Cochrane Collaboration
open learning material
for reviewers

Version 1.1
November 2002

Editors
Phil Alderson, UKCC
Sally Green, ACC
Introduction

Welcome to the Cochrane Collaboration’s Open Learning Material for Cochrane reviewers. Performing a Cochrane review requires skills that may be new to you, including the formulation of appropriate questions, literature-searching, critical appraisal, statistical analysis and interpretation and application of your findings. This material is designed to accompany the Cochrane Reviewers’ Handbook in helping you gain these skills and complete your review. It does not replace the Reviewers’ Handbook, instead it provides a framework to progressing through the Handbook, supplementing it with examples and activities along the way.

Many training events and tools have been developed and published to help reviewers acquire these skills, however not all are accessible to all reviewers. This material is designed to help train reviewers in the methods and processes of performing a Cochrane review. Along with the Cochrane Reviewers’ Handbook, this material will stand alone, offering an alternative to face-to-face training, especially for those reviewers living and working away from easy access to the training offered by Cochrane Centres and Cochrane Collaborative Review Groups. For those able to access this face-to-face training, this material will serve as a useful resource to remind you of what you have learned.

This material will take a step-by-step approach to Cochrane reviews, exploring each step individually, signposting appropriate links and references and providing examples and activities to help you make sense of the information. The material is organised in modules, each module relating to a consecutive section of your review. It is a good idea to complete each module as you start working on the corresponding part of your review. There are also some additional modules relating to issues of reviewing that do not occur in all reviews.

Sections of the Reviewers’ Handbook you need to read, extra references and activities are listed at the beginning of the module so you know what you will need to complete that module, then referenced again throughout the text in the place to which they relate. These references are clearly marked in the margin using the symbols below. In electronic versions, the icons in the margin serve as a link to the Reviewers’ Handbook or other reference material. In paper versions they act to guide you to the appropriate place in the reference.

Handbook reference: Activity: Reading:

Also in the margin are brief summary notes, highlighting the important learning points.
All the references are freely available, either in the text or on the Internet, excepting the optional reference referred to in Module 6 (and useful to have in general):


A note on the other material referred to
This material refers to version 4.1.5 of the Reviewers' Handbook. The Reviewers' Handbook is updated periodically and the next version, 4.2, will have major changes to section 8 covering statistical issues. At the moment, some of the statistical modules refer you to Section 8 of the Reviewers' Handbook. This may refer to the new section 8, and if using a version that does not have this, the section you are told to find may not exist. A new release of RevMan, also called 4.2, is planned for 2003.

The development of this material
This material was developed, based on open learning strategies, in a collaborative project between several Cochrane Centres and the Cochrane Statistical Methods Group. It has been approved by the Handbook Advisory Group of the Cochrane Collaboration and will be regularly updated by the project team.

Editors:
Phil Alderson, UK Cochrane Centre, (palderson@cochrane.co.uk)
Sally Green, Australasian Cochrane Centre, (sally.green@med.monash.edu.au)

Contributors:
Jon Deeks, Centre for Statistics in Medicine, Oxford, UK
Julian Higgins, MRC Biostatistics Unit, Cambridge, UK
Jacob Riis, Nordic Cochrane Centre, Copenhagen, Denmark
Chris Silagy, Australasian Cochrane Centre, Melbourne, Australia

[Chris Silagy wrote his contribution to this material during 2000 while on sabbatical in Oxford. He died in 2001]

Sources of support:
National Health Service Research and Development Programme, UK
Department of Health and Ageing, Australia
Apotekerfonden af 1991 (The 1991 Danish Pharmacy Foundation)

Acknowledgments:
We are grateful to Mike Clarke of the UK Cochrane Centre, and Philippa Middleton of the Australasian Cochrane Centre for their comments on this material, and to Nicola Thornton and Sarahjayne Sierra of the UK Cochrane Centre for their help with copy editing.
Principles underlying use of this material

Use of this material should adhere to the following set of principles.

- This material, developed by the Cochrane Collaboration for training reviewers, should be freely available to those reviewers with a registered Cochrane review title
- Profits generated from training non-Cochrane reviewers with this material should benefit the Cochrane Collaboration
- Organisations utilising this material within their courses should acknowledge its source
- Any suggestions for the improvement and updating of this material should be sent to the editors so that these suggestions can be considered in future revisions of the material

Revision history

24th February 2003  Corrections made to odds ratio calculations in Module 11.

November 2002  Originally published.
Module 1: Introduction to systematic reviews

Learning objectives

- Be familiar with the rationale for systematic reviews
- Understand the concept of bias
- Be familiar with different types of questions that can be posed about the effects of health care interventions
- Consider study designs which are most likely to give a valid answer to these questions
- Be aware of practical problems in finding, appraising and synthesising non-RCT evidence

Relevant sections of the Reviewers’ Handbook

- Section 1

Other relevant material

- *Rationale for systematic reviews* by Cynthia Mulrow (Mulrow CD. Rationale for systematic reviews. BMJ 1994; 309:597-9.)
Too much information, too little time

This is the essence of an important problem facing people who make decisions about health care. There is simply too much information around for people to keep up to date. On top of this, high quality information is often not easy to find.

The rationale for systematic reviews is explained in the article by Cynthia Mulrow, which you should now read. If you don’t have access to it, the main points are summarised below for you.

The full text of the article is available at:

http://www.bmj.com/cgi/content/full/309/6954/597

The rationale for systematic reviews can be summarised like this:

- People making decisions about treatment choices, or other health care interventions, need reliable information
- There is too much information around for decision makers to keep up to date
- Therefore, decision makers need reviews of existing information
- Reviews can be unscientific and biased in the way they collect, appraise and summarise information
- Systematic reviews attempt to minimise these biases to provide a reliable basis for making decisions

Are reviews science?

Systematic reviews aren’t new, and have been used in the natural sciences, such as physics, for some time. The idea behind them is simple. Science is cumulative, with new ideas being based on previous knowledge and observation, and new advances in science should help us make sense of what we already know and have observed. But if we don’t collect previous knowledge and observation in a systematic way, we are unlikely to make progress as quickly as we could. For example, if we are not aware of relevant research done by someone in another country, we cannot use that to help formulate our own research. We might even choose to do exactly the same research, without realising it has already been done. This is wasteful.
You might imagine that researchers do this routinely, but a study of controlled trials in leading medical journals showed that these studies are often published without reference to what we know already, and without discussion of what the new study adds to our knowledge. The reader is expected to go and find out all that for herself. Try the activity on the left.

The need to be more scientific about how science progresses is increasingly being recognised in health care. For instance, the UK Medical Research Council now requires evidence that a systematic review has been prepared before it will commission a new trial. This ensures that the question has not already been answered, and that the results of previous research are used in designing the new trial.

**What is bias?**

You’ve probably already noticed that the word ‘bias’ has been used in this module. This word has a meaning here which is similar to its meaning in common usage. Bias means something that will cause a consistent deviation from the truth. This is different from the play of chance.

For example, if we took a random sample of five people on a shopping street, four of them might be men. Would we then predict that four out of every five people are men? No, the play of chance meant that the sample we took was not representative of the whole population. The next sample of five people may well have had four, or even five, women.

If, however, we tried to estimate the proportion of men in the whole population by taking samples from a football crowd, we would probably find that, even if we took a very large sample, there would be more men than women. This is simply because we are taking a sample from a place which is not typical of the world at large – we have introduced a bias into our sampling which will cause us always to overestimate the proportion of men in the population.

© The Cochrane Collaboration 2002
Minimising bias

This is an expression you will see a lot in relation to systematic reviews. Because the aim is to provide reliable information, we need to do as much as possible to minimise the effects of anything that will cause the results to deviate from the truth. In other words, we need to minimise bias.

There are many possible sources of bias, which we will look at in various modules. For now, they can be grouped into two areas: bias arising from the studies included in the review, and bias arising from the way the review is done.

We’ll come back to bias in the studies collected for a review shortly. The ‘systematic’ part of systematic reviews is all about minimising bias in the way the review is carried out. People have tried to identify the major sources of bias and error in reviews, and to design a system that will minimise them. The process involved in a systematic review goes like this:

- Define the question
- Look for all studies reliably addressing the question
- Sift the studies to select relevant ones
- Assess the quality of the studies
- Calculate results for each study (and combine them if appropriate)
- Interpret results

This general approach is followed in all systematic reviews, although the latter steps depend on finding some suitable studies.

The component studies of a review

There is no point taking a very systematic approach to preparing a review if the individual studies within it are not capable of answering the question your review poses. Just as we want to minimise bias in the review process, so we want to choose component studies that are likely to give us an answer close to the truth. There will always be the effects of chance and we can only overcome this by collecting as much evidence as possible, and taking account of it in the way we analyse and interpret results. But our only means of minimising the effects of bias, is to include only studies that are less likely to be biased.
Choosing studies likely to give a valid answer to the question

These modules are mainly concerned with systematic reviews of health care interventions – questions of the type ‘Does intervention A have different effects to intervention B in this health problem?’ The Cochrane Collaboration does not currently do reviews of diagnostic test accuracy, or other types of health questions such as prevalence, prognosis or genetic predisposition.

Examples of questions about interventions, from published reviews, are:

- Does the application of compression bandages or stockings aid venous ulcer healing?
- What are the effects of topical treatments applied during pregnancy on the later development of stretch marks?

These questions are about comparing different interventions and measuring their effects. This means there should be comparison groups, where one group receives one intervention and the other group receives the alternative.

An important issue in designing studies like this is the generation of the comparison groups. We want the two groups to be identical in every respect, except for the different interventions they get. If the groups are not identical, we are often not sure whether differences in the outcomes between the two groups are due to these differences in the groups, or due to the interventions we wanted to study.

For example, if we wanted to compare the effect of compression bandages for venous ulcers with doing nothing, we would want comparison groups which had similar ages of patients, similar proportions of people with diabetes, similar proportions of smokers, etc. In fact, we would want anything that might affect wound healing to be equal in the two groups, and the only difference to be that one group gets compression bandages and the other doesn’t.

We could try to make sure of this by ‘matching’ patients – for every person in the group getting compression bandages, we could try to find another person with similar age, same sex, diabetes and smoking status, and put them in the group not getting compression bandages.

© The Cochrane Collaboration 2002
Randomisation will ensure, in the long run, that comparison groups are similar.

As well as taking a lot of effort, the problem is that there may be factors influencing wound healing that we haven’t thought about, or don’t know about. We can’t match for these. This is where the power of randomisation lies. If each person coming into the study has an equal chance of going into either group, over the long run all factors, known and unknown, will be equally distributed.

So, randomisation should produce comparable groups, and the groups are more likely to be comparable the more people are randomised. This means we are more certain about concluding that differences in the outcome are due to the treatment. In short, the studies are less likely to be biased, and we are more likely to believe the results. Studies that randomise participants to groups are called randomised controlled trials (RCTs).

There are other practical reasons for focusing on randomised controlled trials. These reflect some of the methodological work done in this area, and some of the work done by the Cochrane Collaboration over the last few years:

- A considerable amount of work has gone into making randomised trials easier to find, so that we don’t find a non-representative, or biased, group of studies
- We are beginning to understand how to tell an unbiased trial from a biased one

The same amount of background work has not been put into the location and utilisation of other study designs. For this reason, concerns about interpreting results will be greater if we choose study designs other than randomised controlled trials. Most Cochrane reviews therefore use the presence of randomisation as a minimum quality criterion when deciding which studies to include and, therefore, only include randomised trials.

**What if there aren’t any trials?**

Randomised controlled trials are more common in some areas of health care than others, for example where the interventions are drugs. What do we do if we have an important question, but there are no randomised controlled trials addressing the question?

Well, there’s absolutely nothing ‘wrong’ with a systematic review that has been done to a high standard, but finds no studies. In fact, these reviews are very useful because they highlight important gaps in our knowledge. Research funders are increasingly looking at the results of systematic reviews to help them decide what studies to commission. So a review that finds no studies can stimulate new research that will be able to answer the question.
See what you think of a review

It’s a good idea to have a look at a systematic review for yourself. Find a recently published systematic review, for example by looking at The Cochrane Library, or searching PubMed ([http://www.ncbi.nlm.nih.gov/entrez/query.fcgi](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi)) using the term ‘systematic review’. Read the review, and ask yourself how well you think it summarises the evidence.

There’s no need to carry out a formal appraisal of it; just think about how relevant it is, and how accurate you think it is. Can you think of any ways in which it could have been done better? Make some notes and keep them somewhere, so that you can come back to them at the end of these modules and think again about the quality of this review, and how you could do it better.
Module 2: Introduction to the Cochrane Collaboration

Learning objectives

- Be familiar with the history, aims and structure of the Cochrane Collaboration
- Be familiar with the types of questions addressed by Cochrane reviews
- Be in contact with the appropriate Cochrane Collaborative Review group

Relevant sections of the Reviewers’ Handbook

Section 1

Other relevant material

Cochrane Collaboration website (www.cochrane.org)
Cochrane Collaboration Brochure, available at (www.cochrane.de/cochrane/cc-broch.htm)
In the previous module, we discussed something of the need for systematic reviews. In this module we’ll look at what has been done to meet this need, and in particular we’ll find out something about the Cochrane Collaboration.

**The Cochrane Collaboration**

This international organization has evolved in response to the need for systematic reviews. It exists with the aim of preparing, maintaining and promoting the accessibility of systematic reviews of the effects of health care interventions.

Most importantly, the Cochrane Collaboration is a network of people – individuals, not institutions. Its success or failure depends on these individuals. Most are working voluntarily on this project in a spirit of collaboration.

It’s also important to note that the Cochrane Collaboration has limited its remit to reviews of the effects of health care interventions, so it is not attempting to address questions about the causes of disease, the identification of disease with diagnostic tests, or the natural history of disease.

For more information about the history, structure and objectives of the Cochrane Collaboration, read the Cochrane Collaboration Brochure.

**The Cochrane Collaboration today**

The Cochrane Collaboration has grown rapidly since its launch in 1993. There are now 14 Cochrane Centres around the world, and 48 review groups which cover most of health care. The infrastructure to accomplish the Collaboration’s mission is now pretty much in place.

Currently, new reviews are being published on *The Cochrane Database of Systematic Reviews* at the rate of about 300 each year. Although this is good, there are plenty more reviews left to do, and we need to make sure the quality remains high and continues to improve over the coming years.
You can find up to date information about the Cochrane Collaboration on one of its websites, and also on *The Cochrane Library*. This is a quarterly electronic publication, containing a wealth of information about the effects of health care interventions. Now would be a good time to find out how to access *The Cochrane Library*, and have a look at the Cochrane Collaboration website. It can be found at:

www.cochrane.org

Now complete the Activity on the left.

Finding your way round the Cochrane Collaboration

In the Cochrane Collaboration Brochure, you read something of the structure of the organisation. At first, the structure can be a bit daunting. This is because the task of co-ordinating an international effort like this is complicated, as you can imagine.

People preparing reviews as part of the Cochrane Collaboration are linked to a Collaborative Review Group, which covers the health problem of interest. For example, if you’re interested in an intervention for people with strokes, you would work with the Cochrane Stroke Group, based in Edinburgh, UK.

So the first step is for you to decide what topics interest you. When you have thought this through, and decided that you really want to work in a particular area, there are two ways to find out which review group you need to talk to. The first is to contact your local Cochrane Centre and discuss it with the staff there. They will be able to put you in touch with review groups, but will not be able to help you decide what you are interested in.

The other way is to look through the list of review groups, either at the website (follow the links ‘contact details for Cochrane groups’ and ‘collaborative review groups’) or by looking at *The Cochrane Library* (in the top part of the main screen, double click on ‘About the Cochrane Collaboration’ and then on ‘Collaborative Review Groups’ to find an alphabetical list of groups with information about each). If using the internet version, the links are in the left hand side of the screen.

© The Cochrane Collaboration 2002
If it is not obvious which review group your topic will belong to, your local Cochrane Centre will help, or you can search for words on *The Cochrane Library* and see which groups are identified in your search. While you are on *The Cochrane Library*, find out which is your local Cochrane Centre. Double click on ‘About the Cochrane Collaboration’ and then on ‘Centres’. It may be obvious from looking through the list which is your local Centre (there may be one in your country). If not, use the Search window and enter the name of your country. Every country is linked to a Cochrane Centre, and you can look to see which Centre includes your country in its entry in *The Cochrane Library*.

You’ll need to get in touch with the appropriate review group who will check how your interests might overlap with other work in the group, to make sure that you aren’t planning work that has already been done. There is usually a bit of discussion about exactly what the topic you want to take on will be, what help you might need, and what help the group can offer you.

**What if I don’t have access to *The Cochrane Library***?

Why not ask your local library to arrange access? Tell them to go to the Cochrane website and click on ‘The Cochrane Library’ to find out how to do it. But if you don’t have access to *The Cochrane Library*, either your review group or your local Cochrane Centre will help you find your way round the Cochrane Collaboration.
Further exploration of *The Cochrane Library* (optional)

If you do have access to *The Cochrane Library*, here are some more things to try.

- If you are interested in how reviews are done, and research about ways to improve reviews, have a look through the list of Methods Groups – groups of people who are interested in particular methodological questions. Do any seem relevant to your interests? If so, you might like to consider joining one.

- Look at the list of Fields. These were explained in the Cochrane Collaboration Brochure. Is there a Field in which you are interested?

- Finally, have a look at the reviews on the *Cochrane Database of Systematic Reviews*. Find a review that is of interest to you, either by scrolling through the list of titles or doing a Search. Have a read of the review (printing it out if you prefer), and ask yourself the same things you did when reading a review in module 1 – how relevant is this review, how well does it summarise the evidence, and how could I do it better? Make some notes so that you can return to them later on, after completing the modules.
Module 3: An introduction to meta-analysis

Learning objectives

- Be able to explain the difference between meta-analyses and systematic reviews
- Understand that meta-analysis is a two stage-process involving (a) computing summary statistics for each trial, (b) averaging the summary statistics
- Understand why simply adding up data from individual studies is inappropriate
- Understand that a full analysis also involves proper consideration of the consistency of trial results
- Be able to name and explain the main characteristics of a forest plot
Systematic reviews and meta-analyses

In Module 1 we summarised the process of preparing a systematic review. Part of that process is to calculate the results of each study identified by the reviewer, and then to calculate an average of those results – if appropriate – in a meta-analysis. Systematic reviews do not have to have a meta-analysis – there are times when it is not appropriate or possible.

To represent this visually, the figure below shows that a meta-analysis may be part of a systematic review. A meta-analysis is also possible without doing a systematic review – you could just find a few studies and calculate a result, with no attempt to be systematic about how the particular studies were chosen.

One slight complication is that these two terms are often used interchangeably, particularly in North America. In this learning material, the term ‘systematic review’ will refer to the entire process of collecting, reviewing and presenting all available evidence, while the term ‘meta-analysis’ will refer to the statistical technique involved in extracting and combining data to produce a summary result.

What is a meta-analysis?

A meta-analysis is a two-stage process. The first stage is the extraction of data from each individual study and the calculation of a result for that study (the ‘point estimate’ or ‘summary statistic’), with an estimate of the chance variation we would expect with studies like that (the ‘confidence interval’).

The second stage involves deciding whether it is appropriate to calculate a pooled average result across studies and, if so, calculating and presenting such a result. Part of this process is to give greater weight to the results from studies which give us more information, because these are likely to be closer to the truth we are trying to estimate. We’ll come back to these topics in later modules.
The results of meta-analyses are often presented in a forest plot. Run through this slide show, which explains the parts of these plots.

A meta-analysis does not just add up the numbers from the trials

One common criticism of meta-analysis is that it somehow simply adds together the results from quite different studies and calculates a summary statistic as if it is one big study. It would be wrong to do this, and this is not what a meta-analysis does. A meta-analysis looks at the results within each study, and then calculates a weighted average.

The reasons for this are explained in a later module. For now, it’s enough to realise that if we just add up the numbers of people and events (such as deaths) from a number of trials, we effectively treat it as one big trial. In effect we will be comparing people in the treatment group of one trial with people in the control group of another trial. This comparison is not randomised, and it is likely that there will be some differences in the way the trials were carried out. This doesn’t make sense when we have gone to a lot of trouble to find randomised comparisons, and it does not make logical sense to do this.
Is it sensible to calculate a pooled estimate?

With the right software, it’s very easy to do a meta-analysis. In fact it’s almost too easy. Before pressing the button to calculate a meta-analysis, it’s important to ask whether it is sensible to do so.

There are two parts to making this decision – clinical and statistical. The first, clinical, part involves asking yourself whether the studies you have found really do all address the same question so that an average of their results would be sensible. There might be differences in the participants, interventions or outcomes that lead you to think that the treatment effect is really very different in the different studies. Here are some examples to think about:

- Would you expect pregnant women to respond differently from teenage boys to interventions designed to help them stop smoking?
- Would you expect the effect of acupuncture to vary depending on who did it?
- Would you expect the outcome of treatments for depression to be similar at 3 days and 3 months?

There is always an element of judgement in these decisions. But if you think there are good reasons why you would expect the effects of an intervention to differ substantially between studies, you should not pool the results.

The second, statistical, way of thinking about the consistency of the results in the studies included in a review is to look for big differences between the results of the trials. This will be covered later, in the module on Heterogeneity. For now, it’s enough to know that you should look for studies where the results don’t seem to fit, and then investigate possible reasons for this.
there's a label to tell you what the comparison is and what the outcome of interest is.
At the bottom there’s a horizontal line. This is the scale measuring the treatment effect. Here the outcome is death and towards the left the scale is less than one, meaning the treatment has made death less likely.

Take care to read what the labels say - things to the left do not always mean the treatment is better than the control.
The vertical line in the middle is where the treatment and control have the same effect - there is no difference between the two.
For each study there is an id

The data for each trial are here, divided into the experimental and control groups

This is the % weight given to this study in the pooled analysis
Each study is given a blob, placed where the data measure the effect. The size of the blob is proportional to the % weight. The horizontal line is called a confidence interval and is a measure of how we think the result of this study might vary with the play of chance. The wider the horizontal line is, the less confident we are of the observed effect.

<table>
<thead>
<tr>
<th>Study</th>
<th>Outcome</th>
<th>Expt n/N</th>
<th>Ctrl n/N</th>
<th>Weight %</th>
<th>RR (95%CI Fixed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexander 1972</td>
<td>16 / 55</td>
<td>22 / 55</td>
<td></td>
<td>6.6</td>
<td>0.73 [0.43, 1.23]</td>
</tr>
</tbody>
</table>

The data shown in the graph are also given numerically.

The label above the graph tells you what statistic has been used.

© The Cochrane Collaboration 2002
The pooled analysis is given a diamond shape where the widest bit in the middle is located at the calculated best guess (point estimate), and the horizontal width is the confidence interval.

**Note on interpretation**

If the confidence interval crosses the line of no effect, this is equivalent to saying that we have found no statistically significant difference in the effects of the two interventions.
Module 4: Planning a review, format of Cochrane reviews, and an introduction to RevMan

In this module we’ll think about all the resources you’ll need to do a Cochrane review, and introduce you to the structure of a review and the Review Manager software (RevMan).

Learning objectives

- Be aware of resource requirements for a review
- Be familiar with the structure of a Cochrane review
- Have loaded RevMan on a computer and be familiar with the tree structure of a review

Relevant sections of the Reviewers’ Handbook

- Section 2
- Appendix 2a
- Appendix 3a

Other relevant material

- RevMan exercise
- RevMan User Guide

© The Cochrane Collaboration 2002
The information contained in this Module relates to Section 2 of the Reviewers’ Handbook, which you should read now.

Planning your review

When thinking about getting on with your review, it is a good idea to be aware of what resources you will need to complete the job. In the next module we’ll think a little bit more about how you can use the protocol for your review to do some detailed planning. In this module, we’ll look at the resources you’re likely to need to complete a systematic review, what a Cochrane systematic review looks like, and get you using RevMan.

Resources for a review

You’ll need a range of knowledge and skills to complete a systematic review, and it pays to think about where you’ll get these early on in the review process.

Here’s a list of the sort of resources you’ll need:

- somebody to do the work
- somebody to get the money
- somebody who is willing to write the review
- an information specialist (a librarian, or someone with in-depth knowledge of how to locate and retrieve studies)
- a methodologist
- content expertise – people who know about the condition from both the clinical and the consumer perspective

Don’t panic if you don’t meet all these criteria yourself! It’s unusual to find all the knowledge, skills and resources required in one person. A systematic review is best prepared by a team of people rather than one individual, and some of these needs could be met by having an advisory group to give you input at intervals. Let’s have a look at these criteria in turn.

It may sound obvious, but there is quite a bit of work involved in a systematic review – make sure someone in your team has the time to do this. Problems finding the time are one of the common reasons reviews don’t progress quickly.

You don’t have to pay anyone to do a Cochrane review, but you will need some money to do a systematic review, if only for postage, access to the internet, paying for library costs, etc. Someone will need to have this, or go and get it from somewhere.

It helps to have someone involved who enjoys writing and is good at it. This often means your review won’t need as many revisions, and the results of your review will be communicated more clearly and effectively.
Advice from an information specialist is invaluable, so look into that. You may need help from someone experienced in the methods of reviews. If you’re not very familiar with the subject matter of the review, you’ll need some input from someone who knows about it (a content expert). Content experts can be clinicians who know a lot about the topic, or consumers who have personal experience of the condition. It’s helpful to have input from both.

Finally, it helps to have someone or a group of people to check on your progress – an advisory group. This can be an efficient way of getting help from some of the experts you’ll be approaching.

The Cochrane review group with which you are preparing the review will be able to provide some help in the last four categories. Many groups have a Trials Search Coordinator, each group should have access to the help of a statistician and be able to suggest people with other types of methodological expertise, and they may be able to help you find a content expert. The editorial team can act, to some extent, like an advisory group.

Make a note of how you plan to cover each of these categories. If you are preparing this review with a Cochrane group, find out what help they are able to offer.

Allowing enough time for your review

How long will your review take? As described in Appendix 3a of the Reviewers’ Handbook, this will vary depending on the content of the review, the time you have to devote to it and your skills. There are also busy times in a review, and quieter times, for instance when you’re waiting to hear back from authors of studies. Appendix 3a contains a time scale that you could adapt to suit your own circumstances.
Format of a Cochrane review

The format of a Cochrane review was devised so that it is flexible enough to accommodate reviews in topics across the whole of health care. It was also designed to give a degree of consistency across reviews, so that readers of reviews can find the same information in the same place in different reviews.

Appendix 2a of the *Reviewers’ Handbook* contains a guide to what goes in each section of a Cochrane review.

Getting started with RevMan

Review Manager, RevMan for short, is the software developed within the Cochrane Collaboration for preparing reviews in a structured way, so that they can then be put together as *The Cochrane Database of Systematic Reviews*.

This software has been developed over a number of years. The current version is RevMan 4.1, and there is not likely to be a major upgrade before about 2004.

You’ll need a computer that can run Windows 95 or above. There is no Mac version of RevMan, so Mac users will need some way of running Windows software.

To get RevMan, there are several options. The website will always have the most up-to-date version of the software:

- visit the RevMan page (www.cochrane-net.org/revman). There you'll find instructions on downloading and installing the software
- if you get the software on CD or floppy disks from your review group or elsewhere, look for a setup file on the CD or disk 1 of the floppies and run it

When you first install RevMan, you are prompted for a ‘reviewer ID’. This is some short code (often your initials) given to you by your Cochrane review group to distinguish you from other reviewers in the group. It is used as part of the code identifying reviews you are working on – this is done automatically by RevMan once you’ve put the ID in.
Once you’re past that stage, you should be at the front page of RevMan, looking something like this:

![Start the RevMan Exercise to familiarise yourself with the program]

A good way to orientate yourself is to press the question mark button in the toolbar, which will tell you about the main window, and some of the features of RevMan. If you’re familiar with Windows, most of this will be fairly familiar.

**RevMan exercise**

Different people like learning software in different ways. Some people like to get to know the whole package in one go, and others prefer to learn it as they’re using it.

David Badger, formerly of the Australasian Cochrane Centre, developed a self-directed exercise for learning RevMan some years ago. It was updated in 2000, and is available from the Cochrane websites in the same place you found the software itself, under the section labelled “training and promotional resources”. It consists of a Word document and some files to use in RevMan.

Use this exercise however you want – it takes about 3 hours to do the whole exercise, which will explore most features of the program. You may wish, instead, just to do bits of the exercise as you work through the modules.
Module 5: Writing your protocol: why we have them, what to put in the Background section, and defining the question

This module discusses why we have protocols and what should go in the first few parts of a protocol.

Learning objectives

- Understand the rationale for having protocols for reviews
- Understand the role of a Background section in a review
- Understand the reasons for breaking down a question into participants, interventions, outcomes and study designs
- Be aware of pitfalls in defining questions
- Be familiar with entering text and references in RevMan and have completed a cover sheet for a review

Relevant sections of the Reviewers’ Handbook

- Section 3
- Section 4
- Appendix 2a.4

Other relevant material

- Cochrane Collaboration Style Guide (under help in RevMan 4.1), Appendix 1

Where does this go in a Cochrane review?

- This module covers the parts of a protocol, and review, with the headings Background, Objectives and Selection Criteria.
Why bother with protocols?

People starting reviews are usually keen to get on with finding studies, reading them and trying to make sense of it all. When you feel like this, having to try to write down exactly what you want to do can seem like a good way to kill that enthusiasm. So we must think protocols are important to expect everyone to produce one. Here are the reasons why.

Planning

An even better way to kill enthusiasm than pausing to write a protocol is to start doing something, only to find out you’ve gone about it the wrong way. Preparing a protocol for your review makes you stop and think about what you’re doing. If you get the protocol right, the rest of the review really does follow from it.

Most reviews have a team of people working on them, and the protocol stage is a good opportunity to focus everyone’s thoughts on the task in hand, debate areas of uncertainty, and make sure everyone knows just what is involved in doing the systematic review.

Once you’ve written down exactly what you plan to do, the protocol can act as a working document for the reviewers. You can use it to assign tasks to people, identify the resources you’ll need, and keep track of progress by setting target dates for parts of the review.

People also use protocols to apply for funding – the information required for a grant application is often similar to what you will put in the protocol.

Tell everyone else what you’re doing

Within the Cochrane Collaboration, protocols for reviews go through some form of peer review (the exact arrangements vary between review groups). Some people also send their protocols to others they know are interested in the review question, as they may have helpful suggestions. Once your review group has approved your protocol, it is published in The Cochrane Library. As well as enabling people to send you comments, this is one way we can try to make sure different people don’t start the same review without realising it, and so avoid duplication of effort.

