Oxford Textbook of Global Public Health (6 ed.)

Edited by Roger Detels, Martin Gulliford, Quarraisha Abdool Karim, and Chorh Chuan Tan

Previous Edition (5 ed.)

Publisher: Oxford University Press
Print ISBN-13: 9780199661756
DOI: 10.1093/med/9780199661756.001.0001
Print Publication Date: Feb 2015
Published online: Feb 2015

Systematic reviews and meta-analysis

Chapter: Systematic reviews and meta-analysis

Author(s): Jimmy Volmink and Mike Clarke

DOI: 10.1093/med/9780199661756.003.0118

Introduction to systematic reviews and meta-analysis

Reviewing and summarizing existing knowledge from studies that have addressed a particular topic (sometimes referred to as research, or evidence, synthesis) has long been recognized as an important scientific activity. At a meeting in 1884, Lord Rayleigh, a Professor of Physics at Cambridge University, noted, 'If, as is sometimes supposed, science consisted in nothing but the laborious accumulation of facts, it would soon come to a standstill, crushed, as it were, under its own weight. . . Two processes are thus at work side by side, the reception of new

material and the digestion and assimilation of the old; and as both are essential we may spare ourselves the discussion of their relative importance' (Rayleigh 1885).

These remarks underscore the importance of understanding the cumulative nature of science, implying that new research should be guided by what is already known and that new research findings should be interpreted in the context of the totality of available evidence (Chalmers et al. 2002). In practice this means that new research should begin and end with an adequate review of other relevant research (Clark and Horton 2020; Clarke et al. 2010).

Reviews also play a critical role in guiding practitioners, policymakers, and consumers to make well-informed decisions and choices. The massive expansion in the volume of biomedical literature, beginning in the middle of the twentieth century, has made it increasingly difficult for healthcare decision-makers to keep up with the findings of individual studies. Reviews of the literature are thus important for taking stock of existing knowledge and making informed choices about healthcare and resource allocation (Mulrow 1995; Sackett et al. 1997; Chalmers and Glasziou 2009).

To the extent that they provide reliable summaries of research evidence, reviews can be extremely valuable. However, poorly conducted reviews drawing incorrect conclusions from the literature have the potential to cause serious harm to patients and populations, and waste much-needed resources (Chalmers 2003).

Rationale for systematic reviews

A number of obstacles to carrying out scientifically sound reviews have been identified. First, research reports are very widely dispersed and relevant papers may be hard to find. In the mid 1990s, it was estimated that some two million articles were published annually in more than 20,000 biomedical journals (Mulrow 1995) and there have been further increases since then. Furthermore, other useful scientific information is distributed in the 'grey literature' (e.g. books, theses, conference abstracts government and company reports) and on the Internet. Added to this, access to research is often biased (Hopewell et al. 2007b, 2009; Dwan et al. 2008). For instance, studies with 'negative' or 'disappointing' findings are less likely to be published. If they are, they are less likely than 'positive' studies to be published in full, in journals that are widely read, or in English. This means that limiting one's efforts to easily accessible studies may result in an over-optimistic view of reality. Another problem is that studies addressing the same question frequently produce conflicting findings which will require explanation, and careful efforts to compare and contrast different studies. Such inconsistencies in findings can be due to factors such as differences in study characteristics (participants, interventions, and outcomes), flaws in study conduct, or

reporting (bias) or random error (the play of chance). To address these challenges, sound methods to identify, assess, and synthesize study findings are needed.

Unfortunately, scientific rigour has not always received sufficient attention in the conduct of literature reviews. A seminal study published in the 1980s demonstrated that reviews published in leading medical journals failed to use scientific methods for synthesizing research evidence (Mulrow 1987). Subsequent research showed that these shortcomings were widespread, often leading to recommendations that were not based on the best available evidence, and to inadequate healthcare being offered with deleterious effects on health outcomes. For example, in the field of cardiovascular medicine it was found that poorly conducted research syntheses had contributed to substantial delays in adopting effective interventions, while allowing other interventions to remain in use long after research had demonstrated that they were ineffective or harmful (Antman et al. 1992). Similar findings have been documented across a range of healthcare fields (Egger et al. 2001; Chalmers 2003; Oxman et al. 2007).

In recent years, the ‘systematic review’ has emerged as a response to these deficiencies (Chalmers and Altman 1995; Glasziou et al. 2001; Higgins and Green 2011; Khan et al. 2011). Systematic reviews use systematic, explicit, and reproducible methods to identify relevant studies, assess the risk of bias in primary studies, extract information, and synthesize findings. Systematic reviews often employ statistical methods to combine data from similar studies (known as meta-analysis) in order to obtain more precise estimates than those derived from the individual studies included in reviews. Table 5.15.1 provides a comparison of the typical characteristics of systematic versus more traditional reviews.

Table 5.15.1 Comparison of systematic and traditional reviews

Systematic review	Traditional review
States the methods of the review	Often does not state the methods of the review
Answers a focused question or tests a hypothesis; usually narrow in scope	May have a clear question but it often involves general discussion with no stated hypothesis; often very broad in scope
Attempts to locate all relevant studies to minimize publication and other biases	Usually does not attempt to locate all relevant studies
Uses explicit eligibility criteria to determine which studies will be included in the review to prevent selection bias on behalf of the reviewer	Usually does not specify why some studies are included and others excluded
Systematically appraises the methods of the included studies to determine possibility of bias in these studies and sources of heterogeneity between the study results	Often does not assess the risk of bias in the studies or acknowledge differences in their methods
Bases results on studies that are most methodologically sound	Often does not distinguish between methodologically sound and unsound studies
May combine study results statistically in a meta-analysis	Often does not pool results in a meta-analysis
Results illustrate true level of heterogeneity of data, allowing quantification of uncertainty	Results often presented in black and white terms without indicating uncertainty or variability

Source: data from Needleman IG, A guide to systematic reviews, *Journal of Clinical Periodontology*, Volume 29, Issue Supplement 3, pp. 6–9, Copyright © 2002 and Petticrew M., Systematic reviews from astronomy to zoology: myths and misconceptions, *British Medical Journal*, Volume 322, pp. 98–101, Copyright © 2001.

In recent years, systematic reviews have gained wide recognition as an essential tool for evidence-based decision-making in healthcare (Petticrew 2001; McMichael et al. 2005; Lavis 2009; Dickersin 2010). For instance, when evaluating the effects of interventions they are considered as the 'gold standard', occupying the top of the 'hierarchy of evidence'. This hierarchy is typically depicted in the form of a pyramid with expert opinion at its base, followed by various types of observational studies, then randomized trials and finally systematic reviews. Because the risk of bias decreases when moving from the bottom to the top of the pyramid, the position of systematic reviews in the hierarchy reflects the fact that they are regarded as the most reliable form of evidence. A further indication of their general acceptance into the mainstream comes from two other developments: funding agencies are beginning to require that investigators provide evidence from up-to-date systematic reviews that a new trial is justified, and some journals now require authors of new trials to report the findings from a systematic review which puts their trial into context (Clark and Horton 2010).

In this chapter, we describe the steps involved in conducting a systematic review: formulation of a review question, identification of relevant studies, extraction of data, assessment of bias in included studies, synthesis of the data, interpreting the evidence, and writing up and updating the review. Further details on each of these steps can be found in published handbooks and guidelines (Liberati et al. 2009; Higgins and Green 2011), as well as in various texts on systematic reviews (Glasziou et al. 2001; Borenstein et al. 2009; Higgins and Green 2011; Khan et al. 2011).

The review protocol

As should be the case for any new piece of scientific research, a systematic review begins with the preparation of a protocol in which the objectives and proposed methods are documented. In a systematic review this is important for reducing bias that may result from reviewers making judgements during the review process that are influenced by their knowledge of individual study findings.

Publication of the review protocol in a publicly accessible register is encouraged (Liberati et al. 2009; *PLoS Medicine* Editors 2011; Stewart et al. 2012). This not only improves transparency and allows detection of any changes to the protocol that may have occurred during the process of the review, but also reduces the likelihood of unnecessary duplication of effort and facilitates peer review of the intended methods (Silagy et al.

