Producing reliable summaries of incontinence research: a 'hands-on' workshop on how to conduct a Cochrane systematic review

W7, 29 August 2011 09:00 - 12:00

Start	End	Topic	Speakers
09:00	09:15	Introduction	• Cathryn Glazener
09:15	10:00	Question formulation and protocol development	• Graham Mowatt
10:00	10:30	Critical appraisal and data extraction	• Mandy Fader
10:30	11:00	Break	None
11:00	11:15	Critical appraisal cont'd	All
11:15	11:45	Data synthesis, including meta analysis	• Jonathan Cook
11:45	12:00	Interpretation and reporting	All

Aims of course/workshop

The course will focus on an aspect of incontinence management chosen in advance. We shall take the participants through the key stages of undertaking a Cochrane-style sytematic review of controlled trials, including demonstrating Review Manager software and description of statistical methods

aims to address the following questions:

- How to formulate a question for a systematic review related to continence care
- How to search for eligible studies
- How to appraise the methodological quality of eligible studies
- How to abstract data from eligible studies
- How to synthesise and analyse data
- How to interpret findings and report them
- The opportunities available for contributing to the Cochrane Incontinence Group

Educational Objectives

- Understanding how a systematic review differs from a traditional review
- Converting a clinical uncertainty into a review question
- Practical experience of the processes involved in undertaking a systematic review
- Acquiring the ability to critically appraise the main components of a systematic review

The following extracts are taken from the Cochrane Handbook for Systematic Reviews of Interventions available online (reference below). Some references to see sections found in full version of handbook.

Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

The Cochrane Incontinence Group

University of Aberdeen, 2nd Floor Health Sciences Building, Foresterhill, Aberdeen, AB25 2ZD, UK

Tel: +44 1224 559244 Fax: +44 1224 554580 E-mail: June Cody j.cody@abdn.ac.uk

- **QUESTION FORMULATION AND PROTOCOL DEVELOPMENT**

Rationale for protocols

Preparing a Cochrane review is complex and involves many judgements. In order to minimize the potential for bias in the review process, these judgements should be made in ways that do not depend on the findings of the studies included in the review. Review authors' prior knowledge of the results of a potentially eligible study may, for example, influence the definition of a systematic review question, the subsequent criteria for study eligibility, the choice of intervention comparisons to analyse, or the outcomes to be reported in the review. Since Cochrane reviews are by their nature retrospective (one exception being prospective meta-analyses, as described in Chapter 19), it is important that the methods to be used should be established and documented in advance. Publication of a protocol for a review prior to knowledge of the available studies reduces the impact of review authors' biases, promotes transparency of methods and processes, reduces the potential for duplication, and allows peer review of the planned methods (Light 1984).

Rationale for well-formulated questions

As with any research, the first and most important decision in preparing a systematic review is to determine its focus. This is best done by clearly framing the questions the review seeks to answer. Well-formulated questions will guide many aspects of the review process, including determining eligibility criteria, searching for studies, collecting data from included studies, and presenting findings (Jackson 1980, Cooper 1984, Hedges 1994). In Cochrane reviews, questions are stated broadly as review 'Objectives', and specified in detail as 'Criteria for considering studies for this review'. As well as focussing review conduct, the contents of these sections are used by readers in their initial assessments of whether the review is likely to be directly relevant to the issues they face.

A statement of the review's objectives should begin with a precise statement of the primary objective, ideally in a single sentence. Where possible the style should be of the form 'To assess the effects of [*intervention or comparison*] for [*health problem*] in [*types of people, disease or problem and setting if specified*]'. This might be followed by one or more secondary objectives, for example relating to different participant groups, different comparisons of interventions or different outcome measures.

