We have used 10 and 30 per 1000 to express different risk strata, respectively.

[3] The confidence interval crosses no difference and does not rule out a small increase.

[4] The measurement of oedema was not validated (indirectness of the outcome) or blinded to the intervention (risk of bias).

[5] If there are very few or no events and the number of participants is large, judgement about the quality of evidence (particularly judgements about imprecision) may be based on the absolute effect. Here the quality rating may be considered 'high' if the outcome was appropriately assessed and the event, in fact, did not occur in 2821 studied participants.

[6] None of the other studies reported adverse effects, apart from four cases of superficial vein thrombosis in varicose veins in the knee region that were compressed by the upper edge of the stocking in one study.

Figure 11.1.a: Example of a 'Summary of findings' table

Summary of findings:

Probiotics compared to no probiotics as an adjunct to antibiotics in children

Patient or population: children given antibiotics

Settings: inpatients and outpatient
Intervention: probiotics

Comparison: no probiotics

Outcomes No of Participants (studies)	Relative effects (95% CI)	Anticipated absolute effects* (95% CI) Without probiotics	 With probiotics	 Difference	Quality of the evidence (GRADE)	What happens
Incidence of Diarrhea: Probiotic dose 5 billion CFU/day Follow-up: 10 days to 3 months Children <5 years 1474 (7 studies)	**RR 0.4**[1] (0.29 to 0.55)	**Children < 5 years** **22.3%**[1]	**8.9%** (6.5 to 12.2)	**13.4% fewer children**[1] (10.1 to 15.8 fewer)	⊕⊕⊕⊖ **moderate**[2] Due to risk of bias	Probably decreases the incidence of diarrhea
Children >5 years 624 (4 studies)	**RR 0.8**[1] (0.53 to 1.21)	**Children > 5 years** **11.2%**[1]	**9%** (5.9 to 13.6)	**2.2% fewer children**[1] (5.3 fewer to 2.4 more)	⊕⊕⊖⊖ **low**[2, 3] Due to risk of bias and imprecision	May decrease the incidence of diarrhea
Adverse events[4] Follow-up: 10 to 44 days 1575 (11 studies)	-	**1.8%**[1]	**2.3%** (0.8 to 3.8)	**0.5% more adverse events**[5] (1 fewer to 2 more)	⊕⊕⊖⊖ **low**[6, 7] Due to risk of bias and inconsistency	There may be little or no difference in adverse events
Duration of diarrhea Follow-up: 10 days to 3 months 897 (5 studies)	-	The mean duration of diarrhea without probiotics was **4 days**	-	**0.6 fewer days** (1.18 to 0.02 fewer days)	⊕⊕⊖⊖ **low**[8, 9] Due to imprecision and inconsistency	May decrease the duration of diarrhea
Stools per day Follow-up: 10 days to 3 months 425 (4 studies)	-	The mean stools per day without probiotics was **2.5 stools per day**	-	**0.3 fewer stools per day** (0.6 to 0 fewer)	⊕⊕⊖⊖ **low**[10, 11] Due to imprecision and inconsistency	There may be little or no difference in stools per day

*The basis for the **risk in the control group** (e.g. the median control group risk across studies) is provided in footnotes. The **risk in the intervention group** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI). **CI:** Confidence interval; **RR:** risk ratio;

EXPLANATIONS

[1] Control group risk estimates come from pooled estimates of control groups. Relative effect based on available case analysis

[2] High risk of bias due to high loss to follow-up.

[3] Imprecision due to few events and confidence intervals include appreciable benefit or harm.
[4] Side effects: rash, nausea, flatulence, vomiting, increased phlegm, chest pain, constipation, taste disturbance, and low appetite

[5] Risks were calculated from pooled risk differences.

[6] High risk of bias. Only 11 of 16 trials reported on adverse events, suggesting a selective reporting bias
[7] Serious inconsistency. Numerous probiotic agents and doses were evaluated amongst a relatively small number of trials, limiting our ability to draw conclusions on the safety of the many probiotics agents and doses administered

[8] Serious unexplained inconsistency (large heterogeneity I^2=79%, P value [P = 0.04], point estimates and confidence intervals vary considerably)
[9] Serious imprecision. The upper bound of 0.02 fewer days of diarrhea is not considered patient important
[10] Serious unexplained inconsistency (large heterogeneity I^2=78%, P value [P = 0.05], point estimates and confidence intervals vary considerably)

[11] Serious imprecision. The 95% confidence interval includes no effect and lower bound of 0.60 stools per day is of questionable patient importance

Figure 11.1.b: Example of alternative 'Summary of findings' table

11.1.4 Producing 'Summary of findings' tables

The GRADE working group's software, GRADEpro or GRADEpro GDT (www.gradepro.org), is available to assist review authors in the preparation of 'Summary of findings' tables. GRADEpro is able to retrieve data from RevMan and to combine this with user-entered control group risks to produce the relative effects and absolute risks associated with interventions. In addition, it leads the user through the process of a GRADE assessment (see context-specific help file in GRADEpro), and produces a table that can be readily imported into RevMan as a 'Summary of findings' table in the standard or alternative format. It can also be used as a standalone interactive 'Summary of findings' table.

11.1.5 Statistical considerations in 'Summary of findings' tables

Here we describe how absolute and relative measures of effect for dichotomous outcomes are obtained. Risk ratios, odds ratios and risk differences are different ways of comparing two groups with dichotomous outcome data (see Chapter 9, Section 9.2.2). Furthermore, there are two distinct risk ratios, depending on which event (e.g. 'yes' or 'no') is the focus of the analysis (see Chapter 9, Section 9.2.2.5). In the presence of a non-zero intervention effect, if there is variation in control group risks across studies, then it is impossible for more than one of these measures to be truly the same in every study. It has long been the expectation in epidemiology that relative measures of effect are more consistent than absolute measures of effect from one scenario to another. There is empirical evidence to support this supposition (Engels 2000, Deeks 2001). For this reason, meta-analyses should generally use either a risk ratio or an odds ratio as a measure of effect (see Chapter 9, Section 9.4.4.4). Correspondingly, a single estimate of relative effect is likely to be a more appropriate summary than a single estimate of absolute effect. If a relative effect is indeed consistent across studies, then different control group risks will have different implications for absolute benefit. For instance, if the risk ratio is consistently 0.75, then the experimental intervention would reduce a control group risk of 80% to 60% in the intervention group (an absolute reduction of 20 percentage points), but would also reduce a control group risk of 20% to 15% in the intervention group (an absolute reduction of 5 percentage points).

'Summary of findings' tables are built around the assumption of a consistent relative effect. It is then important to consider the implications of this effect for different control group risks. For any assumed control group risk, it is possible to estimate a corresponding intervention group risk from the meta-analytic risk ratio or odds ratio. Note that the numbers provided in the 'Corresponding risk' column are specific to the 'Assumed risks' in the adjacent column.

For the meta-analytic risk ratio, RR, and assumed control risk, ACR, the corresponding intervention risk is obtained as:

$$\text{Corresponding intervention risk per 1000} = 1000 \times \text{ACR} \times \text{RR}$$

As an example, in Figure 11.1.a, the meta-analytic risk ratio is for symptomless deep vein thrombosis (DVT) is RR = 0.10 (95% CI 0.04 to 0.26). Assuming a control risk of ACR = 10 per 1000 = 0.01, we obtain:

$$\text{Corresponding intervention risk per 1000} = 1000 \times 0.01 \times 0.10 = 1$$

For the meta-analytic odds ratio, OR, and assumed control risk, ACR, the corresponding intervention risk is obtained as:

$$\text{Corresponding intervention risk, per 1000} = 1000 \times \left(\frac{\text{OR} \times \text{ACR}}{1 - \text{ACR} + (\text{OR} \times \text{ACR})} \right)$$

Upper and lower confidence limits for the corresponding intervention risk are obtained by replacing RR or OR by their upper and lower confidence limits, respectively (e.g. replacing 0.10 with 0.04, then with 0.26, in the example above). Such confidence intervals do not incorporate uncertainty in the assumed control risks.

When dealing with risk ratios, it is critical that the same definition of 'event' is used as was used for the meta-analysis. For example, if the meta-analysis focused on 'death' as the event, then assumed and corresponding risks in the 'Summary of findings' table must also refer to 'death'.

In (rare) circumstances in which there is clear rationale to assume a consistent risk difference in the meta-analysis, in principle it is possible to present this for relevant 'assumed risks' and their corresponding risks, and to present the corresponding (different) relative effects for each assumed risk.

11.1.6 Detailed contents of a 'Summary of findings' table

11.1.6.1 Table title and header

The title of each 'Summary of findings' table should specify the healthcare question, framed in terms of the population and making it clear exactly what comparison of interventions is being made. In Figure 11.1.a, the population is people taking long aeroplane flights, the intervention is compression stockings, and the control is no compression stockings.

The first rows of each 'Summary of findings' table should provide the following 'header' information:

Patients or population: This further clarifies the population (and possibly the sub-populations) of interest and ideally the magnitude of risk of the most crucial adverse outcome at which an intervention is directed. For instance: people on a long-haul flight may be at different risks for DVT; those using selective serotonin reuptake inhibitors (SSRIs) might be at different risk for side effects; while those with atrial fibrillation may be at low (< 1%), moderate (1% to 4%) or high (> 4%) yearly risk of stroke.

Setting: This should specify any specific characteristics of the settings of the healthcare question that might limit the applicability of the summary of findings to other settings; e.g. primary care in Europe and North America.

Intervention: The experimental intervention.

Comparison: The control (comparison) intervention (which might be no specific intervention).

11.1.6.2 Outcomes

The rows of a 'Summary of findings' table should include all desirable and undesirable outcomes (listed in order of importance) that are essential for decision-making, up to a maximum of *seven* outcomes. If there is an excessive number of outcomes in the review, authors will need to omit the less important outcomes. Details of scales and time frames should be provided. Authors should aim to decide which outcomes are important for the 'Summary of findings' table during protocol development and before they undertake the review. Note that authors should list these outcomes in the table *whether data are available or not.* However, review authors should be alert to the possibility that the importance of an outcome (e.g. a serious adverse effect) may only become known after the protocol was written or the analysis was carried out, and should take appropriate actions to include these in the 'Summary of findings' table.

Serious adverse events should be included, but it might be possible to combine minor adverse events, and describe this in a footnote (note that it is not appropriate to add events together unless they are known to be independent). Multiple time points will be a particular problem. In general, to keep the table simple, only outcomes critical to decision making should be presented at multiple time points. The remainder should be presented at a common time point.

Continuous outcome measures can be shown in the 'Summary of findings' table; review authors should endeavour to make these interpretable to the target audience (see Chapter 12, Section 12.6). This requires that the units are clear and readily interpretable, for example, days of pain, or frequency of headache. However, many measurement instruments are not readily interpretable by non-specialist clinicians or patients, for example, points on a Beck Depression Inventory or quality of life score. For these, a more interpretable presentation might involve converting a continuous to a dichotomous outcome, such as > 50% improvement (see Chapter 12, Section 12.6).

11.1.6.3 Illustrative comparative risks 1: Assumed risk (with control intervention)

Authors should provide up to three typical risks for participants receiving the control intervention. It is recommended that these be presented in the form of the number of people experiencing the event per 1000 people (natural frequency). A suitable alternative greater than 1000 may be used for rare events, or 100 may be used for more frequent events. Assumed control intervention risks could be based on assessments of typical risks in different patient groups. Ideally, risks would reflect groups that clinicians can easily identify on the basis of their presenting features. A footnote should specify the source or rationale for each control group risk, including the time period to which it corresponds where appropriate. In Figure 11.1.a, clinicians can easily differentiate individuals with risk factors for deep venous thrombosis from those without. If there is known to be little variation in baseline risk then review authors may use the median control group risk across studies. If typical risks are not known, for a high and low risk population the second highest and second lowest control group risks in the included studies can be chosen.

11.1.6.4 Illustrative comparative risks 2: Corresponding risk (with experimental intervention)

For dichotomous outcomes, a corresponding absolute risk should be provided for each assumed risk in the preceding column, along with a confidence interval. This absolute risk with (experimental) intervention will usually be derived from the meta-analysis result presented in the relative effect column (see Section 11.1.6.5). Formulae are provided in Section 11.1.5. Review authors should present the absolute effect in the same format as assumed risks with control intervention (see Section 11.1.6.3), for example, as the number of people experiencing the event per 1000 people.

For continuous outcomes, a difference in means or standardized difference in means should be presented with its confidence interval. These will typically be obtained directly from a meta-analysis. Explanatory text should be used to clarify the meaning, as in Figure 11.1.a.

11.1.6.5 Relative effect (95% CI)

The relative effect will typically be a risk ratio or odds ratio (or occasionally a hazard ratio) with its accompanying 95% confidence interval, obtained from a meta-analysis performed on the basis of the same effect measure. Risk ratios and odds ratios are similar when the control intervention risks are low and effects are small, but differ considerably as these increase. The meta-analysis may involve an assumption of either fixed or random effects, depending on what the review authors consider appropriate.

11.1.6.6 Number of participants (studies)

This column should include the number of participants assessed in the included studies for each outcome and the corresponding number of studies that contributed these participants.

11.1.6.7 Quality of the evidence (GRADE)

Authors must comment on the quality of the body of evidence (also known as confidence in the effect estimates or certainty in the evidence). Authors should use the specific evidence grading system developed by the GRADE Working Group (GRADE Working Group 2004, Guyatt 2008, Guyatt 2011a), which is described in detail in Section 11.2. The GRADE approach categorizes the quality of a body of evidence as ‘high’, ‘moderate’, ‘low’, or ‘very low’ by outcome. This is a result of judgement, but the judgement process operates within a transparent structure as described in Section 11.2. As an example, the quality would be ‘high’ if the summary were of several randomized trials with low risk of bias, but the rating of quality becomes lower if there are concerns about risk of bias, imprecision, inconsistency, indirectness, or publication bias. Judgements other than of ‘high’ quality should be made transparent using footnotes or the ‘Comments’ column in the ‘Summary of findings’ table (see Figure 11.1.a).

11.1.6.8 Comments

The aim of the ‘Comments’ field is to provide additional comments to help interpret the information or data identified in the row. For example, this may be on the validity of the outcome measure or the presence of variables that are associated with the magnitude of effect. Important caveats about the results should be flagged up here. Not all rows will need comments, it is best to leave a blank if there is nothing warranting a comment.

11.2 Assessing the quality of a body of evidence

11.2.1The GRADE approach

The Grades of Recommendation, Assessment, Development and Evaluation Working Group (GRADE Working Group) has developed a system for grading the quality of evidence (GRADE Working Group 2004, Schünemann 2006, Guyatt 2008, Guyatt 2011a). Over 90 organizations including the World Health Organization (WHO), the American College of Physicians, the American College of Chest Physicians (ACCP), the American Endocrine Society, the American Thoracic Society (ATS), the Canadian Agency for Drugs and Technology in Health (CADTH), BMJ Clinical Evidence, the National Institutes of Health and Care Excellence (NICE) in the UK, and UpToDate® have adopted the GRADE system in its original format or with minor modifications (Schünemann 2006, Guyatt 2008, Guyatt 2011a). The BMJ encourages authors of clinical guidelines to use the GRADE system (http://www.bmj.com/about-bmj/resources-authors/article-types/clinical-management-guidelines).