2002; Straus and Moher 2010; Booth et al. 2011). PROSPERO, developed by the Centre for Reviews and Dissemination at the University of York, England (Booth et al. 2011) and the *Cochrane Database of Systematic Reviews* (http://www.thecochranelibrary.com) are examples of registers of systematic reviews. Certain funders, such as the National Institutes of Health Research in the United Kingdom and the Canadian Institutes of Health Research, now require that all systematic reviews they support are registered (Stewart et al. 2012).

Formulating the review question

The most critical step in preparing a systematic review is the formulation of a clearly defined question, since this influences all subsequent steps of the review. Review questions may address a wide variety of issues, such as the incidence or prevalence of disease (Pendlebury and Rothwell 2009), accuracy of a diagnostic test (Doust et al. 2004), disease aetiology (Renehan et al. 2008), effects of curative or preventive interventions (Siegfried et al. 2009), economic evaluations (Pérez Velasco et al. 2012), or disease prognosis (Damman et al. 2007). These questions typically lead to the search for, and synthesis of, quantitative studies.

Systematic reviews of qualitative studies, despite presenting some unique methodological challenges, are also possible and are growing in popularity (Britten et al. 2002; Munro et al. 2007; Atkins et al. 2008). Such reviews tend to focus on the framing of a health problem (Mays 2005; Lavis et al. 2009) or explore barriers to, and facilitators of, policy or programme implementation (Innvaer 2002; Child 2012). These may be helpful for explaining why certain interventions do or do not work.

The review question may have a broad or narrow scope. For example, when evaluating the effects of interventions to promote adherence to therapy, a review could assess all interventions in any patient using chronic medication. Alternatively, the review question may be limited to a particular kind of intervention (e.g. economic support) in a particular group (e.g. adults) with a specific disease (e.g. tuberculosis) and explore a specific outcome (e.g. cure or treatment completion).

Decisions about a review’s scope will be influenced by considerations such as the purpose, intended relevance and envisaged impact of the review, the theoretical, biological, and epidemiological rationale, and the potential for obtaining valid answers to the review question or for generalizing the findings of the review (O’Connor et al. 2011). Often, the scope will depend on the resources available for conducting the review. Although reviews with a broad scope tend to be more useful than those with a limited scope for generalizing findings across different participants, interventions, and settings, they may present a higher risk of heterogeneity (‘mixing apples and oranges’), which can make it difficult to interpret their findings (O’Connor et al. 2011).

Regardless of how wide the intended scope of a review may be, it is always important to formulate the question as precisely as possible. The reason for this is that the question has a direct impact on all subsequent aspects of the review process, in particular the eligibility criteria for including or excluding studies and the efficiency of the search for relevant studies. A review question that is imprecise or unfocused may result in the reviewer finding a large amount of potentially relevant literature, which will take a long time to screen in order to find the truly relevant articles for answering the question.

A well-structured question for a review of the effects of an intervention will typically specify the following key components: participants (*p*opulation or problem), *i*ntervention (or indicator), *c*omparator (control), and *o*utcome(s) of interest. The acronym 'PICO' is therefore often used as a reminder of these specific aspects of the question and its use can be illustrated as follows: 'In patients receiving treatment for clinical tuberculosis (P) does a policy of direct observation of treatment (I) versus a policy of self-treatment (C) lead to different treatment success rates (O)?'

The PICO approach can also be used for formulating review questions that are not related to the benefits or harms of interventions. Regarding aetiology or risk, one might, for example, ask: 'Are women (P) who smoke during pregnancy (I) compared to women who do not smoke (C) more likely to give birth to children who develop asthma during the first 5 years of life (O)?'

Once a review question has been constructed, it can be converted into one or more specific review objectives. Thus, with reference to the first example of a review question given earlier, the objective could be: 'To compare directly observed therapy with self-administration of treatment on treatment success rates in patients receiving treatment for clinically active tuberculosis.'

Identifying relevant literature and setting eligibility criteria

The most time-consuming step of a systematic review is usually the identification of studies which should be included in the review. This requires the development of an efficient search strategy, a process which begins with the choice of study eligibility criteria.

Building on the defined review question and associated objectives, eligibility criteria for inclusion or exclusion of studies in the review can be specified. These eligibility criteria are usually defined in terms of the type of participants, interventions, and comparisons (and less frequently the outcomes). In addition to these 'PICO' components, the types of study to be included in the review are usually also specified at this stage. Describing each of these components in some detail will assist the

development of the search strategy, as well as the selection of relevant studies. An example of this is provided in Box 5.15.1.

Box 5.15.1 Study eligibility criteria specified in a review of adenoidectomy versus non-surgical management or tympanostomy tubes in children with otitis media

- (P) The participants included children up to 18 years of age diagnosed with otitis media.
- (I/C) Studies that compared adenoidectomy versus non-surgical management only (defined as watchful observation and medical treatment): adenoidectomy and unilateral tympanostomy tube versus unilateral tympanostomy tube only; and adenoidectomy with bilateral tympanostomy tubes versus bilateral tympanostomy tubes only.
- (O) The primary outcome was the proportion of time children had effusion, diagnosed with or without tympanometry. Secondary outcomes were occurrence of acute otitis media, occurrence of otitis media with effusion, and the mean hearing level. Occurrence of the condition was defined as the number of episodes per year, number of days per episode per year, and the proportion of children with recurrent episodes.
- (S) Only randomized trials were considered for inclusion and only those that had follow-up of participants of at least 6 months.

Source: data from Van den Aardweg MTA et al., Adenoidectomy for otitis media in children, *Cochrane Database of Systematic Reviews* 2010, Issue 1, Art. No.:CD0077810, Copyright © 2010 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

Defining the participants or problem (P)

Participants could be described in terms of sociodemographic characteristics, such as age, sex, ethnicity, and educational status, as well as the setting in which the study occurred (e.g. hospital or community based). It may also be important to provide a specific definition of the target health condition under consideration. Where restrictions based on population characteristics (e.g. people living in low- and middle-income countries) are used, these need to be explained clearly. It is often better to include studies covering a wide spectrum of participants, because differences across different groups can always be explored in the review at a later stage. If a narrow range of patients is specified at the outset of

the review, there will be no such variability to explore at later stages of the review.

Defining the types of interventions (I) and comparisons (C)

For reviews of the effects of interventions, it is important to clearly define the intervention being evaluated: what is being delivered, at what intensity, how often, who delivers it, for how long, when it is delivered and what training was given to people involved in the delivery?

Reviewers also need to specify whether the intervention will be compared with an inactive or 'neutral' control intervention (e.g. placebo, standard care, no treatment) or an active control intervention (e.g. different regimen of the same drug or a different kind of therapy).

Defining the types of outcomes (O)

All outcomes that are relevant to end-users of the review, such as patients, healthcare practitioners, and policymakers, should ideally be included. For this reason, outcomes usually do not form part of the eligibility criteria for deciding whether or not an individual study should be included in the review. Reviewers are encouraged to include important outcomes even if they have not been addressed in the original studies, because this could help identify gaps in the evidence base that need to be addressed in future research.

Further, it may be necessary to consider whether the outcomes of interest are measured objectively (e.g. death or number of strokes) or subjectively as rated by clinician, patient, or carer (e.g. disability scales), and whether measuring instruments would have required validation. The timing of outcome measurement may also be important to define, for example, short-term compared to long-term outcomes.

It is advisable to separate outcomes into primary outcomes that are critical for informing decision-making (or that can be measured objectively) and secondary (additional) outcomes that may be of interest for other reasons (or that are less objectively assessed). Reviews should evaluate at least one desirable (beneficial) and one undesirable (adverse) outcome as both are important for making informed healthcare decisions. Various frameworks to evaluate adverse effects in systematic reviews have been proposed (Loke et al. 2007; Guyatt et al. 2011). Reviewers should also consider using pre-defined core outcome sets to help them to decide which outcomes to use in their review (Williamson et al. 2012; Kirkham et al. 2013).

Defining the study design (S)

Based on the review question, a determination needs to be made about the types of study design that are likely to provide the most reliable answers. For instance, it may be appropriate to limit the review to randomized trials in the case of questions about the effects of an

intervention, to cross-sectional studies for reviews of diagnostic accuracy, or to cohort studies if the intention is to answer a question about disease prognosis. Reasons for including and excluding specific study designs need to be made explicit in the review.

