

## Department of Health Sciences

### M.Sc. in Evidence Based Practice, M.Sc. in Health Services Research

#### Meta-analysis: methods for quantitative data synthesis

##### What is a meta-analysis?

Meta-analysis is a statistical technique, or set of statistical techniques, for summarising the results of several studies into a single estimate. Many systematic reviews include a meta-analysis, but not all. Meta-analysis takes data from several different studies and produces a single estimate of the effect, usually of a treatment or risk factor. We improve the precision of an estimate by making use of all available data.

The Greek root 'meta' means 'with', 'along', 'after', or 'later', so here we have an analysis after the original analysis has been done. Boring pedants think that 'metanalysis' would have been a better word, and more euphonious, but we boring pedants can't have everything.

For us to do a meta-analysis, we must have more than one study which has estimated the effect of an intervention or of a risk factor. The participants, interventions or risk factors, and settings in which the studies were carried out need to be sufficiently similar for us to say that there is something in common for us to investigate. We would not do a meta-analysis of two studies, one of which was in adults and the other in children, for example. We must make a judgement that the studies do not differ in ways which are likely to affect the outcome substantially. We need outcome variables in the different studies which we can somehow get in to a common format, so that they can be combined. Finally, the necessary data must be available. If we have only published papers, we need to get estimates of both the effect and its standard error, for example. We discuss this further below.

A meta-analysis consists of three main parts:

- a pooled estimate and confidence interval for the treatment effect after combining all the studies,
- a test for whether the treatment or risk factor effect is statistically significant or not (i.e. does the effect differ from no effect more than would be expected by chance?),
- a test for heterogeneity of the effect on outcome between the included studies (i.e. does the effect vary across the studies more than would be expected by chance?).

Figure 1. Meta-analysis of the association between migraine and ischaemic stroke (Etminan *et al.*, 2005)

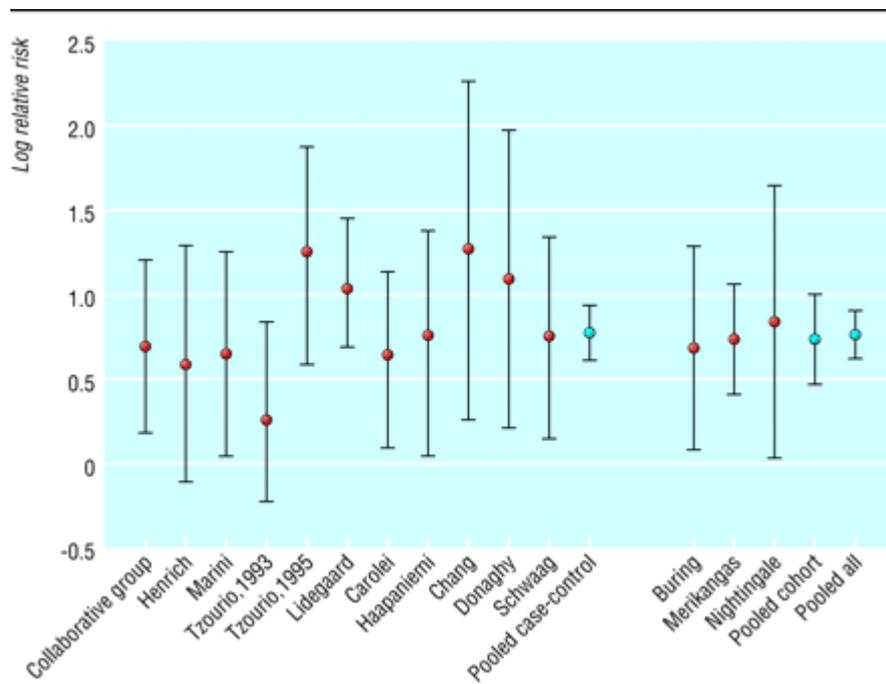
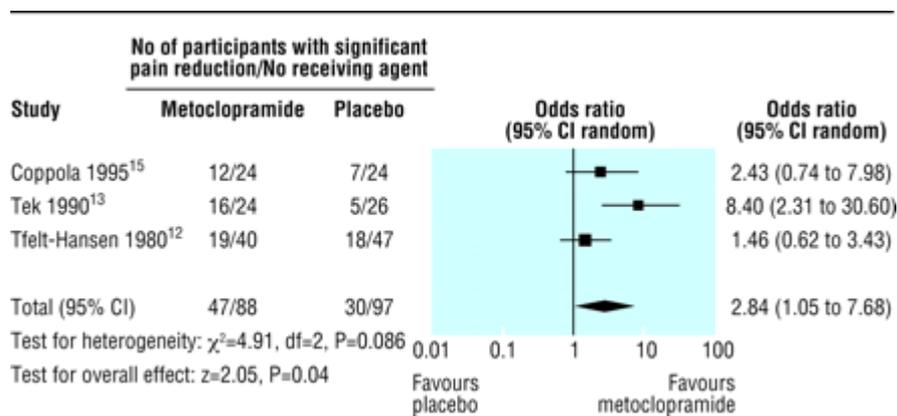


Figure 2. Graphical representation of a meta-analysis of metoclopramide compared with placebo in reducing pain from acute migraine (Colman *et al.*, 2004)