Authors must evaluate the quality of evidence for important outcomes reported in C74
Cochrane Reviews, and must justify and document their assessments. Cochrane has C75
adopted the GRADE approach for evaluating the quality of evidence.

For systematic reviews, the GRADE approach defines the quality of a body of evidence as the extent to which one can be confident that an estimate of effect or association is close to the quantity of specific interest. Assessing the quality of a body of evidence involves consideration of within- and across-study risk of bias (methodological quality), directness of evidence, inconsistency (or heterogeneity), imprecision of the effect estimates and risk of publication bias, as described in Section 11.2.2. The GRADE system entails an assessment of the quality of a body of evidence for each individual outcome. Judgments about the domains that determine the quality of evidence should be described in the results or discussion section or as part of the 'Summary of findings' table.

The GRADE approach specifies four levels of quality (Table 11.2.a). The highest quality rating is for randomized trial evidence when there are no concerns in any of the GRADE factors listed in Table 11.2.b. Review authors can, however, downgrade randomized trial evidence to moderate, low, or even very low quality evidence, depending on the presence of the five factors in Table 11.2.b. Usually, quality rating will fall by one level for each factor, up to a maximum of three levels for all factors. If there are very severe problems for any one factor (e.g. when assessing risk of bias, all studies were unconcealed, unblinded, and lost over 50% of their patients to follow-up), randomized trial evidence may fall by two levels due to that factor alone.

Review authors will generally grade evidence from sound observational studies as low quality. If, however, such studies yield large effects and there is no obvious bias explaining those effects, review authors may rate the evidence as moderate or – if the effect is large enough – even high quality (Table 11.2.c). The very low quality level includes, but is not limited to, studies with critical problems and unsystematic clinical observations (e.g. case series or case reports).

Table 11.2.a: Levels of quality of a body of evidence in the GRADE approach

Underlying methodology	Quality rating
Randomized trials; or double-upgraded observational studies	High
Downgraded randomized trials; or upgraded observational studies	Moderate
Double-downgraded randomized trials; or observational studies	Low
Triple-downgraded randomized trials; or downgraded observational studies	Very low

Table 11.2.b: Factors that may decrease the quality level of a body of evidence

1. Risk of bias

2. Indirectness of evidence (indirect population, intervention, control, outcomes)

3. Unexplained heterogeneity or inconsistency of results (including problems with subgroup analyses)

4. Imprecision of results (wide confidence intervals)

5. High probability of publication bias

Table 11.2.c: Factors that may increase the quality level of a body of evidence

1. Large magnitude of effect

2. All plausible confounding would reduce a demonstrated effect or suggest a spurious effect when results show no effect

3. Dose-response gradient

11.2.2 Factors that decrease the quality level of a body of evidence

We now describe in more detail the five reasons for downgrading the quality of a body of evidence for a specific outcome. In each case, if a reason is found for downgrading the evidence, it should be classified as ‘serious’ (downgrading the quality rating by one level) or ‘very serious’ (downgrading the quality grade by two levels).

1. **Risk of bias or limitations in the detailed design and implementation**: Our confidence in an estimate of effect decreases if studies suffer from major limitations that are likely to result in a biased assessment of the intervention effect. For randomized trials, these methodological limitations include failure to generate a random sequence, lack of allocation sequence concealment, lack of blinding (particularly with subjective outcomes that are highly susceptible to biased assessment), a large loss to follow-up or selective reporting of outcomes. Chapter 8 provides a detailed discussion of study-level assessments of risk of bias in the context of a Cochrane Review, and proposes an approach to assessing the risk of bias for an outcome across studies as ‘low risk of bias’, ‘unclear risk of bias’ and ‘high risk of bias’ (Chapter 8, Section 8.7). These assessments should feed directly into this factor. In particular, ‘low risk of bias’ would indicate ‘no limitation’; ‘unclear risk of bias’ would indicate either ‘no limitation’ or ‘serious limitation’; and ‘high risk of bias’ would indicate either ‘serious limitation’ or ‘very serious limitation’. Authors must use their judgement to decide between alternative categories, depending on the likely magnitude of the potential biases.

   Every study addressing a particular outcome will differ, to some degree, in the risk of bias. Review authors must make an overall judgement on whether the quality of evidence for an outcome warrants downgrading on the basis of study limitations. The assessment of study limitations should apply to the studies contributing to the results in the ‘Summary of findings’ table, rather than to all studies that could potentially be included in the analysis. We have argued in Chapter 8 (Section 8.8.3) that the primary analysis should be restricted to studies at low (or low and unclear) risk of bias.

Table 11.2.d presents the judgements that must be made in going from assessments of the risk of bias to judgements about study limitations for each outcome included in a 'Summary of findings' table. A rating of high quality evidence can be achieved only when most evidence comes from studies that met the criteria for low risk of bias. For example, of the 22 studies addressing the impact of beta-blockers on mortality in patients with heart failure, most probably or certainly used concealed allocation, all blinded at least some key groups and follow-up of randomized patients was almost complete (Brophy 2001). The quality of evidence might be downgraded by one level when most of the evidence comes from individual studies either with a crucial limitation for one criterion, or with some limitations for multiple criteria. For example, we cannot be confident that, in patients with *Plasmodium falciparum* malaria, amodiaquine and sulfadoxine-pyrimethamine together reduce treatment failures compared with sulfadoxine-pyrimethamine alone, because the apparent advantage of sulfadoxine-pyrimethamine was sensitive to assumptions regarding the event rate in those lost to follow-up (> 20% loss to follow-up in two of three studies (McIntosh 2005)). An example of very serious limitations, warranting downgrading by two levels, is provided by evidence on surgery versus conservative treatment in the management of patients with lumbar disc prolapse (Gibson 2007). We are uncertain of the benefit of surgery in reducing symptoms after one year or longer, because the one study included in the analysis had inadequate concealment of the allocation sequence and the outcome was assessed using a crude rating by the surgeon without blinding.

2. **Indirectness of evidence**: Two types of indirectness are relevant. Firstly, a review comparing the effectiveness of alternative interventions (say A and B) may find that randomized trials are available, but they have compared A with placebo and B with placebo. Thus, the evidence is restricted to indirect comparisons between A and B. Secondly, a review may find randomized trials that meet eligibility criteria but that address a restricted version of the main review question in terms of population, intervention, comparator or outcomes. For example, suppose that in a review addressing an intervention for secondary prevention of coronary heart disease, the majority of identified studies happened to be in people who also had diabetes. Then the evidence may be regarded as indirect in relation to the broader question of interest because the population is restricted to people with diabetes. The opposite scenario can equally apply: a review addressing the effect of a preventative strategy for coronary heart disease in people with diabetes may consider studies in people without diabetes to provide relevant, albeit indirect, evidence. This would be particularly likely if investigators had conducted few if any randomized trials in the target population (i.e. people with diabetes). Other sources of indirectness may arise from interventions studied (e.g. if in all included studies a technical intervention was implemented by expert, highly trained specialists in specialist centres, then evidence on the effects of the intervention outside these centres may be indirect), comparators used (e.g. if the control groups received an intervention that is less effective than standard treatment in most settings) and outcomes assessed (e.g. indirectness due to surrogate outcomes when data on patient-important outcomes are not available, or when investigators sought data on quality of life but only symptoms were reported). Review authors should make judgements transparent when they believe downgrading

is justified, based on differences in anticipated effects in the group of primary interest. Review authors may be aided and increase transparency of their judgments about indirectness if they use Table 11.2.e (available in the GRADEpro software (Schünemann 2013)).

3. **Unexplained heterogeneity or inconsistency of results**: When studies yield widely differing estimates of effect (heterogeneity or variability in results), investigators should look for robust explanations for that heterogeneity. For instance, drugs may have larger relative effects in sicker populations or when given in larger doses. A detailed discussion of heterogeneity and its investigation is provided in Chapter 9 (Sections 9.5 and 9.6). If an important modifier exists, with strong evidence that important outcomes are different in different subgroups (which would ideally be pre-specified), then a separate 'Summary of findings' table may be considered for a separate population. For instance, a separate 'Summary of findings' table would be used for carotid endarterectomy in symptomatic patients with high grade stenosis in which the intervention is, in the hands of the right surgeons, beneficial (Cina 2000), and another (if review authors considered it worthwhile) for asymptomatic patients with moderate grade stenosis in which surgery is not beneficial (Chambers 2005). When heterogeneity exists and affects the interpretation of results, but authors fail to identify a plausible explanation, the quality of the evidence decreases.

4. **Imprecision of results**: When studies include few participants and few events, and thus have wide confidence intervals, authors can lower their rating of the quality of the evidence. The confidence intervals included in the 'Summary of findings' table will provide readers with information that allows them to make, to some extent, their own rating of precision. Authors can use the optimal information size (OIS) to make judgments about imprecision. The OIS is calculated on the basis of the number of participants required for an adequately powered individual study. If the 95% confidence interval excludes a risk ratio (RR) of 1.0, and the total number of events or patients exceeds the OIS criterion, precision is adequate. If the 95% CI includes appreciable benefit or harm (an RR of under 0.75 or over 1.25 is often suggested as a rough guide), downgrading for imprecision may be appropriate even if OIS criteria are met (Guyatt 2011b).

5. **High probability of publication bias**: The quality of evidence level may be downgraded if investigators fail to report studies (typically those that show no effect: publication bias) or outcomes (typically those that may be harmful or for which no effect was observed: selective outcome reporting bias) on the basis of results. Selective reporting of outcomes is assessed at the study level as part of the assessment of risk of bias (see Chapter 8, Section 8.14), so for the studies contributing to the outcome in the 'Summary of findings' table this is addressed by factor 1 above (limitations in the design and implementation). If a large number of studies included in the review do not contribute to an outcome, or if there is evidence of publication bias, the quality of the evidence may be downgraded. Chapter 10 provides a detailed discussion of reporting biases, including publication bias, and how it may be addressed in a Cochrane Review. A prototypical situation that may elicit suspicion of publication bias is when published evidence includes a number of small studies, all of which are industry funded

> (Bhandari 2004). For example, 14 studies of flavonoids in patients with haemorrhoids have shown apparent large benefits, but enrolled a total of only 1432 patients (that is, each study enrolled relatively few patients (Alonso-Coello 2006)). The heavy involvement of sponsors in most of these studies raises questions of whether unpublished studies that suggest no benefit exist.

A particular body of evidence can suffer from problems associated with more than one of the five factors listed above, and the greater the problems, the lower the quality of evidence rating that should result. One could imagine a situation in which randomized trials were available, but all or virtually all of these limitations would be present, and in serious form. A very low quality of evidence rating would result.

Table 11.2.d: Further guidelines for factor 1 (of 5) in a GRADE assessment: Going from assessments of risk of bias to judgements about study limitations for main outcomes

Risk of bias	Across studies	Interpretation	Considerations	GRADE assessment of study limitations
Low risk of bias	Most information is from studies at low risk of bias.	Plausible bias unlikely to seriously alter the results.	No apparent limitations.	No serious limitations, do not downgrade.
Unclear risk of bias	Most information is from studies at low or unclear risk of bias.	Plausible bias that raises some doubt about the results.	Potential limitations are unlikely to lower confidence in the estimate of effect.	No serious limitations, do not downgrade.
			Potential limitations are likely to lower confidence in the estimate of effect.	Serious limitations, downgrade one level.
High risk of bias	The proportion of information from studies at high risk of bias is sufficient to affect the interpretation of results.	Plausible bias that seriously weakens confidence in the results.	Crucial limitation for one criterion, or some limitations for multiple criteria, sufficient to lower confidence in the estimate of effect.	Serious limitations, downgrade one level.
			Crucial limitation for one or more criteria sufficient to substantially lower confidence in the estimate of effect.	Very serious limitations, downgrade two levels.

Table 11.2.e: Judgements about indirectness by outcome

Outcome: …					
Domain (original question asked)	Description (evidence found and included, including evidence from other studies) – consider the domains of study design and study execution, inconsistency, imprecision and publication bias	Judgment - Is the evidence sufficiently direct?			
Population:		Yes □	Probably yes □	Probably no □	No □
Intervention:		Yes □	Probably yes □	Probably no □	No □
Comparator:		Yes □	Probably yes □	Probably no □	No □
Direct comparison:		Yes □	Probably yes □	Probably no □	No □

Outcome:		Yes □ Probably yes □ Probably no □ No □
Final judgment about indirectness across domains:		□ **No indirectness** □**Serious indirectness** □**Very serious indirectness**

11.2.3 Factors that increase the quality level of a body of evidence

Although observational studies and downgraded randomized trials will generally yield a low rating for quality of evidence, there will be unusual circumstances in which authors could ‘upgrade’ such evidence to moderate or even high quality (Table 11.2.c).

1. On rare occasions when methodologically well-done observational studies yield large, consistent and precise estimates of the magnitude of an intervention effect, one may be particularly confident in the results. A large effect (e.g. RR > 2 or RR < 0.5) in the absence of plausible confounders, or a very large effect (e.g. RR > 5 or RR < 0.2) in studies with no major threats to validity, might qualify for this. In these situations, while the observational studies may possibly have provided an overestimate of the true effect, the weak study design may not explain all of the apparent observed benefit. Thus, despite reservations based on the observational study design, authors are confident that the effect exists. The magnitude of the effect in these studies may move the assigned quality of evidence from low to moderate (if the effect is large in the absence of other methodological limitations). For example, a meta-analysis of observational studies showed that bicycle helmets reduce the risk of head injuries in cyclists by a large margin (odds ratio (OR) 0.31, 95% CI 0.26 to 0.37 (Thompson 2000)). This large effect, in the absence of obvious bias that could create the association, suggests a rating of moderate-quality evidence.