In the real world, researchers sometimes conduct reviews using study designs that are not ideally suited to a particular question, perhaps because more reliable studies do not exist. This can pose challenges to the interpretation of the evidence, because of inconsistency of the findings, which limits decision-making. The case of male circumcision as a public health intervention for the prevention of HIV acquisition in heterosexual men provides an instructive example of such a scenario. In 1995, a Cochrane Review commissioned by the World Health Organization assessed all 37 observational studies that had studied this relationship (Siegfried et al. 2005). Although many studies showed an association between male circumcision and HIV prevention, the review documented substantial between-study variation in the findings. This was particularly marked in studies conducted in the general population, as opposed to those conducted among high-risk groups. The authors, because of their concerns about residual confounding in these studies, concluded that the evidence available was insufficient to guide policy and that the findings of randomized trials should be awaited. A few years later, a Cochrane Review update (Siegfried et al. 2009) evaluated the results of the three trials which had been completed subsequent to the publication of the original review. It found strong evidence that medical male circumcision reduces the acquisition of HIV by heterosexual men by between 38 per cent and 66 per cent over 24 months, with a low incidence of adverse events. Fig. **5.15.1** presents the findings of the observational studies and trials conducted in the general population, showing the marked differences in consistency in their results, as well as the results of the subsequent randomized trials.

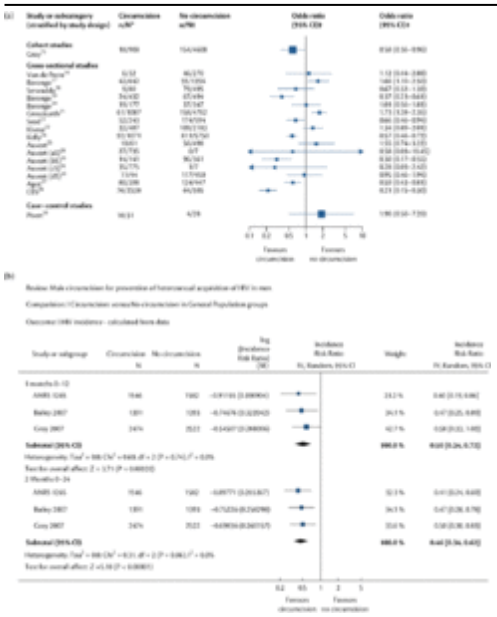

Fig. 5.15.1
The relationship between male circumcision and heterosexual acquisition of HIV in men in studies conducted in the general population. (a) Observational studies. Reprinted from *The Lancet Infectious Diseases*, Volume 5, Issue 3, Siegfried N et al., HIV and male circumcision–a systematic review with assessment of the quality of studies, pp.165–73, Copyright © 2005, with permission from Elsevier, http://www.sciencedirect.com/science/journal/14733099/5/3 (b) Randomized trials. Reproduced with permission from Siegfried N, Muller M, Deeks JJ, Volmink J, *Male circumcision for prevention of heterosexual acquisition of HIV in men. Cochrane Database of Systematic Reviews* 2009, Issue 2. Art. No.: CD003362. Copyright © 2009 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

Search strategy

Once the study eligibility has been determined, the reviewer can proceed to the development of the search strategy. The aim should be to develop a search strategy which has high sensitivity (defined as the number of relevant reports identified divided by the total number of all relevant reports in existence) and high precision (i.e. the number of relevant reports identified divided by the total number of reports identified) (Lefebvre et al. 2011). In order to reduce the impact of reporting biases, it is important to search for articles published in any language, as well as those that are unpublished or available only in the grey literature (Hopewell et al. 2007c).

In the biomedical sciences, there are three key electronic sources of published reports of clinical trials: MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed), EMBASE (http://www.elsevier.com/online-tools/embase), and CENTRAL (Cochrane Central Register of

Controlled Trials) (http://www.thecochranelibrary.com). A comprehensive search strategy requires each of these databases to be searched (Lefebvre et al. 2011) and others as appropriate to the review. As these three databases tend to have a North American and European bias, it might also be worth searching other databases such as LILACS (Latin American Caribbean Health Sciences Literature) (http://lilacs.bvsalud.org/en/) and African HealthLine (http://www.nisc.co.za/databases?id=62). It may also be important to include subject-specific databases, such as PsycINFO (www.apa.org/pubs/databases/psycinfo/index.aspx) (for psychology and related behavioural and social sciences), as well as Internet search engines, such as Google Scholar (http://scholar.google.co.uk/).

Strategies can be used to identify both free-text words in the database, and controlled terms (called MeSH, or Medical Subject Headings, in MEDLINE and EMTREE in EMBASE) which are used as keywords. Search strategies need to include the key terms in the review question, and use the Boolean operators ('AND', 'OR') to produce a search that is most likely to find relevant studies. Because of the complexities inherent in designing search strategies, it is advisable to seek assistance from librarians or information specialists before embarking on a literature search (Lefebvre et al. 2011).

The next step in searching for relevant articles involves screening of reference lists of all studies identified from the electronic search, including previously published reviews, to identify any missed studies (Horsley et al. 2011). It is also legitimate to conduct handsearching of specific journals that are not indexed in electronic databases to make sure that no studies that meet the review inclusion criteria have been missed (Hopewell et al. 2007a). Finally, unpublished studies and 'grey literature' such as conference abstracts, research reports, policy documents, dissertations, book chapters, and personal correspondence, should be sought by contacting authors of identified papers and relevant organizations, including pharmaceutical companies and by searching websites (Hopewell et al. 2007c). It is also important to identify ongoing studies, for example, by searching prospective trials registers, as these may be needed for future updates of the review (Ghersi et al. 2009).

Application of the search strategy typically identifies many (sometimes thousands of) potentially relevant papers. At this stage, two reviewers should read the title and abstract of each report to determine whether the study should be included or excluded based on whether or not it meets the pre-specified eligibility criteria (Edwards et al. 2002). Where there is uncertainty, a copy of the full paper needs to be obtained for further scrutiny. The two reviewers should compare their results, and resolve any differences through consensus or by using a third person to arbitrate when required (Higgins and Deeks 2011). The search strategy should be carefully conducted to ensure reproducibility and its outcome

clearly presented, ideally by means of a flowchart as recommended by PRISMA (Fig. **5.15.2**) (Liberati et al. 2009).

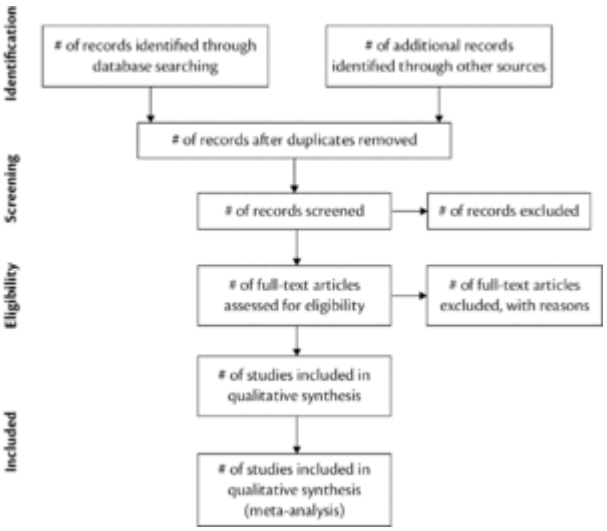

Fig. 5.15.2
Flow diagram of study selection process (from PRISMA guidelines).

Data collection and extraction

Deciding on what data to collect

The specific items of data to be collected from included studies will be influenced by the goals of the systematic review. However, the following items should generally be considered (Higgins and Deeks 2011):

- ◆ Eligibility of the study for inclusion or reason for exclusion.
- ◆ Study methods, such as study design and duration, as well as items that will be used to assess the risk of bias, such as sequence generation, allocation concealment, and blinding.
- ◆ Data regarding participants (e.g. sex, age, ethnicity, and education level) and setting (e.g. institution and geographic location) which may influence the size of the effect estimate or assist in assessing the applicability of results.
- ◆ The number of intervention groups and the details of the specific interventions involved. Any information relating to the fidelity with which the intervention was administered (i.e. the extent to which it

was implemented as planned) should be obtained. This is particularly important in preventive or complex interventions, which are usually implemented under 'real-world' conditions.