Minimising bias

Writing down exactly what you plan to do, in advance, is one of the ways you can minimise bias. A systematic review is retrospective – you’re looking back at a set of studies that have already been done. It’s easy to imagine that if you are familiar with the results of these studies, it might change the way you define the question, set the selection criteria, decide which interventions to compare, and choose which outcomes to look for.
Protocols help to minimise bias

However, most people starting a review know at least some of the studies relevant to their question, simply because they are interested in the topic. So are we being unrealistic to think we can rid ourselves of our biases by writing a protocol? It’s probably impossible to get rid of all bias, which is why we talk about minimising it. Writing a protocol is one way we should be able to reduce bias, even if we can’t eliminate it, by being transparent about what we plan to do in advance.

Can I change my protocol?

It is unreasonable to expect that it will always be possible to prepare a protocol containing a plan for every situation that might arise in a systematic review. Your selection criteria might be worded in a way that just doesn’t work when you try to use them on a set of studies, or the data you are able to collect might need analysing in a way you hadn’t thought of. It’s better to change the protocol than to finish up with a review that’s not helpful for users.

The type of changes you should be particularly wary of are those that mean you will include or exclude different studies than you had originally intended. This is especially important if you know what effect your rule change will have on the results of the review.

What you can, and should, always do is report in your review any changes you made along the way and what effect this had. If the results change, it’s probably sensible to present all the analyses, with and without the changes, so that readers can make up their own minds which results they believe.

Writing your protocol

You will probably have several versions of the protocol as you agree on the content with your co-reviewers and review group. Many people prefer to do this in a word-processing package they know well, rather than put it straight into RevMan. If you plan to do this, it is easy to copy and paste text directly into RevMan after you have written it in another piece of software. However, fancy fonts, formatting, tables, etc don’t paste across well, so keep the text simply formatted.

You should also remember that the target audience for your review is the same as for a general medical journal. So try to imagine the audience when you’re writing, and write for them. If they cannot understand what you are proposing in your protocol, they might not be able to understand the results and conclusions of your review.

The Title

The title of your review is important. When someone using The Cochrane Library is browsing through a list of reviews, it is the first thing they will see. So the title has to provide enough information to help the reader decide if the review is relevant to them.
The reader will need some succinct information about the participants and interventions that are the focus of the review. They won’t want to be distracted with unnecessary words. The Cochrane Collaboration has decided on a standard format for titles which helps to convey information as quickly as possible:

[Intervention] for [problem] in [category]

The last of these three parts is not always needed. Here are some examples:

Acupuncture for chronic asthma  
Antihistamines versus aspirin for itching in late pregnancy  
Case management for people with severe mental disorders

Try writing out your title in this format. Try it out on some people and see if they understand what your review is about. You’ll get further help from your review group in deciding on the title.

The Background section

In the Background section you need to explain to people reading your review why you are asking the review question. Don’t forget to write this section in plain English. Here’s a list of the sort of things you might like to cover:

- How important is the problem?
- Is there any uncertainty about how to deal with the problem?
- Why do people think this intervention might work?
- What is the intervention supposed to achieve?

When you discuss the importance of the health problem addressed by your review, you should cover how common the problem is and what effect it has on people or communities. For example, the common cold affects millions of people, has only minor effects on most people, but is responsible for a lot of time off work which costs the economy money. Multiple sclerosis is not very common, but affects some people very severely. You should try to attach some numbers to these types of statement (and reference them), and also consider how the situation might vary around the world.

Many people undertake reviews because they know there is uncertainty about how to deal with the health problem. If you know of evidence of uncertainty, put it here. For example, you might be aware of a survey showing that different doctors use different drugs in the same situation.

It’s not always obvious why people think an intervention might work, or why people tried it in the first place. So include some explanation of the reasoning behind the use of the intervention. For drugs, this might be some biological information about how the drug acts on a cell; for other interventions it might be psychological or sociological theory. However, don’t get too technical! This section of a review should be understandable to a wide readership. If you think there needs to be some very detailed information for expert readers, you could put a section at the end of the Background with a subheading called ‘technical information’.
Finally, you should explain what the intervention is supposed to achieve. Is it supposed to cure the problem, or help with a specific symptom?

At the end of this section, the reader should understand why you are asking the review question, and the next part of the review shouldn’t come as a surprise. A good way to see if your background contains the right information is to use it to explain to someone what your review is about.

Objectives

In this section you should write down the main questions to be addressed in your review. This should only take a sentence or two. Try it now…..

Defining the question

Getting the question right is the most important step in doing your review. As well as telling others what the review is about, it will guide how you collect studies, how you check whether studies are eligible and how you do the analysis. So take your time over it, discuss it with your co-reviewers and show it to others.

Section 4 of the Reviewers’ Handbook covers this in some detail, so this section just summarises what’s in there and adds a few comments. Defining the question is also the focus of ‘Developing a protocol for a review’ workshops – these can be a good place to try your ideas out on other people and learn from the review questions being thought about by other people. For details of these, and other, workshops contact your review group or look at the dates of workshops on www.cochrane.org.

It’s also worth having a look on The Cochrane Library, or checking with your review group, for reviews of similar topics to your’s to see how others have defined their questions.
Types of participants

Define the health problem. Be careful about using definitions that depend on

- Time
- Place
- Technology

For example, diagnostic criteria that were first developed in the UK in 1994 couldn’t have been used before then, and may not be used by people in other countries at all. Expensive or recent tests may not be available in many countries. So be careful. Don’t risk setting conditions that will force you to exclude studies that would otherwise be relevant to your objective.

Define the population and setting. Think about

- Age
- Sex
- Race
- Other factors that may make some people respond very differently to the intervention
- Where the participants are (hospital, community, etc)

Again, take care with these, as you may find that setting rigid criteria means you have to throw studies out. It is often better to be over inclusive at this stage. If, for example, you restrict the age to people aged eighteen or older, what will you do with a study of hundreds of participants that has three sixteen-year-olds in it?

Types of interventions

Define the interventions. For drugs, think about

- Drug preparation
- Route of administration
- Dose
- Duration
- Frequency

Be careful not to limit the question too much. For example, think about when a drug regimen is so different that it would have a totally different effect, rather than just specifying the commonly used dose.

For non-drug interventions, defining the intervention can be a bit more difficult. You’ll need to think about describing similar bits of interventions or concepts – exactly what was done, how often it was done, who did it, were they trained, etc. It can be particularly tricky where the intervention is complex, for example where a set of interventions is performed by one or more people. You may need to end up with a set of core parts to the intervention, and then some other non-core parts.
A review of assertive community treatment for people with severe mental disorders is a good example of a complex intervention clearly defined, and the reasons why it was defined in that way justified. (Marshall M, Lockwood A. Assertive Community Treatment For People With Severe Mental Disorders. In The Cochrane Library, Issue 4, 2001, Update Software, Oxford.)

“For an intervention to be accepted as ACT it must have been described in the trial report as: Assertive Community Treatment, Assertive Case Management or PACT; or as being based on the Madison, Treatment in Community Living, Assertive Community Treatment or Stein and Test models. Trials of case management that did not meet the criteria for ACT are considered in (a separate review). The review did not consider the use of ACT as an alternative to acute hospital admission. The review also excluded studies of 'Home-Based Care' (which involves a multi-disciplinary team assessing and treating urgent psychiatric referrals at home).

Home-based care is a form of crisis intervention which deals with those who are usually acutely ill, and should not be classified with either ACT or case management as these are long-term interventions for severely and persistently ill people.”

Define the comparisons. Here, you need to decide whether the group you will be comparing the intervention group with have

- A placebo
- Nothing
- Some other treatment

To help you understand this it may be helpful to look at an example, a Cochrane Review of anticonvulsants for women with pre-eclampsia (Duley L, Gulmezoglu AM, Henderson-Smart DJ Anticonvulsants for women with pre-eclampsia. In the Cochrane Library, Issue 4, 2001, Update Software, Oxford).
In this review the part of the question related to types of intervention and comparisons is defined as “comparisons of an anticonvulsant with placebo (or no anticonvulsant), and of one anticonvulsant with another”. From the analysis section of this review (below) you can see how the reviewers have organised their comparisons.

<table>
<thead>
<tr>
<th>Comparison/Outcome</th>
<th>No. of Studies</th>
<th>No. of F...</th>
<th>Statistical Method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total number of included studies: 10</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01 Magnesium sulphate vs placebo/no anticonvulsant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01 Maternal death</td>
<td>1</td>
<td>605</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>02 Convulsion</td>
<td>4</td>
<td>1112</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>03 Pulmonary oedema</td>
<td>1</td>
<td>228</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>04 Renal dialysis</td>
<td>1</td>
<td>228</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>05 Any antihypertensive therapy</td>
<td>1</td>
<td>605</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>06 Rapid acting antihypertensives</td>
<td></td>
<td></td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>07 Toxicity</td>
<td></td>
<td></td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>08 Given calcium gluconate</td>
<td>1</td>
<td>605</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>09 Side effects</td>
<td></td>
<td></td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>10 Caesarean section</td>
<td>4</td>
<td>1112</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>11 Postpartum haemorrhage</td>
<td>1</td>
<td>135</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>12 Placental abruption</td>
<td>1</td>
<td>64</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>13 Mortality for the fetus or infant</td>
<td></td>
<td></td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>03 Magnesium sulphate vs phenytoin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01 Convulsion</td>
<td>2</td>
<td>2241</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>02 Complications of labour</td>
<td></td>
<td></td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>03 Caesarean section</td>
<td>2</td>
<td>2195</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>04 Mortality for the fetus or infant</td>
<td></td>
<td></td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>05 Infant morbidity</td>
<td></td>
<td></td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>04 Magnesium sulphate vs diazepam</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01 Convulsion</td>
<td>2</td>
<td>66</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>02 Caesarean section</td>
<td>2</td>
<td>66</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>03 Mortality for the fetus or infant</td>
<td></td>
<td></td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>05 Magnesium sulphate vs nimodipine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01 Convulsion</td>
<td>1</td>
<td>627</td>
<td>Relative Risk [Fixed]</td>
</tr>
<tr>
<td>02 Antihypertensive drug</td>
<td>1</td>
<td>627</td>
<td>Relative Risk [Fixed]</td>
</tr>
</tbody>
</table>
Types of outcomes

The general rules are to list the outcomes of interest to decision-makers, particularly those who have the health problem of interest. Think about

- How the outcomes might be measured
- When the outcomes should be measured
- Which are the most important outcomes
- Bad outcomes as well as good outcomes

As with defining the participants, be careful about choosing outcomes where the definition changes over time and place.

Even if you believe certain outcomes are unlikely to be reported, you should list them if they are important. Then when you are writing up the review, you can comment that you looked for information on those outcomes but the studies didn’t report it. This will define a gap in the research and hopefully encourage researchers to include that outcome in future trials.

There is more in later modules about the types of data you may encounter, and how you might handle them.

Types of studies

Most Cochrane reviews focus on randomised controlled trials both because of concerns about bias in other types of study design, and bias in retrieving other study types, as explained in module 1.

There are occasions when randomised controlled trials are unlikely ever to be done. For example, if an intervention saves people with a condition where previously everyone died, it would not be appropriate to do a randomised trial in which some of the participants did not get treated. Other examples would be where the intervention is applied at a level that makes randomisation impractical, such as tobacco taxes at a national level. If you think your topic fits into categories like these, you’ll need to discuss it with your review group.

You should not choose to look for other study types, where randomised controlled trials are feasible, simply because you don’t believe trials have been done. There are reviews with no trials – if the question is important enough for a review, then the lack of trials is an important finding.
Building up your table of comparisons

The table below is designed to help you construct your review question and subsequently set your inclusion and exclusion criteria. Using what you have learned and thought about in the above section, complete this table and finalise your question.

| Participants | What are the people (or other participants) receiving the intervention to be like?  
|             | Men or Women?  
|             | All ages or with a cut off?  
|             | How would you define the disease/condition participants should have?  
|             | Who should make the diagnosis?  
|             | Are there any co-morbidities you want to exclude?  
|             | Are there any other types of people who should be excluded from your review (because they are likely to react to the intervention in a different way)?  
|             | In one sentence describe your population, for example “All adults with tennis elbow (pain on lateral aspect of the elbow aggravated by use of the wrist or hand), diagnosed by a health care worker.” |
| Interventions | What is the intervention you are interested in?  
|              | Does it have variations (eg dosage, mode of delivery, who delivers it)?  
|              | Are you going to include all variations or set parameters (for example is there a critical dose below which you think the intervention is not clinically appropriate and so trials of intervention below this dose should not be included in the review)?  
|              | What are you going to do with trials including only part of the intervention?  
|              | What are you going to do with trials including your intervention combined with another intervention (co-intervention)?  
|              | In one sentence describe your intervention. |
| Comparisons  | What are you interested in comparing the intervention to? This depends on the primary question of the review. Are you only interested in whether the intervention offers benefit over the natural course of the disorder (ie a comparison to placebo or no treatment), or are you interested in whether the intervention offers benefit over other interventions.  
|              | List your possible comparisons:  

© The Cochrane Collaboration 2002
### Outcomes

How do you think it is important to measure change with respect to this intervention in this population? List all the outcomes you are including in your review. You should consider all those likely to be important to those suffering the disorder as well as those treating them. Don’t forget to consider if there is a need to include adverse effects.

Divide your outcomes into primary (essential) outcomes and secondary. The main conclusions of your review will be based on the primary outcomes (usually three or fewer) so give this considerable thought.

Finally, give some thought to how your outcomes may be measured, both the type of scale or count likely to be used and the timing of the measurement. Are there any methods of measurement or times you want to exclude (for example there may be a certain duration before which you feel the intervention could not work, and so you may want to exclude trials measuring the intervention before that point. Or, you may feel a particular method of measurement or primary outcome is not valid and you may want to exclude trials measuring the outcome in that way.

List any exclusions you have here:

### Types of study

There may be some study methodology aspects that, if present, you feel renders that study so invalid to your review that it should be excluded. Some more common ones are lack of randomisation, failure to conceal allocation or, in reviews where the outcomes are very subjective (eg global assessment of improvement or levels of depression), blinding of the outcome assessor.

Are there any exclusions related to trial methodology in your review?

### Review title

As we have discussed already, the title of a Cochrane review usually follows the format: *Intervention for Problem in Category*
Getting it into RevMan

This is covered in the first part of the RevMan exercise, which you should now do, if you haven’t done it already.
Module 6: Searching for studies

This module covers finding studies to be considered for inclusion in systematic reviews.

Learning objectives

- Understand the key role of searching for studies in a systematic review
- Be aware of the variety of sources of reports of studies.
- Be able to discuss a logical approach to searching for studies
- Have discussed a search strategy with the appropriate Cochrane review group
- Be aware of the need to manage references and be aware of some ways of doing this

Relevant section of the Reviewers’ Handbook

- Section 5

Other relevant material


Where does this go in a Cochrane review?

- Under the heading ‘Search strategy’ in the text of the protocol and review.
Searching far and wide

How do you go about finding studies that meet the inclusion criteria for your review? At one extreme you could do a very quick search on one electronic database and find a couple of relevant articles. At the other extreme you could try to find every study that’s ever been done addressing your review’s question.

As you might expect, there are problems with both these approaches. If you don’t look very hard, the studies you do find are not likely to be representative of all the studies done. The reasons for this are explained in detail in the module on ‘Publication bias’. For now, you just need to know that studies with dramatic results are much easier to find than studies that don’t have dramatic findings. The other problem with only looking for a few studies is that you end up with less information. This can limit the precision of the results of your review, and restrict the conclusions you can make.

However, is it feasible to find absolutely every relevant study that has ever been done? It’s certainly not easy and might not be possible in most reviews. Many studies are never published, and those that are may not be indexed in places, such as MEDLINE, you’d normally look. At some point, the effort required to find more studies becomes too much, but there is relatively little evidence on exactly when we need to stop. So, for now, most people adopt a pragmatic approach – look as far and as wide as possible, taking care to look in such a way that we take account of what we know about the biases in finding studies.

In the meantime, one of the major, ongoing efforts of the Cochrane Collaboration is to make the reports of relevant studies easier to find.

Developing a logical approach to searching

In developing your search strategy, there are a few principles. Your search should:

- Be sensitive (trying to find as many studies as possible)
- Minimise bias
- Be efficient
You’ll need to look for studies in a number of ways and in a number of places

Start searching where you expect the highest yield

Get advice from an information specialist who is familiar with the searches needed for systematic reviews

To make your search sensitive, you’ll need to look in a number of different places - no single database, journal or book will contain all relevant records. To minimise bias, you will need to think about finding studies that aren’t in the major sources like MEDLINE. For an efficient search, it is usual to start looking in the place you expect to have the highest yield.

To do your search well, you’ll need access to help from an information specialist/librarian, who has a good knowledge of helping people with systematic reviews. It is only recently that people have started going to libraries and asking for help with finding everything on a given topic, rather than asking for only a few bits of relevant information, so these skills are fairly new!

Where to look for studies

The next sections run through the sources you are likely to use in finding studies.

Electronic databases

Cochrane collaborative review group specialised registers

Each Cochrane review group is building up a register of studies relevant to its scope. The idea is that instead of each reviewer searching lots of databases and journals for trials relevant to interventions for, say, asthma, it would be more efficient to search these sources centrally for all trials relevant to asthma. The reviewer can then use the register compiled from this work as their first place to search, and they will effectively be searching several sources at once.

Of course, some Cochrane review groups have been around longer than others, and some have more resources to devote to this. So the completeness of these registers varies. You’ll need to talk to your review group about getting access to the information in their register.

The Cochrane Central Register of Controlled Trials

This register is part of The Cochrane Library. The idea behind this register is that it should be a central place to put all the reports of controlled trials identified through the work of the Cochrane Collaboration. This means that it contains the results of searching MEDLINE, EMBASE, some other databases and a long list of journals, books and conference proceedings. Many of the reports of studies on the register have been included because they might be reports of trials, based on reading the title and abstract (if there was one).

© The Cochrane Collaboration 2002
The content of The Cochrane Central Register of Controlled Trials changes all the time, as does the indexing of entries and retrieval methods. There’s more information about this on The Cochrane Library.

**Other databases**

These are described in section 5.4 of the Reviewers’ Handbook, and in the further reading listed at the start of this module. You’ll need some advice from an information specialist about which ones to search.

A particularly important source might be registers of ongoing and unpublished trials.

**Handsearching**

This means going through journals, books and conference proceedings by hand, looking for relevant studies. Because this takes a long time, the Cochrane Collaboration has encouraged people to register their handsearching and look for all reports of controlled trials. These are then collected and put onto The Cochrane Central Register of Controlled Trials so that no one else has to handsearch that source. The list of journals being searched can be downloaded from the internet at:

http://www.cochrane.us/documents/master.xls

Again, you may need some guidance from an information specialist about which journals may need searching for your review.

**Checking references**

It’s usual to read through the reference lists of any studies you do find, in case the authors have referred to any other relevant studies. It may also be worthwhile looking for previous reviews of the topic and checking their reference lists, too.

**Personal communication**

People who have been working in a particular topic area may know of studies you haven’t yet found. Reviewers commonly send a list of the studies they have found to the authors of those studies, asking if they are aware of any other relevant studies.

Another approach is to write to the manufacturers of relevant drugs or devices and ask if they are aware of any other studies.
Coming up with a search strategy

If you're not an information specialist, you may find the sources listed above a bit daunting. Don’t worry; it doesn’t all have to be done at once. Remember the logical approach of starting to look at the richest source and working down.

Remember that this is something your review group should be able to help you with. Get in touch with them and ask what help they can give you.

Document your search

It’s very important to keep an accurate record of what you’ve searched, when you searched it and how you searched it. It’ll help you avoid having to repeat searches and it will help people using your review to appraise how well they think you’ve minimised bias.

All these details should be documented in the ‘Search strategy’ section of the text box in RevMan. The only exception is that where you’ve used the register of your review group, you don’t need to write down the strategy your group used. You should, however, explain the search used to retrieve studies from your review group’s register.

Keeping it under control

Keeping track of searches can be a challenge. You may find several reports of the same study, and you will probably find the same report of a study in several databases. So you need some way of keeping track of the references you’ve looked at, and then some way of grouping together all the reports of a single study.

You might like to keep a record of where you found each study, so that you can report how useful different sources were.

Some people use reference management software to do all this, such as ProCite, Reference Manager, EndNote or IdeaList. If you like working with databases this is great, and can save time typing in references later on. Other people prefer printing out citations and writing on them. You’re going to need some system for keeping track of which references you think are relevant, which ones you have ordered from the library, which ones you’ve received the paper for, etc. It’s a good idea to keep a note of which studies you have found and rejected. You may well come across them again later and it can be very frustrating to re-read irrelevant records.

If you don't already use one, now might be a good time to learn to use reference management software

© The Cochrane Collaboration 2002
Whichever route you take, you’ll probably end up with a file with a section for each study. In this section you’ll keep the form you use to collect information about the study and all the papers that report that study.

**Tips for saving time and effort**

Here are a few ways you can save yourself some time and effort. At the risk of being repetitive, the main advice is simply to get some help from an expert. But some other little tips are:

- Look at the terms used to index and describe a few studies you already know are relevant to your review, and use these terms in your search strategy
- Add new terms to your search strategy and then pilot them on part of the database to see whether you get relevant material, before you run it on the whole database
- Use date limits for your search if appropriate. For example, if drugs, surgical techniques or diseases have only been around since a certain date, there’s no point searching before then
Module 7: Collecting data from relevant studies

This module covers ways in which we collect the data we need for the review from the included studies. You will work through it in a slightly different way from other modules. At points in the module, you are asked to move to other modules so that you understand the type of information you want as well as the mechanics of collecting it. You will find links to modules 8, 9 and 10.

Learning objectives

- Be able to decide what data to collect from studies
- Devise a method and appropriate form for collecting data from included studies
- Be aware of ways to minimise errors and bias in collecting data
- What to do about obtaining and reporting missing data

Relevant sections of the Reviewers’ Handbook

- Section 7

Where does this go in a Cochrane review?

- You will need to include some description of your methods for collecting data in the Methods section of your protocol and review.
Collecting data you want, in a format you want it

Now that we have identified studies we are going to consider for inclusion in the review, we need to work out how we are going to use them and what information we need to extract from the studies to include in the review. Like all other components of the review we try to do this in a systematic way. At first it may seem odd to have a whole module devoted to the issue of collecting data from the studies you've found. But you've somehow got to manage the process of getting information about a study from a paper publication (or more than one paper, or a letter or a phone call….), organise it and then present it in your review. If you don’t have a system for doing this, it’s easy to forget where information came from, just collect what is easiest to get out of papers, or make mistakes.

In addition to collecting information from the studies, you will need a system for collecting information about the process you follow, for example the number of studies you find, how many of those you include in your review and so on.

What information do we need to collect?

There are four main types or categories of information we need to collect from studies considered for inclusion in our review:

- How the study measures up against our eligibility criteria and whether it should be included in the review (and if not, why not)
- Descriptive information about the study to complete the ‘Table of Included Studies’ or ‘Table of Excluded Studies in RevMan’ in RevMan
- Information about the quality of the study
- Information about the results of the study in the form of data to be used in your review

Why bother with a specific form?

You may be thinking that taking the time and effort to design and test a structured form to extract data from studies is not worth it when you could just sit down with your included studies and enter the information you need straight into RevMan. Experience suggests that it is worth taking time over this step and Section 7.1 of the Reviewers’ Handbook gives you some good reasons why we bother with data collection forms. You should read that now.
In summary we need data collection forms as they:

- Specifically mirror your review question and so allow you to reassess all important components of your question and ensure you extract the relevant data relevant

- Document the history of the process of taking the information from the study into the review and so allow you to backtrack and identify any errors or assumptions you have made in your review

- Record the information you need to generate the results of your review, without any additional, unnecessary information. This will allow more efficient data entry into RevMan and highlight any gaps in the data

- Allow more than one reviewer to extract data from a study and compare the results of their work, increasing the reliability of the data collected for the review. This is an important step in reducing bias.

Format of data collection forms

There are several different approaches to the format of your data collection form and no single style is necessarily the best. Some reviewers prefer an open ended form with a few headings for structure and lots of space to record information in flexible formats. Others prefer to list every item they wish to extract in a highly structured, checklist type approach.

The two examples at the end of this module will give you some idea of the variation in data extraction forms. You should design your form to suit your preferred style of working.
Paper versus electronic forms

Section 7.2 of the Reviewers’ Handbook devotes a little time to discussion of the benefits of paper versus electronic data collection forms. On the surface it may appear that an electronic form where data can be transferred directly from the form into RevMan would save time, however, at this stage that can only occur for transferring numerical data in spreadsheet form.

If however you prefer working with electronic media you may prefer to extract data into an electronic form and enter into RevMan from there. To date, most reviewers have used paper data collection forms.

Some review groups may have a template for a data collection form you can modify to your particular review but in many cases individual reviews are so specific that a form needs to be designed to match that review.

As discussed already there are some key components to data collection forms and these are outlined in Section 7.4 of the Reviewers’ Handbook, which you should read before proceeding. Once you have read this return to this point and progress though each section.

Part One of a Data Collection Form: Review, Reviewer and Study Information

Even if you and your co-reviewers are working on one review at the moment, it is possible in the future you will do more, or your review may be passed on to someone else in your review group for comment or updating. Because of this it is important that it is clear which review this form relates to, as well as which reviewer and which study. The following list of suggestions for data fields may help you in designing Part 1 of your form.

- Title of review
- Review ID (as given by your review group when you register your review)
- Reviewer
- Version or date of designing form
- Date of completing form
- Study ID (to match how you have coded it in revman
- Year of publication of study

Now have a go at designing Part 1 of your data collection form. There may be more information specific to your review to include and you may want to exclude some of the above list.
Complete Module 8 before you progress further.

Example of a study eligibility form

Part Two of a Data Collection Form:
Study Eligibility
Before we design this part of the data collection form we need to understand the process of selecting studies for you review. Skip to Module 8 and complete it, then return to this point to design Part 2 of your form.

To make decisions about including or excluding studies, as we discussed in Module 8, most people use some kind of eligibility form. These forms help you structure the way you go about judging whether a study is eligible for your review. They also help you to find out at a later date why you reached certain judgements.

You will find an example of an eligibility at the end of this module.

You’ll see that it collects some information about the study ID (useful in case the form later gets detached from the study report), and then asks a series of questions about the study design, participants, interventions and outcomes. These will come directly from the inclusion criteria you set up earlier and will be a way of seeing if each study matches your criteria.

Try to design an eligibility form for your review. Use the table you generated in Module 5 as your guide.

One final tip: write on your eligibility form whereabouts in the report the relevant bit of information came from. You could also write on the report of the study which part of the form a particular section helps with. It’s much easier to come back to this and find out why you came to a decision.
Part Three of a Data Collection Form: Qualitative Information about the Study - The Table of Included Studies

Section 7.5 of the *Reviewers’ Handbook* outlines this part of the data collection form, but in summary, the information we need to take from the studies to complete this part of the review is the description of the study, in particular its participants, interventions and outcomes. This information will be used to complete the Table of Included Studies, which has a section also for methods, but it may be best to leave this until after the quality assessment is completed (we will cover that in Part Four). The Table of Included Studies is an important part of a systematic review as it is here that the reader can find detailed information about the included study and interpret the results of the review. It may also provide information about why the results of some included studies differ from others.

It may be worthwhile familiarising yourself with a Table of Included Studies by finding a review on *The Cochrane Library* you are interested in and going to the “Characteristics of Included Studies” section, then follow the link to the Table.

The types of information needed to complete this part of the review are well described in the *Reviewers’ Handbook*, but some hints to make things easier when data are entered later into RevMan are

- Be consistent in the order and style you use to describe the information (for example when extracting information about participants, start with sample size, then diagnostic description, then demographic information etc). This will make it easier to complete the Table of Included Studies, prevent you from overlooking information and make reading of the review easier.
- Highlight any missing information as unclear or not described to make it clear to the reader of your review that the information was not included in the description of the study, not that you forgot to extract it.
- Record any additional, study specific information (for example that the report of the study was translated from a language other than English, or that it was a duplicate publication) in the notes section.
Part Four of a Data Collection Form: Study Quality

Before we discuss the design and use of a data extraction form for recording information about the quality of the included studies, we first need to understand some principles and issues surrounding the assessment of study quality and how it relates to Systematic Reviews. Before progressing further, complete Module 9: Assessing the Quality of Studies, then return to here.

Now that you have a grasp of the process of assessing the quality or validity of your included trials, we need to think about how you are going to do this in your review.

Think about the method you wish to use, describing set criteria such as allocation concealment, blinding and withdrawals; or using a particular quality assessment scale. Look at your review group’s module on The Cochrane Library and see if they have a preferred method.

Once you know how you are going to do your validity assessment, you can design this component of your data collection form. If you are assessing whether a study met certain key criteria you need to list them, using either a checklist or the categorising of risk of bias table as outlined in Section 6.7.2 of the Reviewers’ Handbook. Some examples of criteria you may wish to include are:

- Selection Bias (Was the study randomised? Was allocation concealment adequate?)
- Performance Bias (Were participants and care providers blind to the intervention? Were there any co-interventions?)
- Attrition Bias (Were all participants randomised accounted for in the analysis? Were any withdrawers described?)
- Detection Bias (Was there a blinded assessment of outcome)
- Analysis (Was the analysis appropriate?)

If you are using a specific published validity assessment scale (and there are many) you will need to reproduce it as part of your form.

Design your form now, bearing in mind all the aspects of validity assessment you wish to use. You will need to test out how your form works, but we will discuss that later.
Part Five of a Data Collection Form: Data for Results

The final section of your data collection form is where you record the results of the included studies in a format to allow later entry into RevMan. It is this part of the form that may be transferred electronically into RevMan if it is set up as a spreadsheet, although many reviewers prefer a paper based system with later data entry.

Before we go on to discuss the design and use of this part of your data collection form, you need an understanding of the type of data you will need to extract in your review. Go to Module 10: Planning the Analysis and complete it, along with the Additional Module A1 on continuous data if any of the outcomes in your review are measured on a continuous scale (for example blood pressure, pain on a visual analogue scale), then return to here and complete this module.

