

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 *Cochrane Handbook for Systematic Reviews of Interventions*

- Section 9.2.3: Effect measures for continuous outcomes
- Section 9.2.5: Effect measures for counts and rates

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



*Activity:
From a list of outcomes you intend to address in your review, identify which are, or could be, reported as continuous outcomes*

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.

*Recognising
continuous data*

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.



Read Section 7.7.3.2 and 7.7.3.3 of the Cochrane Handbook for Systematic Reviews of Interventions if you need to calculate a SD from standard error, *p*-value or confidence intervals

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.

Combining continuous outcomes

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.

Combining outcomes measured on different scales

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.

*Interpreting
standardised
mean difference*

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 $0.2 \times 500 = 100\text{g}$, if it is measured in ounces (where the SD is 15oz) it changes by $0.2 \times 15 = 3\text{oz}$.

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 *The Cochrane Library*, the mean difference method is referred to as 'MD' 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.

Detecting skew

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.

Options for dealing with skewed data

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 7.7.3.2 of the *Cochrane Handbook for Systematic Reviews of Interventions*.

(c) The trialists use non-parametric tests (e.g. Mann-Whitney) and

describe averages using medians.

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)

*Change from
baseline data*

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.

	Baseline diastolic BP		Final diastolic BP	
	Mean	(SD)	Mean	(SD)
<90 mmHg (n=6264)	105.4	(3.4)	85.2	(5.1)
<85 mmHg (n=6264)	105.4	(3.4)	83.2	(4.8)

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 16.1.3 of the *Cochrane Handbook for Systematic Reviews of Interventions*. 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 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!