◆ Information regarding the outcomes such as their definitions, time points at which they were measured, and units of measurement should be collected. Usually, this is done only for those outcomes pre-specified in the review protocol. However, listing all other outcomes reported in the included study may enable analysis of risk of bias due to selective outcome reporting.

◆ Results should be collected only for those outcomes pre-specified in the protocol. For a meta-analysis it is also necessary to collect the data for analysis (see next section on data extraction).

◆ Other information to be collected could include funding source, details of ethics approval, name of contact of authors, citations, etc.

Once the data to be collected have been identified, it is necessary to design appropriate data collection forms. These may be electronic or paper-based, depending on the preference of reviewers. Paper-based forms allow extraction of data anywhere, and they are easier to create and use, whereas electronic forms offer the advantages of data extraction and entry in one step, programming of forms, and better storage, sorting, and retrieval of data in reviews with large number of studies (Higgins and Deeks 2011). Usually several forms, serving different purposes, are used.

Data extraction

Data extraction refers to the process of recording all the relevant data from the included studies onto data collection forms. Piloting data collection forms using a representative sample of studies is recommended. This will help to identify data that may not initially have been considered for collection, and to detect coding that might be unclear (Higgins and Deeks 2011).

To reduce the risk of errors during data extraction, it is advisable that at least two reviewers, working independently, are involved in extracting the data. Any disagreements between the reviewers should be resolved through consensus or arbitration by a third person. Blinding data extractors to study details, such as names of authors and results, may decrease the risk of bias (Jadad et al. 1996). However, as the evidence for this practice is not conclusive (Berlin 1997), routine blinding of data extractors is not currently recommended.

It is important to keep in mind that a single study may have been reported in more than one paper. Reviewers should be vigilant to the existence of the often covert practice of reporting the same data in different journals or even in the same journal at different times (Von Elm et al. 2004). The inclusion of duplicate publications with repeated observations in a systematic review will introduce bias on account of undue weight given to the findings concerned (Tramer et al. 1997).

Where multiple papers arising from the same study report different data, it is, however, necessary to include all relevant articles in order to ensure that data extraction is complete.

The type of summary data needed for subsequent meta-analysis will vary according to the type of outcome involved. For dichotomous outcomes, the numbers of participants who did and did not experience the outcome in each intervention group should be extracted. In the case of continuous outcomes, the mean and standard deviation, as well as the number of participants in each group, will be needed.

Where the desired summary measures are not reported, they may be derived from other statistics. For example, numbers of participants may be derived from percentages or effect estimates, such as odds ratios. Sometimes study authors will have to be contacted for clarification.

Assessing risk of bias

Once the data extraction is completed, each study should be carefully assessed for the risk of bias in the design, conduct, analysis, or reporting of a study. The main aim of this exercise is to guide interpretation of the findings of the review. In extreme cases, one may choose to exclude a study which is so fundamentally flawed that its results are not valid.

There are several methods for assessing the risk of bias. The Cochrane Collaboration recommend a ‘domain-based evaluation’ for clinical trials in which judgements are made about the risk of bias in each of the following domains: selection bias, performance bias, detection bias, attrition bias, reporting bias, and other bias (Table 5.15.2).

Table 5.15.2 Domain-based scheme for classification of bias in included studies

Type of bias	Description	Relevant domains
Selection bias	Systematic differences in the 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, personnel and outcome assessors • Other potential threats to validity
Attrition bias	Systematic differences between groups in withdrawals from study	• Incomplete outcome data • Blinding of participants, personnel and outcome assessors
Detection bias	Systematic difference between groups in how outcomes are determined	• Blinding of participants, personnel and outcome assessors • Other potential threats to validity
Reporting bias	Systematic differences between reported and unreported findings	• Selective outcome reporting

Adapted with permission from Higgins JPT, Altman DG, and Sterne JACon behalf of the Cochrane Statistical Methods Group and the Cochrane Bias Methods Group, Chapter 8 'Assessing risk of bias in included studies' in Higgins JPT and Green S (eds), *Cochrane Handbook for Systematic Reviews of Interventions,* Version 5.1.0 updated March 2011, The Cochrane Collaboration, Copyright © 2011, available from http://www.cochrane-handbook.org.

A judgement relating to the risk of bias for each item is assigned, which is supported by a description of what was done in the study. The judgement is indicated by answering a pre-specified question (e.g. 'Was the allocation sequence adequately generated?') which requires a 'yes' representing a low risk of bias, 'no', a high risk of bias, and 'unclear' an unclear or unknown risk of bias. Criteria for judging the risk of bias are provided in The Cochrane Collaboration's 'risk of bias tool'. The final step involves the risk of bias assessments being summarized across all domains for each study, and across all studies by indicating the proportion of domains studies with low, unclear, or high risk of bias.

Risk of bias is assessed based on the information in the reports of trials. Sometimes, the reported information may be insufficient to make critical judgements about the methods of a study. In these cases, it is important to contact the study authors to clarify any issues regarding study design or missing data that are not described clearly in the report (Mullan et al. 2009; Higgins et al. 2011).

Summarizing the evidence

One of the reasons for conducting a systematic review is that the individual studies are too small to provide a reliable and robust answer on their own. The review seeks to bring together the studies of relevance to the question in order to avoid undue emphasis on the findings of any single study, to minimize bias and to maximize statistical power by combining all the relevant data if this would be appropriate. This will help to overcome the biases of focusing on the findings of individual studies, and the effects of chance that may lead to an over- or underestimate of the true effect. Even when a randomized trial is well conducted and has minimized the possibility of biases that might overwhelm any true difference between the interventions being compared in that study, it is still susceptible to the effects of chance and the mathematical testing that is done to assess this cannot provide certainty on whether or not the result was distorted by chance, or on the size or direction of any distortion. However, combining the data from multiple studies in a meta-analysis will serve to minimize the effects of chance, such that any over- or underestimate is likely to be smaller than that in a single study. Even when the mathematical combination of the results of studies is not done, presenting them all in the same place and in the same way helps the reader to compare and contrast the studies in the review.

This provision of a summary of the evidence helps the systematic review fulfil the purpose of placing existing research in context, ensuring that new research is designed and implemented in the most appropriate way. And, when systematic reviews are used as part of the presentation of the findings of a new study, it ensures that those findings are seen in proper context (Clarke et al. 2010).

In deciding on how to summarize the evidence identified for the review, the reviewer might use a quantitative approach seeking to combine the results in a meta-analysis, a qualitative approach in which each study's findings are kept separate, or both. The choice about these techniques depends to some extent on the heterogeneity in the outcomes measured across the studies, since if there is too much variation in the outcomes that were measured, the measurement tools that were used, how or when they were used, or in how the results are made available, it might be impossible to combine the data. Efforts are underway in many areas of healthcare to help overcome this by reaching consensus on standardized core outcome sets, and the COMET Initiative is bringing such examples together and facilitating the development of further core outcome sets (Williamson et al. 2012), but the challenges in the existing literature can be substantial. These challenges, which might prevent the combining of data in the review, can also make it much more difficult to compare and contrast the included studies.

Regardless of whether they will do a meta-analysis, reviewers need to decide how much or how little information to extract and report for each study, and the sources to be used, especially if published reports are inadequate (Clarke and Stewart 1994). In compiling as complete a dataset as feasible and sensible, the principles of minimizing systematic biases and chances effects must be applied. If a meta-analysis is to be done, compatible data should be sought for all relevant trials and, if these data are not available, the proportion of missing data needs to be small enough to allow the reviewer and the reader to feel confident that the review's findings are robust. In particular, the reviewers need to consider whether the unavailability of the data is due to results-related bias. If it is, the meta-analysis of available data will itself be biased, leading to potentially misleading overall findings and conclusions.