To design the rest of your data extraction form you will need to formulate a table to allow recording of data available from the study for each outcome under each comparison. Your table may look something like this:

<table>
<thead>
<tr>
<th>COMPARISON 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention A</td>
</tr>
<tr>
<td>Outcome 1</td>
</tr>
<tr>
<td>Outcome 2</td>
</tr>
<tr>
<td>COMPARISON 2</td>
</tr>
<tr>
<td>Intervention C</td>
</tr>
<tr>
<td>Outcome 1</td>
</tr>
<tr>
<td>Outcome 2</td>
</tr>
</tbody>
</table>

If you are performing subgroup or sensitivity analyses, you may also want to include room in the table to collect information about which group the study belongs to.

You should by now have an idea about the type of data you are likely to be collecting from the results sections of your included studies. This will relate to your outcomes. Think about whether your results are likely to be in the form of dichotomous (for example dead/alive, smoking/not smoking) or continuous (for example blood pressure, pain on a visual analogue scale) outcomes. For dichotomous outcomes you will need to extract the number of participants experiencing the outcome and the total number in the group. For continuous outcomes you will need to extract the number of participants, the mean and the standard deviation for each group.
A useful hint is to allow sufficient space on your data collection form to record any conversion calculations (for example calculating a standard deviation from a standard error) or notes.

Now have a try at designing this section of the form for your review.

**Testing the form**

Now you have designed your form, you need to test it to see if it works. The important features of a data collection form are that it is easy to use, comprehensive (it is very frustrating to get further along your review and find you need a piece of data you didn’t record) and consistent across reviewers. Questions or items need to be phrased in an unambiguous way and you need to check they mean the same thing to all your reviewers.

Section 7.7 to 7.10 of the *Reviewers’ Handbook* discusses the testing of data collection forms and you should read it now.

In summary the main components to testing your data collection form are:

- Pilot testing the form with a sample of studies to ensure it is understandable, easy to complete and comprehensive
- Reliability testing the form in a more formal way by comparing a sample of completed forms by two or more reviewers to ensure they are in agreement.

**Using the form**

Two issues you may need to consider when starting to use your form to extract and record data are whether or not you need to blind reviewers and what to do about missing data. These issues are covered in Section 7.9 and 7.10 of the *Reviewers’ Handbook*. As with the decision about trial inclusion and quality assessment, bias and error are likely to be reduced if more than one reviewer extracts the data, independent of each other. Whether or not bias is further reduced by blinding the reviewer to the source and authors of the study is yet to be determined, but some reviewers have gone to considerable lengths to mask studies of all identifying information.

By the completion of this module you should have designed and tested your data extraction form and will have also completed Modules 8, 9 and 10 as you worked your way through designing your data extraction form. Progress now to Module 11 for some more detailed information about the analysis components of your review.
## Example from a review in schizophrenia

<table>
<thead>
<tr>
<th>Trial ID</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td></td>
</tr>
<tr>
<td>Methods</td>
<td></td>
</tr>
<tr>
<td>Allocation:</td>
<td></td>
</tr>
<tr>
<td>Blindness:</td>
<td></td>
</tr>
<tr>
<td>Duration:</td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td></td>
</tr>
<tr>
<td>Diagnosis:</td>
<td></td>
</tr>
<tr>
<td>N=</td>
<td></td>
</tr>
<tr>
<td>Age:</td>
<td></td>
</tr>
<tr>
<td>Sex:</td>
<td></td>
</tr>
<tr>
<td>History:</td>
<td></td>
</tr>
<tr>
<td>Excluded:</td>
<td></td>
</tr>
<tr>
<td>Interventions</td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
</tr>
<tr>
<td>Outcomes - able to use</td>
<td></td>
</tr>
<tr>
<td>Service outcomes</td>
<td></td>
</tr>
<tr>
<td>Global impression</td>
<td></td>
</tr>
<tr>
<td>Mental state</td>
<td></td>
</tr>
<tr>
<td>Social functioning</td>
<td></td>
</tr>
<tr>
<td>Adverse effects</td>
<td></td>
</tr>
<tr>
<td>Economic data</td>
<td></td>
</tr>
<tr>
<td>Leaving the study early</td>
<td></td>
</tr>
<tr>
<td>Outcomes - unable to use</td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td></td>
</tr>
</tbody>
</table>

© The Cochrane Collaboration 2002
Example from the Effective Practice and Organization of Care CRG

Study design:
- RCT
- CCT
- CBA
- ITS

Interventions:
- **Type of intervention** (for each comparison group the components/categories of the intervention received are recorded, as listed under TOPICS)
  - Professional
  - Financial
  - Organisational
  - Regulatory
- **Controls**
  - Characteristics of the intervention
  - Evidence base
  - Purpose of recommendations (e.g. appropriate management, cost containment)
  - Nature of desired change (initiation of new management, stopping introduction of new management, reduction, increase, cessation or modification of established management)
  - Format
  - Source
  - Based upon implementation of clinical guidelines
  - Guidelines developed through formal consensus process
  - Recipient
  - Deliverer
  - Timing
  - Setting
  - Source of funding
  - Ethical approval
  - Type of targeted behaviour

Participants:
- Characteristics of participating providers
  - Profession
  - Level of training
  - Clinical specialty
  - Age
  - Time since graduation/years in practice
- Characteristics of participating subjects
  - Clinical problem
  - Number of subjects included in the study
  - Episodes of care
  - Patients
  - Providers
  - Practices
  - Hospitals
  - Communities or regions
  - Other characteristics (e.g. age, gender, ethnicity)

Setting:
- Reimbursement system
- Location of care
- Academic status
- Country
- Proportion of eligible providers from the sampling frame

**Methods:**
- Unit of allocation
- Unit of analysis
- Study power
- Methodological quality

RCT/CCT
CBA
ITS

- Consumer involvement
Prospective identification of barriers to change

**Outcomes:**
- Description of the main outcome measures (health professional outcomes/process measures, patient outcomes, costs)
- Length of time outcomes measured after initiation of the intervention
- Length of post-intervention follow-up period
- Possible ceiling effect eg little room for improvement in provider performance because it was adequate without the intervention (based on baseline measurements or control group performance)

**Results:**
- RCTs and CCTs
The results for the main outcomes in natural units
The baseline performance and post-intervention differences between study and control groups (including statistical significance if reported; and indicating if the units of randomisation and analysis were different)
- CBAs
The results for the main outcomes in natural units
The baseline difference in the pre-post intervention change between groups
For each available comparison, the difference across study groups of the pre-post intervention change (including statistical significance if reported; and in all cases reporting a more favourable provider/patient outcome in the intervention group as a positive finding ie where differences in the groups are in the intended direction)
- ITSs
The results for the main outcomes in natural units (in all cases reporting a more favourable provider/patient outcome attributable to the intervention as a positive finding ie where changes in the outcome are in the intended direction)
Amiodarone for the prevention of sudden death in people with heart disease

Study eligibility form

Type of study
Q1. Is the study described as randomised?
    Yes  Unclear  No
    \[Go to next question\]
    \[Exclude\]

Participants in the study
Q2. Did the participants in the study have heart disease?
    Yes  Unclear  No
    \[Go to next question\]
    \[Exclude\]
    \[NB Answer ‘no’ if participants had hypertrophic cardiomyopathy\]

Q3. Were the participants reported to have, or be at risk of, ventricular ectopics (VE or VEA), ventricular tachycardia (VT) or ventricular fibrillation (VF)?
    Yes  Unclear  No
    \[Go to next question\]
    \[Exclude\]
    \[NB Answer ‘no’ if the study reports that participants had, or were at risk of, atrial arrhythmia or had no rhythm problems\]

Interventions in the study
Q4. Was one group given amiodarone by mouth/orally (not intravenously) for at least 3 months?
    Yes  Unclear  No
    \[Go to next question\]
    \[Exclude\]

Q5. Did another group receive the same care, with or without placebo, but without amiodarone? \[NB Answer ‘no’ if this group could receive an implantable defibrillator\]
    Yes  Unclear  No
    \[Go to next question\]
    \[Exclude\]

Outcomes in the study
Q6. Did the study report deaths and follow participants up for at least 3 months?
    Yes  Unclear  No
    \[Include, subject to clarification of ‘unclear’ points\]
    \[Exclude\]

Final decision
Include  Unclear  Exclude

© The Cochrane Collaboration 2002
Module 7  Page 12
Module 8: Selecting studies for your review

This module will discuss how you decide which of the studies your searching found should be included in the review. We will look at ways you can do this as reliably as possible and with the minimum of bias.

Learning objectives

- Understand that selecting studies is a multi-stage process
- Be able to design an eligibility form
- Be aware of the need to pilot inclusion criteria
- Be aware of ways to improve the reliability of study selection
- Be aware of the existence of duplicate publications
- Be aware of ways to obtain sufficient information to make a reliable decision on study eligibility

Relevant section of the Reviewers’ Handbook

- Section 5.7

Other relevant material


Where does this go in a Cochrane review?

- Early on in the Methods section of protocol and review
This module relates to Section 5.7 of the Reviewers’ Handbook and you should read this now.

**Sift and sift again**

We saw in the previous module, *Searching for studies*, that in a systematic review we attempt to find every study that has ever been done addressing our question. We try to do this by running sensitive searches. Inevitably, when we do this, we find lots of reports of studies that *could* be relevant, and we then have to decide which ones *are* relevant to our question.

It’s important to remember that our decisions about which studies to include should be based on the design of those studies, and *not* the results. If we allow ourselves to be swayed by the results of the studies, we might exclude a perfectly eligible study because we don’t like or believe the results.

If you have already run some searches, you will know that these sensitive searches usually turn up hundreds or thousands of records. Most of these will have come from an electronic database, and you’ll only have limited information about the study, like the example below.

ID:CN-00240563  
TI:Antibiotic prophylaxis of wound infections in skin surgery [see comments].  
AU:Bencini, P. L., Galimberti, M., Signorini, M., and Crosti, C.  
SO:Archives of Dermatology  
YR:1991  
VL:127  
NO:9  
PG:1357-60  
AB:A controlled prospective study of 2165 outpatients undergoing skin surgery was performed to evaluate the utility and the effects of several antibiotic schedules for prophylaxis of wound infections. The patients were divided into four groups. Twenty-three of the 541 group A patients, given no antibiotics, had wound infections. Eight of the 542 group B patients, given systemic antibiotics from immediately after surgery until the third day, had wound infections. Four of the 540 group C patients, treated only with local sterile antibiotic powder sprinkled into the wound during surgery, had wound infections develop, and only one infection occurred in the 542 group D patients given systemic antibiotics from 2 days before surgery until the second day after surgery. This last schedule was the best for prophylaxis of wound infections in contamination-prone regions. Local antibiotic administration is a simple method for prevention of infections in routine skin surgery.  
KY:Administration, Cutaneous; Adolescence; Adult; Bandages; Cefazolin; Ad [Administration & Dosage]; *Cefazolin. Tu [Therapeutic Use]; Female; Human; Incidence; Injections, Intramuscular; Italy; Ep [Epidemiology]; Male; Middle Age; Powders; Premedication; Prospective Studies; Skin; Pa [Pathology]; *Skin; Su [Surgery]; Skin Neoplasms; Pa [Pathology]; *Skin Neoplasms; Su [Surgery]; Staphylococcal Infections; Ep [Epidemiology]; Surgical Wound Infection; Ep [Epidemiology]; *Surgical Wound Infection; Pc [Prevention & Control]; Time Factors  
DE:RCT.  
CC:SR-HANDSRCH, SR-SKIN
Activity: write a list of the further information you think you might need to tell whether this study is relevant

The first sift – pre-screening - is to decide which studies to retrieve in full.

The second sift – selection - is to look again at these studies and decide which are to be included in your review

It isn’t usually possible to be certain from this sort of record whether it will turn out be included in the review or not. There usually simply isn’t enough information to be quite sure. Try to list the extra information you might want to help you decide if this was a randomised controlled trial relevant to a review of antibiotic use for preventing wound infection in skin surgery.

Here are a few suggestions

- Was the study randomised and how (ie how were the patients divided into different groups)?
- What type of surgery was performed?
- How was resultant infection determined?
- Were all patients followed up
- Did they have any other type of intervention or co-morbidity that may have influenced the results?

What is normally possible by looking at these records is to tell whether the study might be relevant or is very unlikely to be relevant. You can then order the full paper copy of the ones that you think might be relevant. This saves you getting full paper copies of hundreds of articles that turn out to be irrelevant.

Once you have these hard copies of the probably relevant studies, you can get on with comparing them with your review’s inclusion criteria. This second stage is where you make definite decisions about whether studies are included or excluded from the review.

Practicalities of sifting

A quick note on language

The information you’re using for pre-screening and selection might not be written in your first language. When looking at studies in a language you cannot read, you are likely to need help. How much help you need depends on whether you can identify certain key words to tell you whether a study might be relevant. If not, you will probably need to find someone who is familiar with that language to help. They may just be able to read an abstract and tell you whether it is worth ordering the full article. With a full paper, you may just be able to ask someone to read it and tell you whether it is eligible, or you may need to get certain parts, or all, of the paper translated. You may know people who speak a variety of languages, or your review group may be able to help you find a translator if necessary.
Human error

Both stages of sifting are going to be done by humans. Humans make mistakes and people doing reviews are no different. When looking at a lot of studies we may simply miss some information and mistakenly include or exclude the study. We all have certain prejudices or biases which might make us more or less likely to choose a particular study. Part of the review process is to try to minimise these mistakes and biases.

An important point is that if you exclude a study in the initial sift, it will rarely get another chance to be included. So reviewers usually give studies the benefit of the doubt at this early stage, and go on to obtain the full report.

Should you collect outcome data at the same time as eligibility information?

It might seem that since you are reading these papers anyway for making eligibility decisions, why not collect outcome data at the same time? Why read all the papers again later?

There are two main arguments against this. The first is simple – if you decide to exclude a study, you will have extracted its data unnecessarily, and will have wasted time and effort. We mentioned the second reason earlier on – you should try to avoid letting the results of studies sway your decision. If you look too hard at the results too soon, you might find it difficult not to be swayed.

Reducing mistakes

How can we reduce mistakes? A common sense way to try and do this is to have more than one person make the decisions. Then, if one person makes a mistake, there will be one or more others who might get it right. In one review, pairs of reviewers independently sifted about 11,000 records. It took the reviewers between 11 and 28 hours to go through all these records. The reviewers concluded that having two people check each record was worthwhile. The reference for this study is listed at the start of the module under ‘Other relevant material’. Therefore, assuming that your review has many fewer records than that large review, the time needed for this part of your review might be less than you had thought. So you should think about having more than one of you do this stage.
If you can’t find one or more other people willing to commit the necessary amount of time, there are a couple of compromises

- Have a second person look at a sample of the records
- Have two reviewers each look at separate sets of records, and overlap the sets they look at so that they both do some of the records (for example one looks at years 1966-1988 and the other looks at 1986-2000)

If you find, through either of these ways of working, that you disagree on quite a few records, you may need to look again at the criteria you have set for the sifting of the records of studies. There are no rules about the level of disagreement above which you need to go back and redo the work. It should be low, as mistakes here might mean that you leave out relevant studies, which are unlikely ever to be looked at again.

If you do think you need to go back and recheck this sifting process, the most obvious option is to resift all the records, with two or more reviewers looking at all of them. Another option is for one of the reviewers to sift through all the records they did not look at first time.

For the second stage of the sifting, there will be fewer reports and it seems common sense that at least two reviewers look at each of them. Of course, there may be occasions when you think the decisions are likely to be so straightforward that this isn’t necessary. Whatever you decide to do, report it in the Methods section of your review.

**Reducing bias**

All of us have prejudices that might affect our judgements about whether studies should be included or excluded. Experts may have pre-formed opinions which might affect their assessments of the relevance and validity of studies. On the other hand, it’s difficult to make judgements if you know nothing at all about a topic. Other people might have opinions about the value of research published in particular journals, or research carried out in particular institutions.

One way to minimise the effect of these personal biases is to have reviewers of different backgrounds making judgements about studies, for example an expert and a non-expert. For this to work, however, both reviewers need to be willing to accept that they may have biases and to listen to the other reviewer’s views!
One suggestion for reducing bias is that we should remove as much identifying information from the papers as possible, such as the name of the journal, authors, institutions, etc. A randomised trial has been done to assess this approach (you can find the reference as Berlin 1997 in the Reviewers’ Handbook, section 5). The trialists repeated several meta-analyses, once with reviewers who saw the unblinded papers of the included studies, and once with reviewers blinded to many of the identifying details in the papers. It took many hours to remove as much identifying detail as possible, and it appeared to make little difference to the overall estimate of effect for the meta-analysis. So, unless someone refutes these results, it looks as though we don’t normally need to ‘blind’ the papers, unless we think there are important reasons why not blinding the papers would be a problem for a particular review.

Another interpretation of the above study is that we are just not very good at concealing the identity of the papers. For instance, experts will often have read the paper before and may recognise the text.

**Getting hold of extra information**

You will often find that, even when you have the full report of a study, you don’t have all the information you need. Maybe it doesn’t tell you how the study was randomised, or maybe it doesn’t tell you clearly who the participants were. This may happen later on as well, when you need outcome data.

You may be able to make assumptions about a study that everyone would agree are reasonable. For instance, if a report of a study tells you that it recruited women aged over 60, it is reasonable to assume that the women were all post-menopausal. There isn’t really a need to ask the authors to clarify this. If you cannot make assumptions like this, you may need to create categories for features of trials called ‘unclear’ – it was ‘unclear’ whether the patients were masked to the treatment, etc.

The other way to deal with the situation is to try to get hold of more information. To do this you either need to find other reports of the same study, which contain more information, or try to contact the people who conducted the study. Make sure you’ve looked carefully for other reports – they will probably have come up in your search anyway, but you could do a quick extra search for other publications by the same author.
Resolving disagreements

Just as surely as we all make mistakes, people will disagree about the way they see things. So, when you have two people making decisions about including and excluding studies, there will be times when you come to a different decision about a study.

Some people like to measure and report how often this occurs as part of their review, to give readers an idea of how difficult the decisions were. The Cochrane Collaboration doesn’t insist that this is done.

To resolve differences, you first need to work out why you came to different judgements. It may just be that one person missed a vital sentence. It may be that there wasn’t enough information – in this case you might need to put that study aside and wait till you’ve got the extra information you need. Or maybe one reviewer has special knowledge about a study that isn’t in the report – they may even have been one of the authors. If you really can’t agree, ask another person to help you resolve this. You might choose another content expert or a methodologist, depending on the area of your disagreement.

It is important to plan how you are going to resolve disagreements early on in planning your review and to ensure your review team are happy with the planned process. This will help resolve any possible later disagreement.

Getting your work into RevMan

Once you’ve decided which studies are included, excluded, awaiting assessment (the ones where you needed more information), or ongoing (not yet completed), you’ll need to put them in the appropriate category in RevMan. This is explained in RevMan ‘Help’ under the heading ‘study references’, and is covered in the RevMan exercise.
Pitfalls and problems - duplicate publications

Some studies result in more than one publication. Authors may publish the methods of the study, present preliminary data at a conference resulting in an abstract, then publish some results, and later publish longer term follow-up. There’s nothing wrong with this, as long as you can tell it’s all about the same study.

Sometimes, studies may be published more than once for other reasons – more publicity, more papers on the author’s CV, or to allow different authors to be first authors. Sometimes, it’s not easy to tell whether they are reports of the same study. This can cause problems for reviewers because we might count a study more than once and so give extra weight to it in our review.

This phenomenon was studied by Martin Tramèr and colleagues (Tramèr MR et al BMJ 1997;315:635-640.)

They found that 17% of reports of trials of a drug were duplicates and 28% of the data were duplicated. If this had not been spotted and a review had been done counting studies more than once, the drug would have appeared more effective than it actually was. So, we need to be alert to the possibility of duplicates. Look for

- Same authors in different orders
- Similar study inclusion and exclusion criteria
- Many reports of a study done in the same place at the same time
- Results tables that look familiar

Now you understand the process of selecting studies for your review, return to Module 7 and we will work on designing this Section of your data collection form.
Module 9: Assessing quality of studies

This module will discuss the stage in a review when, having found studies relevant to the question, we assess the quality of studies.

Learning objectives

- Understand what is meant by the ‘quality’ of a trial
- Be aware of empirical studies investigating the relative importance of different aspects of quality
- Be aware of problems in using quality scales to measure quality
- Be aware of ways to reduce bias in quality assessment
- Be aware of methods of incorporating quality in a meta-analysis, for instance use of a threshold for a quality score (requirement of minimal characteristics), quality related sensitivity analyses, and weighting

Relevant section of the Reviewers’ Handbook

- Section 6

Where does this go in a Cochrane review?

- The Methods section of your protocol should describe how you plan to do this, and then you should describe how you did this in your complete review
- The section ‘Methodological quality of included studies’ should summarise the quality of all studies, but not describe the quality of individual studies in great detail, although any important flaws in individual studies should be noted here
- The ‘Characteristics of included studies’ table should include details of the methodological quality of individual studies. A code for the quality of allocation concealment is also entered in this table

© The Cochrane Collaboration 2002
Garbage in, garbage out

This is a phrase that you may have heard in relation to systematic reviews. What people are worried about is that a reviewer might be collecting together poor quality studies and then presenting the results as if they are high quality. This is a real concern – putting together a group of biased studies is likely to give a biased answer.

So, having selected which studies are included in your review, you need to look at the quality of them.

The topic of quality assessment is covered in depth in Section 6 of the Reviewers’ Handbook, so you’ll need your copy close to hand and may want to look at it now.

This module will cover the quality of randomised controlled trials, and not the quality of other study designs.

What do we mean by quality of trials?

Write a list of all the factors that come to mind when you think about the quality of a randomised trial.

You’ve probably come up with quite a long list. Some things on the list will be to do with the design of the study (blinding, sample size), and some to do with the way it was reported (the presentation of tables, who the authors are).

This long list in part explains why there have been many different approaches to measuring quality. So which bits are we interested in for this part of a Cochrane review?

Randomised trials, like systematic reviews, are trying to measure some ‘truth’ about an average effect of an intervention in a group of participants. When we talk about trial quality, we’re usually talking about how well we think the study has measured this ‘true’ effect – this aspect of quality is also called validity.

But since no-one knows what the ‘true’ effect is, there’s an element of guesswork and judgement in knowing which factors are most likely to affect how the study measures it. This is another reason for the many different approaches people have taken.
Measuring validity – measuring what?

The approach taken within the Cochrane Collaboration is first of all to think of factors which might lead to bias in studies. Then we take into account any studies that can tell us which of these are more important. Once we’ve decided which factors we’re interested in, we collect information about them from the study reports.

Now read the first few sections of section 6 of the Reviewers’ Handbook, for an explanation of the sources of bias and which seem to be more important.

In summary, empirical evidence suggests that

- Allocation concealment is very important in protecting against bias, so we should look for information on what was done about this and report what we find
- Blinding of the interventions and the outcome assessment may be important
- Loss to follow-up of study participants may be important

Some people choose to report information on all these, but only assign a score to allocation concealment. Others choose to collect more items, and some use scales to get with overall quality scores. These scores seem attractive in comparing studies with each other, but there are problems with them. Since we have no ‘gold standard’ to check them against, it is difficult to tell how well the scales measure quality. Some scales contain items that are not really to do with validity, and there are worries that adding things up to produce an overall score makes assumptions about the relative importance of different items.

This is a hotly debated topic within the Cochrane Collaboration, and your review group may have a policy on how to assess quality. You could also look for Cochrane reviews that have addressed similar questions to see what those reviewers did. If you do plan to use a scale to measure quality, then you should choose one that has been tested for its reliability and validity, as with any scale.
Practical issues, minimising mistakes and bias

As with eligibility decisions, you’ll want to record the information you collect on quality somewhere. Most people record it on their data collection form (we will cover this when we return to Module 7) and put it into the ‘Table of Included Studies’ under the ‘Methods’ column in RevMan later.

The issues in minimising mistakes and bias while assessing study quality are the same as when we considered these in selecting studies, and you should report how you intend to assess quality, and how you did it:

- How many reviewers will do this?
- What are their backgrounds (for example, are they authors on any of the trials)?
- Will you try and blind the reviewers to details of the papers, such as journals or authors?
- Will you formally assess the reliability of the process

Using information about validity in your review

So what should you do with this information you collected?

There are four approaches

- As a threshold for inclusion of studies
- As a possible explanation for differences in results between studies
- In sensitivity analyses
- As weights in statistical analyses

These are discussed in detail in section 6.10 of the Reviewers’ Handbook, which you should now read.

Everyone uses some sort of threshold for inclusion of studies, and this was set when we chose the selection criteria for the review. Sensitivity analyses are commonly undertaken in reviews to see the effect of aspects of quality. Because of the problems in deciding the weight to allocate to different aspects of quality, this is rarely done in Cochrane reviews. As a minimum, you should discuss the range of quality found in your included studies, and whether you think this may have any influence on the results of these studies, and any meta-analysis you performed. We will return to this topic in a later module, when we look at the investigation of heterogeneity.

Finally, your review group may have a policy on how to incorporate study quality in reviews, so it is worth checking with them.
Activity:
Complete the following activity of assessment of a randomised controlled trial and check how you did against the answers that follow.

Read the trial about a nursing intervention for treating heart failure. You will find the report of the trial at:

http://www.bmj.com/cgi/content/full/323/7315/715?maxtoshow=

Appraise this article as you plan to for your review (either qualitatively or by using a scale). Ensure you extract information about allocation concealment, blinding and loss to follow up.

Test yourself by comparing your answers to the suggested responses on the next page.
**Randomisation and allocation concealment.**
Assignment to groups was determined by central off-site randomisation and so could not be influenced by the investigators, nurses or participants. Allocation concealment was therefore adequate.

**Blinding**
This trial is an example of a complex, procedural intervention where complete blinding is not possible. Attempts to blind participants may be made either by designing an appropriate placebo or partially concealing (within ethical boundaries) the purpose of trial from participants. Neither was done in this trial. Obviously, in a procedural trial it is not possible to blind the health care workers (unlike a drug trial when an identical placebo will blind the carer as well as the participant). Outcome assessment (determining rates of death and rates and cause of admission) was blind to treatment allocation. A potential problem with outcome assessment is the use of hospital and health department records to determine death and readmission, with the possibility of records missed. A more appropriate method of assessing outcome would have been for the investigators to formally follow-up participants at given time points.

Here is a good time to make note of a common problem in assessing the adequacy of blinding within randomised controlled trials. Often you will see a trial, and indeed validity scale, referring to the trial being “single blind”, “double blind” and “triple blind”. The problem with this phrasing is that, while single blind obviously refers to only one party being masked from the intervention, double blind two parties etc, it does not make it clear who it is without the knowledge. A triple blinded trial is usually referring to intervention being masked from participants, care providers and outcome assessors, but what of double blind? Some people interpret this as participants and care providers, others as participants and outcome assessors. And what of single blind? As you can see, a much better way of describing blinding within the methods of a trial, and in any scale you chose to use to assess its validity, is to specify which party or parties were unaware of the intervention.

**Loss to follow up**
165 patients were randomised and 156 completed the trial. A description is provided of those who failed to complete (see Figure 1). From Table 3, we can see intention-to-treat was used as the column heading, showing 81 participants in usual care and 84 in intervention group. This value is the number randomised.

Now return to Module 7 and incorporate what you have learned into your data extraction form.
Module 10: Planning the analysis

This module covers how you go about planning the analysis section of your review.

Learning objectives

- Be aware of the rationale for describing the analysis in advance
- Be able to derive the main comparisons from the review question
- Be aware of the different types of data that might be found in reviews
- Be able to state how a decision will be made on whether to perform a meta-analysis

Relevant sections of the Reviewers’ Handbook

- Section 8.2

Where does this go in a Cochrane review?

- You will need to include a plan of analysis in the Methods section of your protocol.
Specifying your analysis in advance…...as far as possible

The final part of the Methods section in a protocol for a systematic review covers the plan of analysis. Ideally, you would tell the readers exactly what you are planning to do when you have collected all the study results together. This would mean that your analysis could not be influenced by the results you have looked at. In practice, it is not always possible to specify everything you will do in advance.

The compromise is to specify as much as possible in the protocol. Where you make decisions after seeing the study results, you should report exactly what you’ve done in the review, and usually present some analysis to show what would happen if you made a different decision.

Now would be a good time to read Section 8.2 of the Reviewers’ Handbook, which covers some issues in planning the analysis for your review.

Here’s a checklist to use when writing the Methods section, followed by some practical tips

- What are the main comparisons in your review?
- For the outcomes you specified, how will you summarise the result for each study?
- How will you decide whether to combine the results of the separate studies?
- Do you plan any subgroup or sensitivity analyses?

Specifying the comparisons

This should be fairly easy, as you should have thought it through carefully when you were specifying the types of intervention you were interested in.

As an example, imagine you are working on a review of whether corticosteroids (in either oral or intravenous form) were better than non-steroidal anti-inflammatory drugs (NSAIDS) or placebo for early rheumatoid arthritis.

Try writing a list of the possible combinations

A versus B
A versus C
B versus C ……and so on

Now set up a list of comparisons for your review. Draw on the Table you completed in Module 5 to help you do this.

© The Cochrane Collaboration 2002
Summary statistics for individual studies

If you manage to collect useful outcome data from the studies you have included in your review, you will have some choices about how to express the outcome of each study.

These are covered in more detail in subsequent modules. For now, think about the types of data you are likely to have:

- Are they dichotomous data (such as alive/dead, smoking/not smoking)?
- Are they continuous data (such as weight in kilograms or blood pressure)?

These are the main two data types you can analyse using RevMan.

The other two options in RevMan are “Individual patient data” and “Other data”. Individual patient data reviews involve collecting data on each individual participant in the study. This takes a lot of time and effort, but enables more detailed analysis of the relationship between study characteristics (age of the participants, for example) and the effect of the intervention than is possible using summary data for the groups of participants. The “Individual patient data” option allows reviewers who have conducted these types of analyses to put the results into RevMan. More information on the use of individual patient data is given in section 11 of the Reviewers’ Handbook. The option for other data is used to present data in a table without doing any statistical analysis on it. This might be used for qualitative data, or quantitative data where it is inappropriate to calculate pooled averages.

Choosing between different ways of presenting dichotomous and continuous outcomes is discussed in the RevMan exercise, under the heading ‘Issues in analysis’. If you haven’t yet reached that part, now would be a good time to complete the whole exercise.