The detailed specification of the review question requires consideration of several key components (Richardson 1995, Counsell 1997). The 'clinical question' should specify the types of population (participants), types of interventions (and comparisons), and the types of outcomes that are of interest. The acronym PICO (**P**articipants, **I**nterventions, **C**omparisons and **O**utcomes) helps to serve as a reminder of these. Equal emphasis in addressing each PICO component is not necessary. For example, a review might concentrate on competing interventions for a particular stage of breast cancer, with stage and severity of the disease being defined very precisely; or alternately focus on a particular drug for any stage of breast cancer, with the treatment formulation being defined very precisely.

- **CRITICAL APPRAISAL AND DATA EXTRACTION**

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 Collaboration's '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.

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 allocating interventions to participants. Its success in this respect depends on fulfilling several interrelated processes. A rule for allocating interventions to participants must be specified, based on some chance (random) process. We call this **sequence generation**. Furthermore, steps must be taken to secure strict implementation of that schedule of random assignments by preventing foreknowledge of the forthcoming allocations. This process if 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.

For all potential sources of bias, it is important to consider the likely magnitude and the likely 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.

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 compared groups 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.

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.

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.

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).

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.

Sources of bias in non-randomized studies

Bias may be present in findings from NRS in many of the same ways as in poorly designed or conducted randomized trials (see Chapter 8, Section 8.4). For example, numbers of exclusions in NRS are frequently unclear, intervention and outcome assessment are often not conducted according to standardized protocols, and outcomes may not be assessed blind. The biases caused by these problems are likely to be similar to those that occur in randomized trials, and review authors should be familiar with Chapter 8 that describes these issues. None of these problems are any less difficult to overcome in a well-planned non-randomized prospective study than in a randomized trial.

In NRS, use of allocation mechanisms other than concealed randomization means that groups are unlikely to be comparable. These potential systematic differences between characteristics of participants in different intervention 'groups' are likely to be the issue of key concern in most NRS, and we refer to this as selection bias. When selection bias produces imbalances in prognostic factors associated with the outcome of interest then 'confounding' is said to occur. Statistical methods are sometimes used to counter bias introduced from confounding by producing 'adjusted' estimates of intervention effects, and part of the assessment of study quality may involve making judgements about the appropriateness of the analysis as well as the design and execution of the study.

The variety of study designs classified as NRS, and their varying susceptibility to different biases, makes it difficult to produce a generic robust tool that can be used to evaluate risk of bias. Within a review that includes NRS of different designs, several tools for assessment of risk of bias may need to be created. Inclusion of a knowledgeable methodologist in the review team is essential to identify the key areas of weakness in the included study designs.

With randomized trials, assessment of the risk of bias focuses on systematic bias, which is usually assumed to be 'optimistic' in direction. The tendency for researchers to design, execute, analyse and report their primary studies to give the findings that are expected, consciously or subconsciously, is also likely to apply to NRS where researchers have control over key decisions (e.g. allocation to intervention, or selection of centres). In truly observational NRS, bias arising from 'confounding by indication' may not be so consistent; healthcare professionals may have differing opinions about the appropriateness of alternative interventions for their patients, contingent on the patients' presenting severity of illness or co-morbidities. Differences in case-mix between locations that are being compared may be haphazard. Therefore, when reviewing NRS, the variability of biases and the between-study heterogeneity they induce is at least as important as systematic bias.

Summary assessments of risk of bias

The Collaboration's recommended tool for assessing risk of bias in included studies involves the assessment and presentation of individual 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 on the particular scenarios in which biases may be large or small. It may, however, be possible to consider the likely magnitude of bias relative to the estimated magnitude of effect. For example, inadequate allocation sequence concealment and a small estimate of effect might substantially reduce one's confidence in the estimate, whereas minor inadequacies in how incomplete outcome data were addressed might not substantially reduce one's confidence in a large estimate of effect.

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

- **Summarizing risk of bias for a study across outcomes**: Some 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. 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.
- **Summarizing risk of bias for an outcome across studies (e.g. for a meta-analysis)**: These are the main summary assessments that will be made by review authors and incorporated into judgements about the 'quality of evidence' in 'Summary of findings' tables, as described in Chapter 11 (Section 11.5). As explained below, including trial 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. First, 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 of the outcomes that are critical to such an assessment. Second, judgements about which outcomes are critical to a decision may vary from setting to setting, because of differences both in societal values and in 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.