2. On occasion, all plausible biases from observational or randomized studies may be working to underestimate an apparent intervention effect. For example, if only sicker patients receive an experimental intervention or exposure, yet they still fare better, it is likely that the actual intervention or exposure effect is larger than the data suggest. For instance, a rigorous systematic review of observational studies including a total of 38 million patients demonstrated higher death rates in private for-profit versus private not-for-profit hospitals (Devereaux 2004). One possible bias relates to different disease severity in patients in the two hospital types. It is likely, however, that patients in the not-for-profit hospitals were sicker than those in the for-profit hospitals. Thus, to the extent that residual confounding existed, it would bias results against the not-for-profit hospitals. The second likely bias was the possibility that higher numbers of patients with excellent private insurance coverage could lead to a hospital having more resources and a spill-over effect that would benefit those without such coverage. Since for-profit hospitals are likely to admit a larger proportion of such well-insured patients than not-for-profit hospitals, the bias is once again against the not-for-profit hospitals. Since the plausible biases would all diminish the demonstrated intervention effect, one might consider the evidence from these observational studies as moderate rather than low quality. A parallel situation exists when observational studies have failed to demonstrate an association, but all plausible biases would have increased an intervention effect. This situation will usually arise in the exploration of apparent harmful effects. For example, because the hypoglycaemic drug phenformin causes lactic acidosis, the related agent metformin is under suspicion for the same toxicity. Nevertheless, very large observational studies have failed to demonstrate an association (Salpeter 2007). Given the likelihood that clinicians would be more alert to

lactic acidosis in the presence of the agent and over-report its occurrence, one might consider this moderate, or even high quality, evidence refuting a causal relationship between typical therapeutic doses of metformin and lactic acidosis.

3. The presence of a dose-response gradient may also increase our confidence in the findings of observational studies and thereby enhance the assigned quality of evidence. For example, our confidence in the result of observational studies that show an increased risk of bleeding in patients who have supratherapeutic anticoagulation levels is increased by the observation that there is a dose-response gradient between higher levels of the international normalized ratio (INR) and the increased risk of bleeding (Levine 2004).

11.3 Describing the assessment of the quality of a body of evidence using the GRADE framework

Authors should describe the rational for grading the quality of evidence in the results section that refers to the 'Summary of findings' table or the assessment of the quality of a body of evidence. Table 11.3.a provides a framework and examples for how authors can justify their judgements about the quality of evidence.

Table 11.3.a: Framework for describing the quality of evidence and justifying downgrading or upgrading

Criteria for assessing quality of evidence by outcome	Results section	Examples of reasons for lowering or increasing the quality of evidence
Risk of bias	Describe the risk of bias based on the criteria used in the 'Risk of bias' table.	Of ten randomized trials, five did not blind patients and caretakers.
Inconsistency	Describe the degree of inconsistency by outcome using one or more indicators (e.g. I^2 and P value), confidence interval overlap, difference in point estimate, between-study variance.	The proportion of the variability in effect estimates that is due to true heterogeneity rather than chance is not important ($I^2 = 0\%$).
Indirectness	Describe if the majority of studies address the PICO – were they similar to the question posed?	The included studies were restricted to patients with advanced cancer.

Imprecision	Describe the number of events, and width of the confidence intervals.	The confidence intervals for the effect on mortality are compatible with both an appreciable benefit and appreciable harm.
Publication bias	Describe the possible degree of publication bias.	1) The funnel plot of 14 randomized trials indicated that there were several small studies that showed a small positive effect, but small studies that showed no effect or harm may have been unpublished. 2) There are only three small positive studies, it appears that studies showing no effect or harm have not been published. There also is for-profit interest in the intervention.
Large effects (upgrading)	Describe the magnitude of the effect and the widths of the associate confidence intervals.	The RR is 0.3 (95% CI 0.2 to 0.4) with a sufficient number of events.
Dose response (upgrading)	The studies show a clear relation with increases in the outcome of an outcome (e.g. lung cancer) with higher exposure levels.	The dose-response relation shows a relative risk increase of 10% in never smokers, 15% in smokers of 10 pack years, and 20% in smokers of 15 pack years.
Opposing plausible residual bias and confounding (upgrading)	Describe which opposing biases and confounders may have not been considered.	The estimate of effect is not controlled for the following possible confounders: smoking, degree of education, but the distribution of these factors in the studies is likely to lead to an underestimate of the true effect.

11.4 Methodological standards for the conduct of Cochrane Intervention Reviews

No.	Status	Name	Standard	Rationale & Elaboration	Handbook Sections
C74	Mandatory	Assessing the quality of the body of evidence	Use the five GRADE considerations (risk of bias, consistency of effect, imprecision, indirectness and publication bias) to assess the quality of the body of evidence for each outcome, and to draw conclusions about the quality of evidence within the text of the review.	GRADE is the most widely used approach for summarizing confidence in effects of the interventions by outcome across studies. It is preferable to use the GRADEpro tool (as described in the help system of the software). This should help to ensure that author teams are accessing the same information to inform their judgments. Ideally, two people working independently should assess the quality of the body of evidence and reach a consensus view on any downgrading decisions. The five GRADE considerations should be addressed irrespective of whether the review includes a 'Summary of findings' table. It is helpful to draw on this information in the Discussion, in the conclusions and	11.2.1

				to convey the certainty in the evidence in the abstract and Plain Language Summary.	
C75	Mandatory	Justifying assessments of the quality of the body of evidence	Justify and document all assessments of the quality of the body of evidence (for example downgrading or upgrading if using the GRADE tool, GRADEpro).	By adopting a structured approach, transparency is ensured in showing how interpretations have been formulated and the result is more informative to the reader.	11.2.1

11.5 Chapter information

Acknowledgements: Professor Penny Hawe contributed to the text on adverse effects. Jon Deeks provided helpful contributions on an earlier version of this chapter. For details of previous authors and editors of the *Handbook*, please refer to Section 1.4. For details of the Cochrane GRADEing group, see Box 11.5.a; for the Cochrane Statistical Methods Group, see Chapter 9 (Box 9.8.a).

Conflict of interest: Holger Schünemann, Andrew Oxman, Gunn Vist, Paul Glasziou, Elie Akl and Gordon Guyatt are members of the GRADE Working Group from which many of the ideas in this chapter have arisen.

Box 11.5.a: The Cochrane GRADEing

We anticipate continued evolution of the methodologies described in this chapter. The main arenas in which relevant discussions will take place are the Cochrane GRADEing Methods Group and the GRADE Working Group. Both discussion groups welcome new participants with an eagerness to learn more and to contribute to further developments in rating quality of evidence, and in framing issues in the application of Cochrane Reviews.

The Cochrane GRADEing methods group is comprised of individuals with interest and expertise in the interpretation, applicability and transferability of the results of systematic reviews to individuals and groups. The Cochrane GRADEing Methods Group's objective is to explore the process of going from evidence to healthcare recommendations. The ultimate goals are to make this process as rigorous and transparent as possible.

Specific areas currently considered important include:

- evaluating the quality of evidence (www.gradeworkinggroup.org);
- variation of effect with baseline risk;
- prediction of benefit from the patient's expected event rate or severity;
- consideration of how the strength of evidence and the magnitude and precision of the effects bear on the implications; and
- consideration of how people's values bear on the implications when weighing benefits and harms based on individual clinical features.

11.6 References

Alonso-Coello 2006

Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, Mills E, Heels-Ansdell D, Johanson JF, et al. Meta-analysis of flavonoids for the treatment of haemorrhoids. *British Journal of Surgery* 2006; 93: 909-920.

Bhandari 2004

Bhandari M, Busse JW, Jackowski D, Montori VM, Schünemann H, Sprague S, et al. Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. *Canadian Medical Association Journal* 2004; 170: 477-480.

Brophy 2001

Brophy JM, Joseph L, Rouleau JL. Beta-blockers in congestive heart failure. A Bayesian meta-analysis. *Annals of Internal Medicine* 2001; 134: 550-560.

Carrasco-Labra 2016

Carrasco-Labra A, Brignardello-Petersen R, Santesso N, Neumann I, Mustafa RA, Mbuagbaw L, et al. Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format. *Journal of Clinical Epidemiology* 2016; 74: 7-18.

Chambers 2005

Chambers BR, Donnan GA. Carotid endarterectomy for asymptomatic carotid stenosis. *Cochrane Database of Systematic Reviews* 2005, Issue 4. CD001923. DOI: 10.1002/14651858.CD001923.pub2.

Cina 2000

Cina CS, Clase CM, Haynes RB. Carotid endarterectomy for symptomatic carotid stenosis. *Cochrane Database of Systematic Reviews* 2000, Issue 2. CD001081. DOI: 10.1002/14651858.CD001081.

Deeks 2001

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

Devereaux 2004

Devereaux PJ, Choi PT, El-Dika S, Bhandari M, Montori VM, Schünemann HJ, et al. An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. *Journal of Clinical Epidemiology* 2004; 57: 1232-1236.

Engels 2000

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

Gibson 2007

Gibson JN, Waddell G. Surgical interventions for lumbar disc prolapse. *Cochrane Database of Systematic Reviews* 2007, Issue 2. CD001350. DOI: 10.1002/14651858.CD001350.pub4.

GRADE Working Group 2004

GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004; 328: 1490-1494.

Guyatt 2008

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336: 924-926.

Guyatt 2011a

Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology* 2011; 64: 380-382.

Guyatt 2011b

Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of Clinical Epidemiology* 2011; 64: 1283-1293.

Johnston 2011

Johnston BC, Goldenberg JZ, Vandvik PO, Sun X, Guyatt GH. Probiotics for the prevention of pediatric antibiotic-associated diarrhea. *Cochrane Database of Systematic Reviews* 2011: CD004827.

Levine 2004

Levine MN, Raskob G, Beyth RJ, Kearon C, Schulman S. Hemorrhagic complications of anticoagulant treatment: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 2004; 126: 287S-310S.

McIntosh 2005

McIntosh HM, Jones KL. Chloroquine or amodiaquine combined with sulfadoxine-pyrimethamine for treating uncomplicated malaria. *Cochrane Database of Systematic Reviews* 2005, Issue 4. CD000386. DOI: 10.1002/14651858.CD000386.pub2.

Salpeter 2007

Salpeter S, Greyber E, Pasternak G, Salpeter E. Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. *Cochrane Database of Systematic Reviews* 2007, Issue 4. CD002967. DOI: 10.1002/14651858.CD002967.pub2.

Schünemann 2006

Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *American Journal of Respiratory and Critical Care Medicine* 2006; 174: 605-614.

Schünemann 2013

Schünemann HJ, Tugwell P, Reeves BC, Akl EA, Santesso N, Spencer FA, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; 4: 49-62.

Thompson 2000

Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. *Cochrane Database of Systematic Reviews* 2000, Issue 2. Art No: CD001855.

Chapter 12: Interpreting results and drawing conclusions

Authors: Holger J Schünemann, Andrew D Oxman, Gunn E Vist, Julian PT Higgins, Jonathan J Deeks, Paul Glasziou, Elie Akl and Gordon H Guyatt on behalf of the Cochrane Applicability and Recommendations Methods Group.

This chapter should be cited as: Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou P, Akl E, Guyatt GH on behalf of the Cochrane Applicability and Recommendations Methods Group. Chapter 12: Interpreting results and drawing conclusions. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS (editors), *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0 (updated June 2017). Cochrane, 2017. Available from www.training.cochrane.org/handbook.

This extract is from *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0. The previous version of this chapter (5.1.0, 2011) is available online at handbook.cochrane.org.

An earlier version (version 5.0.2, 2008) of the *Handbook* is also published by John Wiley & Sons, Ltd under "The Cochrane Book Series" Imprint, as Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on www.wiley.com.

Key Points

- Methods for computing, presenting and interpreting relative and absolute effects for dichotomous outcome data, including the number needed to treat (NNT), are described in this chapter.
- For continuous outcome measures, review authors can present pooled results for studies using the same units, the standardized mean difference and effect sizes when studies use the same construct but different scales, and odds ratios after transformation of the standardized mean differences.

- Review authors should not describe results as 'not statistically significant' or 'non-significant' or rely unduly on thresholds for P values, but report the confidence interval together with the exact P value.

- Review authors should not make recommendations, but they can – after describing the quality of evidence and the balance of benefits and harms – highlight different actions that might be consistent with particular patterns of values and preferences and other factors that determine decisions, such as cost.

12.1 Introduction

The purpose of Cochrane Reviews is to facilitate healthcare decision-making by patients and the general public, clinicians, administrators, and policy makers. A clear statement of findings, a considered discussion and a clear presentation of the authors' conclusions are important parts of the review. In particular, the following issues can help people make better informed decisions and increase the usability of Cochrane Reviews:

- information on all important outcomes, including adverse outcomes;
- the quality of the evidence for each of these outcomes, as it applies to specific populations, and specific interventions; and
- clarification of the manner in which particular values and preferences may bear on the balance of benefits, harms, burden and costs of the intervention.

A 'Summary of findings' table, described in Chapter 11 (Section 11.5), provides key pieces of information in a quick and accessible format. It is highly desired that review authors include a 'Summary of findings' table in Cochrane Reviews alongside a sufficient description of the studies and meta-analyses to support its contents. This description includes the mandatory rating of the quality of evidence, i.e. the confidence in the estimates of the effects, for each outcome. The 'Discussion' section of the text should provide complementary considerations. Authors should use five subheadings to ensure they cover suitable material in the 'Discussion' section and that they place the review in an appropriate context. These are 'Summary of main results' (benefits and harms); 'Overall completeness and applicability of evidence'; 'Quality of the evidence'; 'Potential biases in the review process'; and 'Agreements and disagreements with other studies or reviews'. 'Authors' conclusions' are divided into 'Implications for practice' and 'Implications for research'. The assessment of the quality of evidence facilitates a structured description of the implications for practice and research that will be described in this chapter.

Because Cochrane Reviews have an international audience, the discussion and authors' conclusions should, so far as possible, assume a broad international perspective and provide guidance for how the results could be applied in different settings, rather than being restricted to specific national or local circumstances. Cultural differences and economic differences may both play an important role in determining the best course of action. Furthermore, individuals within societies have widely varying values and preferences regarding health states, and use of societal resources to achieve particular health states. Even in the face of the same values and preferences, people may interpret the same research evidence differently. For all these reasons, different people will often make different decisions based on the same evidence.

Thus, authors should avoid specific recommendations that inevitably depend on assumptions about available resources, values and preferences and other factors such as feasibility and implementability. The purpose of the review should be to present information and aid interpretation rather than to offer recommendations. The discussion and conclusions should help people understand the implications of the evidence in relation to practical decisions, and to apply the results to their specific situation. Authors can, however, aid decision-making by laying out different scenarios that describe particular value structures.

This chapter provides a more detailed consideration of issues around applicability and around interpretation of numerical results, and provide suggestions for presenting authors' conclusions.

12.2 Issues in applicability

12.2.1 The role of the review author

“A leap of faith is always required when applying any study findings to the population at large” or to a specific person. “In making that jump, one must always strike a balance between making justifiable broad generalizations and being too conservative in one's conclusions” (Friedman 1985). In addition to issues about risk of bias and other factors determining the quality of evidence, this leap of faith is related to how well the identified body of evidence matches the research question posed in terms of participants, interventions, comparisons and outcome (PICO). No individual can be perfectly matched to the population included in research studies. Whenever a decision is made, there will be differences between the study population and the person or population to whom the evidence is applied; sometimes these differences are slight, sometimes large.