The RevMan exercise gives a basic introduction to the issues, which will be covered in more depth in future modules. For your protocol, you need to state how you plan to present these summary statistics for studies. Your review group may have a policy on this, so it’s worth checking.
**Should I combine studies?**

This will be covered in more detail in a future module. For now, you need to make a few decisions about whether it is likely to be appropriate for your review to combine studies in a meta-analysis or not. You’ve already thought through a lot of this when defining the question – were there groups of participants, or categories of intervention that were so different you think the effect of the treatment is likely to be very different? Or do you have reason to believe the effect of treatment will vary over time? If so, it may be inappropriate to combine different studies as they are measuring a different effect.

The other key factor in deciding whether to combine studies is called statistical heterogeneity. This means variation between studies in the measured effect of treatment. This is covered in the module on heterogeneity. For now, you need to know that it is usually inappropriate to calculate an average effect (that is, perform a meta-analysis) if there is a large amount of heterogeneity. You’ll find out more about how the significance of this is investigated later on. Again your review group may have a policy on wording this part of the methods section.

**Subgroups and sensitivity analysis**

If there were some types of participant, intervention or outcome you thought were likely to be quite different to the others, you might plan a subgroup analysis. For example, you may want to know if anti-inflammatory medication reduces pain in shoulder disorders, but not all shoulder disorders will respond the same. You may therefore decide to look at individual disorders (eg rotator cuff tendinitis, frozen shoulder, arthritis) as separate subgroups, then combine them in a pooled analysis to gain overall effect for any shoulder pain. A subgroup analysis will allow you to look at results separately and together for that analysis.

You should specify subgroup analyses in advance. This is because choosing subgroups after you have seen the results of studies may introduce bias.

Sensitivity analyses investigate the effect of our decisions in the way we conduct a review. For example, we choose a threshold for the quality of studies, usually only including randomised controlled trials. A sensitivity analysis would allow the investigation of what would happen to the result of the review if we made this criterion even higher by only accepting randomised controlled trials with excellent allocation concealment.

As with subgroup analyses, you should specify sensitivity analyses in advance, to minimise bias.
Remember, though, that the more subgroup or sensitivity analyses you do, the more likely you are to find a statistically significant result by chance. So you should have good reason for each analysis you do, with, ideally, some independent evidence to support this and to predict the direction of any differences you find in these analyses. Always remember that you may be finding falsely significant results and that all findings of subgroup analyses should usually be seen as hypothesis-generating and not as proof in themselves.

Now return to Module 7 and finish designing your data extraction form, incorporating your planned analysis into the section for extracting the results of studies.
Module 11: Summary statistics for dichotomous outcome data

This module introduces some basic statistics for dealing with dichotomous outcomes. We discuss ways of summarising results and determining treatment effects within single trials, and prepare for meta-analysis.

Learning objectives

- Understand the difference between risk and odds
- Be able to calculate risk ratios and odds ratios from a 2×2 table for a single trial
- Be able to calculate risk differences and numbers needed to treat (NNT) from a 2×2 table for a single trial
- Be able to report and interpret risk ratios, odds ratios, risk differences and NNTs and their confidence intervals
- Be able to choose a suitable summary statistic for the meta-analysis and reporting of dichotomous data

Relevant sections of the Reviewers’ Handbook

- Sections 8.3.1, 8.4.1

Where does this go in a Cochrane review?

The information in this module will be relevant to many parts of your review

- In the data analysis part of the Methods section of a protocol or review, where you will describe what statistical techniques you are planning to use
- When (or if) you actually perform meta-analyses using the analysis part of RevMan or other software
- In the presentation of results in the Results section of the review
- In the interpretation of results, in the Discussion of your review
Different types of data

There are several different types of data you may come across in your included trials. Some of the more common data types are:

- **Dichotomous data** are data from outcomes that can be divided into two categories (e.g. dead or alive, pregnant or not pregnant), where each participant must be in one or other category, and cannot be in both.
- **Counts of events** (for example number of epileptic fits)
- **Short ordinal scales** or scales with a small number of categories where there is a natural order to the categories (for example a pain scale of “none/mild/moderate/severe”)
- **Long ordinal scales** or scales with a large number of categories with a natural order (for example the Short Form-36 scale for assessing quality of life or a depression index)
- **Continuous data** which are data from outcomes measured on a continuous scale (for example blood pressure, range of motion of a knee joint)
- **Censored data or survival data** (such as time to recurrence of cancer, where we will not have a measurement on everyone at the end of the study, because some haven’t had a recurrence of cancer).

This module, and the next, will discuss issues of dichotomous data, as most Cochrane reviews contain data in this form. If you have continuous outcomes in your review then you will need to complete Additional Module 1 after you have completed Modules 11 and 12.

Dichotomising data

We often make our own dichotomous data from outcomes that are not truly dichotomous, so that they are easier to manage and understand. For example, converting blood cholesterol (measured on a continuous scale) to ‘high cholesterol’ or ‘not high cholesterol’ dichotomised around a clinical threshold above which you would consider the cholesterol to be high; or converting pain measured on a short ordinal scale to ‘absent or mild’ or ‘not absent or mild’ (by which we mean moderate or severe). Generally long ordinal scales, or scales with a large number of discrete categories, are treated as continuous data for the purpose of analysis.
Sometimes, censored data are converted into dichotomous data by counting the number of people who have had the event by a particular time (such as the number of people who have a recurrence of cancer within 5 years of an operation). This should only be done when all participants have been followed up to the particular time point.

The benefits of converting non-dichotomous data into dichotomous data relate to ease of analysis and interpretation. Of these, the more important is ease of interpretation. Dichotomous outcomes may be easier for decision makers to understand and make judgements about.

The down side of converting other forms of data to a dichotomous form is that information about the size of the effect may be lost. For example a participant’s blood pressure may have lowered when measured on a continuous scale (mmHg), but if it has not lowered below the cut point they will still be in the ‘high blood pressure group’ and you will not see this improvement. In addition the process of dichotomising continuous data requires the setting of an appropriate clinical point about which to ‘split’ the data, and this may not be easy to determine.

**Summarising dichotomous data**

In studies of treatment interventions we aim to describe what is happening to the participants we are studying so we can predict what is likely to happen to others. In order to do this we take an observation about a particular outcome for each participant. If that outcome is dichotomous, then each participant can be in one of two states. For example if we have a new drug thought to save lives in high risk patients, we might test it first in a group of people representative of these high risk patients and observe the number dead or alive at the end of the intervention. There are a couple of ways of summarising the information we get about the whole of our observed sample in a form that can be applied to others.
Risk

The word *risk* is fundamental to epidemiology and evidence-based health. Risk is the chance, or probability, of having a specific event. As it relates to clinical trials and systematic reviews, risk is not always of a bad event, we can talk about the ‘risk’ of a good outcome (such as cure) as well as the ‘risk’ of a bad outcome (such as death). We can use the word ‘risk’ to describe the chance of the outcome whether it’s good or bad.

Given a single group of people, and knowledge of how many have ‘the event’, we can express the risk of the event by dividing the number with the event by the number of people. For example, of 133 women taking an antibiotic for the treatment of urinary tract infection (UTI), 14 had the event ‘still infected’ after 6 weeks. The risk of remaining infected was $\frac{14}{133} = \approx 0.1$.

Odds

An alternative measure of describing how likely an event is to happen is called odds. The odds of an event is the ratio of events to non-events. Equivalently (and more formally) it’s the risk of having an event divided by the risk of not having it. If we look at the 133 women taking the antibiotic for UTI, the ratio of events (still infected) to non-events (cured) is $\frac{14}{119} = \approx 0.1$. The more formal formula gives $\frac{14}{133}$ (risk of having the event) divided by $\frac{119}{133}$ (risk of not having the event), which also works out as $\frac{14}{119} = \approx 0.1$.

In this example the risk and odds are both similar (approximately 0.1), so why bother to have two alternatives? In this example the 133 women taking antibiotics were the treated group in a clinical trial. In this trial there was also a placebo group with another 148 women. Of the 148 receiving placebos, 128 still had a UTI after 6 weeks. So in this group, what’s the risk of staying infected? It’s $\frac{128}{148} = \approx 0.86$. What are the odds? They are $\frac{128}{20}$ (number still infected) / $20$ (number cured) = 6.4. So in this case the risk and odds are very different.

In fact odds and risk are never identical, but they can be similar. They are similar when they are both small – i.e. when an event is rare. Taking antibiotics, the women rarely stayed infected (the event was rare), so the risk and the odds were similar. Not taking them, they mostly stayed infected (the event was common) so the odds and risk were different.

As values of odds and risks can differ for the same data, it is important to be careful and precise when using statistical summaries of dichotomous outcomes.
Comparing two groups

In the section above we have talked about risk and odds as it applies to a single group of people. In clinical trials, so we can assess the effect of an intervention over and above the natural course of a disease, we usually compare how people respond in an experimental group to how they respond in a control group (i.e. we compare two groups of people). When we are dealing with a dichotomous outcome, we can either compare the risk of having the event between the two groups, or compare odds between the two groups.

Relative Risk or Risk Ratio

Relative risk or risk ratio (they mean the same thing and are both abbreviated as RR) is simply the risk of the event in one group divided by the risk of the event in the other group.

The most common way to go about calculating the risk ratio (and nearly all other statistics from dichotomous data) is to start by presenting your results in a 2x2 table, where each cell in the table contains the number of participants in each category.

<table>
<thead>
<tr>
<th></th>
<th>Event (Still infected)</th>
<th>No event (Not still infected)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>14</td>
<td>119</td>
<td>133</td>
</tr>
<tr>
<td>Antibiotics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>128</td>
<td>20</td>
<td>148</td>
</tr>
<tr>
<td>Placebo</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now, if you think through what you are comparing (risk in the treated group with risk in the control group), the risk ratio is easy to calculate.

RR = risk in the treated group / risk in the control group

\[ RR = \frac{\text{no. with event in treatment group}}{\text{no. in treatment group}} / \frac{\text{no. with event in control group}}{\text{no. in control group}} \]

\[ = \frac{14/133}{128/148} \]

\[ = 0.1/0.86 \]

\[ = 0.12 \]

If an experimental intervention has an identical effect to the control, the risk ratio will be 1. If it reduces the chance of having the event, the risk ratio will be less than 1; if it increases the chance of having the event, the risk ratio will be bigger than 1. The smallest value the risk ratio can take is zero when there are no events in the treated group.
Odds Ratios

Just as odds are an alternative way of expressing how ‘likely’ events are in a single group, odds ratio is an alternative way of comparing how ‘likely’ events are between two groups.

The odds ratio is simply the odds of the event occurring in one group divided by the odds of the event occurring in the other group. If we take the same data from our 2x2 table above,

\[
\text{OR} = \frac{\text{no. with event in treated group}}{\text{no. without event in treated group}} \div \frac{\text{no. with event in control group}}{\text{no. without event in control group}} = \frac{14/119}{128/20} = 0.118/6.40 = 0.018
\]

If an experimental intervention has an identical effect to the control, the odds ratio will be 1. If it reduces the chance of having the event, the odds ratio will be less than 1; if it increases the chance of having the event, the odds ratio will be bigger than 1. The smallest value an odds ratio can take is zero.

The difference between good and bad outcomes

Most dichotomised outcomes will be a dichotomy between a good and a bad event. When we describe risk, it can refer to the risk of having a good event or the risk of having the bad event, so ‘reducing risk’ could be a good or a bad thing. It is important whether we define ‘the event’ as the good outcome or the bad outcome as the results can change if we swap the good and bad outcomes around.

Taking the UTI example again, suppose we decide to define the event as cure. The risk in the antibiotic group is now 119/133 = 0.895 (i.e. 119 women were no longer infected) and in the placebo group it is 20/148 = 0.135 (i.e. 20 women were no longer infected). The risk ratio is therefore 0.895/0.135 = 6.6

Remember we previously calculated the risk ratio for remaining infected as 0.12. By swapping the good and bad outcomes we have changed the risk ratio from 0.12 to 6.6, but there is no simple relationship between these numbers. This makes it difficult to calculate one from the other without going back to the original data.
This means there are essentially two risk ratios: the risk ratio for a good outcome and the risk ratio for a bad outcome. There is quite a lot of work being done on this issue in the Cochrane Collaboration at the moment, but the general rule is that for outcomes which we aim to prevent (e.g. death, recurrence or worsening of symptoms), it is best to report the event as the bad outcome, which is usually the intuitive choice. For outcomes where we are trying to improve health (e.g. healing, resolution of symptoms, clearance of infection), we still do not know which option is best, but you should be very clear in your results section which outcome you are presenting. These rules are based on analysing which statistic is the most consistent – an issue we will discuss in more detail shortly.

For odds ratios, you still need to choose which outcome is the most appropriate to present, but it is easier to convert from ‘good’ to ‘bad’ outcomes or vice versa. From the example above you will see that the odds ratio of the “good event” of cured is

\[
\text{OR} = \frac{\text{odds in the treated group}}{\text{odds in the control group}} = \frac{119/14}{20/128} = 8.50 / 0.156 = 54
\]

From above we know the odds ratio when using the event ‘still infected’ is 0.018, and the odds ratio when using ‘infection cleared’ as the event is 54. These numbers are inversely related (working with more accurate numbers, we find \(0/0.01838 = 54.4\) and \(1/54.4 = 0.01838\)) and this is always the case with odds ratios. So in some senses it doesn’t matter whether we choose good or bad outcomes if we use odds ratios. Whichever we choose, it is vitally important that the results are very clearly reported so that those using the review are clear which outcome they are looking at.

**When do risk ratios and odds ratios differ?**

It is really important to make clear whether the statistic you are presenting is an odds ratio or a risk ratio. As we saw when we looked at using odds and risks to summarise events in a single group, the risk and odds can be very different. So too with risk ratios and odds ratios.

In general an odds ratio will always be further from the point of no effect (where OR=1, RR=1) than a risk ratio. If the event rate increases in the treatment group, the OR and RR will both be greater than 1, but the OR will be bigger than the RR. If the event rate decreases in the treatment group, both the OR and the RR will be smaller than 1, but the OR will be smaller than the RR.
Odds ratios and risk ratios will be similar when the event is rare, but will differ (often by a lot) when the event is common. In situations of common events, the odds and odds ratio can be misleading, because people tend to interpret an odds ratio as if it were a risk ratio. Trials usually study frequent events, so this is a very real issue.

Later on in this module we will discuss how to chose the appropriate statistic in your review.

Another word of caution about both these measures: because the result is expressed as a proportion of the event rate in the control group, it is not possible to determine the actual number of participants who benefited. For example, a RR of 0.5 can mean a risk is decreased from 40% in one group to 20% in the other, or it can mean a 2% in one group and 1% in the other. In both cases the risk is halved by the intervention, but the actual change in the number of events is very different. Because of this, it may also be useful to express results in absolute terms. One way of doing this is to report a risk difference; another is to report the number needed to treat.

**Risk Difference**

As well as comparing risks in relative terms (i.e. risk in one group divided by the risk in the other), we can also compare them in terms of the **absolute** difference between the two groups (i.e. the risk in one group minus the risk in the other). This we call the risk difference, or absolute risk difference (it means the same thing).

Risk difference is calculated as risk in the experimental group minus risk in the control group. For our example this is:

\[
RD = \text{Risk in antibiotic group} - \text{Risk in placebo group}
\]

\[
= 0.10 - 0.86 \\
= -0.76
\]

The risk difference describes the absolute change in risk that is attributable to the experimental intervention. If an experimental intervention has an identical effect to the control, the risk difference will be 0. If it reduces risk, the risk difference will be less than 0; if it increases risk, the risk difference will be bigger than 0. The risk difference cannot be above 1 or below –1. Switching between good and bad outcomes for the risk difference causes a change of sign, from + to – or – to +.

Sometimes it may be useful to present figures for 100 times the RD, or 1000 times the RD, which describe how many people have avoided (or incurred) the event for every 100 or 1000 treated, respectively.
Number needed to treat

Another way of looking at the risk difference is the number needed to treat (NNT). Where we are trying to prevent an event, and the risk difference is less than 0 (i.e. the intervention reduces the risk of the event), NNT is the inverse of the risk difference:

\[ \text{NNT} = \frac{1}{\text{risk difference}} \]

(Where we drop any minus signs from the risk difference). NNT describes the number of patients you would need to treat with the experimental treatment rather than the control treatment in order to prevent a single event. In other words, if the risk difference is 0.76, that means if we treat 100 people, 76 more will benefit when we use the intervention, who would not have benefited if given control. So how many would we need to treat to help one person? 100/76 or 1.3.

We always round up NNT to the next whole number so in this case we need to treat two women with antibiotics to cure one additional woman (over and above those who would have been cured anyway, i.e. those cured in the control group). It is important with NNT to link it to a time frame, so in this case we would need to treat 2 women with antibiotics for 6 weeks to prevent a single extra woman from not being cured.

Where the risk difference is greater than 0 (i.e. the risk of the event we are trying to prevent actually increases) the same calculation produces a number known as the NNH – number need to harm. This is the number of participants treated for a length of time for one extra person to have the event.

While NNTs are easy to interpret, making them popular with consumers and clinicians, they cannot be used for performing a meta-analysis because of their mathematical properties. RR, OR and RD are therefore used for meta-analysis, and all may later be converted to NNTs as a way of communicating results in some Cochrane reviews. In later modules, we’ll look in more depth at interpreting and applying the results of analyses.
Summary to date.

Here is a reminder of the statistics we have covered so far in this module:

- **The risk** describes the number of participants having the event in a group divided by the total number of participants.
- **The odds** describe the number of participants having the event divided by the number of participants not having the event.
- **The risk ratio (relative risk)** describes the risk of the event in the intervention group divided by the risk of the event in the control group.
- **The odds ratio** describes the odds of the event in the intervention group divided by the odds of the event in the control group.
- **The risk difference** describes the absolute change in risk that is attributable to the experimental intervention.
- **The number needed to treat** (NNT) gives the number of people you would have to treat with the experimental intervention (compared with the control) to prevent one event.

Uncertainty

All of these statistics are based on observations in a sample of participants who are randomly split into treatment and control groups. On average randomisation will generate two groups who would have the same event rates if treated identically – so that any observed difference in outcome must be due to the different effects of the treatment and control interventions. However, this comparability is not guaranteed in any particular trial. It is possible that, by chance, the treated group may have a few more people who would naturally do well or badly than the control group, even if they had all received identical treatment.

This means that the observed treatment effect (OR, RR, RD) may actually be an over- or underestimate of the real effect of treatment. A confidence interval (CI) can be calculated as a way of representing the uncertainty in the estimate of treatment effect. The interval contains a range of values above and below the calculated treatment effect within which we can be reasonably certain (usually specified as 95% certain) that the real effect lies. The result is said to be statistically significant if the 95% CI does not include the risk in the two groups being the same (i.e. 1 for risk ratio or odds ratio, 0 for risk difference).
Another way of thinking about CIs is that it gives us an estimate of the range in which the estimate would fall a fixed percentage of times if we repeated the study many times. Picking a 95% CI means that in 5% of all possible trials the effect estimate would fall outside the 95% CI (2.5% above and 2.5% below). In some situations, you may want your CI to include more of the possible trial results, to be more sure that you are quoting an interval that contains the real effect. You can choose, for example, a 99% CI. The interval you come out with then will be wider than for a 95% CI, making your interpretation more conservative.

**Putting these statistics in words**

To make all these numbers useful to decision makers, we have to be able to express them in words. Using the example from the previous section, here are some suggestions of how to express your results.

The RR of ‘still being infected’ on antibiotics relative to no antibiotics was 0.12. We can express this as

- The risk of still being infected on antibiotics was about 12% of the risk on control
- Treatment reduced the risk to 12% of what it would have been
- Treatment reduced the risk by 88% of what it was in the control group

We could reasonably exchange the word ‘risk’ for ‘chance’ or ‘probability’, as they are commonly used to mean the same thing.

Odds ratios are harder. The OR in this example is 0.02. You could express this as:

- Antibiotics reduced the odds of still being infected to about 2% of what they would have been
- Treatment reduced the odds by 98% of what they were in the control group

Note that we have to use the word odds and must not use words like chance, risk or probability.

The risk difference in this example was –0.76. This is best expressed as:

- Antibiotics reduced the risk of still being infected by 76 percentage points.
It is important to be clear about how you express the reduction. If we said the reduction was 76%, it is difficult to know if this reduction is 76% of the risk without treatment (i.e. the control group risk), which in this case was 0.86, or it is a reduction of 76 percentage points. In most cases RD should be expressed as percentage points, as these are the units we need to calculate number needed to treat.

Most people find risk ratio and risk difference (and NNT) easier to interpret than odds ratios. However, communication of the results of your review is only one factor to consider when choosing the best statistic to use.

**Choosing an effect measure**

From what we have seen to date we know that if a randomised trial measures dichotomous outcomes, we can compare the event rates in the two groups using several different summary statistics:

- A risk ratio (of either the good and bad outcome)
- The risk difference
- The NNT
- The odds ratio.

Unfortunately there is no easy way to decide which statistic to use in your review. You will need to make two decisions,

- Which statistic to choose for the analysis
- Which statistic to choose to present the results.

Within RevMan, the choice you make for expressing the results of an individual study will also apply to the meta-analysis if you choose to do one. Whatever statistic you choose for your analysis, you always have the option of re-interpreting the results using another measure in the text of your review. For example, you might perform a meta-analysis by selecting risk ratio in RevMan, then interpret the results by converting it to an NNT in the results section of your review. Remember that NNT is useful for presenting results, but not for analysis purposes. The other three statistics (OR, RR, RD) can be used for either. But also remember that we have two quite different RRs depending on which way the events are coded.

There are three principal issues to consider when choosing a summary statistic:

- Communication, i.e. a straightforward and clinically useful interpretation
- Consistency of the statistic across different studies
- Reasonable mathematical properties.
Communication
We have seen in the section on putting statistics into words that it can be quite hard to explain an odds ratio. Most clinicians and consumers have less difficulty with understanding a risk ratio, however we need to be careful to also give some idea of the absolute difference between the two groups, as relative measures can be misleading. Take the example of buying two lottery tickets instead of one. We could say you are doubling your chances of winning, or we could say your chances of winning have gone up by 1 in 400,000. Both versions give you incomplete information because neither tells us clearly what the chance of winning is in the first place. The statements are likely to be interpreted differently, because many people would think an increase of 1 in 400,000 sounds a lot less attractive than a doubling of the chance of winning.

NNT is a very useful way to express effect in clinical terms (for example, “I have to treat x number of my patients with this treatment for y number of weeks in order to help 1 patient who would not have got better anyway”), as is risk difference (for example, “for every 100 treated, x% will benefit”).

Consistency
The results of your review are probably drawn from many trials, and will be applied in many populations, so it is desirable that the statistic you chose is consistent, i.e. stays nearly the same, or is stable, when applied in different places. In any meta-analysis there is likely to be variation in event rates between trials. The risk ratio, risk difference and odds ratio all vary to some extent in different situations, but one may be more stable than the others. Let’s have a look at a hypothetical example.
In this table we take trial 1 as the reference trial. In trial 2 the events are more common – 36% of all patients have events compared to 20% in trial 1. Despite this, in trial 2A, the effect size is the same as trial 1 if we used odds ratio as the summary statistic. But if we were to use risk ratio (either of the event or the non-event) or if we used risk difference, the effect measure would not be the same. Now consider trial 2B. Here the effect is the same in the 2 trials if we choose risk ratio of the event as the effect measure, but it varies when the other measures are chosen. Trial 2C and 2D show the same phenomenon, with risk difference and risk ratio of the non-event respectively being consistent.

So, as event rates vary, OR, RR and RD may vary to different degrees. When we are choosing which one to use, it would be helpful to choose the one which is most likely to be consistent across the event rates in the studies we have, as it is also the one most likely to be consistent in clinical practice. To try to help make this decision, some researchers looked at many meta-analyses in Cochrane reviews, calculating the results with OR, RR and RD. Of these, RD varied most across trials included in the meta-analyses, and OR and RR less so. This suggests that the relative measures are more likely to be consistent in Cochrane reviews than RD. The RR of the bad event varied less than the RR of the good event.
Mathematical Properties

In order for a summary statistic to be able to be used in a meta-analysis it needs one mathematical property. That is the ability to reliably estimate its variance. This is because the way in which we assign weight to studies within the meta-analysis is inversely proportional to variance (we will cover this in more detail in the next module). We cannot use NNT in a meta-analysis because we don’t have a usable estimate of its variance.

There are two other properties, which are not essential but are mathematically desirable. We’ve already seen that it is easier to switch between odds ratios of ‘good’ and ‘bad’ outcomes, than it is with risk ratios. This is sometimes argued to be a helpful mathematical property of odds ratios.

Another issue that arises when applying results is called ‘bounding’, as we can get predictions outside the bounds of possibility. For example, we calculated earlier on that the risk ratio of cure of UTI for antibiotic use in pregnant women was 6.6. What would happen if we tried to apply this to a group of women where we thought half would get better without antibiotics?

\[ \text{Risk without treatment} \times \text{risk ratio} = \text{risk with treatment} \]

\[ 0.5 \times 6.6 = 3.3 \]

This result is nonsense as it predicts that 330% (i.e. more than all) of the women will be cured!

A similar thing can happen with risk difference. The risk difference we calculated for risk of not being cured of UTI with antibiotics was –0.76. Let’s try to apply that in a situation where we think that, without antibiotics, 30% of people would not be cured.

\[ \text{Risk without treatment} + \text{risk difference} = \text{risk with treatment} \]

\[ 0.3 + -0.76 = -0.46 \]

Again this is nonsense as it predicts that using antibiotics will mean that –46% of women will not be cured.

In practice, this is less important than it may appear, as what we are doing in these examples is applying a result from one situation to a very different situation. In reality, we would not expect to apply results from very high risk populations to very low risk populations or vice versa.
Summary

In summary, risk ratio and odds ratio are better for meta-analysis than the risk difference. Risk ratios are easier to understand than odds ratios, but require some care in choosing whether to analyse the risk ratio of the event or non-event.

This overall view is summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>RR</th>
<th>RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>-</td>
<td>+</td>
<td>++</td>
</tr>
<tr>
<td>Consistency</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Mathematics</td>
<td>++</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Once the analysis is performed the results require careful interpretation. Odds ratios must not be interpreted as if they were risk ratios unless events are very rare. It will be helpful to a reader if any relative effects are re-expressed as absolute effects (RD or NNT) – maybe working out these figures for a range of possible scenarios.

Some review groups will prefer you to use a particular measure to give uniformity across their reviews, and you should check with them or on The Cochrane Library to see if this is the case. If, after reading this module, you don’t agree with their policy, you could always challenge them to justify it.

In the next module we will discuss how results from individual trials are combined to give an overall estimate of treatment effect.
Module 12: Combining studies

In the previous module we discussed choosing an effect measure for presenting the results of studies reporting dichotomous data within your review. If any of the outcomes you are planning to analyse are continuous, you should complete the additional module A1, either before or immediately after completing this module.

This module is about meta-analysis. We will look at the principles for the sensible combination of results from separate studies, and overview the methods that are most commonly used in Cochrane reviews.

Learning objectives

- Understand what a weighted average is, and how it differs from a straightforward average
- Understand the concept of standard error relating to individual trial results
- Be aware that there are alternative methods for meta-analysis (Peto, Mantel-Haenszel) which differ in the way that study weights are calculated
- Be aware that there are a group of methods known as random effects methods which take a different approach to fixed effect meta-analysis

Relevant sections of the Reviewers’ Handbook

- Sections 8.5.1, 8.5.2, 8.5.3

Where does this go in a Cochrane review?

The information in this module will be relevant to many parts of your review:

- In the data analysis part of the Methods section of a protocol or review, where you will describe what statistical techniques you are planning to use
- When (or if) you actually perform meta-analyses using the analysis part of RevMan or other software
- In the presentation of results in the Results section of the review
- In the interpretation of results, in the Discussion of your review

© The Cochrane Collaboration 2002
What is meta-analysis?

**Meta-analysis** is the use of statistical methods to combine results of individual studies. This allows us to make the best use of all the information we have gathered in our systematic review by increasing the power of the analysis. By statistically combining the results of similar studies we can improve the precision of our estimates of treatment effect, and assess whether treatment effects are similar in similar situations. The decision about whether or not the results of individual studies are similar enough to be combined in a meta-analysis is essential to the validity of the result, and will be covered in the next module on heterogeneity. In this module we will look at the process of combining studies and outline the various methods available.

There are many approaches to meta-analysis. We have discussed already that meta-analysis is not simply a matter of adding up numbers of participants across studies (although unfortunately some non-Cochrane reviews do this). This is the ‘pooling participants’ or ‘treat-as-one-trial’ method and we will discuss it in a little more detail now.

**Pooling participants (not a valid approach to meta-analysis).** This method effectively considers the participants in all the studies as if they were part of one big study. Suppose the studies are randomised controlled trials: we could look at everyone who received the experimental intervention by adding up the experimental group events and sample sizes and compare them with everyone who received the control intervention. This is a tempting way to ‘pool results’, but let’s demonstrate how it can produce the wrong answer.

A Cochrane review of trials of daycare for pre-school children included the following two trials. For this example we will focus on the outcome of whether a child was retained in the same class after a period in either a daycare treatment group or a non-daycare control group. In the first trial (Gray 1970), the risk difference is −0.16, so daycare looks promising:

<table>
<thead>
<tr>
<th></th>
<th>Retained</th>
<th>Total</th>
<th>Risk</th>
<th>Risk difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daycare</td>
<td>19</td>
<td>36</td>
<td>0.528</td>
<td>−0.16</td>
</tr>
<tr>
<td>Control</td>
<td>13</td>
<td>19</td>
<td>0.684</td>
<td></td>
</tr>
</tbody>
</table>

© The Cochrane Collaboration 2002
In the second trial (Schweinhart 1993) the absolute risk of being retained in the same class is considerably lower, but the risk difference, while small, still lies on the side of a benefit of daycare:

<table>
<thead>
<tr>
<th>Schweinhart 1993</th>
<th>Retained</th>
<th>Total</th>
<th>Risk</th>
<th>Risk difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daycare</td>
<td>6</td>
<td>58</td>
<td>0.1034</td>
<td>–0.004</td>
</tr>
<tr>
<td>Control</td>
<td>7</td>
<td>65</td>
<td>0.1077</td>
<td></td>
</tr>
</tbody>
</table>

What would happen if we pooled all the children as if they were part of a single trial?