Rationale for data collection forms

The data collection form is a bridge between what is reported by the original investigators (e.g in journal articles, abstracts, personal correspondence) and what is ultimately reported by the review authors. The data collection form serves several important functions (Meade 1997). First, the form is linked directly to the review question and criteria for assessing eligibility of studies, and provides a clear summary of these that can be applied to identified study reports. Second, the data collection form is the historical record of the multitude of decisions (and changes to decisions) that occur throughout the review process. Third, the form is the source of data for inclusion in an analysis.

Given the important functions of data collection forms, ample time and thought should be invested in their design. Because each review is different, data collection forms will vary across reviews. However, there are many similarities in the types of information that are important, and forms can be adapted from one review to the next. Although we use the term 'data collection form' in the singular, in practice it may be a series of forms used for different purposes: for example, a separate form for assessing eligibility of studies for inclusion in the review to facilitate the quick determination of studies that should be excluded.

Design of a data collection form

When adapting or designing a data collection form, review authors should first consider how much information should be collected. Collecting too much information can lead to forms that are longer than original study reports, and can be very wasteful of time. Collection of too little information, or omission of key data, can lead to the need to return to study reports later in the review process.

Here are some tips for designing a data collection form, based on the informal collation of experiences from numerous review authors. The checklist in Table 7.3.a should also be consulted.

- Include the title of the review or a unique identifier. Data collection forms are adaptable across reviews and some authors participate in multiple reviews.
- Include a revision date or version number for the data collection form. Forms occasionally have to be revised, and this reduces the chances of using an outdated form by mistake.
- Record the name (or ID) of the person who is completing the form.
- Leave space for notes near the beginning of the form. This avoids placing notes, questions or reminders on the last page of the form where they are least likely to be noticed. Important notes may be entered into RevMan in the 'Notes' column of the 'Characteristics of included studies' table, or in the text of the review.
- Include a unique study ID as well as a unique report ID. This provides a link between multiple reports of the same study. Each included study must be given a study identifier that is used in RevMan (usually comprising the last name of first author and the year of the primary reference for the study).
- Include assessment (or verification) of eligibility of the study for the review near the beginning of the form. Then the early sections of the form can be used for the process of assessing eligibility. Reasons for exclusion of a study can readily be deduced from such assessments. For example, if only truly randomized trials are eligible, a query on the data collection form might be: 'Randomized? Yes, No, Unclear'. If a study used alternate allocation, the answer to the query is 'No', and this information may be entered into the 'Characteristics of excluded studies' table as the reason for exclusion.
- Record the source of each key piece of information collected, including where it was found in a report (this can be done by highlighting the data in hard copy, for example) or if information was obtained from unpublished sources or personal communications. Any unpublished information that is used should be coded in the same way as published information.
- Use tick boxes or coded responses to save time.
- Include 'not reported' or 'unclear' options alongside any 'yes' or 'no' responses.

- Consider formatting sections for collecting results to match RevMan data tables. However, data collection forms should incorporate sufficient flexibility to allow for variation in how data are reported. It is strongly recommended that outcome data be collected in the format in which they were reported (and then transformed in a subsequent step).
- Always collect sample sizes when collecting outcome data, in addition to collecting initial (e.g. randomized) numbers. There may be different sample sizes for different outcomes because of attrition or exclusions.
- Leave plenty of space for notes.

• DATA SYNTHESIS INCLUDING META-ANALYSIS

Planning the analysis

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

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

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

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

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

Narrative synthesis uses subjective (rather than statistical) methods to follow through questions 1 to 4, for reviews where meta-analysis is either not feasible or not sensible. In a

narrative synthesis the method used for each stage should be pre-specified, justified and followed systematically. Bias may be introduced if the results of one study are inappropriately stressed over those of another.