There are several benefits of doing a meta-analysis if it is appropriate so to do. It provides a more precise estimate of the treatment effect, helping one to be more confident about the size of any effect. There is more statistical power to detect small effects, which may be clinically significant; and it helps in the evaluation of the applicability of the results to other settings. However, one of the important decisions that a reviewer has to take is whether the trials are similar enough that averaging their results is meaningful. This is done by considering whether there is excessive heterogeneity in the design of the studies (including the interventions and participants studied). Statistical tests are also available

to assess whether the results of a series of trials might differ from each other by more than chance (Higgins et al. 2003).

In seeking data and other information from the original researchers to prepare their summary, the reviewers might try to gather aggregate data, for example, by asking the original researchers to complete a table; or they might seek data at a finer level, such as that for individual participants. This gathering of data from the original researchers might make the dataset available for the review more complete, up to date, and accurate than anything that was previously available. It also helps the review to perform standardized analyses across the studies (Stewart et al. 1995).

A variety of techniques for combining results from separate studies in a meta-analysis is available to the reviewer (Cooper and Rosenthal 1980; Deeks 2002). The overriding principle of most of these is that each study is analysed separately and the overall result for the review comes from combining these results from the individual statistics. In this two-step approach, participants in one study are only directly compared with others in the same study and it is now fairly standard for the results of these analyses to be shown as a forest plot (e.g. Fig. **5.15.1**). This allows the reader to see the contribution of each study (Lewis and Clarke 2001; Glasziou and Sanders 2002; Higgins et al. 2003).

In planning and conducting the statistical analyses for any review, careful consideration needs to be given to the types of analyses, the options available for these analyses, and the reliability of the overall average. Once again, the effects of chance and bias should be considered carefully. This is for the overall results and for any subgroup analyses that are conducted; especially if decisions about the analyses to conduct and present are due to fore-knowledge of the likely findings (Counsell et al. 1994). This can be a particular problem for systematic reviews, since they are retrospective research and some or all of the findings of the included studies might already be known to the reviewers before they reach the stage of planning or conducting the analyses.

Subgroup analyses present the reviewer with the dilemma that having brought together the data from multiple studies, they might now wish to break it apart again into subgroups to explore the effects in different patients or settings, or when different interventions were used. This needs to be done with caution because of the possibility that spurious, chance results will be obtained; which will be misinterpreted as being of importance in making decisions about healthcare (Counsell et al. 1994; Clarke and Halsey 2001; Bender et al. 2008). Even if there is an a priori reason to expect a subgroup analysis to show something different to the overall result, this is no guarantee that a statistically significant difference is reliable clinically. This is because the more analyses are done, the more likely it is that statistically significant results will be found, even when there is truly no difference between the subgroups. Subgroup analyses in a systematic review should be regarded as a way of

showing that the direction of effect is the same across different types of patient or as a generator of a hypothesis for testing in future research. Regardless of whether subgroup analyses are done, it is often more reliable to assume that the overall result is as good, and if not a better, estimate of the relative effects of treatments in the particular type of patient than that obtained by looking at the results for just these types of patient in the review. This is because the effect of chance will be smaller for the overall result than it would be on the result in any subgroup.

Systematic reviews might also include sensitivity analyses, which ideally should also be planned in advance. A sensitivity analysis is used to determine how sensitive the results of the systematic review are to the decisions that the reviewer took about how the review was done. They are particularly useful where there is uncertainty about the choices that a reviewer needs to make. For example, sensitivity analyses could be used to determine the effect of including studies published in languages other than English, of using data from studies assessed to be of poor quality, or of choosing one statistical technique over another.

The multiplicity which might arise in a systematic review because of the variety of outcomes, and measures of these outcomes, available is compounded further by the variety of ways to analyse these in order to obtain an estimate of the difference between the intervention and the comparator. If a meta-analysis is done, this multiplicity will arise at the level of the overall estimate of effect. If the results of each study are kept separate, with no combination of these in a meta-analysis, it arises at the study level. Typical effect measures in systematic reviews that use dichotomous, or binary, data are estimates of risk (either a ratio or absolute difference) and odds (almost always a ratio).

The decision on which effect measure to use should, ideally, be taken when the systematic review is being planned. It should be based on the mathematical appropriateness of the effect measure for the type of data to be analysed. As with any of the decisions when conducting a systematic review, changes to this a priori plan should be kept to a minimum, and should be documented and reported. Any changes that are made because of the reviewers' preference for one result over another need to be treated with particular caution, because of the risk of bias. In meta-analyses of dichotomous data, it is most common for the odds ratio, risk ratio, or both to be used (Deeks and Altman 2001).

In some cases, the original intention for the effect measure will turn out to be mathematically inappropriate or suboptimal because of the nature of the data identified for inclusion in the review. This should be dealt with and described in a transparent way by the reviewers. They should avoid conducting multiple exploratory analyses using different effect measures and then focusing on the analysis that generates their preferred result. In circumstances where a variety of effects measures are appropriate, the different analyses may be conducted, and all that are conducted should be reported. The greatest confidence in the findings of the review would

then arise if there is consistency among the results from the different effect measures. If the results differ in important ways, the reviewers and the users of the review need to be cautious.

If the meta-analysis uses continuous data, the typical effect measure will be the mean difference. If the data from the included studies relate to the same outcome but they were measured in different ways, for example, by using a variety of scales to measure satisfaction with the provision of care, this needs to be standardized. The reviewers might also have the option of dividing the continuous data into ‘high’ and ‘low’ in order to dichotomize the outcome and, thereby, use effect measures such as the risk ratio or odds ratio. Particular caution is needed in such circumstances, especially if there is no natural or well-accepted threshold for this splitting of the data.

Writing up the review

In recent years, there has been increasing recognition of the value of guidelines on the reporting of a wide variety of types of research study, including randomized trials (Schulz et al. 2010), observational research (Von Elm et al. 2007), and studies of diagnostic test accuracy (Bossuyt et al. 2003). The principal guide for the reporting of systematic reviews is PRISMA (Moher et al. 2009), which updated the QUOROM statement (Moher et al. 1999).

This section of this chapter outlines some of the key elements in the reporting of a review, which should help both with the preparation of the report and with its use. When preparing a report of a systematic review, though, it is important to bear in mind what the reader will be looking for. The traditional format of scientific papers of Introduction, Methods, Results, and Discussion is used in the following, but the reports of some reviews might change this order, or move some content to appendices or supplementary papers, depending on the interests and needs of their readers.

Title for the review

The title might flow naturally from the question for the review, containing elements relating to the population being studied, interventions or actions, the outcome of most interest, the study designs incorporated into the review, and the fact that the research being reported is a systematic review. The title should help the reader to decide whether the review is potentially relevant to them and might be particularly important in ensuring that the review is retrieved by people searching for research in the particular area. Some titles might be declarative, very briefly summarizing the main finding of their review, but this might not be possible in reviews covering more than one intervention, population, or outcome, where there are multiple important findings.

Question and purpose for the review

One of the distinguishing features of systematic reviews over more traditional reviews is the use of a clear research question at the design stage of the review. Formulating and clarifying the question at the outset of the review and at the start of its report is important for the reviewer, and then for the reader. The question will also help to reveal the purpose of the review, which might have been to provide evidence to guide future practice, or to map out existing research to help design a new, definitive study. For example, the question might be ‘What are the effects of an incentive scheme to encourage exercise in the adult population?’, or ‘What types of incentive scheme to encourage exercise in the adult population have been tested in trials?’, or ‘How have trials of incentive schemes to encourage exercise in the adult population been assessed?’.

A systematic review might contain more than one question and comparison, and this should also be made clear early in the report. For example, the review might cover the comparison of an intervention with usual care, comparisons of different forms of the intervention, or comparisons (direct or indirect) between the intervention and alternative strategies.

Background

Setting the scene for the review will help orientate the reader to the topic and is also an opportunity to explain the importance and relevance of the review. The Background section might provide details on how common the circumstances or condition under investigation are; how the intervention was developed, how it might work and how common it is in practice; and any relevant studies that are already likely to be well known. Information on the consequences and severity of the outcomes being assessed might be helpful. It might also be important to set out the reasons for conducting the review in a particular way, especially if there are alternatives that might be in the minds of the reader. If the review had a pre-prepared protocol or was prospectively registered, this could also be stated in the background section.