<table>
<thead>
<tr>
<th>Pooled results</th>
<th>Retained</th>
<th>Total</th>
<th>Risk</th>
<th>Risk difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daycare</td>
<td>25</td>
<td>94</td>
<td>0.266</td>
<td>+0.03</td>
</tr>
<tr>
<td>Control</td>
<td>20</td>
<td>84</td>
<td>0.238</td>
<td>WRONG!</td>
</tr>
</tbody>
</table>

It suddenly looks as if daycare may be harmful: the risk difference is now bigger than 0. This is called Simpson’s paradox (or bias), and is why we don’t pool participants directly across studies. The first rule of meta-analysis is to keep participants within each study grouped together, so as to preserve the effects of randomisation and compare like with like. Therefore, we must take the comparison of risks within each of the two trials and somehow combine these. In practice, this means we need to calculate a single measure of treatment effect from each study before contemplating meta-analysis. For example, for a dichotomous outcome (like being retained in the same class) we calculate a risk ratio, the risk difference or the odds ratio for each study separately, then pool these estimates of effect across the studies.

Simple average of treatment effects (not used in Cochrane reviews)
If we obtain a treatment effect separately from each study, what do we do with them in the meta-analysis? How about taking the average? The average of the risk differences in the two trials above is \((-0.004 - 0.16) / 2 = -0.082\). This may seem fair at first, but the second trial randomised more than twice as many children as the first, so the contribution of each randomised child in the first trial is diminished. It is not uncommon for a meta-analysis to contain trials of vastly different sizes. To give each one the same influence cannot be reasonable. So we need a better method than a simple average.
**Weighted averages**

The solution is to calculate a weighted average.

A weighted average is an average where the results of some of the studies make a greater contribution to the total than others. All of the methods available for conducting meta-analyses in Cochrane reviews use forms of weighted averages. The various methods do this in different ways, and we will cover these methods in this module. In all methods, the underlying principle is to give more weight to studies that give us more information about the treatment effect.

Sample size is the main factor in determining the weight for a trial. However, the event rate also makes a difference. This is because effects are generally estimated more precisely when there are lots of events. So, trials with higher event rates get more weight. At the extreme, a trial with no events tells us nothing about the effect of the intervention, and so gets no weight at all. The exact relationship between event rates and study weights is complex, and depends on the summary statistic being used.

A statistical concept which takes into account both size of the study’s population, and its event rate, is *variance*. The box below provides an outline of the concept of variance.

---

**What is variance?**

**Standard errors, confidence intervals and variances**

Imagine trying to estimate the proportion of females in the population. If we took a sample of ten people from a list of all the people in a country, we may, by chance, find there were 2, or 7, or even 10 females. We wouldn’t be very confident from this sample to say what the true proportion of females in the population is. Whenever we take samples from populations, there is uncertainty about the estimates we make of the true value in the whole population.

The same is true of trials – each trial involves taking a sample of the possible participants. The basic result of an individual trial is an estimate of treatment effect. The estimate is incomplete without a measure of how certain we can be about it. We’d be much more certain about an estimate from a mega-trial of tens of thousands of patients than we would an estimate from a small trial of less than a hundred. Uncertainty is often described using a confidence interval. For example, a 95% confidence interval gives a range within which we can be 95% confident the true effect lays.

Continued……..
Confidence intervals are calculated from a number known as a standard error. Standard errors are companions of all estimates. They describe the extent to which an estimate might be wrong due to random error. The smaller the standard error the more certain we are about the estimate. To get a feel for standard errors it is helpful to know that 95% confidence intervals are obtained by taking the estimate and creating limits that are 1.96 standard errors below it and 1.96 standard errors above it. Thus an estimate may be wrong by about a standard error, but to be 95% confident about where the true effect lies, we go roughly 2 standard errors either side.

Statistics as a discipline has more uses for the ‘standard error squared’ than for the ‘standard error’, so statisticians have a word for it, the variance. The variance of an estimate is just the square of its standard error. The standard error and variance are interchangeable in terms of the information they convey, but their numerical values are different. It’s the same as describing the size of a square: you could say either that the length of each side is 4 metres or that the area is 16 square metres; you end up with an identical shape. Quoting the length is like using the standard error; quoting the area is like using the variance.

One important point to note is that different treatment effects (OR, RR, RD) calculated for the same trial will have different variances.

We could assume that variance is inversely proportional to importance, i.e. the less variance in the study, the more weight it should contribute. One method, planned for RevMan 4.2 but not in earlier versions of RevMan, the inverse variance method, calculates study weights directly based on this assumption.

There are other methods, called Mantel-Haenszel methods, which attribute weight in a manner closely related to inverse variance. In this module we will expand a little on the various available methods and look at some of the differences between them, finishing with some guidance on which method to use in your review.
Within RevMan, the methods available are:

- For dichotomous data:
  - Fixed effect assumption
    - Mantel-Haenszel risk ratio (RR)
    - Mantel-Haenszel odds ratio (OR)
    - Mantel-Haenszel risk difference (RD)
    - Peto odds ratio (Peto OR)
  - Random effects assumption (DerSimonain and Laird)
    - RR
    - OR
    - RD

- For continuous data
  - Fixed effect inverse variance model
    - Weighted mean difference (WMD)
    - Standardised mean difference (SMD)
  - Random effects assumption (DerSimonian and Laird)
    - WMD
    - SMD

- For generic data (available in Revman 4.2)
  - Fixed effect inverse variance
  - Random effects inverse variance

There is some information about the statistical techniques available in RevMan in Section 8.5 of the Reviewers’ Handbook and you should read it now.

In order to choose the method you are going to use in your meta-analysis, the first concept to understand is the difference between a fixed effect model and a random effects model.

**What does ‘fixed effect’ mean?**

To come up with any statistical model, or method for meta-analysis, we first need to make some assumptions. It is these assumptions that form the differences between all the methods listed above.

A fixed effect model of meta-analysis is based on a mathematical assumption that every study is evaluating a common treatment effect. That means the effect of treatment, allowing for the play of chance, was the same in all studies. Another way of explaining this is to imagine that if all the studies were infinitely large they’d give identical results.
The summary treatment effect estimate resulting from this method of meta-analysis is this one ‘true’ or ‘fixed’ treatment effect, and the confidence interval describes how uncertain we are about the estimate.

Sometimes this underlying assumption of a fixed effect meta-analysis (i.e. that diverse studies can be estimating a single effect) is too simplistic. Therefore, the alternative approaches to meta-analysis are (i) to try to explain the variation or (ii) to use a random effects model.

Random effects meta-analyses (DerSimonian and Laird)

As we discussed above, fixed effect meta-analysis assumes that there is one identical true treatment effect common to every study. The random effects model of meta-analysis is an alternative approach to meta-analysis that does not assume that a common (‘fixed’) treatment effect exists. The random effects model assumes that the true treatment effects in the individual studies may be different from each other. That means there is no single number to estimate in the meta-analysis, but a distribution of numbers. The most common random effects model also assumes that these different true effects are normally distributed. The meta-analysis therefore estimates the mean and standard deviation of the different effects.

By selecting ‘random effects’ in the analysis part of RevMan you can calculate an odds ratio, risk ratio or a risk difference based on this approach.

The Mantel-Haenszel approach

The Mantel-Haenszel approach was developed by Mantel and Haenszel over 40 years ago to analyse odds ratios, and has been extended by others to analyse risk ratios and risk differences. It is unnecessary to understand all the details, but is sufficient to say that the Mantel-Haenszel method assumes a fixed effect and combines studies using a method similar to inverse variance approaches to determine the weight given to each study.
The Peto method

The Peto method works for odds ratios only. Focus is placed on the observed number of events in the experimental intervention. We call this $O$ for ‘observed’ number of events, and compare this with $E$, the ‘expected’ number of events. Hence an alternative name for this method is the ‘$O – E$’ method. The expected number is calculated using the overall event rate in both the experimental and control groups. Because of the way the Peto method calculates odds ratios, it is appropriate when trials have roughly equal number of participants in each group and treatment effects are small. Indeed, it was developed for use in mega-trials in cancer and heart disease where small effects are likely, yet very important.

The Peto method is better than the other approaches at estimating odds ratios when there are lots of trials with no events in one or both arms. It is the best method to use with rare outcomes of this type.

The Peto method is generally less useful in Cochrane reviews, where trials are often small and some treatment effects may be large.

Which method should I use in my review?

We have talked about three methods of combining the results of trials included in a meta-analysis, the fixed effect method (the Mantel-Haenszel approach, which is more specific to dichotomous data and weights studies in a slightly different way), the Peto method which is useful for rare events, and the random effects model, which assumes that all studies are estimating their own true effect, and these effects are normally distributed. The analysis program within RevMan allows us to choose which of these models we want to use in our meta-analysis, and the results of our review will be slightly different depending on which method we use. The table below summarises the summary effect and confidence intervals resulting from selecting each of these methods in the daycare example we used earlier.

<table>
<thead>
<tr>
<th>Method</th>
<th>$RR$ (95% CI)</th>
<th>$OR$ (95% CI)</th>
<th>$RD$ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mantel-Haenszel</td>
<td>0.64 (0.49, 0.82)</td>
<td>0.47 (0.31, 0.73)</td>
<td>-0.14 (-0.23, -0.06)</td>
</tr>
<tr>
<td>Peto Method</td>
<td></td>
<td>0.47 (0.30, 0.72)</td>
<td></td>
</tr>
<tr>
<td>Random effect inverse-variance</td>
<td>0.64 (0.50, 0.82)</td>
<td>0.47 (0.31, 0.72)</td>
<td>-0.13 (-0.25, -0.01)</td>
</tr>
</tbody>
</table>

© The Cochrane Collaboration 2002
As you can see from the table, there is little difference in the results regardless of which method you choose, and the conclusions of your review would certainly not change. So why do we devote so much energy to selecting the summary statistic and devising methods with varying assumptions? There are some cases or circumstances where one method performs better than the others, and if any of these circumstances fit your review you may need to think carefully about which statistic you use.

Some general points about the performance of the various statistics

i. The Mantel-Haenszel methods have been shown to be more reliable when there are not many data (small trials and not many of them). This is why they have been selected as the principle method of meta-analysis in the Cochrane Collaboration. This method (which can be used for OR, RR and RD) is the most appropriate for many Cochrane reviews, and many Cochrane review groups use it as standard. But it should not be used in reviews with sparse data, where lots of trials have zero events in treatment or control groups or both. The choice between OR, RR or RD should be based on the information covered in Module 11.

ii. The Peto method performs well with sparse data and is then the best choice, but when events are common there is usually no preference to use it over the other methods. It is **not** a good idea to use the Peto method when the treatment effect is very large, as the result may be misleading. This method is also unsuitable if there are large imbalances in the size of groups within trials.

iii. A random effects model may be better when there is statistical heterogeneity between the studies in your review (we will discuss this further in Module 13 on Heterogeneity).

Summary

From this module and the one preceding it we can see that there are many choices the reviewer has to make about the way they analyse their dichotomous data in a review. Firstly decisions about which **within trial** statistic to use (OR, RR or RD) need to be made, and then the method of combining the trial data, or **meta-analysis**, needs to be chosen. While there is not consensus about which is the best approach, and you will need to check your review group policy, following the principles set out here should help you make your decision. If you are concerned that your review falls into one of the categories where there are special considerations (rare events with zero cells, very large treatment effects, or large variation in the control event rates between the trials included in your review) you may want to seek the advice of a statistician, and your review group can help you with this.
Module 13: Diversity and heterogeneity

This module will discuss differences between studies. We learn how to recognize and deal with differences between studies in a review. Some statistical methods such as random effects meta-analysis, subgroup analysis and meta-regression can help address these differences, and this module will explain these techniques.

Learning objectives

- Understand that studies usually differ both clinically and methodologically
- Appreciate that differences between studies can result in (statistical) heterogeneity – differences in their results
- Be able to identify heterogeneity
- Know some strategies for dealing with heterogeneity
- Understand the difference between fixed effect and random effects meta-analysis
- Understand when to use subgroup analyses
- Be aware of meta-regression as a tool for exploring differences between studies

Relevant sections of the Reviewers’ Handbook

- Sections 8.7, 8.8

Where does this go in a review?

- In the protocol where you will need to specify factors that you are planning to investigate as potential causes of heterogeneity
- The Methods section of the review should detail what you have done to identify or examine heterogeneity. It is especially important to tell the reader which analyses were pre-specified and which were ‘post hoc’, i.e. designed after collecting all the studies for your review
- The Results section should present results of subgroup analyses and meta-regressions (if used). Remember to interpret the results with caution, and keep in mind the possibility that findings may be spurious if you do more than very few analyses
Variety is the spice of life

Systematic reviews usually bring together studies that were performed

- By different people
- In different settings
- In different countries
- On different people
- In different ways
- For different lengths of time
- To look at different outcomes

... and these aren’t the only differences.

However, while studies are never the same, they may all have similar results. In fact, the purpose of a Cochrane review is to collate studies that are similar. The decision to combine studies in a meta-analysis in your review is a judgement you will have to make, based on your knowledge about how differences between studies might influence how effective a treatment is observed to be. Sometimes studies are similar enough to consider performing a meta-analysis; sometimes they are not. What we can then do is look at the results of the studies we find to see if our judgement was reasonable.

A variety of varieties

We recognise that studies will differ. It is helpful to identify three basic ways in which they differ: clinical diversity, methodological diversity and statistical heterogeneity. Heterogeneity and diversity are words that have pretty much the same meaning. We’ve used different words here as people often mean ‘statistical heterogeneity’ when they just say ‘heterogeneity’.

Clinical diversity

We use the term ‘clinical diversity’ (sometimes called ‘clinical heterogeneity’) to describe clinical differences in the studies to do with the participants, interventions and outcomes. This covers such factors as

- Study location and setting
- Age, sex, diagnosis and disease severity of participants
- Treatments people may be receiving at the start of a study
- Dose or intensity of the intervention
- Definitions of outcomes.
Methodological diversity

‘Methodological diversity’ (sometimes called ‘methodological heterogeneity’) covers differences between how the studies were executed, including such variables as

- A parallel group trial or a crossover trial
- Randomization by clusters (for example, by family or by school) or by individuals, or by body parts (for example, eyes or different parts of the mouth)
- Study quality (for example, the extent to which allocation to interventions was concealed, or whether outcomes were assessed blind to treatment allocation)
- Analysis (for example, performing an intention-to-treat analysis compared with an ‘as treated’ analysis)

The distinction between some aspects of clinical and methodological diversity is not always clear-cut. For example, is the length of a study a feature of the intervention being evaluated, or of the outcome being assessed or of the study design? As long as we remember to assess it, it does not really matter how we classify it.

Before we go on to statistical heterogeneity, try to complete the activity based on your clinical knowledge of how the participants in your included trials may respond differently to the intervention, and your knowledge of the methodology of your included trials. It does not really matter which heading we put it under, as long as we consider it somewhere.

Do you think any of these differences are so great that studies should not be combined?

This is a difficult question to answer. To help you think about it, you can ask yourself the following questions:

- Could any of these differences make the treatment have the opposite effect to the one we want?
- Could any of these differences make the treatment work particularly well?

If you can think of situations in your review where this might be true, and there is good evidence to back up your suspicion, it might not be appropriate to pool all the studies together.
For example, if we look at aspirin as an intervention to prevent death from stroke, are there groups of patients who are more susceptible to the side effect of aspirin induced bleeding, which can actually cause death. In some groups this might outweigh any beneficial effect. Are there groups of patients who might particularly benefit, such as patients at high risk of stroke?

It’s also important to realise that not every factor that influences how well a patient does in general (prognostic factors) will influence the size of the treatment effect. For example, the more severe a head injury is, the more likely you are to die. This doesn’t necessarily mean that we should not combine studies in patients with different severities of head injury. The treatment may work equally well in any severity of head injury.

To summarise, an important decision when performing a systematic review is whether or not to combine studies. This decision needs to be made for each individual outcome of every comparison in your review. It is possible to perform a meta-analysis for some comparisons and not for others; depending on the individual studies you have found addressing this comparison. The decision to combine studies in a meta-analysis should be made based on the setting, participants, interventions and outcomes of the included trials being sensible to combine (i.e. little clinical diversity); and the methods used to perform the trial not varying in a way that is likely to overly influence the results (methodological diversity). To confirm or question your decision, you should consider statistical heterogeneity.

**Statistical heterogeneity**

Having decided that we wish to look at a group of similar studies together, we need some checks to see whether we have made the right judgement. We do this by looking at the estimates of treatment effect of the individual studies. As we are trying to use the meta-analysis to estimate a combined effect from a group of similar studies, we need to check that the effects found in the individual studies are similar enough that we are confident a combined estimate will be a meaningful description of the set of studies.

In doing this, we need to remember that the individual estimates of treatment effect will vary by chance, because of randomisation. So we expect some variation. What we need to know is whether there is more variation than we’d expect by chance alone. When this excessive variation occurs, we call it statistical heterogeneity, or just heterogeneity.
Identifying statistical heterogeneity

You can determine the presence of statistical heterogeneity in two main ways:

- By looking at a forest plot to see how well the confidence intervals overlap. If the confidence intervals of two studies don’t overlap at all, there is likely to be more variation between the study results than what you would expect by chance (unless there are lots of studies), and you should suspect heterogeneity.
- By performing a statistical test, known as a $\chi^2$ (“chi-squared”) test.

The result of this statistical test appears at the bottom of each meta-analysis within the statistical part of RevMan.

The result of the test is (i) a ‘chi-squared’ statistic (ii) a number called the degrees of freedom (which is usually one less than the number of studies, but can be less if some of the studies have no events, as in the example above) and (iii) a ‘p-value’ obtained by referring the first two numbers to statistical tables. A small $p$-value is often used to indicate evidence of heterogeneity (this $p$-value appears in RevMan and the Internet version of The Cochrane Library. It is not yet available in the CD-ROM version of The Cochrane Library).

As it applies to Cochrane reviews, this test is of somewhat limited value. This is because most meta-analyses in Cochrane reviews have very few studies in them. When there are few studies, the test is not very good at detecting heterogeneity if it is present (it has ‘low power’). For this reason, a $p$-value of less than 0.10 is often used to indicate heterogeneity rather than the conventional cutpoint of $p = 0.05$. 

© The Cochrane Collaboration 2002
Conversely, if there are a lot of studies in a meta-analysis, the test can be too good at detecting heterogeneity. Since we have established that heterogeneity is almost certain to be present as studies are rarely identical, the test will detect significant heterogeneity even if it is clinically trivial (the test has too much power). But the basic problem is that the test does not answer a useful question. It asks the question ‘Is there heterogeneity?’ whereas we want to know ‘How much heterogeneity is there?’

A useful way to identify heterogeneity without having to use statistical tables to look up \( p \)-values is to compare the chi-square statistic with its degrees of freedom. If the statistic is bigger than its degrees of freedom then there is evidence of heterogeneity. A visual inspection of the confidence intervals will help get an idea of the amount of statistical heterogeneity, and guide you to think about whether it is reasonable to combine the results of these studies.

**Things you can do with diversity and heterogeneity**

If you identify or suspect that important diversity or heterogeneity is present in your review, there a several options open to you. Don’t forget that one option is that of not performing a meta-analysis. An unwise meta-analysis can lead to highly misleading conclusions. If you have clinical, methodological or statistical heterogeneity it may be better to present your review as a systematic review using a more qualitative approach to combining results, or to combine studies only for some comparisons or outcomes. Studies can always be entered into RevMan and presented on a forest plot with their individual effect sizes and no combined effect. This gives an overall picture of the evidence.

Another alternative if there are subgroups of patients who are likely to respond very differently is to undertake separate reviews. For example, there are separate Cochrane reviews of influenza vaccines in healthy adults, people with cystic fibrosis, people with asthma and people with chronic obstructive pulmonary disease. This sort of decision should, of course, be made at the question formulation stage.

In the remainder of this module we take a brief look at three options for investigating or incorporating heterogeneity in a review:

- Using a different statistical model for combining studies, called a random effects meta-analysis
- Investigate heterogeneity by splitting the studies into subgroups and looking at the forest plot
- Investigating heterogeneity using meta-regression
Fixed and random effects meta-analysis

We briefly discussed the ‘fixed effect’ and ‘random effects’ options for meta-analysis available in RevMan in Module 12.

Fixed and random effects meta-analyses sometimes give you similar results and sometimes give you results that differ. We’ll explain what the technical difference is, then explain what a difference in the results implies.

Fixed effect meta-analysis
Methods of fixed effect meta-analysis are based on the mathematical assumption that a single common (or ‘fixed’) effect underlies every study in the meta-analysis. In other words, if we were doing a meta-analysis of odds ratios, we would assume that every study is estimating the same odds ratio. Under this assumption, if every study were infinitely large, every study would yield an identical result. This is the same as assuming there is no (statistical) heterogeneity among the studies.

Random effects meta-analysis
A random effects analysis makes the assumption that individual studies are estimating different treatment effects. In order to make some sense of the different effects we assume they have a distribution with some central value and some degree of variability. The idea of a random effects meta-analysis is to learn about this distribution of effects across different studies. By convention (but unfortunately) most interest is focused on the central value, or mean, of the distribution of effects. This is what the statistical part of RevMan presents when you select a random effects meta-analysis. It is also important to know the variability of effects.

What are the important differences between fixed and random effects and which one should I choose?
The first point is that you should analyse your review in both ways (i.e. select first one option then the other in RevMan) and see how the results vary. If fixed effect and random effect meta-analyses give identical results then it is unlikely that there is important statistical heterogeneity, and it doesn’t matter which one you present. If however, your results vary a little, you will need to decide which is the better method on which to base your conclusions (usually it will be best to select the most conservative option).
There is a great deal of debate between statisticians about whether it is better to use a fixed or random effect meta-analysis. The debate is not about whether the underlying assumption of a fixed effect is likely (clearly it isn’t) but more about which is the better trade off, stable robust techniques with an unlikely underlying assumption (fixed effect) or less stable, sometimes unpredictable techniques based on a somewhat more likely assumption (random effects).

Sometimes the point estimate of the treatment effect differs between fixed and random effects because of publication or quality related bias. This may indicate that careful investigations are required, perhaps with expert methodological input. If this is the case in your review you should check with your review group.

**Keeping it all in context**
It’s important to remember that whatever statistical model you choose, you have to be confident that clinical and methodological diversity is not so great that we should not be combining studies at all. This is a judgement, based on evidence, about how we think the treatment effect might vary in different circumstances. This judgement is a common source of disagreement about the results of meta-analyses. Make sure you spend enough time considering this judgement in some depth before you worry too much about which statistical model you choose.

**Investigating sources of heterogeneity**
Most meta-analyses aim to summarize the size of an effect across studies or to establish with greater power whether an effect exists. When different studies give different results, an alternative aim is to examine reasons why effects differ across studies. *Subgroup analyses* and *meta-regression* are techniques for trying to work out whether particular characteristics of studies are related to the sizes of the treatment effect. One example may be dose/intensity. In many reviews you might be able to determine some measure of how “intensely” an intervention was given in different studies. For drugs this might be dose; for personal contact therapies this might be the amount of contact time. The ideal way of looking at the effect of dose would be to have randomised trials comparing the doses (head-to-head comparisons), but they often don’t exist. Within your review, it may be of interest to determine whether the dose or intensity of the intervention is related to the extent of benefit of treatment in different studies. You could look at this by meta-regression. An alternative way to get an idea about the effect of the drug when given in different doses is to look at trials using subgroups of varying doses.
Subgroup analyses

Subgroup analyses are meta-analyses on subgroups of the studies. There are problems with subgroup analyses, and they can result in misleading conclusions if not undertaken with care. Some of the most important points are

i) Restrict the number of subgroup analyses to a minimum (to reduce the possibility of finding a “positive” or significant result by chance)

ii) Pre-specify subgroup analyses whenever possible in order to minimise spurious findings (apparent differences between subgroups that are purely due to chance variation). If there were a good clinical reason why a particular group of participants or studies needed to be looked at separately, you should have thought about that in your protocol. Deciding on subgroups after you have the results of the review may lead to bias through putting a subgroup together on the basis of a particular result.

iii) Have a scientific rationale for all subgroup analyses

iv) Remember that a difference between subgroups is based on an observational comparison, and may exist due to confounding by other factors

To help explain subgroup analysis, think of the question of whether training reviewers results in higher quality reviews. Imagine we had 15 trials looking at training versus no training, and, of these, seven used a self-directed learning module such as this one, and in eight the intervention was face-to-face training. You may decide to look at the method of delivery of training (self directed or face-to-face) as separate subgroups. There are good reasons for doing this as the effect of the intervention may differ in these two groups and it may not be appropriate to combine them (there is ‘clinical’ heterogeneity).

Although the use of subgroups in this review will give you some information about the effect of each method of training delivery compared to no training, it does not give you direct information about how each method of training delivery compares to each other. This is because no trial in this example has directly compared self directed to face-to-face training within the same sample. An indirect estimate of the difference between methods can be obtained by comparing the overall effects between the two subgroups. However, differences in the results of the two subgroups compared to no training could be explained by other differences in the trials, not just the intervention. For example, the self-directed training could have been given to people from a different background to those given face-to-face training, and it might be this difference that is really responsible for the observed difference in treatment effects.
One common error in interpreting differences between subgroups is to note that the overall effect in one subgroup is statistically significant whilst the effect in the other subgroup is non-significant, and then to conclude that there is a significant difference between subgroups. The significance of a result depends on both the size of effect and the amount of data. Consider an analysis where subgroup 1 has a statistically significant RR of 2.0, 95%CI (1.5, 3.0) and subgroup 2 has a statistically non-significant RR of 2.0, 95%CI (0.1, 100). The effect in subgroup 2 is not different in magnitude but is obviously based on fewer data. It would be wrong to conclude that there is any difference in treatment efficacy between subgroup 1 and subgroup 2, despite the difference in statistical significance.

**Meta-regression**

Meta-regression can formally test whether there is evidence of different effects in different subgroups of trials. For example, you can use meta-regression to test whether treatment effects are bigger in low quality studies than in high quality studies.

Meta-regression is potentially a very useful technique, however it can’t be done in RevMan and, if used inappropriately, its interpretation can be misleading. This is again because differences between studies, even if they are well-performed randomized trials, are entirely observational in nature and are prone to ‘bias’ and ‘confounding’. If you summarize patient characteristics at a trial level, you run the risk of completely failing to detect genuine relationships between these characteristics and the size of treatment effect. Further, the risk of obtaining a spurious ‘explanation’ for variable treatment effects is high when you have a small number of studies and many characteristics that differ. Meta-regression is rarely performed in Cochrane reviews and not an available option in Cochrane software, so should you have strong reason to include a meta-regression in your review, you will need the help of a statistician.
Summary of this module

- Heterogeneity is simply diversity in characteristics of trials. Not all trials addressing the same question will be identical with respect to clinical components (participants, interventions and outcomes); methodological components (blinding, sample size, method of randomisation) or with respect to their results.

- When trials are ‘too different’, either in clinical, methodological or statistical components, it may be best not to combine them in a meta-analysis, and you need to consider this carefully.

- When doing or interpreting a meta-analysis you can identify heterogeneity graphically and by use of a statistical test.

- When you are combining trials in a meta-analysis there are several methods available to do this. With respect to meta-analysis when there is statistical heterogeneity, there is debate about whether a random or fixed effect analysis is best. The safest option is to look at both sets of results and be conservative in your conclusions. In reality, it is unlikely that trials you consider alike enough in clinical and methodological terms to combine will result in a very different point estimate, regardless of your choice of method.

- Subgroup analysis is a method available in RevMan to look at the results of different subgroups of trials. Subgroup analyses should be planned at the protocol stage, based on good scientific reasoning, and kept to a minimum. Conclusions from subgroup analyses should be drawn cautiously, remembering that these conclusions are based on subdivision of studies and indirect comparisons, and not on formal statistical tests.

- Metaregression is a method that is not available in RevMan to formally test whether there is evidence for different effects related to different characteristics of trials. It needs to be used with great care.
Module 14: Further issues in meta-analysis

This module deals with some important general concepts relevant in all reviews as well as some issues that not all reviewers encounter.

Learning objectives

- Know when and how sensitivity analysis can investigate the robustness of your findings
- Identify types of data not easily analysed using RevMan
- Consider perspectives on dealing with intention-to-treat analyses
- Understand problems associated with indirect comparisons

Relevant sections of the Reviewers’ Handbook

- Section 8.10, subsections of 8.3, 8.4 and 8.5, section 8.14.4

Where does this go in a Cochrane review?

- The protocol should outline where a sensitivity analysis might be appropriate. However, sensitivity analyses are the one form of statistical analysis that can (and should) be done whether or not they were specified in the protocol. They should be reported in the Results and/or Discussion section, perhaps depending on their implications
- Items in this Module should all be described in the Methods section
Sensitivity analysis

The process of undertaking a systematic review and meta-analysis involves many decisions. Ideally, most of these are made while designing the protocol. It is usually necessary, however, to make further decisions in order to deal with the studies subsequently identified: few reviewers correctly anticipate all problems that arise. Many questions that arise during the review process will have obvious answers. Others will not have clear answers, and in some cases our decisions may change the whole conclusion of the review. The role of a sensitivity analysis is to determine whether the assumptions or decisions we have made do in fact have a major effect on the results of the review. You should present your investigations of the effect your assumptions had in the Results section of your review by detailing the range of treatment estimates and confidence intervals resulting from the various sensitivity analyses.

A sensitivity analysis addresses the question ‘Are the findings robust to the method used to obtain them?’ Sensitivity analyses involve comparing the results of two or more meta-analyses calculated using different assumptions. Here are a few examples of the sorts of things performed as sensitivity analyses in Cochrane reviews:

- If a study is of doubtful eligibility for the systematic review, then comparing meta-analyses excluding and including that study might be undertaken as a sensitivity analysis.
- Results may be calculated using all studies and then excluding poorer quality studies.
- Both fixed and random effects meta-analyses might be undertaken to assess the robustness of the results to the method used.
- If a study appears to be an outlier (has results very different from the rest of the studies) then its influence on a meta-analysis might be assessed by excluding it.
- Where missing information is ‘imputed’ (brought in from another source, perhaps by estimating it) then the effect of imputed numbers should be assessed through sensitivity analysis. This would normally take the form of re-analyzing the data using several alternative imputed values. This is frequently necessary when including cross-over trials, cluster randomized trials or change-from-baseline outcomes in meta-analysis.
- To determine whether a meta-analysis result is being heavily determined by a particular trial it might be repeated excluding that trial. The largest trial or the earliest trial could be driving the result, for example.
We take a more detailed look at one particular type of sensitivity analysis below when we address missing outcome data and intention to treat analyses.