For example, Figure 1 shows a graphical representation of the results of a meta-analysis of the association between migraine and ischaemic stroke. In this graph, which is called a **forest plot**, the red circles represent the logarithms of the relative risks for the individual studies and the vertical lines their confidence intervals. It is called a forest plot because the lines are thought to resemble trees in a forest. There are three pooled or meta-analysis estimates: one for all the studies combined, at the extreme right of the picture, and one each for the case-control and the cohort studies, shown as blue or turquoise dots. The pooled estimates have much narrower confidence intervals than any of the individual studies and are therefore much more precise estimates than any one study can give. In this case the study difference is shown as the log of the relative risk. The value for no difference in stroke incidence between migraine sufferers and non-sufferers is therefore zero, which is well outside the confidence interval for the pooled estimates, showing good evidence that migraine is a risk factor for stroke.

Figure 1 is a rather old-fashioned forest plot. The studies are arranged horizontally, with the outcome variable on the vertical axis in the conventional way for statistical graphs. This makes it difficult to put in the study labels, which are too big to go in the usual way and have been slanted to make them legible. The studies with wide confidence intervals are much more visible than those with narrow intervals and look the most important, which is quite wrong. The three meta-analysis estimates look quite unimportant by comparison. These are distinguished by colour, but otherwise look like the other studies. The colour choice is not very good for a colour blind reader and would disappear when printed on a monochromatic printer.

Figure 2 shows the results of a meta-analysis of metoclopramide compared with placebo in reducing pain from acute migraine. This is a combination of three clinical trials. This graph, which is also called a forest plot, has been rotated so that the outcome variable is shown along the horizontal axis and the studies are arranged vertically. The squares represent the odds ratios for the three individual studies and the horizontal lines their confidence intervals. This orientation makes it much easier to label the studies and also to include other information. The size of the squares can represent the amount of information which the study contributes. If they are not all the same size, their area may be proportional to the samples size, the standard error of the estimate, or the variance of the estimate. This means that larger studies appear more important than smaller studies, as they are. On the right hand side of Figure 1 are the individual trial estimates and the combined meta-analysis estimate in numerical form. On the left hand side are the raw data from the three studies. The diamond or lozenge shape represents the common meta-analysis estimate, making it much easier to distinguish from the individual study estimates than in Figure 1. The widest point is the estimate itself and the width of the diamond is the confidence interval. The choice of the diamond is now widely accepted, but other point symbols may be used for the individual study estimates.

The horizontal scale in Figure 2 is logarithmic, labelling the scale with the numerical odds ratio but rather than showing the logarithm itself. We discuss this further below. A vertical line is shown at 1.0, the odds ratio for no effect, making it easy to see whether this is included in any of the confidence intervals.

At the bottom of Figure 2 are two tests of significance. The first is for heterogeneity, which we deal with below. The second is for the overall effect, testing the null hypothesis that there is no difference between the two treatments. In this case the

difference is significant. Individually, only one of the three trials gave a significant improvement and pooling the data from all three enables us to draw a more secure conclusion about the existence of a treatment effect and its magnitude.

Meta-analysis can be done whenever we have more than one study addressing the same issue. The sort of subjects addressed in meta-analysis include:

- interventions: usually randomised trials to give treatment effect,
- epidemiological: usually case-control and cohort studies to give relative risk,
- diagnostic: combined estimates of sensitivity, specificity, positive predictive value.

In this lecture I shall concentrate on studies which compare two groups, but the principles are the same for other types of estimate.

### Using summary statistics

Most meta-analysis is done using the summary statistics representing the effect and its standard error in each study. We use the estimates of treatment effect for each trial and obtain the common estimate of the effect by averaging the individual study effects. We do not use a simple average of the effect estimates, because this would treat all the studies as if they were of equal value. Some studies have more information than others, e.g. are larger. We weight the trials before we average them.

To get a **weighted average** we must define weights which reflect the importance of the trial. The usual weight is

$$\text{weight} = \frac{1/\text{variance of trial estimate}}{1/\text{standard error squared.}}$$

We multiply each trial difference by its weight and add, then divide by sum of weights. If we give the trials equal weight, setting all the weights equal to one, we get the ordinary average.

If a study estimate has high variance, this means that the study estimate contains a low amount of information and the study receives low weight in the calculation of the common estimate. If a study estimate has low variance, the study estimate contains a high amount of information and the study has high weight in the common estimate.

We can summarise the general framework for pooling results of studies as follows:

- the pooled estimate is a summary measure of the results of the included studies,
- the pooled estimate is a weighted combination of the results from the individual studies,
- usually, the weight given to each trial is the inverse of the variance of the summary measure from each of the individual studies,
- therefore, more precise estimates from larger trials with more events are given more weight,
- then find 95% confidence interval and P value for the pooled difference.

There are several different ways to produce the pooled estimate:

- inverse-variance weighting, as described above,
- Mantel-Haenszel method,
- Peto method,
- DerSimonian and Laird method.

Slightly different solutions to the same problem.

## **Heterogeneity**

Studies differ in terms of

- Patients
- Interventions
- Outcome definitions
- Design

These produce **clinical heterogeneity**, meaning that the clinical question addressed by these studies is not the same for all of them. We have to consider whether we should be trying to combine them, or whether they differ too much for this to be a sensible thing to do. We detect clinical heterogeneity from the descriptions of the trial populations, treatments, and outcome measurements..