The terms applicability 'generalizability', 'external validity' and 'transferability' are related, sometimes used interchangeably, and have in common that they lack a clear and consistent definition in the classic epidemiological literature (Schünemann 2013a). However, all of the terms relate to one overarching theme: whether or not available research evidence can be directly utilized to answer the health and healthcare question at hand, ideally supported by a judgement about the degree of confidence in this utilization (Schünemann 2013a). GRADE's quality confidence – or certainty – criteria include a judgment about 'indirectness' to describe all of these aspects, including the concept of direct versus indirect comparisons of different interventions (GRADE Working Group 2004).

In order to address the extent to which a review is relevant for the purpose to which it is being put, there are certain things the review author must do, and certain things the user of the review must do to assess the degree of indirectness. Cochrane and the GRADE Working Group suggest using a very structured framework to address indirectness. This chapter discusses what the review author can do to help the user. Cochrane review authors must be extremely clear about the population, interventions and outcomes that they intend to address. Chapter 11 (Table 11.2.e) emphasizes a crucial step that has not traditionally been part of Cochrane Reviews: the specification of all patient-important outcomes relevant to the intervention strategies under comparison.

With respect to participant and intervention factors, review authors need to make a priori hypotheses about possible effect modifiers, and then examine those hypotheses. If they find apparent subgroup effects, ultimately, they must decide whether or not these effects are credible (Oxman 2002, Sun 2012). Differences between subgroups, particularly those that correspond to differences between studies, need to be interpreted cautiously. Some chance variation between subgroups is inevitable, so unless there is good reason to believe that there is an interaction, authors should not assume that the subgroup effect exists. If, despite due caution, review authors judge subgroup effects in terms of relative effect estimates as credible, they should conduct separate meta-analyses for the relevant subgroups, and produce separate ‘Summary of findings’ tables for those subgroups.

The user of the review will be challenged with ‘individualization’ of the findings. For example, even if relative effects are similar across subgroups, absolute effects will differ according to baseline risk. Review authors can help provide this information by identifying particular groups of people with varying baseline risks in the ‘Summary of findings’ tables, as discussed in Chapter 11 (Section 11.5.5). Users can then identify the patients before them as belonging to a particular risk group, and assess their probable magnitude of benefit or harm accordingly. A description of the identifying prognostic or baseline risk factors in a brief scenario (e.g. age or gender) will help users of a review further.

Another decision that users must make is whether the patients before them are so different from those included in the studies that they cannot use the results of the systematic review and meta-analysis at all. Rather than rigidly applying the inclusion and exclusion criteria of studies, review authors can point out that it is better to ask whether there are compelling reasons why the evidence should not be applied to a particular patient (Guyatt 1994). Authors can sometimes help clinical decision makers by identifying important variation where divergence might limit the applicability of results (Schünemann 2006, Schünemann 2013a), including: biologic and cultural variation, and variation in adherence to an intervention.

In addressing these issues, authors cannot be aware of, or address, the myriad of differences in circumstances around the world. They can, however, address differences of known importance to many people and, importantly, they should avoid assuming that other people's circumstances are the same as their own in discussing the results and drawing conclusions.

12.2.2 Biologic variation

Issues of biologic variation that authors should consider include divergence in pathophysiology (e.g. biologic differences between women and men that are likely to affect responsiveness to an intervention) and divergence in a causative agent (e.g. for infectious diseases such as malaria).

12.2.3 Variation in context and culture

Some interventions, particularly non-pharmacological interventions, may work in some contexts but not in others; the situation has been described as ‘program by context interaction’ (Hawe 2004). Context factors might pertain to the host organization in which an intervention is offered, such as the expertise, experience and morale of the staff expected to carry out the intervention, the competing priorities for the staff’s attention, the local resources such as service and facilities made available to the programme and the status or importance given to the programme by the host organization. Broader context issues might include aspects of the system within which the host organization operates, such as the fee or payment structure for healthcare providers. Context factors may also pertain to the characteristics of the target group or population services (such aspects include the cultural and linguistic diversity, socioeconomic position, rural/urban setting), which may mean that a particular style of care or relationship evolves between service providers and consumers that may or may not match the values and technology of the programme. For many years these aspects have been acknowledged (but not clearly specified) when decision makers have argued that results of evidence reviews from other countries do not apply in their own country.

Whilst some programmes/interventions have been transferred from one context to another and benefits have been observed, others have not (Resnicow 1993, Lumley 2004). Authors should take caution when making generalizations from one context to another. Authors should report on the presence (or otherwise) of context-related information in intervention studies, where this information is available (Hawe 2004).

12.2.4 Variation in adherence

Variation in the adherence of the recipients and providers of care can limit the applicability of results. Predictable differences in adherence can be due to divergence in economic conditions or attitudes that make some forms of care not accessible or not feasible in some settings, such as in low- or middle-income countries (Dans 2007). It should not be assumed that high levels of adherence in closely monitored randomized studies will translate into similar levels of adherence in normal practice.

12.2.5 Variation in values and preferences

Decisions between healthcare management strategies and options involve trade-offs between different benefits and different downsides. The right choice may differ for people with different values and preferences (i.e. the importance people place on the outcomes and interventions), and it is up to the clinician to ensure that decisions are consistent with patients' values and preferences. Section 12.6 describes how the review author can help this process.

12.3 Interpreting results of statistical analyses

12.3.1 Confidence intervals

Results for both individual studies and meta-analyses are reported with a point estimate together with an associated confidence interval. For example, "The odds ratio was 0.75 with a 95% confidence interval of 0.70 to 0.80". The point estimate (0.75) is the best guess of the magnitude and direction of the experimental intervention's effect compared with the control intervention. The confidence interval describes the uncertainty inherent in this estimate, and describes a range of values within which it is reasonably certain that the true effect actually lies. If the confidence interval is relatively narrow (e.g. 0.70 to 0.80), the effect size is known precisely. If the interval is wider (e.g. 0.60 to 0.93) the uncertainty is greater, although there may still be enough precision to make decisions about the utility of the intervention. Intervals that are very wide (e.g. 0.50 to 1.10) indicate that there is little knowledge about the effect and this imprecision affects our certainty in the evidence, and further information is needed.

A 95% confidence interval is often interpreted as indicating a range within which it is possible to be 95% certain that the true effect lies. This statement is a loose interpretation, but is useful as a rough guide. Strictly speaking, the correct interpretation of a confidence interval is based on the hypothetical notion of considering the results that would be obtained if the study were repeated many times. If a study were repeated infinitely, and on each occasion a 95% confidence interval calculated, then 95% of these intervals would contain the true effect.

The width of the confidence interval for an individual study depends to a large extent on the sample size. Larger studies tend to give more precise estimates of effects (and hence have narrower confidence intervals) than smaller studies. For continuous outcomes, precision depends also on the variability in the outcome measurements (the standard deviation of measurements across individuals); for dichotomous outcomes it depends on the risk of the event, and for time-to-event outcomes it depends on the number of events observed. All these quantities are used in computation of the standard errors of effect estimates from which the confidence interval is derived.

The width of a confidence interval for a meta-analysis depends on the precision of the individual study estimates and on the number of studies combined. In addition, for random-effects models, precision will decrease with increasing heterogeneity, and confidence intervals will widen correspondingly (see Chapter 9, Section 9.5.4). As more studies are added to a meta-analysis, the width of the confidence interval usually decreases. However, if the additional studies increase the heterogeneity in the meta-analysis and a random-effects model is used, it is possible that the width of the confidence interval will increase.

Confidence intervals and point estimates have different interpretations in fixed-effect and random-effects models. While the fixed-effect estimate and its confidence interval address the question 'what is the best (single) estimate of the effect?', the random-effects estimate assumes there to be a distribution of effects, and the estimate and its confidence interval address the question 'what is the best estimate of the average effect?'

A confidence interval may be reported for any level of confidence (although they are most commonly reported for 95%, and sometimes 90% or 99%). For example, the odds ratio of 0.80 could be reported with an 80% confidence interval of 0.73 to 0.88; a 90% interval of 0.72 to 0.89; and a 95% interval of 0.70 to 0.92. As the confidence level increases, the confidence interval widens.

There is logical correspondence between the confidence interval and the P value (see Section 12.3.2). The 95% confidence interval for an effect will exclude the null value (such as an odds ratio of 1.0 or a risk difference of 0) if – and only if – the test of significance yields a P value of less than 0.05. If the P value is exactly 0.05, then either the upper or lower limit of the 95% confidence interval will be at the null value. Similarly, the 99% confidence interval will exclude the null if – and only if – the test of significance yields a P value of less than 0.01.

Together, the point estimate and confidence interval provide information to assess the clinical usefulness of the intervention. For example, suppose that an intervention that reduces the risk of an event is being evaluated, and it is decided that it will be useful only if it reduces the risk of an event from 30% by at least five percentage points to 25% (these values will depend on the specific clinical scenario and outcome). If the meta-analysis yields an effect estimate of a reduction of 10 percentage points with a tight 95% confidence interval, say, from 7% to 13%, it would be possible to conclude that the intervention was useful, since both the point estimate and the entire range of the interval exceed the criterion of a reduction of 5% for clinical usefulness. However, if the meta-analysis reported the same risk reduction of 10% but with a wider interval, say, from 2% to 18%, although the conclusion would still be that the best estimate of the effect of the intervention is that it is useful, confidence in the result would be reduced, as the possibility that the effect could be between 2% and 5% would not be excluded. If the confidence interval were wider still, and included the null value of a difference of 0%, the possibility that the intervention has any effect whatsoever would not be excluded, and conclusions would need to be even more sceptical.

Confidence intervals with different levels of confidence can demonstrate that there is differential evidence for different degrees of benefit or harm. For example, it might be possible to report the same analysis results: 1) with 95% confidence that the intervention does not cause harm; 2) with 90% confidence that it has some effect; and 3) with 80%

confidence that it has a patient-important benefit. These elements may suggest both usefulness of the intervention and the need for additional research.

Review authors may use the same general approach to conclude that an intervention is *not* useful. Continuing with the above example where the criterion for a minimal patient-important difference is a 5% risk difference, an effect estimate of 2% with a confidence interval of 1% to 4% suggests that the intervention is not useful.

12.3.2 P values and statistical significance

A P value is the probability of obtaining the observed effect (or larger) under a ‘null hypothesis’, which in the context of Cochrane Reviews is either an assumption of ‘no effect of the intervention’ or ‘no differences in the effect of intervention between studies’ (no heterogeneity). Thus, a P value that is very small indicates that the observed effect is very unlikely to have arisen purely by chance, and therefore provides evidence against the null hypothesis. It has been common practice to interpret a P value by examining whether it is smaller than particular threshold values. In particular, P values less than 0.05 are often reported as ‘statistically significant’, and interpreted as being small enough to justify rejection of the null hypothesis. However, the 0.05 threshold is an arbitrary one that became commonly used in medical and psychological research largely because P values were determined by comparing the test statistic against tabulations of specific percentage points of statistical distributions. RevMan, like other statistical packages, reports precise P values. If review authors decide to present a P value with the results of a meta-analysis, they should report a precise P value, together with the 95% confidence interval.

In RevMan, two P values are provided. One relates to the summary effect in a meta-analysis and is from a Z test of the null hypothesis that there is no effect (or no effect on average in a random-effects meta-analysis). The other relates to heterogeneity between studies and is from a Chi^2 test of the null hypothesis that there is no heterogeneity (see Chapter 9, Section 9.5.2). C72

For tests of a summary effect, the computation of P involves both the effect estimate and the sample size (or, more strictly, the precision of the effect estimate). As sample size increases, the range of plausible effects that could occur by chance is reduced. Correspondingly, the statistical significance of an effect of a particular magnitude will be greater (the P value will be smaller) in a larger study than in a smaller study.

P values are commonly misinterpreted in two ways. First, a moderate or large P value (e.g. greater than 0.05) may be misinterpreted as evidence that the intervention has *no effect*. There is an important difference between this statement and the correct interpretation, which is that there is *no strong evidence* that the intervention has an effect. To avoid such a misinterpretation, review authors should always examine the effect estimate and its 95% confidence interval, together with the P value. In small studies or small meta-analyses it is common for the range of effects contained in the confidence interval to include both no intervention effect and a substantial effect. Review authors must not interpret results as ‘not statistically significant’ or ‘non-significant’.

The second misinterpretation is to assume that a result with a small P value for the summary effect estimate implies that an intervention has an important benefit. Such a misinterpretation is more likely to occur in large studies, such as meta-analyses that accumulate data over dozens of studies and thousands of participants. The P value addresses the question of whether the intervention effect is precisely nil; it does not examine whether the effect is of a magnitude of importance to potential recipients of the intervention. In a large study, a small P value may represent the detection of a trivial effect. Again, inspection of the point estimate and confidence interval helps to correct interpretations (see Section 12.3.1).

12.4 Interpreting results from dichotomous outcomes (including numbers needed to treat)

12.4.1 Relative and absolute risk reductions

Clinicians may be more inclined to prescribe an intervention that reduces the relative risk of death by 25% than one that reduces the risk of death by one percentage point, although both presentations of the evidence may relate to the same benefit (i.e. a reduction in risk from 4% to 3%). The former refers to the *relative* reduction in risk and the latter to the *absolute* reduction in risk. As described in Chapter 9 (Section 9.2.2), there are several measures for comparing dichotomous outcomes in two groups. Meta-analyses are usually undertaken using risk ratios (RR), odds ratios (OR) or risk differences (RD), but there are several alternative ways of expressing results.

Relative risk reduction (RRR) is a convenient way of re-expressing a risk ratio as a percentage reduction:

$$\text{RRR} = 100\% \times (1 - \text{RR}).$$

For example, a risk ratio of 0.75 translates to a relative risk reduction of 25%, as in the example above.

The risk difference is often referred to as the **absolute risk reduction** (ARR), and may be presented as a percentage (e.g. 1%), as a decimal (e.g. 0.01), or as counts, (e.g. 10 out of 1000). A simple transformation of the risk difference known as the number needed to treat (NNT) is a common alternative way of presenting the same information. NNTs are discussed in Section 12.4.2, and different choices for presenting absolute effects are considered in Section 12.4.3. Computations for obtaining these numbers from the results of individual studies and of meta-analyses are also described.

12.4.2 More about the number needed to treat (NNT)

The **number needed to treat** (NNT) is defined as the expected number of people who need to receive the experimental rather than the comparator intervention for one additional person either to incur or to avoid an event in a given time frame. Thus, for example, an NNT of 10 can be interpreted as 'it is expected that one additional (or one fewer) person will incur an event for every 10 participants receiving the experimental intervention rather than control over a given time frame'. It is important to be clear that:

- since the NNT is derived from the risk difference, it is still a *comparative* measure of effect (experimental versus a certain control) and not a general property of a single intervention; and
- the NNT gives an 'expected value'. For example, NNT = 10 does not imply that one additional event *will* occur in each and every group of ten people.