Inclusion and exclusion criteria

In providing the eligibility criteria for a review, it may be important to be clear that these are what the reviewers sought, which will not necessarily be matched completely by what they found and were able to include in the review. For example, the review might have looked for a variety of interventions to protect factory workers from noise but found that only a few of the commonly used strategies had been tested in research (Verbeek et al. 2012). For most reviews, the eligibility criteria can be set out under domains, as discussed earlier in the chapter in relation to the use of PICO. This would identify the types of participants, interventions, actions or strategies, outcome measures, and types of study design.

The types of participants should include information on the types of people that would be eligible. This might contain information about personal characteristics such as sex, age, socioeconomic group, and diagnosis, as well as information about their setting. The types of interventions, actions, and strategies covers what was to happen to the participants, and needs to contain information on any interventions, actions, and strategies that were used for comparative purposes. Types of outcome measures might be divided into primary and secondary outcome measures. It may also be helpful to distinguish ‘outcomes’ (e.g. depression) from ‘outcome measures’ (e.g. the instrument used to measure depression and the timing and application of the instrument). If there is a core outcome set relevant to the circumstance of the review, this might be referred to in this section (Williamson et al. 2012). Finally, the types of studies sets out which research designs were eligible for review.

Within the eligibility criteria, providing both inclusion and exclusion criteria can help clarify what would definitely be eligible for the review, and what would not be. This should not simply include ‘mirror images’ of the inclusion criteria in the exclusion criteria. For example, if the inclusion criteria require that people were living in residential care homes, it is superfluous to mention the exclusion of people who live in other settings. Rather, the exclusion criteria might be used to note that a subpopulation of people who would otherwise be eligible are not eligible, such as people in the care homes who are under the age of 60 years.

Methods

The authors of the review should describe what they had intended to do, and what they were able to do given the material that they found for their review. If they had a protocol for the review, this could be referred to here. As a minimum, the Methods section should provide details of the search strategy, how eligibility was determined, methods for data extraction or collection, and the techniques used to summarize the findings, which—as discussed earlier—might be quantitative, qualitative or both. Strategies used within the review to minimize bias, such as blinding during data extraction, checking of reference lists and obtaining unpublished data, and to minimize error, such as independent assessment for eligibility should be described. Changes to the methods between the original plans or the protocol for the review, and the final review should be noted and explained.

Results

The Results section might begin with details of what was found for the review, separating studies that had been judged to be eligible, ineligible, and eligible but ongoing (with results not available). This might include the PRISMA flow diagram to show the attrition at each stage in the process from original search to final inclusion in the review’s analyses. Details should be provided on each included study, which might vary in

depth from a thorough description of several features of each study, to a brief summary and details of sources of more information, such as the citation for a published report, link to a study website or to the entry for the study in a research register.

Studies that were judged eligible but were excluded, and studies which readers might expect to have been included but weren't, should be reported along with a reason for their exclusion. This reassures the reader that these studies were identified by the reviewers. Failure to mention them might raise concerns in the reader's mind about, for example, the robustness of the searches for the review or the potential for reporting bias.

Whether or not the review includes quantitative or qualitative synthesis of the included studies, if the reviewer presents sufficient detail for each study and its findings, this would allow the reader to repeat such syntheses for themselves, perhaps changing how the studies are grouped or leaving out some studies. The use of forest plots for each meta-analysis is a fairly straightforward way to both show the proportional contribution of each study and present its numerical results (Lewis and Clarke 2001).

Conclusions

The next section in this chapter goes into more detail on the interpretation of the findings of a systematic review but some of the things for the reviewers to consider in their Discussion section are outlined here. They need to be cautious about introducing new evidence at this point to answer the question if, for example, there was insufficient information available in the studies included in the review. Having been so careful in planning the systematic review itself, it may be undermined by non-systematic approaches to introducing alternative sources of evidence for the review's primary question in the Conclusion section. Where relevant, the reviewers' conclusions might be separated into their opinions on the implications of their review for practice and for further research. The latter might include specific suggestions on the types of study that are needed now, given the existing research that has been summarized in the review (Brown et al. 2006; Clarke et al. 2007).

In a complex or large review of the effects of interventions, the reviewers might choose to summarize their conclusions by highlighting where they have found convincing evidence of which interventions are beneficial, harmful, are lacking in evidence and need more research, or are lacking in evidence from the studies in their review but which probably do not warrant more research. If the review is not about interventions but concerns, for example, the relationship between characteristics of people and their behaviour, a similar categorization of the conclusions on the basis of confidence about a positive or negative relationship, one that is unproven and worthy of more research, and one that is unproven but does not justify more research, might be helpful. This categorization of

the interventions might help policymakers to focus in on those that might be introduced or reinforced, abandoned (Garner et al. 2013), or restricted to use in research.

Interpreting the findings

When interpreting the findings of a review, either as the person preparing the review or someone reading it, it is important to remember some of the key principles underpinning evidence-based healthcare (Sackett et al. 1997). The systematic review is a source for one part of the evidence needed for well-informed decision-making. In the context of systematic reviews of the effects of interventions, this will be an estimate of the difference between the alternative interventions. In the context of, for example, associations between the presence of a characteristic, such as social deprivation, and an outcome, such as diabetes, it will be the estimate of the strength of this association. Other types of evidence are also needed to make a well-informed decision, including information on diagnosis, expertise, feasibility, values, and preferences. Some of this evidence might be provided by additional systematic reviews of other types of research, which could be incorporated into the main review, but some will come from the target population to whom the review will be applied. This makes it important for the review to provide the information needed to use its findings in practice and to interpret the findings and recommendations cautiously.

One way to decide on the potential relevance of the review to settings other than those in the included studies, is to consider whether there are any strong reasons why the participants in the studies included in the review, the interventions assessed in these studies, and the outcome measures reported are not sufficiently similar to the people, interventions, and outcome measures of relevance in other circumstances and the values and preferences of the population for whom the results will be applied.

Some examples of the decisions that might need to be made about the possible applicability of the findings of the review are if the evidence in the review is from research done in academic research hospitals in large cities but it will be used in decisions for clinics in rural settings; the trials were done in North America and the burden of the problem is greatest in Africa; or the patients in the trials were from a particular, high socioeconomic group but the findings will need to be applied to other groups.

Earlier in this chapter, the challenges of subgroup analyses were noted and one possibility for applying the results of a heterogeneous review to a target population is to consider the results within the target population subgroup. As noted, this needs to be done with caution because of the problems of lessening statistical power, multiplicity, and bias. In fact, the more reliable guide to the effects for a patient in the subgroup of interest might be the overall average from the meta-analysis, rather than the specific result from the subgroup analysis. One way to assess this and to decide between the overall result and the result of a specific subgroup is to consider whether the rationale for the subgroup analysis is to identify different effects among the subgroups or to identify similarities. For

instance, do the subgroup analyses provide reassurance that the difference between the treatments is similar regardless of specific patient characteristics (such as age, setting, or socioeconomic group)? Or, do they show that one intervention will be beneficial in one type of patient but harmful in another? This might happen if a treatment has a beneficial effect for one outcome but increases the risk for another. The treatment might be regarded as beneficial, on average, for people with few risk factors for the latter; but harmful in those at high risk of it.

As well as considering whether or not there are substantial differences between the participants in the studies in the review and people for whom it will be applied, it is also important to consider whether the interventions in the review are similar enough to the ones that would be used in practice. This might be straightforward if the interventions were drugs that can be administered in a standard way but it might be more difficult for interventions that are more skill based, such as surgery or psychotherapy; or for complex interventions, such as a community intervention strategy to reduce criminal behaviour or a training programme for nurses. An additional challenge here is that information on how the drug should be administered might be well described in the included studies and the review, while the details on a complex intervention might be inadequate for implementation.