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

1. The most straightforward Cochrane review assembles studies that make one particular comparison between two treatment options, for example, comparing kava extract versus placebo for treating anxiety (Pittler 2003). Meta-analysis and related techniques can be used if there is a consistent outcome measure to:
    - establish whether there is evidence of an effect;
    - estimate the size of the effect and the uncertainty surrounding that size; and
    - investigate whether the effect is consistent across studies.

2. Some reviews may have a broader focus than a single comparison. The first is where the intention is to identify and collate studies of numerous interventions for the same disease or condition. An example of such a review is that of topical treatments for fungal infections of the skin and nails of the foot, which included studies of any topical treatment (Crawford 2007). The second, related aim is that of identifying a 'best' intervention. A review of interventions for emergency contraception sought that which was most effective (while also considering potential adverse effects). Such reviews may include multiple comparisons and meta-analyses between all possible pairs of treatments, and require care when it comes to planning analyses (see Section 9.1.6 and Chapter 16, Section 16.6).
3. Occasionally review comparisons have particularly wide scopes that make the use of meta-analysis problematic. For example, a review of workplace interventions for smoking cessation covered diverse types of interventions (Moher 2005). When reviews contain very diverse studies a meta-analysis might be useful to answer the overall question of whether there is evidence that, for example, work-based interventions can work (but see Section 9.1.4). But use of meta-analysis to describe the size of effect may not be meaningful if the implementations are so diverse that an effect estimate cannot be interpreted in any specific context.
4. An aim of some reviews is to investigate the relationship between the size of an effect and some characteristic(s) of the studies. This is uncommon as a primary aim in Cochrane reviews, but may be a secondary aim. For example, in a review of beclomethasone versus placebo for chronic asthma, there was interest in whether the administered dose of beclomethasone affected its efficacy (Adams 2005). Such investigations of heterogeneity need to be undertaken with care (see Section 9.6).

Why perform a meta-analysis in a review?

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

1. To increase power. Power is the chance of detecting a real effect as statistically significant if it exists. Many individual studies are too small to detect small effects, but when several are combined there is a higher chance of detecting an effect.
2. To improve precision. The estimation of an intervention effect can be improved when it is based on more information.

3. To answer questions not posed by the individual studies. Primary studies often involve a specific type of patient and explicitly defined interventions. A selection of studies in which these characteristics differ can allow investigation of the consistency of effect and, if relevant, allow reasons for differences in effect estimates to be investigated.
4. To settle controversies arising from apparently conflicting studies or to generate new hypotheses. Statistical analysis of findings allows the degree of conflict to be formally assessed, and reasons for different results to be explored and quantified.

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

When not to use meta-analysis in a review

If used appropriately, meta-analysis is a powerful tool for deriving meaningful conclusions from data and can help prevent errors in interpretation. However, there are situations in which a meta-analysis can be more of a hindrance than a help.

- A common criticism of meta-analyses is that they 'combine apples with oranges'. If studies are clinically diverse then a meta-analysis may be meaningless, and genuine differences in effects may be obscured. A particularly important type of diversity is in the comparisons being made by the primary studies. Often it is nonsensical to combine all included studies in a single meta-analysis: sometimes there is a mix of comparisons of different treatments with different comparators, each combination of which may need to be considered separately. Further, it is important not to combine outcomes that are too diverse. Decisions concerning what should and should not be combined are inevitably subjective, and are not amenable to statistical solutions but require discussion and clinical judgement. In some cases consensus may be hard to reach.
- Meta-analyses of studies that are at risk of bias may be seriously misleading. If bias is present in each (or some) of the individual studies, meta-analysis will simply compound the errors, and produce a 'wrong' result that may be interpreted as having more credibility.
- Finally, meta-analyses in the presence of serious publication and/or reporting biases are likely to produce an inappropriate summary.

What does a meta-analysis entail?

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

1. Which comparisons should be made?
2. Which study results should be used in each comparison?
3. What is the best summary of effect for each comparison?
4. Are the results of studies similar within each comparison?
5. How reliable are those summaries?