Other types of data

At the beginning of Module 11 we listed a range of different types of outcome data. We dealt with dichotomous data in that module and in Module 12. Continuous data (including long ordinal scales) are addressed in Additional Module 1. Here we address the remaining types of data that you may plan as outcomes, or might come across in your included studies.

Counts of events

Count data are counts of occurrences measured on individuals. Examples include number of lesions, number of pregnancies, number of cigarettes smoked, number of strokes, or number of days in hospital. In Additional Module 2 on unit of analysis issues, you will see that one source of errors in meta-analysis is treating count data directly as dichotomous data. We cannot enter ‘12 strokes out of 28 people’ as dichotomous data if any of the 28 people had more than one stroke.

How you might deal with count data depends on how common they are, in two respects:

- Do most participants have events? (Will most people have at least one?)
- Do participants have lots of events? (Will individual people tend to have high counts?)

Answering each question with a ‘yes’ or a ‘no’ gives us four approximate classifications. Two of them have fairly obvious solutions; the other two do not.

1. **Most people have counts that are mostly high.**

   An example of this might be number of days on which a person with arthritis has pain, or pulse rate (number of pulses in a given time period). These count data can usually be treated as continuous data, although the distribution may be skewed. See Additional Module 1.

2. **Only some people have the event and counts are mostly low.**

   Admissions to hospital and strokes may come into class. A convenient way to analyze these data is to dichotomize people into those that have at least one event and those that have no events. The data can then be entered into RevMan. Alternatively, you might use the data as rate data (see below).
3. Most people have counts that are mostly low. Number of days off work with "flu might fall into this class. These data can be awkward. They are most like ordinal data, so look in the section below for ideas.

4. Only some people have counts that are mostly high. The number of cigarettes smoked in a smoking cessation trial will likely be of this sort. This is another awkward type of count data. It may be wise to dichotomize people, for example, as smokers or non-smokers. If you are tempted to treat the high counts as continuous data, then remember that if substantial numbers of people have counts of zero, then the distribution of the outcome will be severely skewed.

When counts per unit time are of interest, then counts should be treated as rate data. A rate is simply a count per unit time, for example, 2 relapses per year. Rate data are extracted for a whole treatment group. To allow for the different times individual participants are followed up for, we come up with a total follow up time for all participants, by adding up the time for each participant. Data corresponding to the rate of 2 relapses per year might be extracted from a placebo group as ‘1042 relapses during 6247 person-months of follow-up’. These may be compared to ‘983 relapses during 6229 person-months’ in an intervention group. The rate ratio from these numbers is

\[
\frac{983}{6229} = 0.158 \quad \text{and} \quad \frac{1042}{6247} = 0.167
\]

A standard error for the (log) rate ratio is available, and the generic inverse-variance method for meta-analysis (not available in RevMan 4.1) may be used to combine rate ratios across studies. The facility for combining rate data in meta-analysis is planned for RevMan 4.2, due for release in 2003, but in earlier versions the best solution is to include the results as individual studies in an additional table.

**Short ordinal scales**

Disease severity is commonly classified as ‘none’, ‘mild’, ‘moderate’ or ‘severe’. Many assessment scales have only a few categories, say a score between 1 and 5. We often refer to such data as ordinal data. There are numerous approaches to their analysis. The simplest and usually the best is to try and find a cutpoint in the scale and to create a dichotomous outcome. For example, ‘none or mild’ versus ‘moderate or severe’. Sometimes, reviewers present more than one cutpoint, such as also giving ‘none’ versus ‘mild or moderate or severe’ in a sensitivity analysis to investigate whether choice of cutpoint affects conclusions.
Another approach is to treat ordinal data as continuous data. Thus we could assign ‘none’ = 1, ‘mild’ = 2, ‘moderate’ = 3 and ‘severe’ = 4 and take the mean and standard deviation. This is rarely a reasonable approach because it assumes that the assigned numbers represent a real measure of the outcome (i.e. that the difference between mild and moderate is exactly the same as the difference between moderate and severe), when in fact they are arbitrary.

There are more sophisticated methods for analysing these data, which avoid these assumptions, but they are not widely used, and not available in RevMan.

**Censored data or survival data**

In many situations, health care interventions aim to affect the time until an event happens. For example, we may aim to prolong disease-free survival in cancer, or extend the time to the next fit in epilepsy, or time to heart attack or stroke in people who have just had their first heart attack. The outcome that is measured on each patient in studies of such treatments may be a **time until the event**. When interest is focussed on time to the event rather than simply whether the event happens, we have **survival data**. Although time is a continuous outcome, survival data cannot be analysed in the same way as continuous data because we usually have some patients who have not yet experienced the event by the end of the study. For example, although everybody dies eventually, many patients will survive beyond the end of follow up in a randomized trial. Patients that don’t experience the event have survival times that are **censored**. All we know about these patients is that they survived at least until the time when they were last observed.

One way to deal with survival data is to select particular points in time and determine whether each participant had experienced the event by each time. The resulting data are dichotomous and can be analyzed as such. For example, many infectious diseases lead to fever, and one marker of a successful treatment is a reduction in the length of fever. Trials of such treatments tend to assess fever at a specific time point, for example after five days. This avoids the problem of censoring and the analysis is straightforward, as long as you have all the data. In longer-term trials, one might create a series of dichotomous outcomes for, say, mortality such as (i) death within 3 months; (ii) death within 1 year; (iii) death within 3 years, and so on. However, this approach can only be used when all participants have been followed up to or beyond the time point used for the analysis, i.e. all participants have been in the study for at least as long as the time point.
In many specialties, the tradition is to analyze survival data using special methods that account for censoring. This is especially the case in cancer research. These methods include ‘log-rank’ tests, and ‘proportional hazards regression’ (or ‘Cox regression’). These results can be used in meta-analyses. If the only data you can obtain are of this sort, then statistical expertise will be needed.

**Intention to treat issues**

Intention-to-treat (ITT) analyses are widely recommended as the preferred approach to the analysis of most clinical trials. Systematic reviewers often wish to practice this recommendation and plan to conduct meta-analyses according to the ITT principle. But what does this mean, and how might it be achieved?

*The ITT principle*

The basic intention-to-treat principle is that participants in trials should be analysed in the groups to which they were randomized, regardless of whether they received or adhered to the allocated intervention. Two issues are involved here. The first issue is that participants who strayed from the protocol (for example by not adhering to the prescribed intervention, or by being withdrawn from active treatment) should still be kept in the analysis. An extreme variation of this is participants who receive the treatment from the group they were not allocated to, who should be kept in their original group for the analysis. This issue causes no problems provided that, as a systematic reviewer, you can extract the appropriate data from trial reports.

The rationale for this approach is that, in the first instance, we want to estimate the effects of allocating an intervention in practice, not the effects in the subgroup of participants who adhere to it.

The second issue in ITT analyses is the problem of loss to follow-up. People are lost from clinical trials for many reasons. They may die, or move away; they may withdraw themselves or be withdrawn by their clinician, perhaps due to adverse effects of the intervention being studied.

If participants are lost to follow-up then the outcome may not be measured on them. But the strict ITT principle suggests that they should still be included in the analysis. There is an obvious problem – we often do not have the data that we need for these participants. In order to include such participants in an analysis, we must either find out whether outcome data are available for them by contacting the trialists, or we must ‘impute’ (i.e. make up) their outcomes. This involves making assumptions about outcomes in the ‘lost’ participants.
Several methods of filling in missing outcomes can be compared in a sensitivity analysis.

There are many ‘formal’ approaches to imputing missing outcomes in clinical trials. A review of these is beyond the scope of this course. We shall look at one particular situation that arises commonly and consider some alternative approaches that might be compared in a sensitivity analysis.

Consider the following trial of a Larium-Qinghaosu combination versus Qinghaosu alone for treating malaria that was included in a Cochrane review. Although 20 were randomized to the former and 18 to the latter, results were available only for the 34 people that did not drop out. These were the findings regarding the presence of parasitic infection after four weeks:

<table>
<thead>
<tr>
<th></th>
<th>Clear</th>
<th>Not clear</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Larium+Qinghaosu</td>
<td>17</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>Qinghaosu</td>
<td>10</td>
<td>7</td>
<td>17</td>
</tr>
</tbody>
</table>

There were two other similar trials, also with missing data. In order to perform an ITT analysis, these data needed to be imputed. Four particular strategies for doing this are:

1. (Assume the best) assume everybody missing was clear of infection
2. (Assume the worst) assume everybody missing was not clear of infection
3. (Best-case scenario for combination treatment) assume everybody missing on the combination treatment was clear and everybody missing on Qinghaosu was not clear
4. (Best-case scenario for single treatment, or worst-case scenario for combination treatment) assume everybody missing on the combination treatment was not clear and everybody missing on Qinghaosu was clear

Strategies 1 and 2 are attempts to fill in the missing data in a realistic manner. One of these may be an obvious contender for your situation. For example, in clinical trials of smoking-cessation treatments it may be reasonable to assume that dropouts continue to smoke.

Strategies 3 and 4 put bounds on the possible results of the trial had all participants been observed. Fortunately in the malaria review the results were not sensitive to any of these strategies. When dropout rates are higher the results may not be robust and caution may be needed in interpreting the findings.

If your outcome is a continuous measure, imputation of missing data for ITT purposes is more difficult, as there are more than two different possibilities for each participant.
Because imputation of missing data in order to perform a full ITT analysis is controversial, it may be best to present only the results for available participants. If you do this, you should also consider the possible effects of the missing participants, either through sensitivity analyses as described here or by discussing the implications in the Discussion of your review. An alternative approach may be to only analyse the data available, but to consider drop out rate as a marker of trial quality. Whichever approach you use, ensure that it is described in the methods section of the review and that the numbers of participants with missing data are described in the results section and the characteristics of included studies table.

Indirect comparisons

We talked a little about indirect comparisons when we considered subgroup analysis in the module on heterogeneity. We frequently wish to compare interventions across studies. Here are two examples.

- In Module 13 we thought about a collection of trials comparing training of systematic reviewers with no training of systematic reviewers. Some trials used self-directed learning and others used face-to-face training. None compared the two types of training, although we really want to know which one works best.

- A Cochrane review of pharmacological interventions for dyspepsia included many comparisons between different drugs and between drugs and placebo. Among them were comparisons of histamine H2 antagonists (H2RAs) versus placebo, comparisons of prokinetics versus placebo and comparisons of H2RAs versus prokinetics. There were substantially more placebo-controlled comparisons, yet an important clinical question is whether H2RAs are more effective than prokinetics. Could the placebo-controlled studies reliably tell the reviewers more about this comparison?

It is clear that the best way to compare treatments is to seek direct comparisons in randomized trials. Sometimes there aren’t enough of these studies to draw reliable conclusions, and sometimes there aren’t any such studies at all. Here we address some of the issues involved in making indirect comparisons across different studies.

The first question is whether we should compare the treatment arms in the different studies, or compare the treatment effect estimates from the different studies.
Suppose we had a single trial of an H2RA versus placebo and a single trial of a prokinetic versus placebo. Can we reliably compare the patients given the H2RA in one trial with the patients given the prokinetic in the other trial? The answer is clearly no. These groups are unlikely to be comparable, because there are likely to be differences in the case-mix, outcome assessments and other aspects of trial design between the trials. Although the groups have been generated by randomisation, the randomisation acted within each trial (H2RA versus placebo, prokinetics versus placebo). Participants were not randomized to be in one trial or the other, i.e. there was no randomisation to prokinetics versus H2RA.

A more reasonable approach is to exploit the randomization within each study, and compare treatment effect estimates. This does not make the comparison randomized, but makes it considerably more reliable since we now relate the effects of the two active drugs to a common reference. For example, if we have one trial comparing Drug A to placebo, and one comparing Drug B to placebo, we could compare the treatment effects measured in the two studies as an indirect way of comparing Drug A to Drug B.

We have determined that indirect comparisons should exploit the comparisons within randomized trials. But when is it reasonable to make indirect comparisons like that suggested above? We need to ask ourselves, ‘are the studies sufficiently similar to provide a meaningful result?’.

Remember, this comparison between the results of two studies is still a non-randomized comparison. The problem with making indirect comparisons is that there are often systematic differences between the types of trials addressing one intervention and the types of trials addressing the other. For example, trials of self-directed learning for systematic reviewers may be undertaken in resource-limited countries, and trials of face-to-face training in resource-rich countries. There may be reasons, other than the approach to learning, for differences in the effectiveness of training in these two situations.

Methods are available for undertaking indirect comparisons if it is reasonable to do so. Meta-regression is one approach. Another approach is to formally compare the estimates and confidence intervals for the direct comparisons. Neither of these methods is available in RevMan.

If we are going to use indirect comparisons, we need to be very cautious about their interpretation.
Summary

We have covered several issues in this module.

Sensitivity analyses are a way of investigating the importance of some of the assumptions and decisions we make during a systematic review.

We have briefly covered some types of data encountered during reviews, and looked at the difficult issue of the value of indirect comparisons.

In the next modules, we start to look at some important issues in the interpretation of the data we have collected.
Module 15: Publication Bias

Learning objectives

- Understand what publication bias is and describe how it could arise
- Be familiar with the design and interpretation of funnel plots
- Be aware of other relationships which can cause funnel plot asymmetry
- Be aware of other forms of bias in the reporting of studies, especially duplicate publication, selective reporting of outcomes and subscales, and subgroups of data.
- Consider methods of reducing the potential impact of publication bias on the results of a systematic review
- Be aware of ongoing research in this field

Other relevant material

- Presentation on reading a funnel plot

Relevant sections of the Reviewers’ Handbook

- Section 8.13

Where does this go in a Cochrane review?

- Ensure your search strategy is as sensitive as it can be to reduce the chances of missing negative studies, possibly attempt to measure publication bias as part of the analysis of your review in RevMan and consider the likelihood of publication bias in your review’s Discussion.
What is publication bias?

Systematic reviews aim to find and assess for inclusion all high quality studies addressing the question of the review. But finding all studies is not always possible and we have no way of knowing what we have missed. Does it matter if we miss some of the studies? It will certainly matter if the studies we have failed to find differ systematically from the ones we have found. Not only will we have less information available than if we had all the studies, but we might come up with the wrong answer if the studies we have are unrepresentative of all those that have been done.

We have good reason to be concerned about this, as many researchers have shown that those studies with significant, positive, results are easier to find than those with non-significant or ‘negative’ results. The subsequent over-representation of positive studies in systematic reviews may mean that our reviews are biased toward a positive result.

Publication bias is just one type of a group of biases termed reporting bias. We have quite a lot of evidence that these biases exist, so it is fair to assume that most systematic reviews will be subject to reporting bias to some extent.

Publication bias and other related biases can be summarised as statistically significant, ‘positive’ results being:

- More likely to be published (publication bias)
- More likely to be published rapidly (time lag bias)
- More likely to be published in English (language bias)
- More likely to be published more than once (multiple publication bias)
- More likely to be cited by others (citation bias)

All of these reporting biases make positive studies easier to find than those with non-significant results, something that we can try to minimise by extensive searching.

Managing publication bias

If we accept that your review will almost certainly be subject to publication bias to some extent, we are left with the problem of estimating how big a problem it is in your review, and what to do about it. There are several methods for getting an idea about how much of a problem this may be, and the method available in RevMan is the funnel plot. This is available in RevMan as a tool for reviewers, but is not on The Cochrane Library. This means you should use the funnel plot option to investigate the presence of publication bias in your review and then discuss this in the Discussion section of the text of your review. If you suspect there may be a problem in your review, you need to bear this in mind when making your conclusions and recommendations. The likeliest scenario is that the results of your review are biased to the positive.
Interpreting funnel plots

If publication bias is not present, you would expect your funnel plot to be roughly symmetrical, as in the example below:

As the studies become less precise (i.e. higher standard error), you would expect the results (given here as a log odds ratio) of the studies to be more variable, scattered to both sides of the more precise larger studies.

When you plot your studies onto a funnel plot, you may find it is not symmetrical and does not resemble an inverted funnel. This may be due to publication bias, however there are other factors leading to an asymmetrical plot.
The next funnel plot is from a review of prevention for chronic non-steroidal anti-inflammatory medication induced gastro-intestinal toxicity.

![Funnel plot example](image)

An asymmetrical funnel plot may be due to study factors other than publication bias.

As you can see it is not symmetrical, although this impression is mainly caused by one small study to the left of the most common effect. This may indicate publication bias, but there are other possible explanations. The small study may be of lesser quality, and poor quality studies, especially those failing to conceal allocation, often result in exaggerated treatment effect sizes. Or this small study may have been performed in a particularly high risk population where the effect is large. In looking at this plot, we can only report that there may be publication bias.

Look at the plot below from a review of Aversive Smoking for smoking cessation. The outcome is risk of quitting, so the larger the OR the better aversive smoking works.

![Funnel plot example](image)

Does this look symmetrical? At first look it appears that the smaller, less precise studies are all much more positive than the larger, more precise studies, and there are no smaller studies to the left (negative) side of the graph. This appears to be a good example of publication bias.
If however, we add the control event rates (quit rate in the control group) to the plot, the interpretation may be different.

The trials with the lowest control event rates demonstrate the most positive results. We could convince ourselves when looking at this that the pattern of greater effect with lower control event rates represented a true relationship, adverse smoking works better in those more addicted people less likely to give up anyway (i.e. without the experimental intervention). Or it could be publication bias. There are lots of possible explanations for this pattern. The point is that from the funnel plot it is impossible to know.
Another possible type of funnel plot is a hollow plot, like this one from a review of dieting to reduce body weight for controlling hypertension in adults.

Here there are some trials to the right of the point of no effect, indicating that dieting increases blood pressure (measured as the mean difference on a continuous scale) and some to the left, indicating that dieting reduces blood pressure. There are no trials around no difference. This is possibly publication bias of the type where significant studies (i.e. those showing the intervention is significantly beneficial and those showing the intervention is significantly harmful) are published or found systematically more than those showing no difference.

From these examples, we can see that a funnel plot is not a very reliable method of investigating publication bias, although it does give us some idea of whether our study results are scattered symmetrically around a central, more precise effect. Funnel plot asymmetry may be due to publication bias, but it may also result from clinical heterogeneity between studies (for example different control event rates) or methodological heterogeneity between studies (for example failure to conceal allocation).

Even if there is publication bias in a review, it may not result in an asymmetrical funnel plot, for example when the plot is hollow.

There are some statistical tests for detecting funnel plot asymmetry (Egger’s linear regression test and Begg’s rank correlation test) but these have low power and are rarely used in Cochrane reviews. If you would like to use them, you should discuss this with a statistician.
**Correcting for publication bias.**
From what we have seen in this module so far, we know that the methods we have for detecting the possibility of publication bias in systematic reviews are not very good. Any methods for attempting to correct for this perceived bias are therefore also not ideal, but the following methods have been suggested. These are rarely used in Cochrane reviews but are included here for completeness.

**“Trim and fill method”**
In this method the tail of the side of the funnel plot with the smaller studies is chopped off to make the funnel plot symmetrical, and it is then replicated and added back to both sides so the plot becomes symmetrical. The centre and variability of the filled funnel plot are then estimated (there are complicated statistical methods to do this formally).

**Fail safe N**
Here, the number of null studies (of similar size) which would be required to remove an observed significant effect is estimated. This method may give you an idea of the likely importance of any publication bias present. For example, if it tells you that several large negative trials would need to exist to overturn your positive result, you may decide it is quite unlikely that these were missed. This remains, however, a judgement.

**Modelling**
Models for the probability that studies with particular results do or do not get published can be designed and used to investigate possible publication bias.

**Summary**
There is quite a lot of work being undertaken, both in the form of trials registers and more intensive searching to try to help reviewers identify all trials, and methodological research to advance the ways we measure and account for publication bias in systematic reviews. Currently however, the only thing we know for certain about publication bias is that it exists. Our methods for assessing its presence can only provide suggestions, not definite answers.

The main purpose of including issues to do with publication bias in your review is to ensure that you, and readers of your review, are aware of the fact that publication bias is possible, and to attempt, at least in part, to estimate how big an impact it might have on the results of your review.
There should be a label telling you what the comparison is.
The vertical axis is some measure of the precision of the estimate of the treatment effect. So the smaller the confidence interval, the more precise the study, and the further up the study is placed.

Here, the measure of precision is the standard error of the log RR. Elsewhere, you may see sample size or weight used.
The horizontal axis measures the treatment effect - here it is the relative risk, on a log scale so that the distance from 0.1 to 1 is the same as from 1 to 10.
The point estimate from each study is then plotted, and a vertical line added (this line isn’t added in all packages) where the pooled estimate from the meta-analysis lies.
We would expect less precise studies (with fewer participants and events) to be more affected by the play of chance, and so more widely scattered about the pooled estimate.

As studies get bigger with more events, we expect them to be closer to the pooled estimate. Overall, this should produce a triangular shape, or inverted funnel (depending on how the axes are plotted).
Module 16: Strength and relevance of the evidence.

This module will cover how to discuss your results both in terms of how effective (or ineffective) an intervention may be and how the results of your review may be applied to individual health care situations. We will discuss interpretation of the results of your review as it applies to writing your results and discussion section, not with respect to using the review to making an actual decision about health care.

Learning objectives

- Be able to recognize the components of evidence in health care interventions and understand their relative contributions to the strength of evidence of a systematic review (study design, number of studies, study quality, statistical significance, clinical importance, biological plausibility, and consistency of results)
- Be able to interpret the available evidence as strong, weak or inconclusive and be able to justify these ratings
- Judge whether a particular systematic review is likely to be of limited or wide applicability
- Recognize the features of a particular systematic review that may limit or widen its applicability (considering degree of mismatch between trial characteristics and settings to which the results may be applied)
- Understand the purpose, calculation and application of NNTs to assist assessment of appropriate application

Relevant sections of the Reviewers’ Handbook

- Sections 9.1 and 9.2

Where does this go in a Cochrane review?

- The information in this Module relates to the Results and Discussion section of your review.
The forest plots and summary statistics you generate in RevMan form the results of your review, but it is essential that you summarize them in the Results section. Make sure you include all the important results, not just those that are statistically significant.

The purpose of the Discussion and Reviewers’ Conclusions section of your review is for you to help the reader interpret your results. There are three main things you need to cover:

- Any limitations of your review and the assumptions you have made
- The strength of the evidence
- The relevance of the evidence

**Limitations and assumptions of the review**

From all the previous modules you have seen that when preparing a systematic review we make many choices, most based on assumptions. All the choices you made that may have affected the results of your review should be outlined in the Methods section of your review, and preferably tested with sensitivity analyses. There may be some aspects of your review (for example that most included trials had methodological flaws, or were small; or that you haven’t been able to get the data for some of the included trials, or there is funnel plot asymmetry) that you want to highlight in your Discussion section, along with some thoughts about how these issues may have affected the results of your review.

**Strength of evidence**

Section 9.1 of the Reviewers’ Handbook discusses some things you may like to consider when discussing the strength of evidence in your review and you should read it now.

The phrase ‘strength of evidence’ applies to more than just the results of your review, although your review will obviously contribute greatly to the overall body of knowledge on its topic. It is how strong the overall case for the use or cessation of use for the intervention is. The strength of evidence relating to your review question is determined by factors both within your review and external to your review.
Internal factors

Some factors internal to your review which you may want to consider when drawing conclusions about the strength of evidence are:

- Methodological issues relating to both the review and the included trials as outlined in earlier modules. If most included trials were methodologically sound, with adequate allocation concealment, careful control for confounding and little missing data you may be more confident regarding the strength of your conclusions.
- The number of studies in your review and the number of participants in the studies. If your data are sparse, the evidence is less strong and you should be careful about what you conclude.

For example, compare the strength of evidence from the forest plots below.

### Comparison: 01 Operatively single agent and post-operatively single agent, single dose Antibiotics vs placebo

<table>
<thead>
<tr>
<th>Outcome: 01 Wound Infection</th>
<th>Peto OR (95% CI)</th>
<th>Weight %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>Treatment n/N</td>
<td>Control n/N</td>
</tr>
<tr>
<td>01 Appendicectomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biener 1985 A</td>
<td>3/39</td>
<td>8/39</td>
</tr>
<tr>
<td>Biener 1985 B</td>
<td>0/44</td>
<td>0/39</td>
</tr>
<tr>
<td>Vidalis 1975</td>
<td>3/47</td>
<td>4/46</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>3/132</td>
<td>25/124</td>
</tr>
<tr>
<td>Test for heterogeneity: chi-square p=0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: z=0.85 p=0.19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 trials with total of 256 participants versus 21 trials with total of 2343 participants, the strength of evidence in the second example is greater.

### Comparison: 01 Systemic Antibiotics vs Placebo (Clinical)

<table>
<thead>
<tr>
<th>Outcome: 01 Wound Infection</th>
<th>Peto OR (95% CI)</th>
<th>Weight %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>Treatment n/N</td>
<td>Control n/N</td>
</tr>
<tr>
<td>01 Appendicectomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergman 1995</td>
<td>5/60</td>
<td>8/65</td>
</tr>
<tr>
<td>Brown 1988 A</td>
<td>3/70</td>
<td>8/79</td>
</tr>
<tr>
<td>Brogden 1988 B</td>
<td>0/44</td>
<td>5/39</td>
</tr>
<tr>
<td>Cohn 1983 A</td>
<td>4/100</td>
<td>17/100</td>
</tr>
<tr>
<td>Cohn 1983 B</td>
<td>10/100</td>
<td>17/100</td>
</tr>
<tr>
<td>Cohn 1983 C</td>
<td>7/100</td>
<td>17/100</td>
</tr>
<tr>
<td>Corti 1979</td>
<td>4/152</td>
<td>9/153</td>
</tr>
<tr>
<td>Cramer 1970 A</td>
<td>1/16</td>
<td>3/17</td>
</tr>
<tr>
<td>Creutzig 1980 G</td>
<td>5/2</td>
<td>3/2</td>
</tr>
<tr>
<td>Fodor 1979</td>
<td>1/70</td>
<td>3/68</td>
</tr>
<tr>
<td>Geenland 1990</td>
<td>1/70</td>
<td>6/69</td>
</tr>
<tr>
<td>Geenland 1979</td>
<td>1/49</td>
<td>12/15</td>
</tr>
<tr>
<td>Gift 1975</td>
<td>0/14</td>
<td>4/14</td>
</tr>
<tr>
<td>Kater 1993</td>
<td>3/30</td>
<td>4/29</td>
</tr>
<tr>
<td>Loke 1995</td>
<td>6/100</td>
<td>17/100</td>
</tr>
<tr>
<td>Marshall 1992</td>
<td>2/10</td>
<td>4/10</td>
</tr>
<tr>
<td>Marshall 1982</td>
<td>1/12</td>
<td>8/10</td>
</tr>
<tr>
<td>Mehal 1970</td>
<td>0/10</td>
<td>4/10</td>
</tr>
<tr>
<td>Pelletier 1983</td>
<td>1/70</td>
<td>6/69</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>3/112</td>
<td>17/101</td>
</tr>
<tr>
<td>Test for heterogeneity: chi-square p=0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: z=0.19 p=0.19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© The Cochrane Collaboration 2002
• The size of the treatment effect. If your summary treatment effect is large, and the confidence intervals fall within a range that would be considered clinically significant, the strength of evidence is greater.

For example, compare the strength of evidence from the forest plots below.

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment</th>
<th>Control</th>
<th>Odds Ratio</th>
<th>95% CI</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leith 1970</td>
<td>6/100</td>
<td>17/100</td>
<td>2.8</td>
<td>0.24-14.09</td>
<td>0.24</td>
</tr>
<tr>
<td>O'Sullivan 1982</td>
<td>2/145</td>
<td>5/145</td>
<td>0.9</td>
<td>0.44-10.62</td>
<td>0.46</td>
</tr>
<tr>
<td>Rodger 1979</td>
<td>2/42</td>
<td>7/46</td>
<td>0.1</td>
<td>0.04-0.96</td>
<td>0.56</td>
</tr>
<tr>
<td>Whalley 1982</td>
<td>7/192</td>
<td>5/185</td>
<td>1.9</td>
<td>0.83-20.32</td>
<td>0.83</td>
</tr>
<tr>
<td>Wyllie 1975</td>
<td>0/40</td>
<td>9/46</td>
<td>1.1</td>
<td>0.10-10.41</td>
<td>0.25</td>
</tr>
<tr>
<td>Williams 1983</td>
<td>0/51</td>
<td>5/52</td>
<td>1.1</td>
<td>0.10-10.41</td>
<td>0.74</td>
</tr>
<tr>
<td>Subtotal (N=49)</td>
<td>52/1162</td>
<td>171/1181</td>
<td>0.31</td>
<td>0.10-0.92</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Test for heterogeneity: chi-square = 18.07, df = 8, p = 0.047
Test for overall effect: z = -3.12, p = 0.001

OR of 0.31 with upper confidence interval of 0.42 indicates confidence intervals fall within a clinically significant range compared to OR of 0.52 with upper CI of 0.93, upper confidence interval may not be clinically significant. The first example is stronger evidence for an important effect.
- The precision of the treatment effect. If the confidence intervals around your summary estimate are narrow and so you are more sure that the ‘true’ result lies within the range bordered by the upper and lower confidence interval (and is clinically significant) you can be more confident about the strength of evidence.

For example, compare the strength of evidence from the forest plots below.