We may also have variation between studies in the true treatment effects or risk ratios, either in magnitude or direction. If this is greater than the variation between individual subjects would lead us to expect, we call this **statistical heterogeneity**. We detect statistical heterogeneity on purely statistical grounds, using the study data.

Statistical heterogeneity may be caused by clinical differences between studies, i.e. by clinical heterogeneity, by methodological differences, or by unknown characteristics of the studies or study populations. Even if studies are clinically homogeneous there may be statistical heterogeneity.

To identify statistical heterogeneity, we can test the null hypothesis that the studies all have the same treatment (or other) effect in the population. The test looks at the differences between observed treatment effects for the trials and the pooled treatment effect estimate. We square these differences, divide each by variance of the study effect, and then sum them. This gives a chi-squared test with degrees of freedom = number of studies – 1.

In the metoclopramide trials in Figure 2, the test for heterogeneity gives  $\chi^2 = 4.91$ ,  $df = 2$ ,  $P=0.086$ .

If there is significant heterogeneity, then we have evidence that there are differences between the studies. It may therefore be invalid to pool the results and generate a single summary result. We should try to describe the variation between the studies and investigate possible sources of heterogeneity. We should not just ignore it, but try to account for the heterogeneity in some way. If we can explain the heterogeneity, we may be able to produce a final estimate of the effect which adjusts for it. If not, we can also carry out meta-analysis which allows for heterogeneity, called **random effects** analyses. We shall discuss these methods in more detail in the next lecture.

If the heterogeneity not significant, we have little or no statistical evidence for differences between studies. However, the test for heterogeneity has low power. The number of studies is usually low and the test may fail to detect heterogeneity as statistically significant when it exists. As with any significance test, we cannot interpret a not significant result as evidence of homogeneity. To compensate for the low power of the test some authors accept a larger P value as being significant, often using  $P < 0.1$  rather than  $P < 0.05$ .

## Types of outcome measure

The choice of the measure of treatment or other effect depends on the type of outcome variable used in the study. These might be:

- **dichotomous**, such as dead/alive, success/failure, yes/no, we use a relative risk or risk ratio (RR), odds ratio (OR), absolute risk difference (ARD),
- **continuous**, e.g. weight loss, blood pressure, we use the mean difference (MD), or standardised mean difference (SMD),
- **time-to-event or survival time**, e.g. time to death, time to recurrence, time to healing, we use the hazard ratio,
- **ordinal** (very rare), an outcome categorised with an ordering to the categories, e.g. mild/moderate/severe, score on a scale, we may dichotomise, treat as continuous, or use advanced methods specially developed for this type of data.

## Dichotomous outcome variables

For a dichotomous outcome measure we present the treatment effect as a relative risk or risk ratio (RR), odds ratio (OR), or absolute risk difference (ARD). Both relative risk and odds ratio are analysed and presented using logarithmic scales. Why is this? For example, in a trial of two treatments for ulcer healing (Fletcher *et al.*, 1997) two groups were compared

elastic bandage: 31 healed out of 49 patients  
inelastic bandage: 26 healed out of 52 patients.

The risk ratio can be presented in two ways:

$$RR = (31/49)/(26/52) = 1.27 \text{ (elastic over inelastic)}$$

$$RR = (26/52)/(31/49) = 0.79 \text{ (inelastic over elastic)}$$

We want a scale where 1.27 and 0.79 are equivalent. They should be equally far from 1.0, the null hypothesis value. We use the logarithm of the risk ratio:

$$\log_{10}(1.273) = 0.102, \log_{10}(0.790) = -0.102$$

$$\log_{10}(1) = 0 \text{ (null hypothesis value)}$$

If we invert a ratio, we change the sign of the logarithm. For example,  $\log_{10}(1/2) = -0.301$  and  $\log_{10}(2) = +0.301$ . The no difference value for a ratio is 1.00, and the log of this is zero. It is also easy to calculate standard errors and confidence intervals for the log of the ratio.

Results are often shown on a logarithmic scale, i.e. one where the scale intervals are logarithms, but the numbers given are the actual ratios. Figure 3 shows an example. The distance on the horizontal scale between 0.1 and 1 is the same as the distance between 1 and 10, because the ratio  $1/0.1$  is the same as the ratio  $10/1$ .