Assessing the quality of a systematic review

There are a number of guides to help readers assess the quality of systematic reviews (Katrak et al. 2004). Three of the commonest ways to assess the quality of systematic reviews are tools based on the Overview Quality Assessment Questionnaire (OQAQ) developed by Andy Oxman and Gordon Guyatt (1991), the PRISMA checklist (Liberati et al. 2009; Moher et al. 2009), and the AMSTAR tool (Shea et al. 2007, 2009). As part of the appraisal of the systematic review and when deciding upon its relevance to a health or social care decision, the user should also consider whether the review is sufficiently up to date. To be truly up to date, a systematic review would need to have been completed just before its findings were to be used and would need to include all the relevant research available at that time. This is not practical. It would require reviews to be revised as soon as new evidence became available. Although this might be possible if updating the findings of a systematic review simply involved the insertion of data from the latest evidence into a meta-analysis, the separate chapter on systematic reviews shows that the review process is more complex than the statistical combination of the results of the included studies. Reviewers need to apply their eligibility criteria, appraise the quality of the studies, make decisions about the suitability of the outcome data, and draw conclusions about their findings. This will take time. It also needs to be preceded by searching for studies and the assessment of the retrieved records for potential relevance. In considering how up to date a review is, it is also important to consider whether the review is so early in the

evaluation of an intervention that the availability of studies for it was influenced by time-lag bias (Hopewell et al. 2007b), meaning that the conclusions are likely to change as this bias is corrected by increased availability of studies with less favourable findings.

Updating systematic reviews

Unless reviews are regarded as historical documents summarizing the evidence up to the time point of the review, reviewers and commissioners of reviews need to consider how the review might be kept up to date, so that it remains a valid source of knowledge for decision-makers in the future. For example, the intention for Cochrane Reviews is that these will be updated periodically and, at least, every 2 years or annotated to explain why this has not been done. The ideal, of course, might be that a review includes all relevant research available at the time that it is being used to inform a decision but this is not practical without a process for new evidence to be continually incorporated into existing reviews. The updating process takes time and can also serve to maintain the contemporary relevance of the review. This may be especially important if the review uses information that changes over time, such as economic costs, the organizational structures for delivering healthcare or the processes in which decisions are made about healthcare.

Acknowledgement

We thank Solange Durao of the South African Cochrane Centre for her help in preparing this chapter.

References

Address by the Rt. Hon. Lord Rayleigh (1885). In *Report of the Fifty-Fourth Meeting of the British Association for the Advancement of Science*, held at Montreal in August and September 1884. London: John Murray. Available at: http://www.jameslindlibrary.org.

Antman, E.M., Lau, J., Kupelnick, B., Mosteller, F., and Chalmers, T.C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *Journal of the American Medical Association*, 268, 240–8.

Atkins, S., Lewin, S., Smith, H., Engel, M., Fretheim, A., and Volmink, J. (2008). Conducting a meta-ethnography of qualitative literature: lessons learnt. *BMC Medical Research Methodology*, 8(21), 1471–2288.

Bender, R, Bunce, C., Clarke, M., et al. (2008). Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology*, 61, 857–65.

Berlin, J.A. (1997). Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. *Lancet*, 350, 185–6.

Booth, A., Clarke, M., Ghersi, D., Moher, D., Petticrew, M., and Stewart, L. (2011). An international registry of systematic-review protocols. *Lancet*, 377, 108–9.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (eds.) (2009). *Introduction to Meta-Analysis*. Hoboken, NJ: Wiley and Sons.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., et al. (2003). The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine*, 138(1), W1–12.

Britten, N., Campbell, R., Pope, C., Donovan, J., Morgan, M., and Pill, R. (2002). Using meta ethnography to synthesise qualitative research: a worked example. *Journal of Health Services & Research Policy*, 7(4), 209–15.

Brown, P., Brunnhuber, K., Chalkidou, K., et al. (2006). How to formulate research recommendations. *British Medical Journal*, 333, 804–6.

Chalmers, I. (2003). Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations. *Annals of the American Academy of Political and Social Science*, 589, 22–40.

Chalmers, I. and Altman, D.G. (eds.) (1995). *Systematic Reviews*. London: BMJ Publishing Group.

Chalmers, I. and Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *Lancet*, 374(9683), 86–9.

Chalmers, I., Hedges, L.V., and Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the Health Professions*, 25, 12–37.

Child, S., Goodwin, V., Garside, R., Jones-Hughes, T., Boddy, K., and Stein, K. (2012). Factors influencing the implementation of fall-prevention programmes: a systematic review and synthesis of qualitative studies. *Implementation Science*, 7, 91.

Clark, S. and Horton, R. (2010). Putting research into context—revisited. *Lancet*, 376(9734), 10–11.

Clarke, L., Clarke, M., and Clarke, T. (2007). How useful are Cochrane reviews in identifying research needs? *Journal of Health Services Research and Policy*, 12, 101–3.

Clarke, M. and Halsey, J. (2001). DICE 2: a further investigation of the effects of chance in life, death and subgroup analyses. *International Journal of Clinical Practice*, 55, 240–2.

Clarke, M., Hopewell, S., and Chalmers, I. (2010). Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. *Lancet*, 376, 20–1.

Clarke, M.J. and Stewart, L.A. (1994). Systematic reviews: obtaining data from randomised controlled trials: how much do we need for reliable and informative meta-analyses? *British Medical Journal*, 309, 1007–10.

Cooper, H.M. and Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442–9.

Counsell, C.E., Clarke, M.J., Slattery, J., and Sandercock, P.A.G. (1994). The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *British Medical Journal*, 309, 1677–81.

Damman, K., Navis, G., Voors, A.A., et al. (2007). Worsening renal function and prognosis in heart failure: systematic review and meta-analysis. *Journal of Cardiac Failure*, 13(8), 599–608.

Deeks, J.J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, 21, 1575–600.

Deeks, J.J. and Altman, D.G. (2001). Effect measures for meta-analysis of trials with binary outcomes. In M. Egger, G. Davey Smith, and D.G. Altman (eds.) *Systematic Reviews in Health Care: Meta-Analysis in Context*, pp. 313–35. London: BMJ Books.

Dickersin, K. (2010). To reform US health care, start with systematic reviews. *Science*, 329, 516–17.

Doust, J.A., Glasziou, P.P., Pietrzak, E., and Dobson, A.J. (2004). A systematic review of the diagnostic accuracy of natriuretic peptides for heart failure. *Archives of Internal Medicine*, 164(18), 1978–84.

Dwan, K., Altman, D.G., Arnaiz, J.A., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*, 3(8), e3081.

Edwards, P., Clarke, M., DiGuiseppi, C., Pratap, S., Roberts, I., and Wentz, R. (2002). Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in Medicine*, 21, 1635–40.

Egger, M., Davey Smith, G., and Altman, D.G. (eds.) (2001). *Systematic Reviews in Health Care: Meta-Analysis in Context* (2nd ed.). London: BMJ Publishing Group.

Garner, S., Docherty, M., Somner, J., et al. (2013). Reducing ineffective practice: challenges in identifying low-value health care using Cochrane

systematic reviews. *Journal of Health Services Research and Policy*, 18, 6–12.

Ghersi, D. and Pang, T. (2009). From Mexico to Mali: four years in the history of clinical trial registration. *Journal of Evidence Based Medicine*, 2(1), 1–7.

Glasziou, P., Irwig, L., Bain, C., and Colditz, G. (2001). *Systematic Reviews in Healthcare: A Practical Guide*. Cambridge: Cambridge University Press.

Glasziou, P.P. and Sanders, S.L. (2002). Investigating causes of heterogeneity in systematic reviews. *Statistics in Medicine*, 21, 1503–11.

Guyatt, G.H., Oxman, A.D., Kunz, R., et al. (2011). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology*, 64, 395–400.

Higgins, J.P.T., Altman, D.G., Sterne, J.A.C., on behalf of the Cochrane Statistical Methods Group and the Cochrane Bias Methods Group. (2011). Assessing risk of bias in included studies. In J.P.T. Higgins and S. Green (eds.) *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* [updated March 2011]. The Cochrane Collaboration. Available at: http://www.cochrane-handbook.org.

Higgins, J.P.T. and Deeks, J.J. (2011). Selecting studies and collecting data. In J.P.T. Higgins and S. Green (eds.) *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* [updated March 2011]. The Cochrane Collaboration. Available at: http://www.cochrane-handbook.org.

Higgins, J.P.T. and Green, S. (eds.) (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* [updated March 2011]. The Cochrane Collaboration. Available at: http://www.cochrane-handbook.org.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327, 557–60.