<table>
<thead>
<tr>
<th>Comparison:</th>
<th>Systemic Antibiotics vs Placebo (Clinical)</th>
<th>Outcomes:</th>
<th>51 Weeds infection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>Treatment (no.)</td>
<td>Control (no.)</td>
<td>Pons Off (%)</td>
</tr>
<tr>
<td>Biber 1985</td>
<td>17/45</td>
<td>10/63</td>
<td>0.07</td>
</tr>
<tr>
<td>Biber 1986</td>
<td>15/50</td>
<td>11/50</td>
<td>0.10</td>
</tr>
<tr>
<td>Biber 1987</td>
<td>18/60</td>
<td>12/60</td>
<td>0.12</td>
</tr>
<tr>
<td>Biber 1988</td>
<td>19/62</td>
<td>13/62</td>
<td>0.14</td>
</tr>
<tr>
<td>Biber 1989</td>
<td>20/64</td>
<td>14/64</td>
<td>0.16</td>
</tr>
<tr>
<td>Biber 1990</td>
<td>21/66</td>
<td>15/66</td>
<td>0.18</td>
</tr>
<tr>
<td>Biber 1991</td>
<td>22/68</td>
<td>16/68</td>
<td>0.20</td>
</tr>
<tr>
<td>Biber 1992</td>
<td>23/70</td>
<td>17/70</td>
<td>0.22</td>
</tr>
<tr>
<td>Biber 1993</td>
<td>24/72</td>
<td>18/72</td>
<td>0.24</td>
</tr>
<tr>
<td>Biber 1994</td>
<td>25/74</td>
<td>19/74</td>
<td>0.26</td>
</tr>
<tr>
<td>Biber 1995</td>
<td>26/76</td>
<td>20/76</td>
<td>0.28</td>
</tr>
<tr>
<td>Biber 1996</td>
<td>27/78</td>
<td>21/78</td>
<td>0.30</td>
</tr>
<tr>
<td>Biber 1997</td>
<td>28/80</td>
<td>22/80</td>
<td>0.32</td>
</tr>
<tr>
<td>Biber 1998</td>
<td>29/82</td>
<td>23/82</td>
<td>0.34</td>
</tr>
<tr>
<td>Biber 1999</td>
<td>30/84</td>
<td>24/84</td>
<td>0.36</td>
</tr>
<tr>
<td>Biber 2000</td>
<td>31/86</td>
<td>25/86</td>
<td>0.38</td>
</tr>
<tr>
<td>Biber 2001</td>
<td>32/88</td>
<td>26/88</td>
<td>0.40</td>
</tr>
<tr>
<td>Biber 2002</td>
<td>33/90</td>
<td>27/90</td>
<td>0.42</td>
</tr>
<tr>
<td>Biber 2003</td>
<td>34/92</td>
<td>28/92</td>
<td>0.44</td>
</tr>
<tr>
<td>Biber 2004</td>
<td>35/94</td>
<td>29/94</td>
<td>0.46</td>
</tr>
<tr>
<td>Biber 2005</td>
<td>36/96</td>
<td>30/96</td>
<td>0.48</td>
</tr>
<tr>
<td>Biber 2006</td>
<td>37/98</td>
<td>31/98</td>
<td>0.50</td>
</tr>
<tr>
<td>Biber 2007</td>
<td>38/100</td>
<td>32/100</td>
<td>0.52</td>
</tr>
<tr>
<td>Biber 2008</td>
<td>39/102</td>
<td>33/102</td>
<td>0.54</td>
</tr>
<tr>
<td>Biber 2009</td>
<td>40/104</td>
<td>34/104</td>
<td>0.56</td>
</tr>
<tr>
<td>Biber 2010</td>
<td>41/106</td>
<td>35/106</td>
<td>0.58</td>
</tr>
<tr>
<td>Biber 2011</td>
<td>42/108</td>
<td>36/108</td>
<td>0.60</td>
</tr>
<tr>
<td>Biber 2012</td>
<td>43/110</td>
<td>37/110</td>
<td>0.62</td>
</tr>
<tr>
<td>Biber 2013</td>
<td>44/112</td>
<td>38/112</td>
<td>0.64</td>
</tr>
<tr>
<td>Biber 2014</td>
<td>45/114</td>
<td>39/114</td>
<td>0.66</td>
</tr>
<tr>
<td>Biber 2015</td>
<td>46/116</td>
<td>40/116</td>
<td>0.68</td>
</tr>
<tr>
<td>Biber 2016</td>
<td>47/118</td>
<td>41/118</td>
<td>0.70</td>
</tr>
<tr>
<td>Biber 2017</td>
<td>48/120</td>
<td>42/120</td>
<td>0.72</td>
</tr>
<tr>
<td>Biber 2018</td>
<td>49/122</td>
<td>43/122</td>
<td>0.74</td>
</tr>
<tr>
<td>Biber 2019</td>
<td>50/124</td>
<td>44/124</td>
<td>0.76</td>
</tr>
<tr>
<td>Biber 2020</td>
<td>51/126</td>
<td>45/126</td>
<td>0.78</td>
</tr>
<tr>
<td>Biber 2021</td>
<td>52/128</td>
<td>46/128</td>
<td>0.80</td>
</tr>
<tr>
<td>Biber 2022</td>
<td>53/130</td>
<td>47/130</td>
<td>0.82</td>
</tr>
<tr>
<td>Biber 2023</td>
<td>54/132</td>
<td>48/132</td>
<td>0.84</td>
</tr>
<tr>
<td>Biber 2024</td>
<td>55/134</td>
<td>49/134</td>
<td>0.86</td>
</tr>
<tr>
<td>Biber 2025</td>
<td>56/136</td>
<td>50/136</td>
<td>0.88</td>
</tr>
<tr>
<td>Biber 2026</td>
<td>57/138</td>
<td>51/138</td>
<td>0.90</td>
</tr>
<tr>
<td>Biber 2027</td>
<td>58/140</td>
<td>52/140</td>
<td>0.92</td>
</tr>
<tr>
<td>Biber 2028</td>
<td>59/142</td>
<td>53/142</td>
<td>0.94</td>
</tr>
<tr>
<td>Biber 2029</td>
<td>60/144</td>
<td>54/144</td>
<td>0.96</td>
</tr>
<tr>
<td>Biber 2030</td>
<td>61/146</td>
<td>55/146</td>
<td>0.98</td>
</tr>
<tr>
<td>Biber 2031</td>
<td>62/148</td>
<td>56/148</td>
<td>1.00</td>
</tr>
</tbody>
</table>

CIs of 0.29 to 0.38 compared to 0.25 to 0.73
The first example is more precise, hence greater strength of evidence.
• The consistency of the results. If the results of all or most of the trials in your review are in the same direction (i.e. demonstrating an effect) the evidence is stronger (although ensure you also discuss the possibility of publication bias).

For example, compare the strength of evidence from the forest plots below.

Consistency of effect versus disagreement between trials (i.e. all effect estimates in example 1 to the left of the line, example 2 differs in direction). The first example offers greater strength of evidence.
• The consistency of outcomes. If the intervention is showing similar effects on all related outcomes (for example if an exercise program both significantly reduces pain and increases function) you may be more confident to conclude it is effective.

For example, compare the strength of evidence from the forest plots below.
- Apparent dose response relationship. It may be that if an intervention is significantly beneficial (or harmful), the more of the intervention you have, the better (or worse) you will do. This is known as a dose-response relationship and, in studies of causation, the presence of a dose-response relationship (i.e. the more of a harmful agent you are exposed to the greater your chance of developing the outcome), the stronger the evidence of association. In some reviews however dose-response association may not be important, for example if there is no threshold dose.

**External factors**

- Biological plausibility. Your conclusion will be strengthened if the effect makes sense. There may well be laboratory based research to explain the effect demonstrated by your results, for example studies analyzing the biomechanics of hemiplegic gait may help explain evidence for the effect of a physiotherapy intervention following stroke. Similarly, the effect on intermediate outcomes, such as physiological markers, may help to explain and so strengthen the effect demonstrated by your review.
• Other evidence. There may be non-randomised studies, such as cohort studies or case series which are not included in your review but support your conclusions. This may strengthen the evidence. It is important to bear in mind however that these studies were excluded from your review for a reason (probably methodological) and so not too much weight should be given to their conclusions. In addition, you are unlikely to have systematically searched for these non-randomised studies and so you may not have all available information.

• Concordance with related reviews. If your review results are similar to other related reviews and the intervention appears effective (or not effective) in other comparable situations, this will strengthen your conclusions. For example, if an analgesic medication demonstrated similar pain relieving properties in a review of populations with acute pain and a review of populations with chronic pain, the strength of evidence for the intervention could be interpreted as stronger.

All of the above points need to be considered when helping the reader interpret the results of your review. Your conclusions about the effect of the intervention should reflect the strength of evidence as determined by these internal and external factors.

**Applicability and Relevance of the evidence**

Even if the evidence is strong, with significant, precise, consistent and plausible results, it may not be useful to all those reading your review. Clinicians and consumers, looking for evidence about the best way to deal with an individual health care situation, will not only need evidence which is strong, they will need to know if it applies to them and is relevant to their clinical need.

Section 9.2 of the *Reviewers Handbook* discusses Applicability, or Relevance, and you should read it now.

Don’t forget that the results of your review may be used in many different populations and settings around the world. Care needs to be taken not to interpret your review only as it applies to your own setting. Reviewers should try to help the reader in applying the results of their review to individual settings.
A useful way of doing this is to ask yourself if there are any reasons why the results of your review may be different if the intervention was used in a different setting. Some examples may be:

- **Biologic and cultural variation**
  There may be genetic reasons why an intervention may have varying effects in different populations with varying risks and co-morbidities. There may be gender or age differences in response to the treatment, or the intervention may simply not be feasible or acceptable in a given setting.

- **Variation in baseline risk**
  If all the trials in your review include populations with a similar baseline risk of the outcome (i.e. control event rate), while this may strengthen the evidence for the effect of the intervention in such populations, it may limit the applicability of the results. In asking whether the review is relevant, the issue is whether or not the evidence matters. In some cases interventions are assessed only by their effect on outcomes that are not important to the people with the disorder. Take the example of antiviral regimes for people living with HIV. If a review were only to report the effect of the regime on CD-4 count, and not include relevant outcomes such as health related quality of life, adverse drug effects or survival, the review may not be relevant to people living with HIV, although it may be helpful to their physician.

A relevant review is one that asks a sensible question. A review, or indeed a trial, that assesses the effect of an intervention compared to placebo when there is a proven effective treatment which can be used is clinically irrelevant. What we need to know is whether the new intervention is better than the existing proven intervention. You may chose to include trials comparing the new intervention to placebo in your review in order to draw an indirect comparison (as discussed in Module 14, Further Considerations In Meta-Analysis), however the primary question of your review should be to determine effect compared to the existing intervention.

**Summary**

In conclusion, when writing the Discussion section of your review you need to consider not only how convincing the evidence is in terms of the effect of the intervention, but also how this evidence will help in the many clinical settings in which people will wish to apply it.

The following Modules will discuss making conclusions from your review with respect to trade-offs between benefit and harm, and ways of expressing uncertainty.
Module 17: Applying the results – trade-offs, adverse effects and outcomes

This module will discuss the stage in a review when, having analysed and presented the results, we are trying to interpret their application to clinical practice and policy.

Learning objectives

- Understand the necessary (but 'insufficient') role of systematic reviews in making decisions about health care
- Appreciate that the balance between benefit and risk may be different for different groups and different individuals
- Understand the limitations of systematic reviews in identifying all adverse effects and be aware of other sources of information which are more likely to be able to identify possible adverse effects
- Be aware of the problems in interpreting subjective outcomes
- Appreciate the need for caution in interpreting and applying results from surrogate outcomes

Relevant sections of the Reviewers’ Handbook

- Sections 9.2 - 9.5

Where does this go in a Cochrane review?

- The Inclusion Criteria section of the review should describe all the outcomes (both benefits and harms) that you will consider in the review
- The Discussion section of the review should describe how you have interpreted the results, in particular the strength of evidence about all the outcomes, clarification of any important trade-offs between the expected benefits, harms and costs of the intervention, as well as any other considerations that might be relevant to someone making a decision

© The Cochrane Collaboration 2002
Evidence of effectiveness alone is not enough for making decisions

Think about the decision of whether or not to use beta-interferon for people with multiple sclerosis. This is a relatively new drug for a condition which can progress, and for which few other treatments are available. Try to write down a list of all the factors that come into your mind when you are trying to make a health care decision like this.

Your list probably has many factors, among which will (hopefully) be evidence about whether the treatment improves the outcome for people with multiple sclerosis. But I imagine your list would probably include things like any special needs of the user, their priorities and values, as well as the available resources. You may have many other things on your list as well.

The key point is that research evidence is only one factor, albeit an important one, that needs to be taken into account during the decision-making process. Other types of evidence, such as the needs of the user, their priorities and values, as well as the available resources also need to be considered in the decision-making process.

It is not the job of the reviewer to try to weigh up all those factors, because a reviewer can never take account of all the variation in all the factors in the list in all the places that the review will be used. It is, however, possible to assist the user of your review in how to interpret and apply the evidence.

In the previous module we began the task of interpreting the results by considering the strength and relevance of the evidence. Even with this information the ultimate decision as to how to apply the evidence will be up to the users of the review. We need to remember that users of reviews will often come from very different settings around the world with very different circumstances and backgrounds. In this module we explore some of the other things that we can do to help users decide how to apply the evidence.

Identifying all the possible outcomes

A useful starting point is to identify all the outcomes that were considered in the review. These should have been listed with the inclusion criteria for studies in the review. It is often a good idea to have a further check that you haven’t missed any important outcomes. For example, reviewers often unintentionally focus on the positive effects of treatment and forget the possible adverse effects.
In many reviews it may be important to consider the cost of treatment as one of the outcomes. Undertaking formal economic evaluations of the costs and benefits of an intervention are beyond the scope of the standard review. However, increasingly systematic reviews are being used as part of economic analyses.

**Drawing up a balance sheet**

A useful tip is to draw up a balance sheet where you list all the “beneficial” (or positive) effects associated with the intervention on one side, and all the potential “harms” (or negative) outcomes on the other side.

Take the example of whether or not to treat a sore throat with antibiotics (when we don’t know for sure what is causing it). The potential positive effects of treatment include:

- Reduction in severity of illness
- Reduction in pain
- Reduction in the duration of illness
- Reduction in infections following on (such as sinusitis, and acute otitis media)
- Reduction in subsequent noninfective complications (such as rheumatic fever and acute glomerulonephritis (an acute inflammation of the kidney)).

The potential negative effects of treatment include:

- Adverse effects of antibiotics (such as diarrhoea, thrush etc).
- The cost to society in terms of antibiotic prescription and overuse

Sometimes the labeling of outcomes can be subjective, depending on the perspective you are coming from. Take the example of an intervention for Alzheimer’s disease. If the intervention improves mobility, this might be seen as a positive effect from a patient’s perspective, but a negative effect from a carer’s perspective.

**What if there are no data available on some outcomes?**

It may be that there are no data available on some of the outcomes. For example, some of the slow onset or more rare beneficial or adverse effects associated with treatments are unlikely to be seen in trials with relatively short-term follow-up or with small numbers of participants. It is often necessary to consider other types of data (such as from long term cohort studies or post marketing surveillance studies) in order to obtain reliable data on potential harms. To do this thoroughly, however, adds a whole new dimension to a systematic review, as we would have to search for these studies and then appraise and combine them. At the moment few Cochrane reviews do this.
If you have insufficient information about particular outcomes, this should be clearly stated in your review. If you feel these outcomes are important you may suggest that they are included in future trials by discussing them in the “Recommendations for Future Research” section of your review.

**What do users of reviews need to know about these outcomes?**

We have seen, in Module 11, the importance of knowing the absolute change (not just the relative change) in the probability of each outcome as a result of the intervention. However, in order for users in different settings to decide how to apply this information, they also need to know how common the outcome is in their particular setting. To come back to the example we used for the balance sheet, in most developed countries the risk of rheumatic fever or acute glomerulonephritis as a complication of sore throat is extremely low, however, in many developing countries the risk is much greater to begin with, and even a small reduction in the absolute probability of this outcome may be very important.

In other instances, it may be important for the user to know the natural history of the condition. For example, with many self-limiting illnesses, a one day reduction in symptom duration, while it may be statistically significant, needs to be set against an illness which may only last for four or five days at most.

**Try to avoid value judgments**

Value judgments are any statements where you make assumptions about the value placed on particular outcomes. An easy way to spot them is to ask yourself “Would everyone agree with this statement?” If the answer is probably not, then it may be best to not make that judgment in your review. It’s better to present the data in a way that will allow the user to make a balanced objective “trade-off” given their own personal circumstances. A one day reduction in the duration of an illness may mean a lot to a particular user, where others will be willing to accept that extra day’s illness in return for reducing their chance of some other outcome (for example, the side effects of the intervention).

**Ask the question: Can the results apply in my situation?**

People in many different situations will use your review. It is important that you write the review in a way which will allow individuals to decide whether your review applies to them.
First, a user has to decide whether the review provides valid information about the potential benefits and harms that may be important to them. Then they need to decide whether the participants and settings in the included studies are reasonably similar to their own situation. It is often helpful for users to consider asking themselves whether there are any good reasons why the evidence should not be applied in their situation. Some of these reasons might be related to:

- Biological and cultural variation
- Variation in compliance
- Variation in the baseline risk

These are explained further in sections 9.2.1 – 9.2.3 of the Reviewers’ Handbook, which you should read now.

Now try to think of some examples where these factors might influence the applicability of the findings from your own review and write these down. An example would be that if the trials in your review are done in a developed country with good access to diagnostic equipment, and the reader of your review is a doctor trying to decide whether to implement the treatment in a resource poor country where the equipment available to determine diagnosis is not available, and there are no trials in your review performed in participants similar to the people she treats, the results of your review may not help her make this decision.

If you have come up with any factors that might limit the applicability of your own review these might be worth highlighting in the Discussion.
How consistent are the results?

There are many reasons why the results of studies included within a review may vary (even if there is no heterogeneity detected with statistical testing). For example, in a review of giving smoking cessation advice, we found a considerable variation in the results depending on the intensity of the advice and follow-up provided. Other reasons why results may vary include differences in the participants (for example, their age, gender, or presence of some biochemical marker) or differences in their underlying disease status. You should have an idea of the possible reasons why the results of the trials in your review might vary from thinking about it in the module on heterogeneity. If these differences might be clinically relevant they would be worth highlighting when discussing the evidence in your review.

A word of caution

Some chance variation between different subgroups is inevitable. It is often a trap to try to explore variation in results by undertaking separate analyses of different sub-groups. You should avoid doing this unless there was a good prior reason to believe that a particular sub-group might respond differently to the intervention.

Other useful information

Often there may be some other information that might be useful to consider when discussing the results of a review. For example, it might help to include some information about the size or frequency of a particular health care problem that the review is addressing. However, you need to be quite careful because this information may be very context specific and might be better addressed at a more local level. Much as it would be good if we could provide users of a review with all the information they require to make a decision, this is rarely (if ever) possible.

There are more formal ways of using the evidence and considering benefits, harms, patient preferences and other factors specific to your particular setting. These tools, often termed decision analyses or decision trees, are outside the scope of this material, but often draw on evidence provided by Cochrane Reviews.
A final check ....
At the end of the discussion section it is worth re-reading and asking yourself:

- Have all the main outcomes been considered?
- Have data been presented about the absolute change as a result of the intervention for all possible outcomes?
- Have I considered any factors that might limit the application of these results in different situations?
- Are these results consistent across the included studies or do they vary for some reason?
- Have I avoided making value judgments about how to interpret my findings?
Module 18: Applying the results - inferences and uncertainty

This module will discuss the stage in a review when, having analysed and discussed the findings, we try to draw some conclusions.

Learning objectives

- Recognise inferences which may be misleading
- Be able to draw appropriate conclusions in the absence of evidence
- Be aware of uncertainty in clinical practice and health policy decision-making

Relevant sections of the Reviewers’ Handbook

- Sections 9.6 – 9.7

Where does this go in a Cochrane review?

- In the Discussion section of the review you will usually try to make some inferences from the findings
- The Reviewers’ Conclusions section should succinctly summarise these inferences and any implications from the review for future practice
- The Reviewers’ Conclusions section should also identify any implications from the review for future research
Starting at the end …

We know that many people who read a review begin at the end, by looking at the conclusions and, often, the brief conclusions in the abstract. It is probably human nature as much as anything else to want a ‘bottom line’ from a research study or a review. Unfortunately, it is often very difficult or indeed impossible to provide this bottom line.

Whose conclusions?

As we have already seen in previous modules, users in different settings will often interpret the same evidence from the same review quite differently.

Within a review, any conclusions drawn are only those of the reviewers. For this reason Cochrane reviews specifically refer to Reviewers’ Conclusions in order to highlight where these conclusions come from.

Usually there are three aspects to drawing conclusions:

- Does the intervention work at all? (as covered in module 16)
- What have we learned from this review that can be applied to clinical practice?
- What have we learned from this review about the need for further evaluation and research?

What are the implications for clinical practice

The first component of a conclusion is a bit like putting all the evidence you have obtained from the review on a balance scale and trying to see which way it tips and how confident you are about this interpretation. In essence, you need to decide, after taking all the possible outcomes into account, which of the alternatives (on balance) is likely to do more good than harm (and should be applied in practice) or more harm than good (in which case it should be abandoned from practice). However, more often than not your conclusion will probably be somewhere in between these two scenarios. It may be that there are important trade-offs between different outcomes (such as in the example of treating sore throats with antibiotics that we used in Module 17) and your conclusion needs to reflect this.
Uncertainty reigns

Alternatively, you may find that there is insufficient evidence to arrive at a firm conclusion of any kind. In situations where there is still uncertainty (for example not enough or no trials, trials of poor quality or contradictory evidence from a group of trials), you may be wise to conclude that there is a need for further research in order to gain sufficient evidence to assess the effects of an intervention.

What are the implications for further research

While there may sometimes be a need for more research in order to gather sufficient evidence to address the original objectives of the review, it is often the case that the further research needs to be targeted to specific issues that have arisen out of the review. For example, in a review of nicotine replacement that has been maintained regularly since 1994, it was established that different nicotine replacement delivery systems (such as nicotine gum, patch, and inhaler) were more effective than placebo. However, further primary research was required to establish the effectiveness of a combination of different delivery formats. This was subsequently done and has now been included in updates of the review.

Each time a review is updated some questions that previously required further research might have been answered, and new questions may arise. This is why updating reviews regularly is so important.

Categorising interventions according to evidence

There have been several attempts to try to categorise interventions according to whether there is sufficient evidence to provide clear guidelines for clinical practice and, if not, whether further evaluation is required. One of these was developed several years ago as part of a successful pilot project in Pregnancy and Childbirth. It has since been further developed by the Cochrane Pregnancy and Childbirth Group. Stop and read it now in Section 9.6 of the Reviewers’ Handbook.

Then try to classify the interventions in your review according to these categories.
Traps for the unwary

There are several common mistakes frequently made by reviewers:

i. Evidence of no effect versus no evidence of effect

Often, when there is inconclusive evidence we confuse ‘no evidence of effect’ with ‘evidence of no effect’. These two statements are not the same thing.

Often when there is insufficient evidence, a summary estimate will have wide confidence intervals that will include both the null and the possibility of quite big effects (either helpful or harmful). As an example, let’s consider a summary relative risk of 1.02 with 95% confidence intervals that extend from 0.74 to 1.40. In this situation, there is no (or insufficient) evidence of effect, We can’t say with any confidence that there is no effect, and it could be in either direction.

What we need is more evidence to increase our capability (or power) of detecting an important effect. If we wrongly concluded from the available evidence that the intervention had no effect when it really does, this would be referred to as Type II error (i.e. false negative, or missing an effect that is actually there). For this example it may be better to conclude along the lines of “the results of this review are consistent with the intervention being effective or not effective and further research is needed”.

As a second example take the identical relative risk, but with much more precise 95% confidence intervals from 0.98 to 1.06. This time we can say there is evidence that, if there is an effect, it is small and might be too small to be important to users of health care. Statements like this will depend on what size of effect is clinically important. A 2% change in risk might be clinically important in some situations, but not others.

Irrespective of where the boundaries of the 95% confidence interval lie, we need to remember there is a 5% possibility that the true summary effect lies outside of these boundaries, and our result may be incorrect. If we wanted to be more certain, we could calculate the 99% confidence intervals around the summary estimate.
ii. **Evidence-conclusion mismatches**  
Another error is a tendency to try and draw conclusions beyond the available evidence. It is very important that every statement you make in your review conclusions is backed up by your results, and all important results are dealt with in your conclusions. A trend towards an effect cannot be interpreted as a positive effect. It is important that you don’t become guilty of reporting bias by only including the outcomes of review for which there was a significant effect in the conclusion and abstract of your review. As a minimum, results for the important outcomes should be reported in the conclusions and abstract.

iii. **“Further research is needed”**  
A third common mistake is to be non-specific about the need for further research. A statement like “further research is needed” is not particularly helpful. The research that is needed should be described.

While it is hard to be prescriptive about how to draw inferences and conclusions from a review, try and practice by using these common mistakes as a form of check-list against which to try and improve conclusions from your own review. In doing a review you also need to remind yourself that no matter how well it is done, the review itself cannot make a health care decision, it can only assist in the process. Furthermore, once someone has used a review to make a decision there is no guarantee that they will derive the predicted benefits or avoid the potential harms. There never can be certainty in how individuals or groups of individuals will respond to an intervention. All we can hope to do is provide them with an objective summary of the best available evidence to inform their own treatment choices.

**Try it for yourself**

If you have access to The Cochrane Library, choose a review, read the results, without looking at the conclusions, and then try to write some conclusions. Then compare your conclusions with those of the authors. Are they the same? If not, why do you think they differ?
Module 19: Maintaining your review

In this module we discuss why it is important to keep your review up to date, some issues about the timing of updating and the process of updating your review.

Learning objectives

- Understand the rationale for updating reviews
- Be able to plan when and how to update a review
- Responding to comments and criticisms

Relevant sections of the Reviewers’ Handbook

- Section 10.10-10.11

Where does this go in your Cochrane review?

- This will depend on what components of your review need updating. It may be that there is little to change, but if you identify new evidence on your search, updating may involve altering many of the sections of your review.
Why update a review?

The main aim of a Cochrane review is to provide the ‘best available’ and most up-to-date evidence which can be used by consumers, clinicians and policy makers to bring about improvements in health care. Clinical practice guidelines are often based on Cochrane or other systematic reviews and increasingly funding agencies internationally are requiring systematic reviews to justify the need for new clinical trials (and other forms of research). As a result, in the last decade the number of systematic reviews published either electronically (through the Cochrane Collaboration) or in paper journals has greatly increased.

However, since evidence on a given subject is generally dynamic and continually evolving as new research becomes available, systematic reviews run the risk of becoming out of date and even misleading. To date, there is little empirical data available to allow informed decisions about what is a reasonable and efficient approach to updating evidence in Cochrane reviews, although some guidelines do exist.

We do know that incorporating additional studies as they become available may, on occasions, change the results of a review. However, there is a risk that updating evidence too frequently may result in introducing a different form of bias related to the slower publication (or even non-publication) of studies with negative and inconclusive results.

You may be able to gauge from doing the review whether the research relevant to your review is being published frequently, suggesting your review should be updated sooner, or if you should wait a while.

In addition, other developments in your research field may result in you needing to update your review. Some examples may be better tools or markers for characterising sub-groups, new treatment regimens and/or comparisons, or new outcome measures (or refined measurement methods of existing outcomes).

In contrast, there may be areas where new evidence or data are not emerging and a review prepared many years earlier is still current and valuable. In these cases updating your review may be unnecessary and wasteful.

Cochrane reviews are unique in that reviewers are committed to not only preparing systematic reviews of evidence but also to maintaining (and updating) these reviews on a regular basis. The current recommendation is that your review should be updated at least every two years, or else a commentary added to it to explain why the update has not been done.

© The Cochrane Collaboration 2002
How to update your review

Section 10.10 of the Reviewers’ Handbook discusses some of the issues involved in updating your review and you should read it now.

The process of updating a Cochrane review will depend on the structure of the original review and the amount of new data available since the previous publication of the review. Probably the key factor in updating your review is to periodically update your literature search to assess if there are any new studies relevant to your review question. Some review groups will be able to give you more help with this than others. Some intermittently send new trials to reviewers, and all groups will maintain their own trial register. At the least you should keep an eye on the Cochrane Central Register of Controlled Trials, published on The Cochrane Library, and your review group's trial register.

If new trials are identified, they need to be assessed for inclusion in your review using the same process (and form) as your original review. If they are to be included, data are extracted and entered into RevMan as in your original review. Don’t forget to update your results text, discussion, conclusions, and abstract accounting for the new evidence, and highlight any differences it may have made in the What’s New section of your review.

Sometimes updating your review results in new outcomes or comparisons being added. If this is the case, your original data extraction forms may need to be altered or extended, and piloted again. If this is the case, it might be necessary to go back to the original included trials and check that they did not include any information relevant to this new outcome or comparison.

Another possible update to a review is more methodological. As systematic review methodology advances, and the techniques available in RevMan are extended, it may be possible to perform additional analyses or strengthen existing ones not available in an earlier version of your review.

There is variability between review groups about when and if your updated review should again go through the process of editorial review. If your update involves no further analysis or change of result, it may not need to be refereed, however if there are new analyses and changes to conclusion, the same pre-publication process that was followed for your original review is likely to be repeated. You will need to check with your review group regarding their editorial policy for updated reviews.
If the update of your review leads to changes that you and your review group decide are important enough that anyone who read the review in the past should read it again, you should mark the review as a ‘substantive update’ when making the changes in RevMan.

Sometimes a search of the new evidence reveals there is no need to update your review. If this is the case, you should still update the date of the search strategy and indicate in the What’s New section of your review that a search has been done, resulting in no additions or changes to your review.

Comments and Criticisms

The ability to comment on a review allows users of Cochrane reviews to contribute to the updating process. Section 10.11 of the Reviewers’ Handbook discusses responding to criticisms and you should read it now.

A strength of Cochrane reviews is that they can be updated every quarter with publication of the new issue of The Cochrane Library. This means that you can alter or correct your review in response to valid criticism or comment, which is posted to your review group’s Criticism Editor via the ‘Feedback’ or ‘Comments/criticisms’ on The Cochrane Library, or via the Cochrane website. Sometimes these criticisms take the form of letting reviewers know of a new or missed trial to be assessed for inclusion in an update, and sometimes they point out a potential error in the review which can be corrected, or the criticism responded to as appropriate.

Summary

While time consuming and ongoing, the process of updating Cochrane reviews is one of the key factors setting them apart from almost all other reviews. You need to carefully plan how and when you are going to update your review, and who will take responsibility to ensure it happens.
Additional Module 1: Meta-analysis of continuous data

This module will discuss analysing data from continuous outcomes. We address issues of data extraction and meta-analysis of continuous data, and consider some of the problems reviewers frequently encounter.

Learning objectives

- Identify continuous data
- Identify scales and ordinal outcomes that can reasonably be treated as continuous
- Understand the concepts of mean and standard deviation
- Be aware that you can convert between standard deviations, standard errors and confidence intervals
- Understand the use of differences between means and standardized differences between means as measures of treatment effect
- Be aware of problems when dealing with skewed data and non-parametric summaries
- Understand the methods of combining continuous data in meta-analysis
- Be aware of issues in choosing between change scores and final values

Relevant sections of the Reviewers’ Handbook

- Sections 8.3.2, 8.4.2, 8.5.4, 8.9.3

Where does this go in a Cochrane review?