Figure 3. Interventions for the prevention of falls in older adults, pooled risk ratio of participants who fell at least once (Chang *et al.*, 2004)

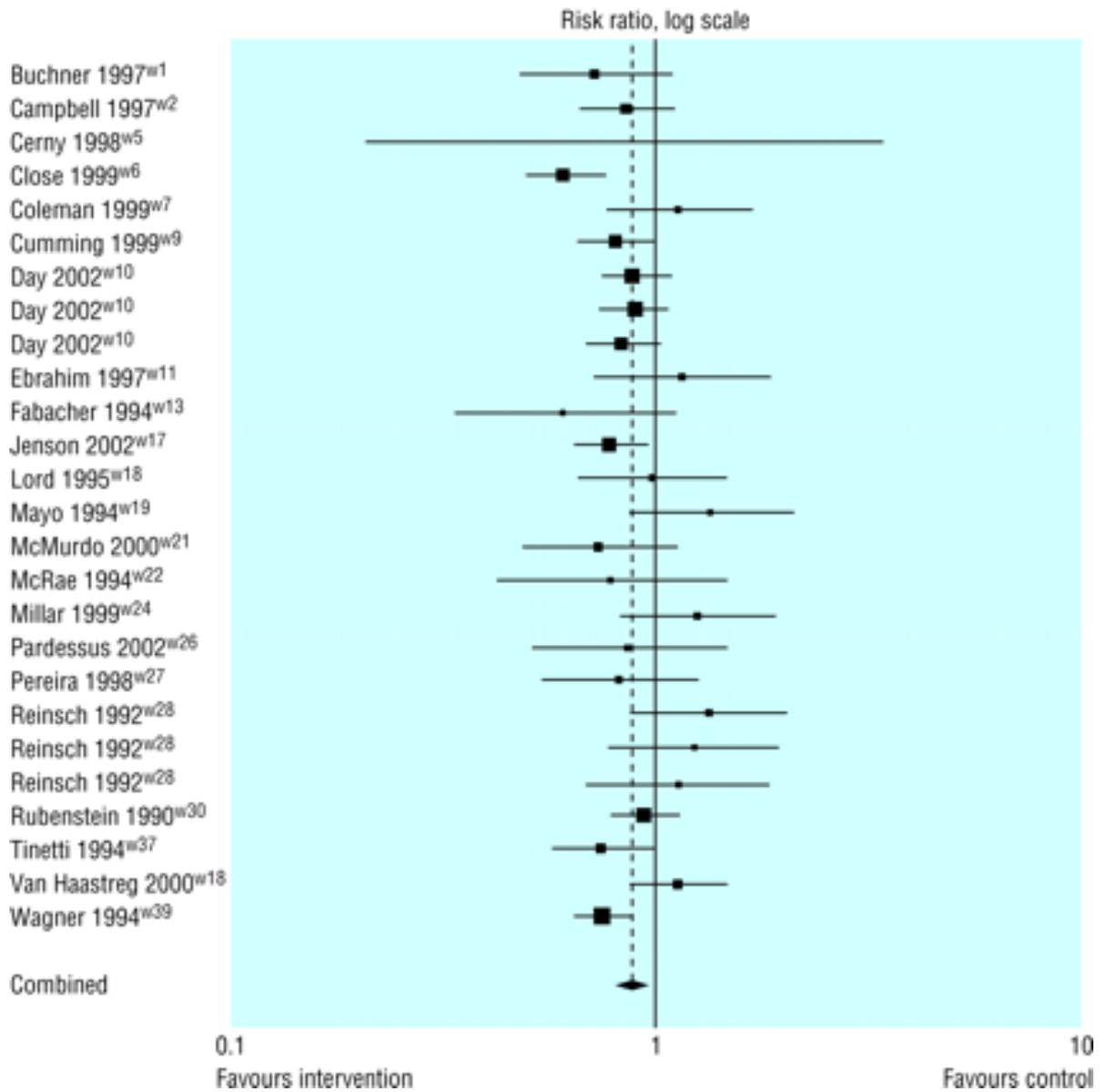


Figure 4. Rates of Caesarean section in trials of nulliparous women receiving epidural analgesia or parenteral opioids Liu EHC and Sia ATH. (2004)