The first step in addressing these questions is to decide which comparisons to make (see Section 9.1.6) and what sorts of data are appropriate for the outcomes of interest (see Section 9.2). The next step is to prepare tabular summaries of the characteristics and results

of the studies that are included in each comparison (extraction of data and conversion to the desired format is discussed in Chapter 7, Section 7.7). It is then possible to derive estimates of effect across studies in a systematic way (Section 9.4), to measure and investigate differences among studies (Sections 9.5 and 9.6) and to interpret the findings and conclude how much confidence should be placed in them (see Chapter 12).

Which comparisons should be made?

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

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

Principles of meta-analysis

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

1. Meta-analysis is typically a two-stage process. In the first stage, a summary statistic is calculated for each study, to describe the observed intervention effect. For example, the summary statistic may be a risk ratio if the data are dichotomous or a difference between means if the data are continuous.
2. In the second stage, a summary (pooled) intervention effect estimate is calculated as a weighted average of the intervention effects estimated in the individual studies. A weighted average is defined as:

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

   where Y_i is the intervention effect estimated in the *i*th study, W_i is the weight given to the *i*th study, and the summation is across all studies. Note that if all the weights are the same then the weighted average is equal to the mean intervention effect. The bigger the weight given to the *i*th study, the more it will contribute to the weighted average. The weights are therefore chosen to reflect the amount of information that each study contains. For ratio measures (OR, RR, etc), Y_i is the natural logarithm of the measure.
3. The combination of intervention effect estimates across studies may optionally incorporate an assumption that the studies are not all estimating the same

intervention effect, but estimate intervention effects that follow a distribution across studies. This is the basis of a **random-effects meta-analysis** (see Section 9.5.4). Alternatively, if it is assumed that each study is estimating exactly the same quantity a **fixed-effect meta-analysis** is performed.

4. The standard error of the summary (pooled) intervention effect can be used to derive a confidence interval, which communicates the precision (or uncertainty) of the summary estimate, and to derive a P value, which communicates the strength of the evidence against the null hypothesis of no intervention effect.
5. As well as yielding a summary quantification of the pooled effect, all methods of meta-analysis can incorporate an assessment of whether the variation among the results of the separate studies is compatible with random variation, or whether it is large enough to indicate inconsistency of intervention effects across studies (see Section 9.5).

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.

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

- the risk ratio (RR) (also called the relative risk);
- the odds ratio (OR);
- the risk difference (RD) (also called the absolute risk reduction); and
- the number needed to treat (NNT).

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

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

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

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

	Event ('**S**uccess')	No event ('**F**ail')	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:

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

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

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

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 an odds is more complicated than for a risk. The simplest way to ensure that the interpretation is correct is to first convert the odds into a risk. For example, when the odds are 1:10, or 0.1, one person will have the event for every 10 who do not, and, using the formula, the risk of the event is 0.1/(1+0.1) = 0.091. In a sample of 100, about 9 individuals will have the event and 91 will not. When the odds is equal to 1, one person will have the event for everyone 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 first example above where a risk of 0.091 was seen to be similar to an odds of 0.1). When events are common, as is often the case in clinical trials, the differences between odds and risks are large. For example, a risk of 0.5 is equivalent to an odds of 1; and a risk of 0.95 is equivalent to odds of 19.

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

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 a treatment implies that events with treatment are three times more likely than events without treatment. Alternatively we can say that treatment increases the risk of events by 100 × (RR – 1)% = 200%. Similarly a risk ratio of 0.25 is interpreted as the probability of an event with treatment being one-quarter of that without treatment. This may be expressed alternatively by saying that treatment decreases the risk of events by 100 × (1 –RR)% = 75%. This is known as the relative risk reduction (see also Chapter 12, Section 12.5.1). The interpretation of the clinical importance of a given risk ratio cannot be made without knowledge of the typical risk of events without treatment: a risk ratio of 0.75 could correspond to a clinically important reduction in events from 80% to 60%, or a small, less clinically important reduction from 4% to 3%.