NNTs can be computed for both beneficial and detrimental events, and for interventions that cause both improvements and deteriorations in outcomes. In all instances NNTs are expressed as positive whole numbers, all decimals being rounded up. Some authors use the term 'number needed to harm' (NNH) when an intervention leads to a deterioration rather than improvement in outcome. However, this phrase is unpleasant, misleading and inaccurate (most notably, it can easily be read to imply the number of people who will experience a harmful outcome if given the intervention), and it is strongly recommended that 'number needed to harm' and 'NNH' are avoided. The preferred alternative is to use phrases such as 'number needed to treat for an additional beneficial outcome' (NNTB) and 'number needed to treat for an additional harmful outcome' (NNTH) to indicate direction of effect.

As NNTs refer to events, their interpretation needs to be worded carefully when the binary outcome is a dichotomization of a scale-based outcome. For example, if the outcome is pain measured on a 'none, mild, moderate or severe' scale it may have been dichotomized as 'none or mild' versus 'moderate or severe'. It would be inappropriate for an NNT from these data to be referred to as an 'NNT for pain'. It is an 'NNT for moderate or severe pain'.

12.4.3 Expressing absolute risk reductions

Users of reviews are liable to be influenced by the choice of statistical presentations of the evidence. Hoffrage et al. suggest that physicians' inferences about statistical outcomes are more appropriate when they deal with 'natural frequencies' – whole numbers of people, both treated and untreated (e.g. the intervention results in a drop from 20 out of 1000 to 10 out of 1000 women having breast cancer) – than when effects are presented as percentages (e.g. 1% absolute reduction in breast cancer risk) (Hoffrage 2000). Probabilities may be more difficult to understand than frequencies, particularly when events are rare. While standardization may be important in improving the presentation of research evidence (and participation in healthcare decisions), current evidence suggests that the presentation of natural frequencies for expressing differences in absolute risk is best understood by consumers of healthcare information (Akl 2011a). This evidence provides the rationale for presenting absolute risks in ‘Summary of findings’ tables as numbers of people with events per 1000 people receiving the intervention.

Risk ratios and relative risk reductions remain crucial because relative effect tends to be substantially more stable across risk groups than does absolute benefit. Review authors can use their own data to study this consistency (Cates 1999, Smeeth 1999). Risk differences are least likely to be consistent across baseline event rates; thus, they are rarely appropriate for computing numbers needed to treat in systematic reviews. If a relative effect measure (OR or RR) is chosen for meta-analysis, then a control group risk needs to be specified as part of the calculation of an ARR or NNT. It is crucial to express absolute benefit for each clinically identifiable risk group, clarifying the time period to which this applies. Studies in patients with differing severity of disease, or studies with different lengths of follow-up will almost certainly have different control group risks. In these cases, different control group risks lead to different ARRs and NNTs (except when the intervention has no effect). A recommended approach is to re-express an odds ratio or a risk ratio as a variety of NNTs across a range of assumed control risks (ACRs) (McQuay 1997, Smeeth 1999, Sackett 2000). Review authors should bear these considerations in mind not only when constructing their ‘Summary of findings’ table, but also in the text of their review.

For example, a review of oral anticoagulants to prevent stroke presented information to users by describing absolute benefits for various baseline risks (Aguilar 2005). They presented their principal findings as “The inherent risk of stroke should be considered in the decision to use oral anticoagulants in atrial fibrillation patients, selecting those who stand to benefit most for this therapy” (Aguilar 2005). Among high-risk atrial fibrillation patients with prior stroke or transient ischaemic attack who have stroke rates of about 12% (120 per 1000) per year, warfarin prevents about 70 strokes yearly per 1000 patients, whereas for low-risk atrial fibrillation patients (with a stroke rate of about 2% per year or 20 per 1000), warfarin prevents only 12 strokes. This presentation helps users to understand the important impact that typical baseline risks have on the absolute benefit that they can expect.

12.4.4 Computations

Direct computation of an absolute risk reduction (ARR) or a number needed to treat (NNT) depends on the summary statistic (odds ratio, risk ratio or risk differences) available from the study or meta-analysis. When expressing results of meta-analyses, authors should use, in the computations, whatever statistic they determined to be the most appropriate summary for pooling (see Chapter 9, Section 9.4.4.4). Here calculations to obtain ARR as a reduction in the number of participants per 1000 are presented. For example, a risk difference of –0.133 corresponds to 133 *fewer* participants with the event per 1000.

ARRs and NNTs should not be computed from the aggregated total numbers of participants and events across the studies. This approach ignores the randomization within studies, and may produce seriously misleading results if there is unbalanced randomization in any of the studies.

When computing NNTs, the values obtained are by convention always rounded up to the next whole number.

12.4.4.1 Computing NNT from a risk difference (RD)

NNTs can be calculated for single studies as follows. Note that this approach, although applicable, should only very rarely be used for the results of a meta-analysis of risk differences, because meta-analyses should usually be undertaken using a relative measure of effect (RR or OR).

A NNT may be computed from a risk difference as

$$\text{NNT} = \frac{1}{\text{absolute value of risk difference}} = \frac{1}{|\text{RD}|},$$

where the vertical bars (‘absolute value of’) in the denominator indicate that any minus sign should be ignored. It is convention to round the NNT up to the nearest whole number. For example, if the risk difference is –0.12 the NNT is 9; if the risk difference is –0.22 the NNT is 5. Cochrane review authors should be specific about whether the NNT is one that provides benefit (improvement) or harm by denoting the NNT as NNTB or NNTH, respectively.

12.4.4.2 Computing absolute risk reduction or NNT from a risk ratio (RR)

To aid interpretation, review authors may wish to compute an absolute risk reduction or NNT from the results of a meta-analysis of risk ratios. In order to do this, an assumed control risk (ACR) is required. It will usually be appropriate to do this for a range of different ACRs. The computation proceeds as follows:

$$\text{number fewer per 1000} = 1000 \times \text{ACR} \times (1 - \text{RR}),$$

$$\text{NNT} = \left|\frac{1}{\text{ACR} \times (1 - \text{RR})}\right|$$

As an example, suppose the risk ratio is RR = 0.92, and an assumed control risk of ACR = 0.3 (300 per 1000) is assumed. Then the effect on risk is 24 fewer per 1000:

$$\text{number fewer per 1000} = 1000 \times 0.3 \times (1 - 0.92) = 24$$

The NNT is 42:

$$\text{NNT} = \left|\frac{1}{0.3 \times (1 - 0.92)}\right| = \left|\frac{1}{0.3 \times 0.08}\right| = 41.67$$

12.4.4.3 Computing absolute risk reduction or NNT from an odds ratio (OR)

Review authors may wish to compute an absolute risk reduction or NNT from the results of a meta-analysis of odds ratios. In order to do this, an assumed control risk (ACR) is required. It will usually be appropriate to do this for a range of different ACRs. The computation proceeds as follows:

$$\text{number fewer per 1000} = 1000 \times \left(\text{ACR} - \frac{\text{OR} \times \text{ACR}}{1 - \text{ACR+OR} \times \text{ACR}} \right)$$

$$\text{NNT} = \frac{1}{\left| \text{ACR} - \frac{\text{OR} \times \text{ACR}}{1 - \text{ACR+OR} \times \text{ACR}} \right|}$$

As an example, suppose the odds ratio is OR = 0.73, and a control risk of ACR = 0.3 is assumed. Then the effect on risk is 62 fewer per 1000:

$$\begin{aligned} \text{number fewer per 1000} &= 1000 \times \left(0.3 - \frac{0.73 \times 0.3}{1 - 0.3 + 0.73 \times 0.3} \right) \\ &= 1000 \times \left(0.3 - \frac{0.219}{1 - 0.3 + 0.219} \right) = 1000 \times (0.3 - 0.238) = 61.7 \end{aligned}$$

The NNT is 17:

$$\text{NNT} = \frac{1}{\left| \left(0.3 - \frac{0.73 \times 0.3}{1 - 0.3 + 0.73 \times 0.3} \right) \right|} = \frac{1}{\left| 0.3 - \frac{0.219}{1 - 0.3 + 0.219} \right|} = \frac{1}{|0.3 - 0.238|} = 16.2$$

12.4.4.4 Computing risk ratio from an odds ratio (OR)

Because risk ratios are easier to interpret than odds ratios, but odds ratios have favourable mathematical properties, a review author may decide to undertake a meta-analysis based on odds ratios, but to express the result as a summary risk ratio (or relative risk reduction). This requires an assumed control risk (ACR). Then

$$\text{RR} = \frac{\text{OR}}{1 - \text{ACR} \times (1 - \text{OR})}$$

It will often be reasonable to perform this transformation using the median control group risk from the studies in the meta-analysis.

12.4.4.5 Computing confidence limits

Confidence limits for ARRs and NNTs may be calculated by applying the above formulae to the upper and lower confidence limits for the summary statistic (RD, RR or OR) (Altman 1998). Note that this confidence interval does not incorporate uncertainty around the control group risk (CGR).

In the case of what conventionally are considered non-statistically significant results (e.g. the 95% confidence interval of OR or RR includes the value 1), one of the confidence limits will indicate benefit and the other harm. Thus, appropriate use of the words 'fewer' and 'more' is required for each limit when presenting results in terms of events. For NNTs, the two confidence limits should be labelled as NNTB and NNTH to indicate the direction of effect in each case. The confidence interval for the NNT will include a 'discontinuity': within the interval there will be an infinitely large NNTB, which will switch to an infinitely large NNTH.

12.5 Interpreting results from continuous outcomes (including standardized mean differences)

12.5.1 Meta-analyses with continuous outcomes

When outcomes are continuous, review authors have a number of options for presentation of pooled results. If all studies have used the same units, a meta-analysis may generate a pooled estimate in those units, as a difference in mean response (see, for instance, the row summarizing results for oedema in Chapter 11, Figure 11.5.a). The units of such outcomes may be difficult to interpret, particularly when they relate to rating scales. 'Summary of findings' tables should include the minimum and maximum of the scale of measurement, and the direction (again, see the Oedema column of Chapter 11, Figure 11.5.a). Knowledge of the smallest change in instrument score that patients perceive is important – the minimal important difference – and can greatly facilitate the interpretation of results. Knowing the minimal important difference allows authors and users to place results in context, and authors should state the minimal important difference – if known – in the Comments column of their 'Summary of findings' table.

When studies have used different instruments to measure the same construct, a standardized mean difference (SMD) may be used in meta-analysis for combining continuous data (see Chapter 9, Section 9.2.3.2). For clinical interpretation, such an analysis may be less helpful than dichotomizing responses and presenting proportions of patients who benefit. Methods are available for creating dichotomous data out of reported means and standard deviations, but require assumptions that may not be met (Suissa 1991, Walter 2001).

The SMD expresses the intervention effect in standard units, rather than the original units of measurement. The SMD is the difference in mean effects in the experimental and control groups divided by the pooled standard deviation of participants' outcomes (see Chapter 9, Section 9.2.3.2). The value of a SMD thus depends on both the size of the effect (the difference between means) and the standard deviation of the outcomes (the inherent variability among participants).

Without guidance, clinicians and patients may have little idea how to interpret results presented as SMDs. There are several possibilities for re-expressing such results in more helpful ways, as follows.

12.5.2 Re-expressing SMDs using rules of thumb for effect sizes

Rules of thumb exist for interpreting SMDs (or 'effect sizes'), which have arisen mainly from researchers in the social sciences. One example is as follows: 0.2 represents a small effect, 0.5 a moderate effect, and 0.8 a large effect (Cohen 1988). Variations exist (e.g. < 0.40 = small, 0.40 to 0.70 = moderate, > 0.70 = large). Review authors might consider including a rule of thumb in the Comments column of a 'Summary of findings' table. However, some methodologists believe that such interpretations are problematic because the importance to *patients* of a finding is context-dependent and not amenable to generic statements.

12.5.3 Re-expressing SMDs by transformation to odds ratio

A transformation of a SMD to a (log) odds ratio is available, based on the assumption that an underlying continuous variable has a logistic distribution with equal standard deviation in the two intervention groups (Furukawa 1999, Chinn 2000). The assumption is unlikely to hold exactly and the results must be regarded as an approximation. The log odds ratio is estimated as

$$\text{lnOR} = \frac{\pi}{\sqrt{3}}\text{SMD}$$

(or approximately 1.81 × SMD) The resulting odds ratio can then be combined with an assumed control group risk to obtain an absolute risk reduction as in Section 12.4.4.3. These control group risks refer to proportions of people who have improved by some (unspecified) amount in the continuous outcome ('responders'). Table 12.5.a shows some illustrative results from this method. These NNTs may be converted to people per thousand by using the formula 1000/NNT.

Table 12.5.a: NNTs equivalent to specific SMDs for various given 'proportions improved' in the control group

Control group proportion improved	10%	20%	30%	40%	50%	60%	70%	80%	90%

SMD = 0.1	57	33	26	23	23	24	28	37	66
SMD = 0.2	27	16	13	12	12	13	15	20	36
SMD = 0.5	9	6	5	5	5	6	7	10	18
SMD = 0.8	5	4	3	3	4	4	5	7	14
SMD = 1.0	4	3	3	3	3	4	5	7	13

12.5.4 Re-expressing SMDs using a familiar instrument

The final possibility for interpreting the SMD is to express it in the units of one or more of the specific measurement instruments. Multiplying a SMD by a typical among-person standard deviation for a particular scale yields an estimate of the difference in mean outcome scores (experimental versus control) on that scale. The standard deviation could be obtained as the pooled standard deviation of baseline scores in one of the studies. To reflect among-person variation better in practice, it may be preferable to use a standard deviation from a representative observational study. The pooled effect is thus re-expressed in the original units of that particular instrument and the clinical relevance and impact of the intervention effect can be interpreted. However, authors should be aware that such back-transformation of effect sizes can be misleading if it is applied to individual studies rather than for a summary measure of effect (Scholten 1999). Consider two studies that *did* use the same instrument and observed the same effect, but observed different among-participant variability (perhaps due to different eligibility criteria). Then back-transformations using the different standard deviations from these studies would yield different sizes of effect for *the same scale* and *the same effect*.

12.6 Drawing conclusions

12.6.1 Conclusions sections of a Cochrane Review

Authors' conclusions from a Cochrane Review are divided into 'Implications for practice' and 'Implications for research'. In deciding what these implications are, it is useful to consider four factors: the quality of evidence, the balance of benefits and harms, values and preferences, and resource utilization (Eddy 1990). Considering these factors involves judgements and effort that go beyond the work of most review authors.

12.6.2 Implications for practice

Drawing conclusions about the practical usefulness of an intervention entails making trade-offs, either implicitly or explicitly, between the estimated benefits, harms and the estimated costs. Making such trade-offs, and thus making specific recommendations for an action, goes beyond a systematic review and requires additional information and informed judgements that are typically the domain of clinical practice guideline developers. Authors of Cochrane Reviews must avoid making recommendations for practice.

