

Module 13: Diversity and heterogeneity

CAUTION!! This module is not up to date.

Authors should refer to ‘Section 9.5: Heterogeneity’ in the *Cochrane Handbook for Systematic Reviews of Interventions* for current information.

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 *Cochrane Handbook for Systematic Reviews of Interventions*

- Section 9.5: Heterogeneity

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

There will be many differences between the studies included in your review

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'.

Heterogeneity can result from clinical or methodological diversity

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.



Activity: list the important sources of clinical and methodological diversity in your review

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.

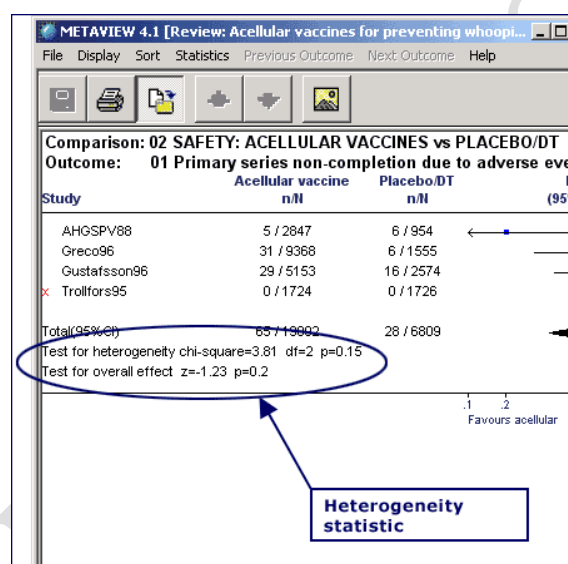
Identifying statistical heterogeneity

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

To identify heterogeneity you can visually assess the forest plot or perform a statistical test

- 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 χ^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.

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 $p = 0.05$.

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 are 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

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.

How to choose between fixed and random effects meta-analyses

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

You can investigate heterogeneity with sub-group analyses or meta-regression

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.