- Details of how continuous data are analyzed should be given in the Methods section. When presenting results of analyses involving continuous data, it should always be clear what method has been used, what the units of measurement are and, for scale based outcomes, a description of what the measurement scale means, noting whether the scale goes up or down with improving outcome. If assumptions are made when performing analyses, these should be assessed and if potentially important, addressed in the Discussion
What are continuous outcome data?

The strict definition of a continuous outcome is one measured on a scale that is continuously variable, i.e. for any two valid continuous measurements there is always one in between. Thus the number of hairs on one’s head is not strictly continuous since it can only be a whole number, but the length of a hair is continuous since it can have any number of decimal places. However, such subtle differences do not really matter in practice and we often treat outcomes that don’t quite meet this strict definition as if they were continuous. This includes outcomes that are:

- Numerical
- Made up of many ordered categories

This covers a lot of potential outcomes, including things like weight loss, dimensions (e.g. length of a hospital visit, area of scar, or volume of a tumour), concentrations, costs, and scores on psychometric scales. That doesn’t mean that we can automatically enter results for all such outcomes as continuous data in RevMan. We shall see that there are several issues to think about before we do so.

Can I analyse this outcome as a continuous outcome?

Here are two things to consider if you have an outcome that you think may be treated as a continuous variable.

Is an increase in 1 unit in one region equivalent to an increase of 1 unit in another region?

There is an implicit assumption that the answer is “yes” when we analyse continuous data in RevMan. Consider weight as a simple outcome. Is an increase in weight of 1kg from 50kg to 51kg the same as an increase from 80kg to 81kg? One might question whether these are clinically equivalent, but there is general consensus that we are talking about the same quantity.

Now consider a pain scale. Is a change from 5 (maximum pain) to 4 the same as a change from 1 to 0? It’s impossible to say: maybe the reduction at the more severe end of the scale is greater than a change from 1 to 0, maybe not.

Many health measurement scales are constructed by counting the number of positive responses to a set of questions or criteria. So, in terms of their psychometric properties they meet this criterion, although it may be difficult to argue that an increase in one unit has the same clinical meaning at all points on the scale.
As a general rule, short scales (those with not many categories) such as the pain scale tend to be unsuitable for the methods described in this module, and are usually analysed as dichotomous data, as discussed at the beginning of Module 11.

Is it reasonable to summarize a group of people using a mean and standard deviation?

Methods for meta-analysis of continuous data are derived assuming the data have a Normal distribution, and revolve around means and standard deviations. A mean is the ‘average’ (i.e. sum of the observations divided by the number of observations). The standard deviation is a measure of how variable the observations are around the mean. A small standard deviation indicates that the observations are all near the mean; a large standard deviation indicates that the observations vary a lot.

A key fact about means is that they can be sensitive to extreme values. For example, the mean of the numbers 1, 2 and 3 is 2, which is a fair single summary of the three numbers. The mean of 1, 2, 3 and 50 is 14, which seems a rather less satisfactory summary. When the mean is influenced by an extreme value, we have skew, and the observations have an asymmetrical distribution. When outcomes have an asymmetrical or skewed distribution, the mean (and hence the standard deviation) are not very useful ways to summarize the data. This may lead to analyses reaching spurious conclusions, especially when sample sizes are small. In practice it is not essential that the data have a perfect Normal distribution, but analyses may become misleading if the distribution of data is severely skewed.

Summary

We can treat outcomes as continuous data if they have an approximately symmetrical distribution and if realistic differences in the outcome can be interpreted similarly from different starting points. This may be the case for dimensions, counts of common events, and scales with many categories. It is less likely to be the case for costs, concentrations and counts of rare events (which all tend to be skewed) or short scales.

Skewed data are not bad data, they are just more difficult to analyse. We will look at ways you might make use of them later in the module.
What information do I need?

In order to perform meta-analyses using continuous data, we require three numbers from each treatment group. These are

- The sample size
- The mean
- The standard deviation

making six numbers in total for a two-group trial.

Sometimes we can readily extract these numbers from tables or the text of a report; sometimes we can’t. Often it is possible to derive them from other statistics. Standard deviations are the most likely statistic to be missing in a trial report, but the Handbook includes details of how standard deviations can be obtained from standard errors, confidence intervals, t-statistics and p-values. If you have any need to perform these conversions, you should read it now. If other statistics are reported, such as medians, ranges and non-parametric tests (for example, a ‘Mann-Whitney’ test), then this is an indication that the outcome may have a skewed distribution. In some cases trials report nothing that will allow you to obtain a mean or a standard deviation. If this is the case you should attempt to contact the trialist and obtain the missing data.

Measuring the effect of treatment

Meta-analyses involving continuous outcomes are based on comparing means. The basic way of comparing outcomes from two treatment groups is to look at the difference between the mean of each group. This difference between means, and its standard error, can be calculated from the six numbers listed above. Given this standard error we can award each trial a weight and use the inverse-variance method of meta-analysis to obtain a summary or combined mean difference and its confidence interval. Fixed effect and random effects methods for achieving this are available using RevMan, where you need only enter the six basic numbers from each study.

The meta-analysis of differences between means from different trials relies on the outcome being measured in the same units in every trial: we can’t combine a difference in mean weight loss in kilograms with a difference in mean weight loss in pounds. If you know the multiplication factor to convert from one scale to another (for example how many pounds there are in a kilogram), then you should directly convert all the data to the same units. However, we can’t combine two different psychometric scales even if they both measure depression as the multiplication factor is not known. A way around this is to compare standardized mean differences, rather than actual means.
The standardized mean difference is the difference in means divided by a standard deviation. This standard deviation is the pooled standard deviation of participants’ outcomes across the whole trial. Note that it is not the standard error of the difference in means (a common confusion).

The standardized mean difference has the important property that its value does not depend on the measurement scale. For example, consider a trial evaluating an intervention to increase birth weight. The mean birth weights in intervention and control groups were 2700g and 2600g with an average SD of 500g. The SMD will be

\[(2700 - 2600)/500 = 0.2\]

If the trial had measured birth weight in ounces, the results would be means of 95oz and 92oz with an average SD of 15oz. The SMD will be

\[(95-92)/15 = 0.2\]

- the same number from the analysis based on grammes.

So, if we have several trials assessing the same outcome, but using different scales, we use a standardised mean difference to convert all outcomes to a common scale, measured in units of standard deviations. But what is the interpretation of the standardized mean difference? That is a good question, and one that troubles statisticians and health care decision makers. What it actually measures is the number of standard deviations between the means. This quantity is not directly useable.

Let us consider the birth weight example. We can view the number of standard deviations’ difference as a ‘standardized’, or dimensionless, form of the actual findings. The value of 0.2 is the number of SDs by which the intervention changes outcome – if it is measured in grammes (where the SD is 500g) it changes by 0.2 x 500 = 100g, if it is measured in ounces (where the SD is 15oz) it changes by 0.2 x 15 = 3oz.

In practice, of course, we would not want to use the SMD method to analyse birth weight as we are able to convert between units of measurement using an exact conversion factor. However, we commonly have to use it when different measurement tools (e.g. scales) are used to measure the same clinical outcome.
For example, suppose a potential treatment for depression in the elderly achieves an average improvement of 2 points on the Hamilton Rating Scale for depression (HAMD). And suppose that the pooled standard deviation of HAMD scores is 8. Then the standardized mean difference is $2/8 = 0.25$. If a similar treatment effect was to be observed on an alternative depression scale, say the Geriatric Depression Scale (GDS) which has a standard deviation of 5 points, then a standardized mean difference of 0.25 is equivalent to an improvement of 1.25 points on the GDS.

We must be careful with using the standardized mean difference, however. First, we must be sure that the different measurement scales are indeed measuring the same clinical outcome. Second, problems can arise through the use of the pooled standard deviation for the standardizing. To illustrate the latter, let us return to our study with a 2-point improvement in HAMD score (pooled SD = 8). Imagine a second study in the same meta-analysis that also used the HAMD, but had more restrictive inclusion criteria. The tight inclusion criteria meant that participants were more similar to each other, and their pooled standard deviation in HAMD scores was only 5. Imagine further that the drug was equally effective in this study in that it also achieved a 2-point average improvement in HAMD score. The standardized mean difference for this study is $2/5 = 0.4$. Therefore the same effect of treatment gives a different standardized mean difference just because of the tighter inclusion criteria. This is an unfortunate implication of using standardized mean differences. Nevertheless, if studies do use different scales, there are usually few alternatives to using the standardized mean difference to combine results in a meta-analysis.

Finally, we should point out that in RevMan and commonly in The Cochrane Library, the mean difference method is referred to as ‘WMD’ (weighted mean difference) and the standardized mean difference method as ‘SMD’.

**On skew**

As we have said above, skewed data are not bad data. They are simply data that create a few complications because the distribution of likely measurements is asymmetrical and less convenient for statistical analysis. The main problem is that means and standard deviations are not very useful summaries of skewed data. Having said this, many investigators still report means and standard deviations even when data are skewed.
There is a handy trick to check the results of your included studies to see if they are skewed, even if they present a mean and standard deviation. This is often used in Cochrane reviews. The trick works if (i) you have a mean and standard deviation and (ii) there is an absolute minimum possible value for the outcome. Consider blood concentrations. These cannot be less than zero, so have an absolute minimum. Weight also has an absolute minimum possible value, as do scores on most psychometric scales. But weight loss and change-from-baseline measures can be negative and usually don’t have an absolute minimum, so the trick won’t work on these. Here’s the trick. Divide the mean by the standard deviation. If this is less than 2 then there is some indication of skewness. If it is less than 1 (i.e. the standard deviation is bigger than the mean) then there is almost certainly skewness.

There are a number of ways to deal with skewed data, but unfortunately few of them tend to be useful in meta-analysis. It is worth remembering that methods for meta-analysis (being based on t-tests) are quite robust to a little bit of skewness, especially if sample sizes are large.

The strategies that you might consider using with skewed data depend on the way the original trialists analyse and report results. The options you might encounter include:

(a) The trialists have ignored (or not noticed) the skewness and simply report means, standard deviations, and sample sizes.
This appears to be the simplest situation, as you can directly enter these numbers into RevMan. However, as we have noted, there is a possibility that these ‘improperly’ analysed data may be misleading. So, we will be unsure of the validity of our findings.

(b) The trialists have log-transformed the data for analysis, and report geometric means.
When a positively skewed distribution is log-transformed the skewness will be reduced. This is a recommended method of analysis for skewed data. In some fields, such as analysing antibody concentrations after vaccination, this approach is the norm. The data we wish to analyze in RevMan should also be on the log scale: the mean of the logged data will be the log of the geometric data. The standard deviation can be obtained from the confidence interval for the geometric mean, as described in section 8.4.2 of the Reviewers’ Handbook.
Non-parametric tests are a satisfactory alternative for analysing skewed data in trials. But as we cannot obtain means and standard deviations, we cannot include results of such analyses directly in a meta-analysis. This is, of course, unsatisfactory, especially when the inappropriately analysed results described in (a) can be used. One suggestion is that results of all studies are reported in a table in your review, regardless of the method of analysis used in the trials. This means that such data will not be lost from the review, and their results can be considered when drawing conclusions, even if they cannot be formally pooled.

Statistical methods do exist for combining p values from non-parametric tests, but not for estimating effects or detecting heterogeneity.

**Fixed effect and random effects for continuous data**

In Module 11 we covered differences between fixed effect and random effects meta-analysis of dichotomous data, and the issues are similar in continuous data. In a fixed effect inverse variance meta-analysis, the assumption is that all included studies are estimating one true or fixed effect and so variations between studies are due to random error. Studies are weighted according to the inverse of their variance, determined by the standard deviation. A potential problem therefore is that studies with restrictive eligibility criteria will have less variance (smaller standard deviation) and so will be given greater weight.

A random effects meta-analysis of continuous data assumes that all studies are estimating different effects (as they will all have differences to do with population, setting etc.) and these different effects are distributed according to a particular pattern. A random effects meta-analysis and fixed effect meta-analysis will therefore approximate each other in the absence of heterogeneity. Weight is attributed slightly differently when we use a random effects meta-analysis, however again studies with restrictive eligibility criteria will be given greater weight.

**Deciding on a change (from baseline)**

Another problem in meta-analysis of continuous data is change-from-baseline outcomes. As an example, consider the following results from the Hypertension Optimal Treatment (HOT) trial. This trial was published in *The Lancet* in 1998. Two of the treatment groups presented were attempts to reduce diastolic blood pressure in hypertensive participants to targets of less than 90 mmHg and less than 85 mmHg respectively.
Should the analysis focus on the final BP or the change-from-baseline? Does it matter?

We can work out that the average (mean) change-from-baseline is $85.2 - 105.4 = -20.2$ for the first group and $83.2 - 105.4 = -22.2$ for the second group. The difference between these is the same as the difference between the final means, that is $2.0$. As a general rule, the two estimates of treatment effect (i.e. differences between the two groups) should not be too different in properly conducted randomized trials where the two groups are similar at baseline. Indeed in this example they are identical because the trial is so large that the average baseline BPs were identical in the two groups. In most randomized trials, this won’t quite be the case. In some trials, especially small or poorly conducted trials, the difference can appear quite profound. The choice of whether to use final value or change score in your meta-analysis is a difficult one. There are two issues to consider.

First, there is a statistical argument to prefer change-from-baseline outcomes. This is closely related to the arguments in favour of crossover trials. Repeated measurements made on the same participants (at baseline and after treatment) tend to be correlated. This leads to smaller standard errors, and hence smaller confidence intervals, for the estimate of treatment effect when using change-from-baseline.

Second, there is a very real practical problem that can make the use of change-from-baseline very difficult. In order to use change-from-baseline outcomes in a meta-analysis we need their standard deviations. Notice in the table above that we have given standard deviations for the baseline measures and the final measures, but not the changes. What are the standard deviations for these changes? The answer is that we can’t possibly know from the information in the table. It could be that every participant in the $<90$ mmHg group reduced their BP by exactly 20.2 and every participant in the $<85$ mmHg group reduced theirs by exactly 22.2. In this case the standard deviations of the changes will both be 0. That would be very strong evidence of a difference between the groups. Or it could be that BP reductions in the two groups were highly variable (some increase, some decrease), with a large standard deviation, and the difference in means of 2.0 would then look quite unimportant.
So how do we find out the standard deviations of the changes? If you are lucky you will find them explicitly presented in the trial report. In fact, the report of the HOT trial does give them. The BP reductions are, 20.3 (SD 5.6) in the <90 mmHg group and 22.3 (SD 5.4) in the <85 mmHg group. Note that in this case the standard deviations of change are actually larger than the standard deviations of final values – so there is no benefit in this study in terms of power in using change scores.

Many studies however will not give you the standard deviation of the change, and often reviewers face the situation of several included studies, some presenting final value mean and standard deviation, and some reporting mean and standard deviation of the change. In this situation, you can follow one of two alternatives:

(a) You can derive the standard deviation of change and estimates of mean change

If initial and final mean values are given, the mean change in each group is the difference between these values. The standard deviation of the change depends on the correlation between initial and final values, which is unlikely to be reported. If the correlation can be obtained, or perhaps imputed, methods for calculating the standard deviation are given in Section 8 of the Reviewers’ Handbook. If data are imputed, the effect of uncertainty in the correlation should be investigated in a sensitivity analysis. If initial values aren’t given, this approach cannot be used.

(b) Combine final values and change scores in the same analysis

The data we quoted from the HOT trial demonstrated that both the difference in mean final values and the difference in mean changes both estimate the same treatment effect. Because of this we can combine trials reporting mean changes with trials reporting mean final values in the same meta-analysis. Often the change scores will be less variable than the final values – combining the data in a weighted mean difference analysis will give appropriate weights both to change scores and final values, as study weights are related to the standard deviations of the outcomes. So, in many circumstances it is not necessary to get very concerned about having a mixture of final values and change scores from your trials.
However, there are two points of concern. The first is the confusion you may cause in a reader by mixing change scores and final values in a review. For example, the final values in the data from the HOT trial were around 85mmHg, the change scores were around –20mmHg. It will be clearer to a reader if you present the change scores as one subgroup, and the final values as another subgroup in RevMan, and then combine the two in an overall analysis.

The second concern is that this approach will not work when you have different measurement scales, when you would want to use the standardised mean difference – this method cannot mix change and final values.

Summary

To perform meta-analysis of continuous data you will need to extract or calculate means and standard deviations from the reports of your included trials. This is often more difficult to do than extracting event rates for dichotomous outcomes as the information you need is not always present, or in a standard form. Some things to check are:

- Are these data symmetrically distributed or skewed? If skewed, you may need to present the results in the Additional Tables and not perform a meta-analysis.
- Is the presented measure of variation a standard deviation? It may be a standard error (check if it looks too small), or something else. If so, convert it before you enter it in RevMan.
- Do your included studies all measure outcome using the same scale? If not, you will need to convert to standard units (if you can) or use a standardised mean difference.
- Should you use a random effects or fixed effect meta-analysis? Whether this makes a difference will depend on the amount of heterogeneity present.
- Should you enter final value or change scores? This will be partly determined by what is reported in your included studies, and it is possible to mix the two in the same analysis. If you have to impute a standard deviation, you should perform a sensitivity analysis and see how it affects your results. If they change, draw your conclusions with care!
Additional Module 2: Issues related to the unit of analysis

In this module we look at issues related to the unit of analysis, including the incorporation of crossover trials and cluster randomized trials into a Cochrane review, or meta-analysis. The main aim of the module is to help you to recognize when unit of analysis errors can occur, rather than to learn how to deal with every eventuality. Many of the methods for dealing with unit of analysis errors are tricky, and if you do have to deal with these issues, you should ensure that you have access to appropriate methodological expertise.

Learning objectives

- Identify a variety of sources of unit-of-analysis errors
- Identify a crossover trial and understand when it is an appropriate design to use
- Understand that the correct analysis of a crossover trial is a paired analysis, in which within-patient differences are the focus
- Identify a cluster randomized trial
- Understand that the appropriate analysis of a cluster randomized trial should recognize that individuals are members of clusters
- Identify repeated outcome measurements, and be aware of strategies for incorporating them into meta-analyses
- Be aware that the inverse variance method for meta-analysis can be used for crossover, cluster randomized trials and trials with other designs

Relevant sections of the Reviewers’ Handbook

- Sections 8.12.1, 8.12.2

Where does this go in a Cochrane review?

- Think carefully about possible unit of analysis issues when writing your protocol. Methods for dealing with them should be described in the Methods section

© The Cochrane Collaboration 2002
For starters

The simplest trial design to analyse is called a parallel group design, where participants are randomized to receive one of two interventions, and outcome measurements are made at one time point at the end of the treatment period.

In this module we will look at some other types of trial design, which depart from the simple trial design above. These departures lead, among other problems, to mismatches between numbers of randomized units and the number of included participants (such as twenty participants but forty eyes) or outcome observations. These can lead to confusion, and errors. We call these errors ‘unit of analysis errors’, and the issues discussed in this module are generally known as ‘unit of analysis issues’.

To determine which issues are relevant to your review, consider whether you have, or anticipate, any trials of the following sort:

- Crossover trials, in which patients are randomized to a sequence of interventions. In crossover trials, all patients might receive both (or more, if there are more than two) interventions.
- Cluster randomized trials, in which groups of participants are randomized.
- ‘Body-part randomization’ designs, in which different parts of a person are randomized to be treated differently. For example, a split-mouth design in dentistry might randomize the left and right halves of the mouth to different topical applications. Or, a physiotherapy trial might randomize two limbs of the same individual to different types of exercise.

Four other related topics can occur with standard parallel group designs:

- ‘Body part analyses’ of standard trial designs, in which a person is randomized to an intervention, but outcomes are measured for several different body parts. For example, in dentistry, a person might be randomized to use a particular toothpaste, but outcomes are collected on every tooth (i.e., cavity: yes/no for each tooth). Or eczema severity may be measured separately on both arms, even though they both received the same treatment.
- Trials with three or more treatment groups. An experimental intervention may be compared with both a standard intervention and with a placebo. Alternatively, two or more experimental interventions might be tested against a standard intervention.
- Trials with the same outcome assessed at several time points.
• Trials where an event outcome may occur more than once in each participant. A common example is the number of adverse events – participants may have more than one adverse event. Another example, in sub-fertility studies, a woman might have more than one pregnancy during a period of treatment. Or, in cardiovascular studies, a patient might experience more than one stroke.

Crossover trials

In a crossover trial, participants are randomized to a sequence of treatments. We shall consider the simplest design, in which there are two treatments. As an example, a trial was conducted in patients with asthma to compare laser acupuncture with a sham (‘placebo’) procedure. Patients were randomized either to acupuncture for five weeks (first period) then sham for five weeks (second period), or sham then acupuncture. One of the outcomes was a symptom score.

This gives us data for each patient both when they were on acupuncture, and when they were on sham treatment. We can compare these for each patient to assess the effect of acupuncture within each patient. This is a very efficient approach to analysis, because when making the comparison between treatment and control we do not have to allow for all the variation that occurs between patients, which we have to deal with in a parallel group trial.

In practice this means that, for the same number of participants, a crossover design is likely to be more powerful. However, crossover trials are not always appropriate. The most important consideration is whether the patients start the second period in a similar state to how they started the first period. If the characteristics of the participant have changed in some way by the time the second period starts, then the comparison of treatments is not fair, and there will be within patient variation. Some questions you should ask yourself are as follows.

• Is the condition of the patients chronic and stable? Crossover trials are common for conditions such as asthma, osteoarthritis and epilepsy. Crossover trials may not be appropriate for progressive diseases or acute conditions that will worsen or improve by the second period. If patients vary from one period to the other there is said to be a ‘period effect’.

• Does the intervention provide temporary relief, and not permanent change? For example, surgical interventions are unsuitable for crossover trials if the surgery permanently alters the condition.
• Can the outcome be repeated in the second period if it occurs in the first? For example, crossover trials are certainly unsuitable when the primary outcome is mortality, or pregnancy in infertility studies.

• Might the effect of the first intervention last into the second treatment period? In the acupuncture trial, a three-week ‘washout’ period was built into the trial between the two treatment periods. This is a common method to minimize ‘carry-over’ effects and ensure the participant is in the same state at the beginning of each period, though it is not always sufficient.

• Does the trial go on long enough for drugs to have effects and outcomes to occur? For example, a trial in epilepsy with the outcome of number of fits, needs to observe the patients for long enough to make sure that we haven’t, by chance, just picked a particularly good or bad time in their illness.

If you have crossover trials in your review, you will need to decide on the following points:

• Should I include crossover trials in the review?
• Should I include crossover trials in any meta-analyses?
• Should I combine crossover trials with other types of trials?
• How should I include crossover trials in a meta-analysis?

The brief answer to the first three questions is:

• There is no reason to exclude crossover trials solely because they are crossover trials.

Of course, there may be other reasons why crossover trials might be excluded, for example, if they have treatment periods that are too short or do not have an appropriate wash out period. If you are anticipating cross over studies in your review you will need to set these inclusion criteria in your protocol.

Analysis of crossover trials should exploit the fact that each patient acts as his or her own control. The appropriate analysis is a ‘paired’ analysis. See Section 8.12.2 of the Reviewers’ Handbook for a more extensive discussion of how you can include crossover trials in a Cochrane review or a meta-analysis. Properly analysed crossover trials may be incorporated into meta-analyses using the generic inverse-variance method, available in RevMan 4.2.
Cluster randomized trials

A cluster randomized trial is a trial in which individuals are randomized in groups (i.e. the group is randomized, not the individual). For example, in a rural area with an endemic disease, we might randomise whole villages to have the intervention or not, rather than individual people. We then say that the village is the unit of randomization. In other situations, general practices, hospitals, families or school classrooms may be randomized. Reasons for performing cluster randomized trials vary. Sometimes the intervention can only be administered to the group, for example an addition to the water supply or a public education campaign; sometimes the motivation is to avoid contamination (all participants in the trial are affected by the intervention, even if it is only given directly to some of them); sometimes the design is simply more convenient or economical.

A simple approach to dealing with cluster randomized trials is to assess outcomes only at the level of the group thereby keeping the unit of analysis the same as the unit of randomisation. One might measure a dichotomous outcome of whether the practice/classroom/village was a ‘success’ or a ‘failure’, or a continuous outcome of the percentage of individuals in the group who benefited. In this way, we obtain one outcome measurement from each randomized unit, and the analysis can proceed as if the groups were individuals – that is, using the techniques described elsewhere in Modules 11 and 12. It will probably strike you that there are limitations to this approach. First, cluster randomized trials are likely to randomize fewer groups than most simple trials randomize individuals. For example, a trial might randomise ten villages with a total of 15,000 inhabitants. Analysing by village, we would end up with only ten observations. So, we would end up with fewer data (and hence less statistical power) than a simple trial involving substantially fewer participants analysed as individuals. Second, not all groups will be the same size, and we would give the same weight to a village of 10,000 inhabitants as a village of 150 inhabitants.

An alternative possibility is to ignore the groupings and compare all the individuals in intervention groups with all those in control groups. This has been a common approach both to analysing individual cluster randomized trials and to representing them in systematic reviews. But it is problematic because it ignores the fact that individuals within a particular group tend to be more similar to each other than to members of other groups. Such analyses can spuriously overestimate the significance of differences, and should be avoided.

© The Cochrane Collaboration 2002
Think of the example where we randomize villages. Residents in one village may share the same climate, nutrition, education and health care, which make their outcomes more similar to each other than to residents in a different village. We use the term *intra-cluster correlation coefficient* to describe the extent to which two members of one cluster are more similar than two people from different clusters.

There are statistical techniques for appropriate analyses of cluster randomized trials. We can recognize that clusters are made up of individuals and that there may be more individuals in one cluster than in another. The intra-cluster correlation coefficient plays an important role in these techniques. Further details can be found in Section 8.12.1 of the *Reviewers’ Handbook*. Cluster randomized trials may be incorporated into meta-analyses using the generic inverse-variance method.

‘Body-part randomization’ and ‘body-part analysis’

We use the terms ‘body-part randomization’ and ‘body-part analysis’ to distinguish between two different types of study design involving parts of the body.

By ‘body-part randomization’ designs we mean those in which similar body parts are randomized to different interventions. For example, a person’s arms may be randomized so that each gets a different cream applied. Other examples include studies for eyes and teeth. One issue to think about is contamination – could a treatment in one part affect what happens in another? If so this raises the question of whether such designs are appropriate in the first place. You may notice a similarity between this design and the crossover design described above. In both designs, each person receives both interventions. In fact, the analysis of a body-part randomization trial should proceed in the same way as the analysis of a crossover trial, involving a paired analysis.

By ‘body-part analysis’ we mean a particular approach to the analysis of a standard parallel group design trial. Suppose a (whole) individual is randomized to receive a surgical intervention for cataracts. If he or she has cataracts in both eyes you might collect outcomes for the vision in each eye separately, and you might want the patient to therefore contribute two measurements to the data analysis. This is rather like a cluster randomized trial, where the person is the cluster and the eyes are the individuals within the cluster. In fact, the analysis of these types of trials should proceed in the same way as the analysis of a cluster randomized trial. However, if there are only one or two measurements for each individual, it may be preferable and simpler to select only one measurement per individual.
More than two treatment groups

Many trials are designed to compare more than two treatments. However, there may be two or more experimental interventions, for example a drug at different doses, or variations on a counseling intervention. Alternatively, there may be more than one control group, perhaps an established treatment and a placebo.

The most common problem reviewers experience is trials with several experimental groups. If you are comparing each of the treatment groups with placebo in two separate meta-analyses (i.e. as two separate comparisons in RevMan), then the study can be treated as two separate trials (intervention 1 versus placebo and intervention 2 versus placebo). If however, you are putting all three arms of the study into the same meta-analysis it can be tempting to, for example, enter the data as if it were two trials, one comparing high dose with placebo and one comparing low dose with placebo. But when you then pool these results in a meta-analysis you will be counting the placebo patients twice. This approach should not be followed.

There are two main approaches to dealing with trials like this. The first is to break up the control group into several parts, so that the total numbers add up to the original size of the group. The second is to group together all the experimental groups and compare them collectively with the control group. There’s no single right answer, since both approaches have advantages and disadvantages.

Repeated measurements

Repeated measurements refer to measurements made at different points in time. Thus, a dietary trial may report weight loss at 4 weeks, at 8 weeks and at 6 months. We can’t include all of these in the same meta-analysis since again we’d be counting the same person more than once and we’d have a unit-of-analysis error. The problem of repeated measurements can partly be overcome by specifying in the protocol which time-points are of interest, and discarding the rest. It may be helpful to classify outcomes as ‘short-term’, ‘medium-term’ and ‘long-term’, and to perform separate meta-analyses for these different outcomes including only one time-point from each trial in each analysis. Alternatively, you may only be interested in a single time-point, or the longest available follow-up. The disadvantages of opting for longest available follow-up are that more patients may have been lost to follow-up, and it may vary considerably between studies introducing heterogeneity.
Who is having the events?

Our final topic here can cause confusion. Consider an outcome that is an event: a stroke, for example. What if the event can happen more than once? A person can have more than one stroke during a period of follow-up. We have to be very careful that we know exactly the nature of the data being reported. For an outcome to be considered as a dichotomous outcome there must be an all-or-nothing distinction between the “yes” and “no” classifications. Thus, if we count that 245 people had at least one stroke and 765 people had no strokes, we have dichotomous data. If, on the other hand, we just know that there were 312 strokes among 1010 people then we may not know who had them. If some people could have had more than one of the strokes we do not have dichotomous data. We cannot treat the data as dichotomous data when entering them into RevMan, unless we can distinguish the number of people having events from the number of events.

The same difficulty occurs when we are analysing adverse events. Suppose we know that there were 120 adverse events among 250 participants in a trial. It may be that 20 people had two events each, accounting for 40 of the 120 events. If the other 80 adverse events occurred as one in each of 80 people, that would mean a total of 100 people had the 120 events. Unless we are told about how often people had different numbers of events, we cannot calculate the number of participants who had any adverse event.

So what can we do? If you think this poses a problem, consult Module 14 or section ‘Counts and rates’ in sections 8.3, 8.4, and 8.5 of the Reviewers’ Handbook.