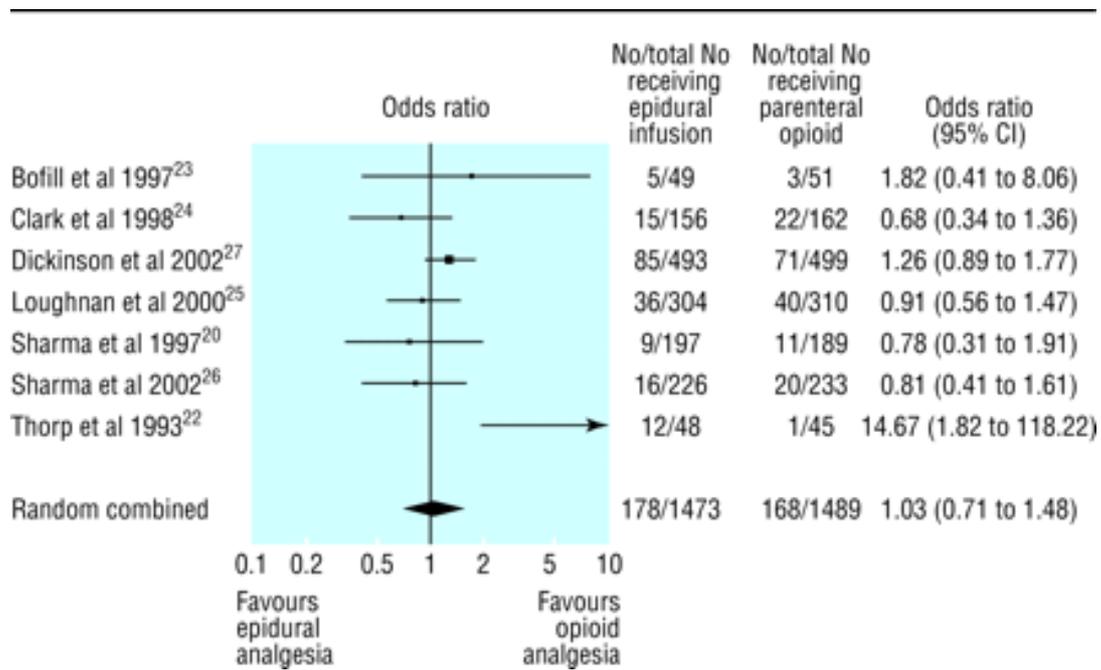
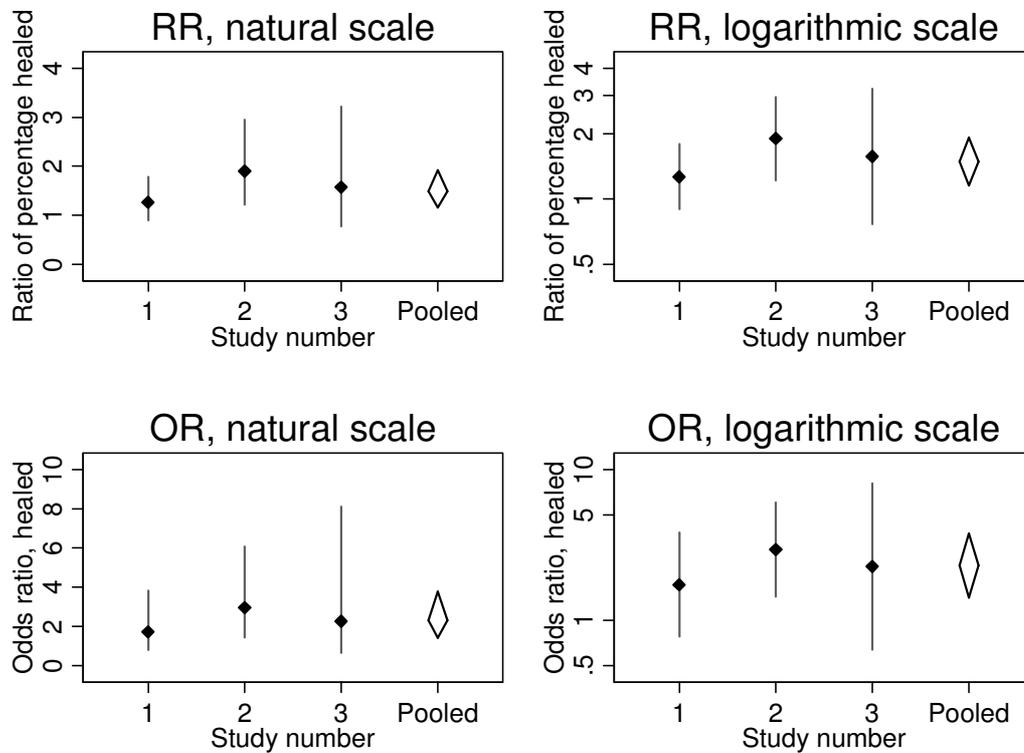


Figure 5. Forest plots for risk ratio and odds ratio on the natural and logarithmic scales (data of Fletcher *et al.*, 1997)



For both relative risk and odds ratio we find the standard error of the log ratio rather than the ratio. The log ratio also tends to have a Normal distribution. On the logarithmic scale, confidence intervals are symmetrical. Figure 4 shows a forest plot using odds ratios rather than relative risks. One small trial has such a large odds ratio with a very wide interval and is off the scale, its presence merely indicated by an arrow. If they had wanted to include this confidence interval the rest of the information would have squeezed into a very narrow area of the graph, making it difficult to read.

Figure 5 shows forest plots on the natural and logarithmic scales for risk ratio and odds ratio for the venous ulcer trial data. The confidence intervals are asymmetrical on the natural scales, symmetrical on the logarithmic scales.

## **Continuous outcome variables**

There are two main measures of treatment or other effect for a continuous outcome variable, weighted mean difference and standardised mean difference.

The weighted mean difference takes the difference in effect, measured in the units of the original variable, and weights them by the variance of the estimate. It is in the same units as the observations, which makes it easy to interpret. It is useful when the outcome is always the same measurement. These are usually physical measurements. For example, Figure 6 shows the results of a meta-analysis where the outcome variable is blood pressure measured in mm Hg.

The **standardised mean difference** is found by turning the individual study effect estimates into standard deviation units. We divide the estimate by the standard deviation of the measurement, either using the common standard deviation within groups for the study, as found in a two-sample t test, or the standard deviation in the control group. This is also called the effect size. We also divide the standard error of the difference by this standard deviation. We then find the weighted average as above. This is useful when the outcome is not always the same measurement. It is often used for psychological scales. Figure 7 shows an example of the use of standardised mean difference, the outcome variables being various pain scales used to measure the outcome of trials of non-steroidal anti-inflammatory drugs.