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

Odds ratios, like odds, are more difficult to interpret (Sinclair 1994, Sackett 1996). Odds ratios describe the multiplication of the odds of the outcome that occur with use of the intervention. To understand what an odds ratio means in terms of changes in numbers of events it is simplest to first convert it into a risk ratio, and then interpret the risk ratio in the context of a typical control group risk, as outlined above. The formula for converting an odds ratio to a risk ratio is provided in Chapter 12 (Section 12.5.4.4). Sometimes it may be sensible to calculate the RR for more than one assumed control group risk.

Warning: OR and RR are not the same

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

Measure of absolute effect: the risk difference

The **risk difference** is the difference between the observed risks (proportions of individuals with the outcome of interest) in the two groups (see Box 9.2.a). The risk difference can be calculated for any study, even when there are no events in either group. The risk difference is straightforward to interpret: it describes the actual difference in the observed risk of events

between experimental and control interventions; for an individual it describes the estimated difference in the probability of experiencing the event. However, the clinical importance of a risk difference may depend on the underlying risk of events. For example, a risk difference of 0.02 (or 2%) may represent a small, clinically insignificant change from a risk of 58% to 60% or a proportionally much larger and potentially important change from 1% to 3%. Although the risk difference provides more directly relevant information than relative measures (Laupacis 1988, Sackett 1997) it is still important to be aware of the underlying risk of events and consequences of the events when interpreting a risk difference. Absolute measures, such as the risk difference, are particularly useful when considering trade-offs between likely benefits and likely harms of an intervention.

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

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

What is the event?

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

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

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

By contrast, switching the outcome can make a substantial difference for risk ratios, affecting the effect estimate, its significance, and the consistency of intervention effects across

studies. This is because the precision of a risk ratio estimate differs markedly between situations where risks are low and situations where risks are high. In a meta-analysis the effect of this reversal cannot easily be predicted. The identification, before data analysis, of which risk ratio is more likely to be the most relevant summary statistic is therefore important and discussed further in Section 9.4.4.4.

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

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

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.*

The standardized mean difference

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

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

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

However, the method assumes that the differences in standard deviations among studies reflect differences in measurement scales and not real differences in variability among study populations. This assumption may be problematic in some circumstances where we expect real differences in variability between the participants in different studies. For example, where pragmatic and explanatory trials are combined in the same review, pragmatic trials may include a wider range of participants and may consequently have higher standard deviations. The overall intervention effect can also be difficult to interpret as it is reported in units of standard deviation rather than in units of any of the measurement scales used in the review, but in some circumstances it is possible to transform the effect back to the units used in a specific study (see Chapter 12, Section 12.6).

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

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

- **INTERPRETATION AND REPORTING**

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.
- 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. Review authors are encouraged to include such tables in Cochrane reviews, and to ensure that there is sufficient description of the

studies and meta-analyses to support their contents. 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'.

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, 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 apply the results to their specific situation. Authors should avoid specific recommendations that depend on assumptions about available resources and values. Authors can, however, aid decision-making by laying out different scenarios that describe certain value structures.

In this chapter we address first one of the key aspects of interpreting findings that is also fundamental in completing a 'Summary of findings' table: the quality of evidence related to each of the outcomes. We then provide a more detailed consideration of issues around applicability and around interpretation of numerical results, and provide suggestions for presenting authors' conclusions.

The 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 2006b, Guyatt 2008a, Guyatt 2008b). Over 20 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 Institute for Health and Clinical Excellence (NICE) in the UK, and UpToDate® have adopted the GRADE system in its original format or with minor modifications (Schünemann 2006b, Guyatt 2006a, Guyatt 2006b). The BMJ encourages authors of clinical guidelines to use the GRADE system (www.bmj.com/advice/sections.shtml). The Cochrane Collaboration has adopted the principles of the GRADE system for evaluating the quality of evidence for outcomes reported in systematic reviews. This assessment is being phased in together with the introduction of the 'Summary of findings' table (see Chapter 11, Section 11.5).