Hopewell, S., Clarke, M.J., Lefebvre, C., and Scherer, R.W. (2007a). Handsearching versus electronic searching to identify reports of randomized trials. *Cochrane Database of Systematic Reviews*, 2, MR000001.

Hopewell, S., Clarke, M.J., Stewart, L., and Tierney, J. (2007b). Time to publication for results of clinical trials. *Cochrane Database of Systematic Reviews*, 2, MR000011.

Hopewell, S., Loudon, K., Clarke, M.J., Oxman, A.D., and Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or

direction of trial results. *Cochrane Database of Systematic Reviews*, 1, MR000006.

Hopewell, S., McDonald, S., Clarke, M.J., and Egger, M. (2007c). Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database of Systematic Reviews*, 2, MR000010.

Horsley, T., Dingwall, O., and Sampson, M. (2011). Checking reference lists to find additional studies for systematic reviews. *Cochrane Database of Systematic Reviews*, 8, MR000026.

Innvaer, S., Vist, G., Trommald, M., and Oxman, A. (2002). Health policy-makers' perceptions of their use of evidence: a systematic review. *Journal of Health Services Research & Policy*, 7(4), 239–44.

Jadad, A.R., Moore, R.A., Carroll, D., et al. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clinical Trials*, 17, 1–12.

Katrak, P., Bialocerkowski, A., Massy-Westropp, N., Kumar, V.S., and Grimmer, K. (2004). A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology*, 4, 22.

Khan, K., Kunz, R., Kleijnen, J., and Antes, G. (2011). *Systematic Reviews to Support Evidence-based Medicine* (2nd ed.). New York: CRC Press, Taylor & Francis Group.

Kirkham, J.J., Gargon, E., Clarke, M., and Williamson, P.R. (2013). Can a core outcome set improve the quality of systematic reviews? A survey of the Co-ordinating Editors of Cochrane review groups. *Trials*, 14, 21.

Lavis, J.N. (2009). How can we support the use of systematic reviews in policy making? *PLoS Medicine*, 6(11), e1000141.

Lefebvre, C., Manheimer, E., Glanville, J., on behalf of the Cochrane Information Retrieval Methods Group (2011). Searching for studies. In J.P.T. Higgins and S. Green (eds.) *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* [updated March 2011]. The Cochrane Collaboration. Available at: http://www.cochrane-handbook.org.

Lewis, S. and Clarke, M. (2001). Forest plots: trying to see the wood and the trees. *British Medical Journal*, 322, 1479–80.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., and Gøtzsche, P.C. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine*, 6(7), e1000100.

Loke, Y.K., Price, D., Herxheimer, A., for the Cochrane Adverse Effects Methods Group (2007). Systematic reviews of adverse effects: framework for a structured approach. *BMC Medical Research Methodology*, 7, 32.

Mays, N., Pope, C., and Popay, J. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research & Policy*, 10(Suppl. 1), 6–20.

McMichael, C., Waters, E., and Volmink, J. (2005). Evidence-based public health: what does it offer developing countries? *Journal of Public Health*, 27(2), 215–21.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., and Stroup, D.F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet*, 354(9193), 1896–900.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., and The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7), e1000097.

Mullan, R.J., Flynn, D.N., Carlberg, B., et al. (2009). Systematic reviewers commonly contact study authors but do so with limited rigor. *Journal of Clinical Epidemiology*, 62, 138–42.

Mulrow, C.D. (1987). The medical review article: state of the science. *Annals of Internal Medicine*, 106, 485–8.

Mulrow, C.D. (1995). Rationale for systematic reviews. In I. Chalmers and D.G. Altman (eds.) *Systematic Reviews*, pp. 1–8. London: BMJ Publishing Group.

Munro, S.A., Lewin, S.A., Smith, H.J., Engel, M.E., Fretheim, A., and Volmink, J. (2007). Patient adherence to tuberculosis treatment: a systematic review of qualitative research. *PLoS Medicine*, 4(7), e238.

Needleman, I.G. (2002). A guide to systematic reviews. *Journal of Clinical Periodontology*, 29(Suppl. 3), 6–9.

O'Connor, D., Green, S., and Higgins, J.P.T. (2011). Defining the review question and developing criteria for including studies. In J.P.T. Higgins and S. Green (eds.) *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* [updated March 2011]. The Cochrane Collaboration. Available at: http://www.cochrane-handbook.org.

Oxman, A.D. and Guyatt, G.H. (1991). Validation of an index of the quality of review articles. *Journal of Clinical Epidemiology*, 44, 1271–8.

Oxman, A.D., Lavis, J.N., and Fretheim, A. (2007). Use of evidence in WHO recommendations. *The Lancet* 369, 1883–9.

Pendlebury, S.T. and Rothwell, P.M. (2009). Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis. *The Lancet Neurology*, 8(11), 1006–18.

Pérez Velasco, R., Praditsitthikorn, N., Wichmann, K., et al. (2012). Systematic review of economic evaluations of preparedness strategies and interventions against influenza pandemics. *PLoS ONE*, 7(2), e30333.

Petticrew, M. (2001). Systematic reviews from astronomy to zoology: myths and misconceptions. *British Medical Journal*, 322, 98–101.

PLoS Medicine Editors (2011). Best practice in systematic reviews. The importance of protocols and registration. *PLoS Medicine*, 8(2), e1001009.

Renehan, A.G., Tyson, M., Egger, M., Heller, R.F., and Zwahlen, M. (2008). Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *The Lancet*, 371(9612), 569–78.

Sackett, D.L., Richardson, W.S., Rosenberg, W., and Haynes, B.R. (1997). *Evidence-Based Medicine: How to Practice & Teach EBM*. New York: Churchill Livingstone.

Schulz, K.F., Altman, D.G., and Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *PLoS Medicine*, 7(3), e1000251.

Shea, B.J., Grimshaw, J.M., Wells, G.A., et al. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, 7, 10.

Shea, B.J., Hamel, C., Wells, G.A., et al. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology*, 62, 1013–20.

Siegfried, N., Muller, M., Deeks, J., et al. (2005). HIV and male circumcision—a systematic review with assessment of the quality of studies. *Lancet Infectious Diseases*, 5(3), 165–73.

Siegfried, N., Muller, M., Deeks, J.J., and Volmink, J. (2009). Male circumcision for prevention of heterosexual acquisition of HIV in men. *Cochrane Database of Systematic Reviews*, 2, CD003362.

Silagy, C.A., Middleton, P., and Hopewell, S. (2002). Publishing protocols of systematic reviews: comparing what was done to what was planned. *Journal of the American Medical Association*, 287, 2831–4.

Stewart, L., Clarke, M., for the Cochrane Collaboration Working Group on meta-analyses using individual patient data (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*, 14, 2057–79.

Stewart, L., Moher, D., and Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. *Systematic Reviews*, 1, 7.

Straus, S. and Moher, D. (2010). Registering systematic reviews. *Canadian Medical Association Journal*, 182, 13–14.

Tramer, M.R., Reynolds, D.J., Moore, R.A., and McQuay, H.J. (1997). Impact of covert duplicate publication on meta-analysis: a case study. *British Medical Journal*, 315, 635.

Van den Aardweg, M.T.A., Schildre, A.G.M., Herket, E., Boonacher, C.W.B., and Rovers, M.M. (2010). Adenoidectomy for otitis media in children. *Cochrane Database of Systematic Reviews*, 1, CD0077810.

Verbeek, J.H., Kateman, E., Morata, T.C., Dreschler, W.A., and Mischke, C. (2012). Interventions to prevent occupational noise-induced hearing loss. *Cochrane Database of Systematic Reviews* 10, CD006396.

Von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gotzsche, P.C., and Vandenbroucke, J.P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Annals of Internal Medicine*, 147(8), 573–7.

Von Elm, E., Poglia, G., Walder, B., and Tramer, M.R. (2004). Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *Journal of the American Medical Association*, 291, 974–80.

Williamson, P., Altman, D., Blazeby, J., Clarke, M., and Gargon E. (2012). Driving up the quality and relevance of research through the use of agreed core outcomes. *Journal of Health Services Research and Policy*, 17(1), 1–2.