If authors feel compelled to lay out actions that clinicians and patients could take, they should – after describing the quality of evidence and the balance of benefits and harms – highlight different actions that might be consistent with particular patterns of values and preferences. Other factors that might influence a decision should also be highlighted, including any known factors that would be expected to modify the effects of the intervention, the baseline risk or status of the patient, costs and who bears those costs, and the availability of resources. Authors should ensure they consider all patient-important outcomes, including those for which limited data may be available. In the context of public health reviews the focus may be on population-important outcomes as the target may be an entire (non-diseased) population. This process implies a high level of explicitness about judgements about values or preferences attached to different outcomes. The highest level of explicitness would involve a formal economic analysis with sensitivity analysis involving different assumptions about values and preferences; this is beyond the scope of most Cochrane Reviews (although they might well be used for such analyses) (Mugford 1989, Mugford 1991); this is discussed in Chapter 15.

A review on the use of anticoagulation in cancer patients to increase survival provides an example for laying out clinical implications for situations where there are important trade-offs between desirable and undesirable effects of the intervention (Akl 2011b): “The decision for a patient with cancer to start heparin therapy for survival benefit should balance the benefits and downsides and integrate the patient’s values and preferences ((Haynes 2002, Schünemann 2013b). Patients with a high preference for a potential survival prolongation, limited aversion to potential bleeding, and who do not consider heparin (both UFH or LMWH) therapy a burden may opt to use heparin, while those with aversion to bleeding may not.”

12.6.3 Implications for research

Review conclusions should help people make well-informed decisions about future healthcare research. The ‘Implications for research’ section should be structured to comment on the need for further research, and the nature of the further research that would be most desirable, including population, intervention, comparison, outcome (PICO), and type of study. One format that has been proposed for reporting research recommendations (‘EPICOT’) follows (Brown 2006).

- E (Evidence): what is the current evidence?
- P (Population): diagnosis, disease stage, co-morbidity, risk factor, sex, age, ethnic group, specific inclusion or exclusion criteria, clinical setting.
- I (Intervention): type, frequency, dose, duration, prognostic factor.
- C (Comparison): placebo, routine care, alternative intervention/management.

- O (Outcome): which clinical or patient-related outcomes will the researcher need to measure, improve, influence or accomplish? Which methods of measurement should be used?
- T (Time stamp): date of literature search or recommendation.

Other factors that might be considered in recommendations include the disease burden of the condition being addressed, the timeliness (e.g. length of follow-up, duration of intervention), and the study type that would best suit subsequent research (Brown 2006).

While Cochrane review authors will find the EPICOT criteria helpful in ensuring that they include the PICO aspects, the criteria of the GRADE framework can help to describe further the additional research that will improve the confidence or certainty in the available evidence. Table 12.6.a shows how review authors may be aided in their interpretation of the body of evidence and drawing conclusions about future research and practice.

The review of compression stockings for prevention of deep vein thrombosis in airline passengers described in Chapter 11 (Section 11.1.6) provides an example where there is some convincing evidence of a benefit of the intervention: “This review shows that the question of the effects on symptomless DVT of wearing versus not wearing compression stockings in the types of people studied in these studies should now be regarded as answered. Further research may be justified to investigate the relative effects of different strengths of stockings or of stockings compared to other preventative strategies. Further randomized studies to address the remaining uncertainty about the effects of wearing versus not wearing compression stockings on outcomes such as death, pulmonary embolus and symptomatic DVT would need to be large.” (Clarke 2006).

A review of therapeutic touch for anxiety disorder provides an example of the implications for research when no eligible studies are found: “This review highlights the need for randomized controlled trials to evaluate the effectiveness of therapeutic touch in reducing anxiety symptoms in people diagnosed with anxiety disorders. Future trials need to be rigorous in design and delivery, with subsequent reporting to include high quality descriptions of all aspects of methodology to enable appraisal and interpretation of results” (Robinson 2007).

Table 12.6.a: Interpretation of the quality of a body of evidence according to individual GRADE criteria

By outcome	Implications for research	Examples	Implications for practice
Risk of bias	Need for methodologically better designed and executed studies	All studies suffered from lack of blinding of outcome assessors. Studies where outcome assessors are blinded are required.	The estimates of effect may be biased because of a lack of blinding.
Inconsistency	Unexplained inconsistency: need for individual participant data meta-analysis; need for studies in relevant subgroups	Studies in patients with small cell lung cancer are needed to understand whether the effects differ from those in patients with pancreatic cancer.	Unexplained inconsistency: consider and interpret overall effect estimates as for the overall quality of a body of evidence Explained inconsistency (if results are not presented separately for different scenarios): consider and interpret effect estimates by subgroup
Indirectness	Need for studies that fit the PICO question of interest better	Studies in patients with early cancer are needed because the evidence is from studies with advanced cancer.	It is uncertain whether the results apply directly to the patients or the way that the intervention is applied in a particular setting.
Imprecision	Need for more studies with more participants to reach optimal information size	Studies with approximately 200 more events in the intervention and control group are required.	Consider and interpret overall effect estimates as for quality of a body of evidence
Publication bias	Need to investigate and identify		Consider and interpret overall effect

	unpublished data; large studies might help resolve this issue		estimates as for quality of a body of evidence
Large effects	No implications	No implications	The effect is large in the populations that were included in the studies. The effect is going to be in the vicinity of the observed effect.
Dose effects	No implications	No implications	The greater the reduction in the exposure the larger is the expected benefit (or harm).
Opposing bias and confounding	Studies controlling for the residual bias and confounding are needed.	Studies controlling for possible confounding by smoking, and degree of education are required.	The effect could be even larger than the one that is observed in the studies presented here.

12.6.4 Common errors in reaching conclusions

A common mistake when there is inconclusive evidence is to confuse ‘no evidence of an effect’ with ‘evidence of no effect’. When there is inconclusive evidence, it is wrong to claim that it shows that an intervention has ‘no effect’ or is ‘no different’ from the control intervention. It is safer to report the data, with a confidence interval, as being compatible with either a reduction or an increase in the outcome. When there is a ‘positive’ but statistically non-significant trend authors commonly describe this as ‘promising’, although a ‘negative’ effect of the same magnitude is not commonly described as a ‘warning sign’; such language may be harmful.

Another mistake is to frame the conclusion in wishful terms. For example, authors might write “the included studies were too small to detect a reduction in mortality” when the included studies showed a reduction or even increase in mortality that failed to reach conventional levels of statistical significance. One way of avoiding errors such as these is to consider the results blinded; i.e. consider how the results would be presented and framed in the ‘Conclusions’ if the direction of the results had been reversed. If the confidence interval for the estimate of the difference in the effects of the interventions overlaps the null value, the analysis is compatible with both a true beneficial effect and a true harmful effect. If one of the possibilities is mentioned in the conclusion, the other possibility should be mentioned as well.

Another common mistake is to reach conclusions that go beyond the evidence. Conclusions must be based only on findings from the synthesis (quantitative or narrative) of studies included in the review. Often additional information or judgement is incorporated implicitly in reaching conclusions. Even when additional information and explicit judgements support conclusions about the implications of a review for practice, review authors rarely conduct systematic reviews of the additional information. Furthermore, implications for practice are often dependent on specific circumstances and values that must be taken into consideration. As noted, authors should always be cautious when drawing conclusions about implications for practice, and they should not make recommendations. Table 12.6.a provides guidance about the interpretation of evidence for research and practice.

12.7 Methodological standards for the conduct of Cochrane Intervention Reviews

No.	Status	Name	Standard	Rationale & elaboration	Handbook sections
C72	Mandatory	Interpreting results	Interpret a statistically non-significant P value (e.g. larger than 0.05) as a finding of uncertainty unless confidence intervals are sufficiently narrow to rule out an important magnitude of effect.	Authors commonly mistake a lack of evidence of effect as evidence of a lack of effect.	12.4.2, 12.7.4

12.8 Chapter information

Authors: Holger J Schünemann, Andrew D Oxman, Gunn E Vist, Julian PT Higgins, Jonathan J Deeks, Paul Glasziou, Elie Akl and Gordon H Guyatt on behalf of the Cochrane Applicability and Recommendations Methods Group.

Acknowledgements: Jonathan Sterne, Michael Borenstein and Rob JM Scholten contributed text to the prior version of this chapter.

Declarations of interest: Holger Schünemann, Andrew Oxman, Gunn Vist, Paul Glasziou, Elie Akl and Gordon Guyatt have, to varying degrees, taken leadership roles in the GRADE Working Group from which many of the ideas in this chapter have arisen.

12.9 References

Aguilar 2005

Aguilar MI, Hart R. Oral anticoagulants for preventing stroke in patients with non-valvular atrial fibrillation and no previous history of stroke or transient ischemic attacks. *Cochrane Database of Systematic Reviews* 2005, Issue 3. CD001927. DOI: 10.1002/14651858.CD006186.pub2.

Akl 2011a

Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, et al. Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database of Systematic Reviews* 2011, Issue 3. CD006776. DOI: 10.1002/14651858.CD006776.pub2.

Akl 2011b

Akl EA, Gunukula S, Barba M, Yosuico VE, van Doormaal FF, Kuipers S, et al. Parenteral anticoagulation in patients with cancer who have no therapeutic or prophylactic indication for anticoagulation. *Cochrane Database of Systematic Reviews* 2011, Issue 4. CD006652. DOI: 10.1002/14651858.CD006652.pub3.

Altman 1998

Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998; 317: 1309-1312.

Brown 2006

Brown P, Brunnhuber K, Chalkidou K, Chalmers I, Clarke M, Fenton M, et al. How to formulate research recommendations. *BMJ* 2006; 333: 804-806.

Cates 1999

Cates C. Confidence intervals for the number needed to treat: Pooling numbers needed to treat may not be reliable. *BMJ* 1999; 318: 1764-1765.

Chinn 2000

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

Clarke 2006

Clarke M, Hopewell S, Juszczak E, Eisinga A, Kjeldstrøm M. Compression stockings for preventing deep vein thrombosis in airline passengers. *Cochrane Database of Systematic Reviews* 2006, Issue 2. CD004002. DOI: 10.1002/14651858.CD004002.pub2.

Cohen 1988

Cohen J. *Statistical Power Analysis in the Behavioral Sciences*. Hillsdale (NJ): Lawrence Erlbaum Associates, Inc., 1988.

Dans 2007

Dans AM, Dans L, Oxman AD, Robinson V, Acuin J, Tugwell P, et al. Assessing equity in clinical practice guidelines. *Journal of Clinical Epidemiology* 2007; 60: 540-546.

Eddy 1990

Eddy DM. Clinical decision making: from theory to practice. Anatomy of a decision. *JAMA* 1990; 263: 441-443.

Friedman 1985

Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*. Littleton (MA): John Wright PSG, Inc., 1985.

Furukawa 1999

Furukawa TA. From effect size into number needed to treat. *The Lancet* 1999; 353: 1680.

GRADE Working Group 2004

GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004; 328: 1490-1494.

Guyatt 1994

Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA* 1994; 271: 59-63.

Hawe 2004

Hawe P, Shiell A, Riley T, Gold L. Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. *Journal of Epidemiology and Community Health* 2004; 58: 788-793.

Haynes 2002

Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *ACP Journal Club* 2002; 136: A11-A14.

Hoffrage 2000

Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Medicine. Communicating statistical information. *Science* 2000; 290: 2261-2262.

Lumley 2004

Lumley J, Oliver SS, Chamberlain C, Oakley L. Interventions for promoting smoking cessation during pregnancy. *Cochrane Database of Systematic Reviews* 2004, Issue 4. CD001055. DOI: 10.1002/14651858.CD001055.pub2.

McQuay 1997

McQuay HJ, Moore A. Using numerical results from systematic reviews in clinical practice. *Annals of Internal Medicine* 1997; 126: 712-720.

Mugford 1989

Mugford M, Kingston J, Chalmers I. Reducing the incidence of infection after caesarean section: implications of prophylaxis with antibiotics for hospital resources. *BMJ* 1989; 299: 1003-1006.

Mugford 1991

Mugford M, Piercy J, Chalmers I. Cost implications of different approaches to the prevention of respiratory distress syndrome. *Archives of Disease in Childhood* 1991; 66: 757-764.

Oxman 2002

Oxman A, Guyatt G. When to believe a subgroup analysis. In: Guyatt G, Rennie D, editor(s). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago (IL): AMA Press, 2002.

Resnicow 1993

Resnicow K, Cross D, Wynder E. The Know Your Body program: a review of evaluation studies. *Bulletin of the New York Academy of Medicine* 1993; 70: 188-207.

Robinson 2007

Robinson J, Biley FC, Dolk H. Therapeutic touch for anxiety disorders. *Cochrane Database of Systematic Reviews* 2007, Issue 3. CD006240. DOI: 10.1002/14651858.CD006240.pub2.

Sackett 2000

Sackett DL, Richardson WS, Rosenberg W, Haynes BR. *Evidence-Based Medicine: How to Practice and Teach EBM*. Edinburgh (UK): Churchill Livingstone, 2000.

Scholten 1999

Scholten RJPM. From effect size into number needed to treat [letter]. *The Lancet* 1999; 453: 598.

Schünemann 2006

Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 13. Applicability, transferability and adaptation. *Health Research Policy and Systems* 2006; 4: 25.

Schünemann 2013a

Schünemann HJ, Tugwell P, Reeves BC, Akl EA, Santesso N, Spencer FA, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; 4: 49-62.

Schünemann 2013b

Schünemann HJ, Guyatt G. Clinical Epidemiology and Evidence-based medicine In: Pigeot I, Ahrens W, editor(s). *Handbook of Epidemiology*. New York (NY): Springer Verlag, 2013.

Smeeth 1999

Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses - sometimes informative, usually misleading. *BMJ* 1999; 318: 1548-1551.

Suissa 1991

Suissa S. Binary methods for continuous outcomes: a parametric alternative. *Journal of Clinical Epidemiology* 1991; 44: 241-248.

Sun 2012

Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*; 344: e1553.

Walter 2001

Walter SD. Number needed to treat (NNT): estimation of a measure of clinical benefit. *Statistics in Medicine* 2001; 20: 3947-3962.

Chapter 21: Reviews in public health and health promotion

Editors: Elizabeth Waters†, Rebecca Armstrong, Jodie Doyle and Helen Morgan.

This chapter should be cited as: Waters E, Armstrong R, Doyle J, Morgan H (editors). Chapter 21: Reviews in public health and health promotion. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS (editors), *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0 (updated June 2017), Cochrane, 2017. Available from www.training.cochrane.org/handbook.

This extract is from *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0. The previous version of this chapter (5.1.0, 2011) is available online at handbook.cochrane.org.

An earlier version (version 5.0.2, 2008) of the *Handbook* is also published by John Wiley & Sons, Ltd under “The Cochrane Book Series” Imprint, as Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on www.wiley.com.