For purposes of 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. Quality of a body of evidence involves consideration of within-study risk of bias (methodological quality), directness of evidence, heterogeneity, precision of effect estimates and risk of publication bias, as described in Section 12.2.2. The GRADE system entails an assessment of the quality of a body of evidence for each individual outcome.

The GRADE approach specifies four levels of quality (Table 12.2.a). The highest quality rating is for randomized trial evidence. 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 12.2.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 limitations in design and implementation, 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 12.2.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).

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.

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 should not make recommendations.

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. 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 (Akl 2007) provides an example for laying out clinical implications for situations where there are important trade-offs between desirable and undesirable effects of the intervention: "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). Patients with a high preference for survival prolongation (even though that prolongation may be short) and limited aversion to bleeding who do not consider heparin therapy a burden may opt to use heparin, while those with aversion to bleeding and the related burden of heparin therapy may not."

Implications for research

Review conclusions should help people make well-informed decisions about future healthcare research. The 'Implications for research' should comment on the need for further research, and the nature of the further research that would be most desirable. A format has been proposed for reporting research recommendations ('EPICOT'), as 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 treatment/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).

Cochrane review authors should ensure that they include the PICO aspects of this format. It is also helpful to note the study types, as well as any particular design features, that would best address the research question.

A review of compression stockings for prevention of deep vein thrombosis in airline passengers 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 trials 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 trials 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 had been found: “This review highlights the need for randomised 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).

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’, whereas 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 had the direction of the results 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. Often this is done implicitly, without referring to the additional information or judgements that are used in reaching conclusions about the implications of a review for practice. 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 we have noted, authors should always be cautious when drawing conclusions about implications for practice and they should not make recommendations.

References

Adam 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.

Akl 2007. Akl EA, Kamath G, Kim SY, Yosuico V, Barba M, Terrenato I, Sperati F, Schünemann HJ. Oral anticoagulation for prolonging survival in patients with cancer. *Cochrane Database of Systematic Reviews* 2007, Issue 2. Art No: CD006466.

Brown 2006. Brown P, Brunnhuber K, Chalkidou K, Chalmers I, Clarke M, Fenton M, Forbes C, Glanville J, Hicks NJ, Moody J, Twaddle S, Timimi H, Young P. How to formulate research recommendations. *BMJ* 2006; 333: 804-806.

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.

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. Art No: CD004002.

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

Cooper 1984. Cooper HM. The problem formulation stage. In: Cooper HM, editor. Integrating Research. A Guide for Literature Reviews. Newbury Park: Sage Publications, 1984; 19-37.

Councell 1997. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. *Annals of Internal Medicine* 1997; 127: 380-387.

Eddy 1990. Eddy DM. Clinical decision making: from theory to practice. Anatomy of a decision. *JAMA* 1990; 263: 441-443.

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.

Hedges 1994. Hedges LV. Statistical considerations. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1994; 30-3.

Jackson 1980. Jackson GB. Methods for integrative reviews. Rev Educ Res 1980; 50:438-60.

Laupacis 1988. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. New Engl J Med 1988; 318: 1728-1733.

Light 1984. Light RJ, Pillemer DB. Organizing a reviewing strategy. In: Summing Up: The Science of Reviewing Research. Cambridge, Massachusetts: Harvard University Press, 1984; 13-31.

Meade 1997. Meade MO, Richardson WS. Selecting and appraising studies for a systematic review. *Annals of Internal Medicine* 1997; 127: 531-537.

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

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.

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

Richardson 1995. Richardson WS, Wilson MS, Nishikawa J, Hayward RSA. The well-built clinical question: a key to evidence based decisions. *ACP Journal Club* 1995: A12-A13.

Robinson 2007. Robinson J, Biley FC, Dolk H. Therapeutic touch for anxiety disorders. *Cochrane Database of Systematic Reviews* 2007, Issue 3. Art No: CD006240.

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: Churchill Livingstone, 1997.

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.