The data required for meta-analysis of a continuous outcome variable are, for each study, the difference between means and its standard error. If these are not given in the paper, provided we have the mean, standard deviation, and sample size for each group, we then find the difference between means and its standard error in the usual way. For standardised differences, we need either the standardised difference and its standard error or the standard deviation. In the latter case we can divide the difference between means by the standard deviation. Everything is then in the same units, i.e. standard deviation units.

Figure 6. Example of weighted mean difference: blood pressure control by home monitoring (Cappuccio *et al.*, 2004)

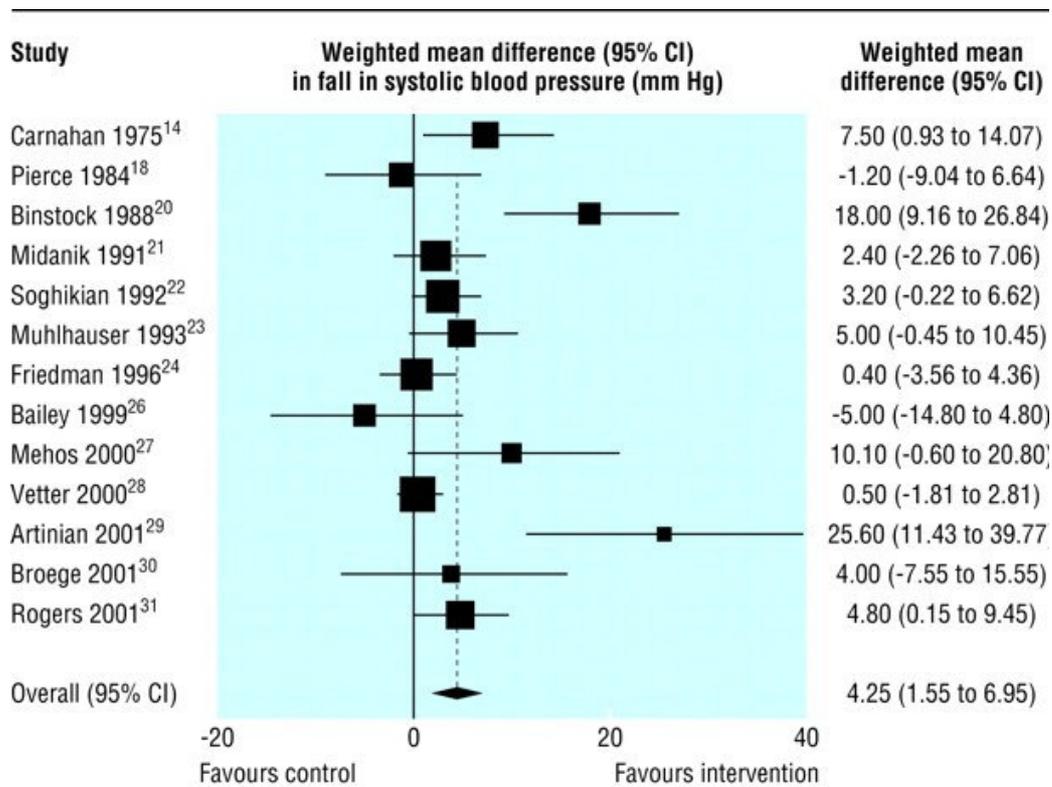
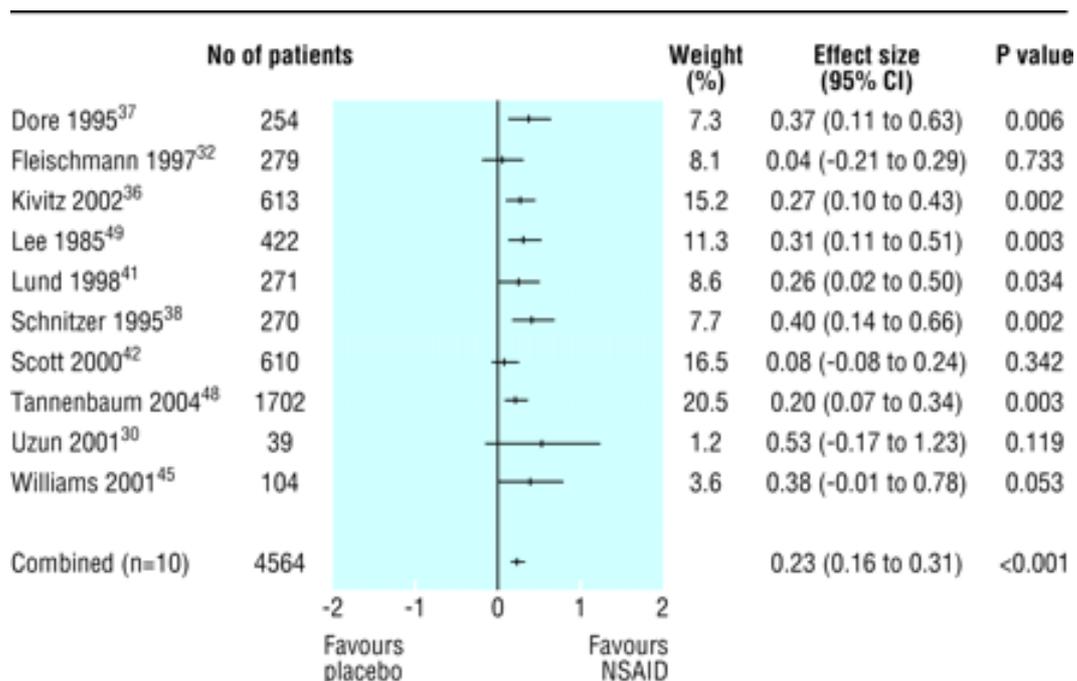