Key Points

- The scope of public health topics is hugely varied, with perceptions of what constitutes public health influenced by different global perspectives and paradigms. However, some features often characterize public health topics: they consider population level in breadth, they are complex in content and they are focused on improving inequity and reducing inequalities, and on addressing the causes or determinants of health problems, the responsibility of which may often lie outside of the health sector.
- Public health, health promotion, prevention, equity and social determinants of health interventions are evaluated using a wide variety of approaches and study designs. The ‘intervention’ may operate at various levels within the health sector (legislative,

†Deceased 22 September 2015

community and individual) and/or it may involve other sectors (for example, transport, finance, agriculture or education).

- The study design options available have increased in diversity and include randomized studies, cohort studies, modelling and, invariably, designs that incorporate measures of objective outcome as well as the views and experiences of participants and stakeholders (ie mixed methods that incorporate qualitative and quantitative research methods).
- Identifying public health, health promotion, prevention and complex intervention literature requires authors to use methods beyond database searching to retrieve studies.
- An important theoretical and conceptual consideration is the disentanglement of public health intervention effects from the influence of the context in which the intervention is implemented. Information should also be sought on contextual factors and on implementation characteristics that may explain the extent to which the intervention or outcomes are sustained.

21.1 Introduction

This chapter provides an overview of issues specific to public health that are not discussed elsewhere in the *Handbook*. The complete version of *Guidelines for Health Promotion and Public Health Systematic Reviews* can be accessed at the Cochrane Public Health website: www.ph.cochrane.org.

21.2 Study designs to include

There are a wide range of questions that are important for decision makers and researchers within the context of public health. Those that relate to intervention effectiveness in public health, health promotion, prevention and population health can be evaluated using a wide variety of approaches and study designs and no single method can be used to answer all relevant questions. This is for a number of reasons related to the sector and discipline involved and the level at which the intervention is implemented (the scope may include legislation and regulation, organizational changes, setting-based policies and individual behaviour change).

If a review question has been specified clearly then understanding of the types of study designs needed to answer it should automatically follow (Petticrew 2003). A preliminary scoping search will also help to identify the types of study designs that may have been used to study the intervention. The criteria used to select studies should primarily reflect the question or questions being answered in the review, rather than any predetermined hierarchy (Glasziou 2004). The decisions about which type(s) of study design to include will influence subsequent phases of the review, particularly searching, assessment of risk of bias and analysis (especially meta-analysis).

Randomized trials continue to provide a useful study design to explore the various contexts within which they can be applied. They provide an exceptionally useful design to understand the differences between the population, group or arm who have been exposed to an intervention in comparison to those who have not or have been less exposed, and they have a long legacy in education and social policy research. Concerns have often been expressed in relation to their use within public health and health promotion, sometimes justified and sometimes paradigmatically; however, when these concerns are unpacked most of them centre around limitations to generalizability if studies have been conducted in small samples, in constrained contexts or are heavily resourced and cannot be replicated (Black 1996). Cluster-randomized studies, stepped wedge designs, cohort studies, complex modelling and other study types are increasingly being proposed within the field of public health.

For some questions, non-randomized studies may represent the best available evidence (of effectiveness). Reviewing non-randomized evidence can give an estimate of the nature, direction and size of effects. Demonstrating the patterns of evidence drawn from different study designs may lead to the development of subsequent study designs (including randomized trials) to test the intervention. Studies generating qualitative data may also be relevant to other kinds of questions beyond effectiveness questions. For example, data may be gathered on the preferences of the likely recipients of the interventions and the factors that constrain or facilitate the successful outcome of particular interventions. There are programmes of research in Europe and the UK on randomized and non-randomized studies of public health and health promotion interventions. Chapter 13 discusses general issues on the inclusion of non-randomized studies in Cochrane Reviews and Chapter 20 addresses qualitative studies.

21.3 Searching

Finding studies on public health interventions is much more complicated than retrieving medical studies because the literature is widely scattered (Peersman 2001). The multidisciplinary nature of public health and health promotion means that studies can be found in a number of different areas and through a wide range of electronic databases (Beahler 2000, Grayson 2003). Difficulties also arise because the terminology is imprecise and constantly changing (Grayson 2003). In public health examples of publication bias, including database bias, language bias and grey literature bias, mean that review findings will be compromised when the results in the ‘difficult to locate’ sources are systematically different from those found in the easily accessible sources (Howes 2004). Thorough searching for public health and health promotion literature is therefore essential, albeit often a very time-consuming task and one that requires authors to use retrieval methods in addition to comprehensive searches of bibliographic databases to locate studies.

Non-randomized trials are not adequately indexed in bibliographic databases and therefore we do not currently recommend that study design filters be applied for this type of literature. We also recognize that pragmatic decisions may often need to be taken when balancing the time and other resources required to conduct comprehensive searches against the ratio of relevant to non-relevant studies identified. Researchers may decide that they need to apply study design filters and, if so, they need to report this when describing their search strategies to make the potential limitations of the searches clear.

Table 21.3.a lists some electronic databases relevant to a variety of public health and health promotion topics.

Table 21.3.a: Electronic databases relevant to public health and health promotion

Field	Resources
Generic health	CENTRAL, CINAHL (Cumulative Index to Nursing and Allied Health Literature), Embase, MEDLINE, Scopus, Web of Science, NHS Evidence
Education	ERIC (Educational Resources Information Center), Database of Education Research, Current Educational Research in the United Kingdom (CERUK)
Health promotion and public health	BiblioMap, TRoPHI (Trials Register of Promoting Health Interventions)
International development	LILACS (Latin American and Caribbean Health Sciences Literature), 3ie database
Physical activity	SPORTDiscus
Psychology	PsycINFO
Sociology	ASSIA (Applied Social Sciences Index and Abstracts), Sociological Abstracts, Social Science Citation Index (included in Web of Science), Social Policy and Practice
Transport	NTIS (National Technical Information Service), TRID (integrated transport research database)
Other	Enviroline (environmental health), TOXLINE (toxicology), EconLit (economics)
Systematic reviews	CDSR (Cochrane Database of Systematic Reviews), DoPHER (Database of Promoting Health Effectiveness Reviews), Health-Evidence.org
Qualitative	UK Data Service, DIPEX (Database of Interviews on Patient Experience)

21.4 Assessment of study quality and risk of bias

Assessing the quality of public health and health promotion studies, and their resulting risk of bias, requires authors to consider the criteria to be used at the planning stage of the review. Appraisal criteria will depend on the type of study included in the review. Authors should be guided by the Cochrane Review Group (CRG) editing their review and the appraisal tools they use. However, the following describes tools that may be useful for assessing studies of public health and health promotion interventions.

- The risk of bias in randomized trials should be assessed using the Cochrane 'Risk of bias' tool, described in Chapter 8 (Section 8.5).
- Issues around risk of bias assessment for cluster-randomized studies are discussed in Chapter 16 (Section 16.3.2).
- For risk of bias in non-randomized studies authors should consult Chapter 13 (Section 13.5).
- The results of uncontrolled studies (also called before-and-after studies without a control group) should be treated with caution. The absence of a comparison group makes it almost impossible to know what would have happened without the intervention. There are particular problems with interpreting data from uncontrolled studies, including susceptibility to problems with confounding (including seasonality) and regression to the mean.

21.5 Considering equity in reviews

Public health, health promotion, health improvement and population health interventions are largely intended to improve the health of populations. Systematic reviews are an extremely useful methodology to determine the overall effectiveness of interventions in achieving population-level outcomes.

However, there are some specific ethical considerations that should be taken into account in reviewing the effectiveness of public health, health promotion, health improvement and population-level interventions. Effectiveness is typically measured in terms of the total number (population) who benefit from the intervention, or on the mean effect across a population. This consequentialist approach (i.e. the end justifying the means) takes no account of the distribution of benefits (Hawe 1995), and therefore does not address issues of health equity. Overall improvements in health behaviours or health outcomes may actually mask the differences in health outcomes between groups (Macintyre 2003), thus closer scrutiny is essential. Interventions that work for those in the middle and upper socio-economic positions may not be as effective for those who are disadvantaged. Well-intentioned interventions may actually increase inequalities if factors affecting uptake, context or systems are not adequately planned for or theorized in the development of the intervention. Health differentials that exist between groups at the start of and following an intervention may also be due to complex interactions between many of the factors relating to disadvantage (Jackson 2003).

An important way to improve current effectiveness evidence for health and social interventions is therefore to assess, and report on, the impact of interventions on health equity (Petticrew 2009). Systematic review methodology has the potential to investigate differential outcomes for groups with varying levels of disadvantage. This is important because identifying the effect of interventions on disadvantaged groups can inform strategies aimed at reducing health inequalities and health inequities. Health inequalities are "differences, variations, and disparities in the health achievements of individuals and groups" (Kawachi 2002). Health equity is an ethical concept referring to the fairness or unfairness of particular health inequalities. The International Society for Equity in Health defines equity in health as: "the absence of potentially remediable, systematic differences in one or more aspects of health status across socially, economically, demographically, or geographically defined populations or subgroups" (Macinko 2002). Turning this around, health inequities are those health inequalities that are unfair or unjust, or stem from some kind of injustice (Kawachi 2002). Reviews of the effectiveness of public health and health promotion interventions can provide information about the differential effects of interventions on health and analysis of the intervention components against inequalities can help to identify solutions (Waters 2011).

Disadvantage may be considered in terms of place of residence, race or ethnicity, occupation, gender, religion, education, socio-economic position (SES) and social capital, known by the PROGRESS acronym (Evans 2003). Authors should carefully consider which of these are relevant to their population of interest; data should then be extracted on these factors. The Campbell and Cochrane Equity Methods Group (http://methods.cochrane.org/equity) provides author resources and guidance on including an equity lens in Cochrane reviews, including an Equity Checklist (http://methods.cochrane.org/equity/resources-review-authors).

Systematic reviews rely upon there being sufficient detail in study data to allow for identification of relevant subgroups for analysis in relation to health inequalities. This requires attention not only to levels of benefit or harm, but also to the distributions of these: who is benefiting, who is harmed, who is excluded?

Reviews of the effectiveness of interventions in relation to health inequalities require three components for calculation:

- a valid measure of health status (or change in health status);
- a measure of socio-economic position (or disadvantage); and
- a statistical method for summarizing the magnitude of health differences between people in different groups.

Review authors should decide which indicator(s) of disadvantage or status (refer again to the PROGRESS acronym) are relevant to the review topic. Practitioners and/or policy makers should be consulted if the authors are not familiar with the topic under review to help identify the most appropriate factors.

Conducting reviews addressing inequalities is complicated not only by limited collection of information about differences between groups, but also by the fact that there is limited participation of disadvantaged groups in research. Despite these barriers, systematic reviews can play an important role in raising awareness of health inequalities.

To locate studies that examine inequalities, review authors will need to cast the net broadly when performing searches and contact authors for further information regarding socio-economic data. This latter task may be necessary because primary studies often fail to present information on the socio-economic composition of participants (Oakley 1998, Jackson 2003, Ogilvie 2004). Once studies have been appraised and data have been extracted, studies need to be classified as to whether they are effective for reducing health inequalities. An effective intervention to reduce inequity is generally one that is more effective for disadvantaged groups or individuals. The judgement becomes more difficult when the intervention is targeted only at disadvantaged individuals or groups. It is impossible to determine differential effectiveness if studies comprise mixed levels of advantage and disadvantage but do not include results that can be broken down by socio-economic (or similar) grouping.

21.6 Context

The type of interventions implemented and their subsequent success or failure are highly dependent on the social, economic and political context in which they are developed and implemented (see example in Figure 21.6.a). A problem in reviewing public health and health promotion interventions is how to disentangle ‘intervention’ effects from effects that should be more appropriately called ‘program by context interactions’ (Hawe 2004). Traditionally, outcomes have been attributed to the intervention. However, the outcomes noted in studies may in fact be due to pre-existing factors of the context into which the intervention was introduced. Hence, context should be considered and measured as an effect modifier in studies (Eccles 2003, Hawe 2004). Such contextual factors might relate to aspects of the program’s ‘host organization’. Broader aspects of context might include aspects of the *system within which the host organization operates*. Some investigators would also argue that context factors also pertain to the *characteristics of the target group or population.* For many years these aspects have been acknowledged (but not clearly specified) when decision makers have argued that the results of evidence reviews from other countries do not apply in their own country.

Use of the term ‘context evaluation’ became more prevalent in health promotion after the review by Israel and colleagues (Israel 1995). However, the systematic investigation of context-level interactions as part of the design of randomized trials of community or organizational-level interventions is almost unknown (Eccles 2003, Hawe 2004). Instead, aspects of context have been explored as part of the more developed field of sustainability research or research on program institutionalization: see Section 21.7. A related and growing multidisciplinary research field is the implementation and integration sciences, which are leading researchers further into the complexity of the change processes that interventions represent (Ottoson 1987, Bauman 1991, Scheirer 1994).

Systematically disentangling context effects from intervention effects in anything other than a study set up for this purpose is extremely difficult. However, if intervention reviews are to be useful to decision makers, contextual and implementation information are essential and non-negotiable elements of the review (Waters 2011). Whilst some programs have been transferred from one context to another and benefits have been observed (Resnicow 1993), others have not (Lumley 2004). Cluster-randomized designs may be expected (in theory) to even out important aspects of context, provided that the sample size is sufficient. However, few investigators at present measure or report on any aspect of context that might be important to our assessment. A stronger focus on external validity has long been called for (Glasgow 2006, Green 2006). Working together, journal editors and researchers are encouraging more examination of, and reporting on, aspects of intervention context (Armstrong 2008). This should be reflected in the content of Cochrane Reviews.

Figure 21.6.a: Example of intervention success as dependent on the context in which it is implemented (Frommer 2003)

Media-based intervention to promote the consumption of fruit and vegetables

↓ *Dependent on the following contextual factors:*

Availability and relative price of fruit and vegetables

↓ *Dependent on the following contextual factors:*

Geographic factors, food distribution systems and retail prices

21.7 Sustainability

Sustainability in the context of interventions is very broadly defined in the literature, referring to the general phenomenon of the continuation of an intervention or its effects (Shediac-Rizkallah 1998, Swerissen 2004). Sustainability of interventions should be an important consideration in systematic reviews. Whilst the realities of primary intervention research and implementation funding mean the production and assessment of relatively short-term outcomes, there is a need to identify, at the very least, indicators of longer-term effects. Attention to the long-term viability of health interventions is likely to increase as policy makers, practitioners and funders become increasingly concerned with allocating scarce resources effectively and efficiently (Shediac-Rizkallah 1998). Users of reviews are interested in knowing whether the health benefits, such as reductions in specific diseases or improvements in health (or otherwise), are going to be sustained beyond the life of the interventions, as well as an intervention's ability to sustain its activities over time.

Unfortunately, collection of data on the extent to which the intervention and outcomes are sustained is often not carried out, which limits the extent to which long-term impacts can be assessed. Careful consideration in Cochrane Reviews of how previous studies have (or have not) addressed issues of sustainability will increase our understanding in this area and hopefully also stimulate improved designs for the assessment of sustainability in future studies.