Figure 7. Example of standardised mean difference: pain scales used to measure the outcome of trials of non-steroidal anti-inflammatory drugs in osteoarthritic knee pain (Bjordal *et al.*, 2004)



Unfortunately, the required data are not always available for all published studies. Studies sometimes report different measure of variation. These might be:

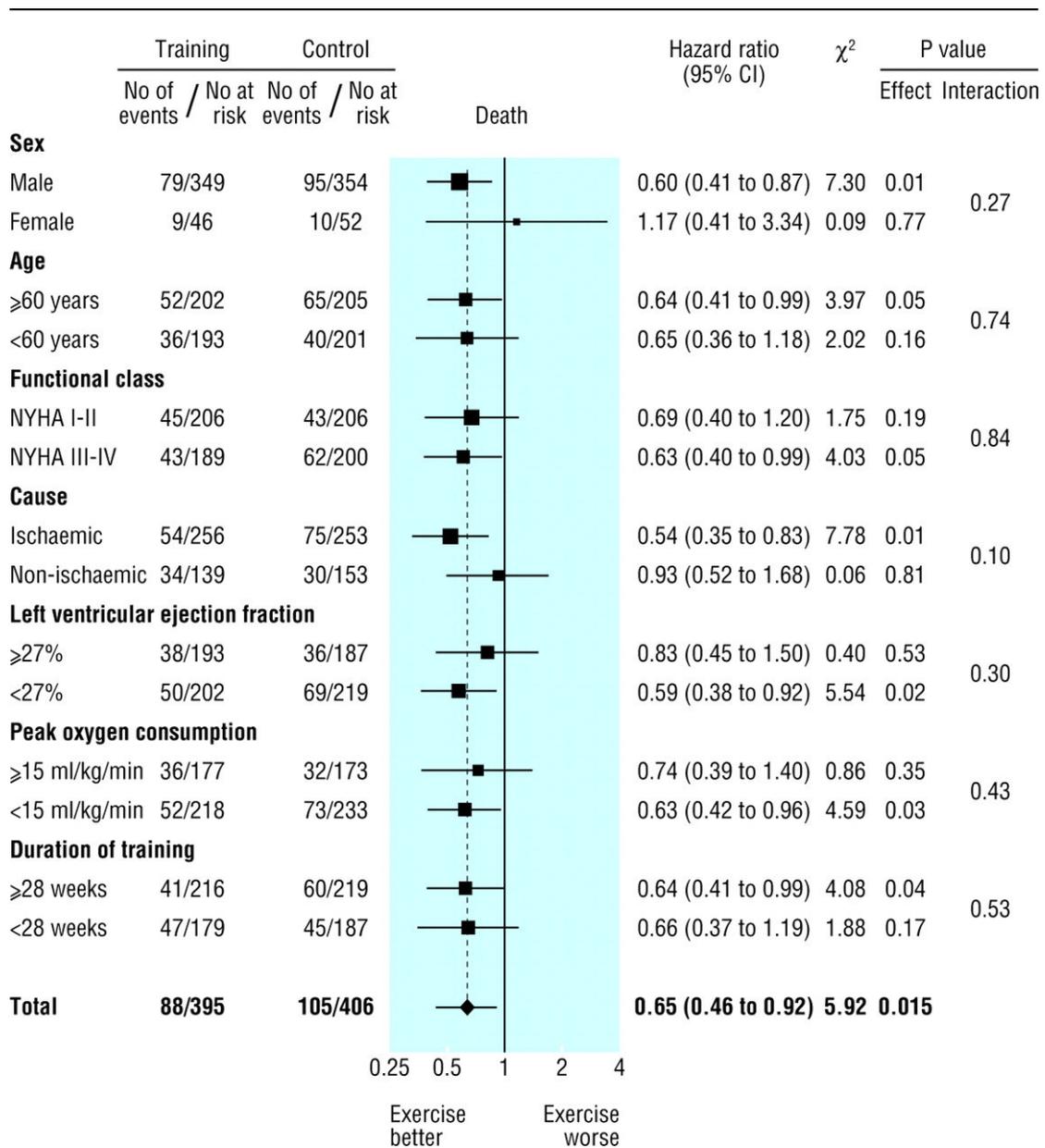
- standard errors
- confidence intervals
- reference ranges
- interquartile ranges
- range
- significance test
- P value
- ‘Not significant’ or ‘ $P < 0.05$ ’.

We need to extracting the information required from what is available.

- standard errors — this is straightforward, as we know the formula for the standard error and so provided we have the sample sizes we can calculate standard deviation,
- confidence intervals — this is also straightforward, as we can work back to the standard error,
- reference ranges — again straightforward, as the reference range is four standard deviations wide,
- interquartile ranges — here we need an assumption about distribution; provided this is Normal we know how many standard deviations wide the IQR should be, but of course this is often not the case,
- range — this is very difficult, as not only to we need to make an assumption about the distribution but the estimates are unstable and affected by outliers,
- significance test — sometimes we can work back from a t value to the standard error, but not from some other tests, such as the Mann Whitney U test,
- P value — if we have a t test we can work back to a t value hence to the standard error, but not for other tests, and we need the exact P value.
- ‘Not significant’ or ‘ $P < 0.05$ ’ — this is hopeless.



Figure 11. Results of a meta-analysis of trials of exercise training in patients with chronic heart failure, time to death (ExTraMATCH Collaborative 2004)



## **Time to event outcome variables**

Time-to-event data arise whenever we have subjects followed over time until some event takes place. Such data are often called survival data, because the early applications were often in time to death. These techniques are also used for time to recurrence of disease, time to discharge from hospital, time to readmission to hospital, time to conception, time to fracture, etc. The usual problem with such data is that not all subjects have an event, so we know only that they were observed to be event-free up to some point, but not beyond it. Also, usually some of those observed not to have an event were observed for a shorter time than some of those who did have an event. A special body of statistical techniques, survival analysis, have been developed for such data.

The main effect measure is the hazard ratio. This is the standard outcome measure in survival analysis. It is the ratio of the risk of having an event at any given time in one group divided by the risk of an event in the other.

For example, Maier *et al.* (2005) analysed the time to visual field loss or deterioration of the optic disc, or both, in patients with ocular hypertension (Figure 9). The patients were randomised to pressure lowering treatment or to no treatment. A hazard ratio which is equal to one represents no difference between the groups. The hazard ratio is active treatment divided by no treatment, so if the hazard ratio is less than one, this means that the risk of visual field loss is less for patients given pressure lowering treatment. As for risk ratios and odds ratios, hazard ratios are analysed by taking the log and the results are shown on a logarithmic scale.

## **Individual patient data meta-analysis**

In this kind of meta-analysis, we get the raw data from each study. We may then combine them into a single data set and analyse them like a single, multicentre clinical trial. Alternatively, we may use the individual data to extract the corresponding summary statistics from each study then proceed as we would using summary statistics from published reports.

An example was the ExTraMATCH study (ExTraMATCH Collaborative 2004), a meta-analysis of trials of exercise training in patients with chronic heart failure. Nine trials identified and principal investigators provided a minimum data set in electronic form. Because in this study the trials were pooled to form one data set, individual study results are not given. The outcome was time to death or time to death or admission to hospital. Figure 10 shows the Kaplan Meier survival curves for the exercise and control groups, pooled across the studies. The Kaplan Meier survival curve shows the estimated proportion of subjects who have not yet experienced the event at each time.

Figure 11 shows more results from the ExTraMATCH study. This looks like a forest plot as in Figures 1-9, But it is different. It shows the estimated treatment effect for the subjects as they are grouped by different prognostic variables. It is to show that the effects of treatments are not explained by differences in prognostic variables between the groups, highly unlikely in these randomised trials, and also to suggest where there might be interactions between treatment and prognostic variables.

## **And finally . . .**

Meta-analysis is straightforward if the data are straightforward and all available.

It depends crucially on the data quality and the completeness of the study ascertainment.

Martin Bland

16 February 2006

## References

Bjordal JM, Ljunggren AE, Klovning A, Slørdal L. (2004) Non-steroidal anti-inflammatory drugs, including cyclo-oxygenase-2 inhibitors, in osteoarthritic knee pain: meta-analysis of randomised placebo controlled trials. *BMJ*, **329**, 1317.

Cappuccio FP, Kerry SM, Forbes L, Donald A. (2004) Blood pressure control by home monitoring: meta-analysis of randomised trials. *British Medical Journal*, **329**, 145.

Chang JT, Morton SC, Rubenstein LZ, Mojica WA, Maglione M, Suttrop MJ, Roth EA, Shekelle PG. (2004) Interventions for the prevention of falls in older adults: systematic review and meta-analysis of randomised clinical trials. *British Medical Journal*, **328**: 680-3.

Colman I, Brown MD, Innes GD, Grafstein E, Roberts TE, Rowe BH. (2004) Parenteral metoclopramide for acute migraine: meta-analysis of randomised controlled trials. *British Medical Journal*, **329**, 1369.

Etminan M, Takkouche B, Isorna FC, Samii A. (2005) Risk of ischaemic stroke in people with migraine: systematic review and meta-analysis of observational studies. *British Medical Journal*, **330**, 63.

ExTraMATCH Collaborative. (2004) Exercise training meta-analysis of trials in patients with chronic heart failure (ExTraMATCH). *British Medical Journal*, **328**, 189.

Fletcher A, Nicky Cullum N, Sheldon TA. (1997) A systematic review of compression treatment for venous leg ulcers. *British Medical Journal*, **315**, 576-580.

Liu EHC and Sia ATH. (2004) Rates of caesarean section and instrumental vaginal delivery in nulliparous women after low concentration epidural infusions or opioid analgesia: systematic review. *British Medical Journal*, **328**, 1410-12.

Maier PC, Funk J, Schwarzer G, Antes G, Falck-Ytter YT. (2005) Treatment of ocular hypertension and open angle glaucoma: meta-analysis of randomised controlled trials. *British Medical Journal*, **331**, 134.