In addition, sustainability is often documented at the program or project level only (Harris 2013) and in reference to program continuity, duration and institutionalization (Savaya 2012). A sustained or sustainable program does not necessarily result in sustained outcomes and not all interventions need to be sustained in order to be useful or effective (Shediac-Rizkallah 1998). A more contemporary understanding of sustainability would include concepts such as practice, capacity and outcomes (Swerissen 2004, Wiltsey Stirman 2012, Chambers 2013, Harris 2013).

Also, review authors should consider whether the sustainability of the outcomes is relevant to the objectives of the intervention. If this is the case, authors should consider what outcomes have (or should have) been measured, over what period and what the pattern of outcomes is over time.

Information should be sought on both contextual factors and intervention characteristics that may explain the extent to which the interventions or outcomes are sustained. Where sustainability of outcomes has not been measured, authors should explore the *potential* of the intervention outcomes to be sustained. Six frameworks (presented below) are available and may assist in determining sustainability. Different methods may be necessary to assess the sustainability of different types of interventions found in primary research (Scheirer 2013).

1. Luke, Schell and colleagues developed a tool that may be used to assess program sustainability (https://sustaintool.org) from an extensive literature review and concept mapping process (Schell 2013, Luke 2014). The tool is particularly suited to public health and health promotion interventions aimed at the community level. The nine core domains that affect a program's capacity for sustainability, which are included in the tool, are: Political Support, Funding Stability, Partnerships, Organizational Capacity, Program Evaluation, Program Adaptation, Communications, Public Health Impacts and Strategic Planning.
2. Bossert lists the following five factors that influence sustainability (Bossert 1990):
    - the economic and political variables surrounding the implementation and evaluation of the intervention;
    - the strength of the institution implementing the intervention;

- the full integration of activities into existing programs/services/curriculum/etc;
- whether the program includes a strong training component (capacity building); and
- community involvement/participation in the program.

3. The framework developed by Swerissen and Crisp guides decisions about the likely sustainability of interventions and effects at different levels of social organization (Swerissen 2004). This framework outlines the relationships between intervention level, strategies and the likely sustainability of interventions and effects.
4. Shediac-Rizkallah and Bone present a useful framework for conceptualizing sustainability (Shediac-Rizkallah 1998). In this framework, key aspects of program sustainability are defined as 1) maintenance of health benefits from the program; 2) institutionalization of a program within an organization; and 3) capacity building in the recipient community. Key factors influencing sustainability are defined as 1) factors in the broader environment; 2) factors within the organizational setting; and 3) project design and implementation factors.
5. The Centre for Health Promotion, University of Toronto, has also produced a document outlining four integrated components of sustainability.
6. Finally, Cochrane Public Health has contributed to a scoping review of the literature that aimed to understand key elements of sustainability in public health and health promotion interventions, using community-based obesity prevention as an example (Whelan 2014). This review revealed advances in how sustainability is defined, conceptualized and understood. Ten key elements were distilled from recent developments in the literature and checked for consistency with existing frameworks (such as those described above). While the key elements were synthesized to guide decision-making in intervention planning and practice, they may be useful in considering the extent to which primary studies address the sustainability of interventions. These elements could be used during data collection, reported within the intervention description sections or the ‘Characteristics of included studies’ table, and discussed in the review findings (refer to Table 1 in (Whelan 2014).

21.8 Applicability and transferability

Applicability needs to be considered when deciding how to translate the findings of a given study or review to a specific population, intervention or setting, see Chapter 12 (Section 12.3). *Transferability* or the *potential for translation* are similar and appropriate terms. Applicability is closely related to integrity, context and sustainability, as discussed in previous sections of this chapter.

Systematic reviews of public health and health promotion interventions encompass several issues that make the process of determining applicability even more complex than in the clinical trials literature. First, a number of public health interventions do not involve randomization. Although not an inherent characteristic of non-randomized designs, these studies may have less well-defined eligibility criteria, settings and interventions, making determinations of applicability more difficult. Then again, results from randomized trials may be less generalizable due to unrepresentative intervention providers or study participants not being typical of the target group (Black 1996). Second, public health and health promotion interventions tend to have multiple components. This makes it difficult to 1) determine which specific intervention component had the noted effect, and 2) assess the synergy between components. Third, in community interventions, implementation and adherence may be much more difficult to achieve and to measure. This also makes it harder to interpret and apply the findings. Fourth, in public health and health promotion interventions the underlying socio-cultural characteristics of communities are complex and difficult to measure. Thus it is difficult to define to whom and to what degree the intervention was applied, complicating determinations of applicability. On the other hand, this heterogeneity may increase applicability, as the original populations, settings and interventions may be quite diverse, increasing the likelihood that the evidence can be applied broadly.

Review authors are ideally positioned to summarize the various aspects of the evidence that are relevant to potential users. This enables users to compare their situation or setting to that presented in the review and to note the similarities and differences. Users can then be explicit about the relationship between the body of evidence and their specific situation.

The following questions may help authors to consider issues of applicability and transferability relevant to public health and health promotion (Wang 2006).

Applicability

- Does the **political environment** of the local society allow this intervention to be implemented?
- Is there any political barrier to implementing this intervention?
- Would the general public and the targeted (sub) population accept this intervention? Does any aspect of the intervention go against local **social norms**? Is it ethically acceptable?
- Can the contents of the intervention be tailored to suit the local culture?
- Are the essential **resources** for implementing this intervention available in the local setting? (A list of essential resources may help to answer this question.)
- Does the target population in the local setting have a sufficient **educational** level to comprehend the contents of the intervention?
- Which organization will be responsible for the provision of this intervention in the local setting?
- Is there any possible barrier to implementing this intervention due to the **structure of that organization**?

- Does the provider of the intervention in the local setting have the **skill** to deliver this intervention? If not, will training be available?

Transferability

- What is the **baseline prevalence** of the health problem of interest in the local setting? What is the difference in prevalence between the study setting and the local setting?
- Are the **characteristics of the target population** comparable between the study setting and the local setting? With regard to the particular aspects that will be addressed in the intervention, is it possible that the characteristics of the target population, such as ethnicity, socio-economic status, educational level, etc, will have an impact on the effectiveness of the intervention?
- Is the **capacity to implement** the intervention comparable between the study setting and the local setting in such matters as political environment, social acceptability, resources, organizational structure and the skills of the local providers?

21.9 Chapter information

Editors: Elizabeth Waters, Rebecca Armstrong, Jodie Doyle and Helen Morgan.

Contributing authors: Rebecca Armstrong, Elizabeth Waters, Nicki Jackson, Sandy Oliver, Jennie Popay, Jonathan Shepherd, Mark Petticrew, Laurie Anderson, Ross Bailie, Ginny Brunton, Penny Hawe, Elizabeth Kristjansson, Lucio Naccarella, Susan Norris, Elizabeth Pienaar, Helen Roberts, Wendy Rogers, Amanda Sowden and Helen Thomas.

Acknowledgements: Thanks to Ruth Turley for helpful comments.

21.10 References

Armstrong 2008

Armstrong R, Waters E, Moore L, Riggs E, Cuervo LG, Lumbiganon P, et al. Improving the reporting of public health intervention research: advancing TREND and CONSORT. *Journal of Public Health* 2008.

Bauman 1991

Bauman LJ, Stein RE, Ireys HT. Reinventing fidelity: the transfer of social technology among settings. *American Journal of Community Psychology* 1991; 19: 619-639.

Beahler 2000

Beahler CC, Sundheim JJ, Trapp NI. Information retrieval in systematic reviews: challenges in the public health arena. *American Journal of Preventive Medicine* 2000; 18: 6-10.

Black 1996

Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996; 312: 1215-1218.

Bossert 1990

Bossert TJ. Can they get along without us? Sustainability of donor-supported health projects in Central America and Africa. *Social Science & Medicine* 1990; 30: 1015-1023.

Chambers 2013

Chambers D, Glasgow R, Stange K. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. *Implementation Science*; 8: 117.

Eccles 2003

Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. *Quality and Safety in Health Care* 2003; 12: 47-52.

Evans 2003

Evans T, Brown H. Road traffic crashes: operationalizing equity in the context of health sector reform. *International Journal of Injury Control and Safety Promotion* 2003; 10: 11-12.

Frommer 2003

Frommer M, Rychetnik L. From evidence-based medicine to evidence-based public health. In: Lin V, Gibson B, editor(s). *Evidence-based Health Policy: Problems and Possibilities*. Melbourne (Australia): Oxford University Press, 2003.

Glasgow 2006

Glasgow RE, Green LW, Klesges LM, Abrams DB, Fisher EB, Goldstein MG, et al. External validity: we need to do more. *Annals of Behavioral Medicine* 2006; 31: 105-108.

Glasziou 2004

Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. *BMJ* 2004; 328: 39-41.

Grayson 2003

Grayson L, Gomersall A. *A Difficult Business: Finding the Evidence for Social Science Reviews*. London (UK): ESRC UK Centre for Evidence Based Policy and Practice, 2003.

Green 2006

Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. *Evaluation and the Health Professions* 2006; 29: 126-153.

Harris 2013

Harris N, Sandor M. Defining sustainable practice in community-based health promotion: A Delphi study of practitioner perspectives. *Health Promotion Journal of Australia*; 24: 53-60.

Hawe 1995

Hawe P, Shiell A. Preserving innovation under increasing accountability pressures: the health promotion investment portfolio approach. *Health Promotion Journal of Australia* 1995; 5: 4-9.

Hawe 2004

Hawe P, Shiell A, Riley T, Gold L. Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. *Journal of Epidemiology and Community Health* 2004; 58: 788-793.

Howes 2004

Howes F, Doyle J, Jackson N, Waters E. Evidence-based public health: The importance of finding 'difficult to locate' public health and health promotion intervention studies for systematic reviews. *Journal of Public Health* 2004; 26: 101-104.

Israel 1995

Israel BA, Cummings KM, Dignan MB, Heaney CA, Perales DP, Simons-Morton BG, et al. Evaluation of health education programs: current assessment and future directions. *Health Education Quarterly* 1995; 22: 364-389.

Jackson 2003

Jackson T, Aldrich R, Dixon J, Furler J, Turrell G, Wilson A, et al. *Using Socioeconomic Evidence in Clinical Practice Guidelines*. Canberra (Australia): National Health and Medical Research Council, 2003.

Kawachi 2002

Kawachi I, Subramanian SV, Almeida-Filho N. A glossary for health inequalities. *Journal of Epidemiology and Community Health* 2002; 56: 647-652.

Luke 2014

Luke DA, Calhoun A, Robichaux CB, Elliott MB, Moreland-Russell S. The Program Sustainability Assessment Tool: A New Instrument for Public Health Programs. *Preventing Chronic Disease*; 11: E12.

Lumley 2004

Lumley J, Oliver SS, Chamberlain C, Oakley L. Interventions for promoting smoking cessation during pregnancy. *Cochrane Database of Systematic Reviews* 2004, Issue 4. CD001055. DOI: 10.1002/14651858.CD001055.pub2.

Macinko 2002

Macinko JA, Starfield B. Annotated Bibliography on Equity in Health, 1980-2001. *International Journal for Equity in Health* 2002; 1: 1.

Macintyre 2003

Macintyre S. Evaluating the evidence on measures to reduce inequalities in health. In: Oliver A, Exworthy M, editor(s). *Health Inequalities: Evidence, Policy and Implementation. Proceedings from a meeting of the Health Equity Network*. London (UK): The Nuffield Trust, 2003.

Oakley 1998

Oakley A, Peersman G, Oliver S. Social characteristics of participants in health promotion effectiveness research; trial and error? *Education for Health* 1998; 11: 305-317.

Ogilvie 2004

Ogilvie D, Petticrew M. Reducing social inequalities in smoking: can evidence inform policy? A pilot study. *Tobacco Control* 2004; 13: 129-131.

Ottoson 1987

Ottoson JM, Green LW. Reconciling concept and context: theory of implementation. In: Ward WB, editor(s). *Advances in Health Education and Promotion Volume 2*. Greenwich (CT): JAI Press, 1987.

Peersman 2001

Peersman G, Oakley A. Learning from research. In: Oliver S, Peersman G, editor(s). *Using Research for Effective Health Promotion*. Buckingham (UK): Open University Press, 2001.

Petticrew 2003

Petticrew M, Roberts H. Evidence, hierarchies, and typologies: horses for courses. *Journal of Epidemiology and Community Health* 2003; 57: 527-529.

Petticrew 2009

Petticrew M, Tugwell P, Welch V, Ueffing E, Kristjansson E, Armstrong R, et al. Better evidence about wicked issues in tackling health inequities. *Journal of Public Health* 2009; 31: 453-456.

Resnicow 1993

Resnicow K, Cross D, Wynder E. The Know Your Body program: a review of evaluation studies. *Bulletin of the New York Academy of Medicine* 1993; 70: 188-207.

Savaya 2012

Savaya R, Spiro SE. Predictors of Sustainability of Social Programs. *American Journal of Evaluation*; 33: 26-43.

Scheirer 1994

Scheirer MA. Designing and using process evaluations. In: Wholey JS, Hatry HP, Newcomer KE, editor(s). *Handbook of Practical Program Evaluation*. San Francisco: Jossey Bass, 1994.

Scheirer 2013

Scheirer MA. Linking Sustainability Research to Intervention Types. *American Journal of Public Health* 2013; 103: e73-e80.

Schell 2013

Schell S, Luke D, Schooley M, Elliott M, Herbers S, Mueller N, et al. Public health program capacity for sustainability: a new framework. *Implementation Science*; 8: 15.

Shediac-Rizkallah 1998

Shediac-Rizkallah MC, Bone LR. Planning for the sustainability of community-based health programs: conceptual frameworks and future directions for research, practice and policy. *Health Education Research* 1998; 13: 87-108.

Swerissen 2004

Swerissen H, Crisp BR. The sustainability of health promotion interventions for different levels of social organization. *Health Promotion International* 2004; 19: 123-130.

Wang 2006

Wang S, Moss JR, Hiller JE. Applicability and transferability of interventions in evidence-based public health. *Health Promotion International* 2006; 21: 76-83.

Waters 2011

Waters E, Hall BJ, Armstrong R, Doyle J, Pettman TL, de Silva-Sanigorski A. Essential components of public health evidence reviews: capturing intervention complexity, implementation, economics and equity. *Journal of Public Health*; 33: 462-465.

Whelan 2014

Whelan J, Love P, Pettman T, Doyle J, Booth S, Smith E, et al. Cochrane Update: Predicting sustainability of intervention effects in public health evidence: identifying key elements to provide guidance. *Journal of Public Health*; 36: 347-351.

Wiltsey Stirman 2012

Wiltsey Stirman S, Kimberly J, Cook N, Calloway A, Castro F, Charns M. The sustainability of new programs and innovations: a review of the empirical literature and recommendations for future research. *Implementation Science*; 7: 17.