# 15 Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis

JONATHAN J DEEKS, DOUGLAS G ALTMAN, MICHAEL J BRADBURN

## Summary points

- Meta-analysis is a two-stage process involving the calculation of an appropriate summary statistic for each of a set of studies followed by the combination of these statistics into a weighted average.
- Methods are available for combining odds ratios, risk ratios and risk differences for binary data, and hazard ratios for time to event data.
- Continuous data can be combined either as differences in means, or as standardised differences in means when a mixture of measurement scales has been used.
- Fixed effect models average the summary statistics, weighting them according to a measure of the quantity of information they contain. Several methods are available (inverse variance, Mantel–Haenszel and Peto) which differ mainly in the computations used to calculate the individual study weights.
- Random effects models incorporate an estimate of between study variation (heterogeneity) into the calculation of the common effect. One simple method is readily available (DerSimonian and Laird); other methods require more complex statistical computations.
- Selection of a meta-analysis method for a particular analysis should reflect the data type, choice of summary statistic (considering the consistency of the effect and ease of interpretation of the statistic), observed heterogeneity, and the known limitations of the computational methods.

An important step in a systematic review is the thoughtful consideration of whether it is appropriate to combine all (or perhaps some) of the studies in a meta-analysis, to yield an overall statistic (together with its confidence interval) that summarises the effectiveness of the treatment (see Chapter 2). Statistical investigation of the degree of variation between individual study results, which is known as heterogeneity, can often contribute to making decisions regarding the "combinability" of results. In this chapter we consider the general principles of meta-analysis, and introduce the most commonly used methods for performing meta-analysis and examining heterogeneity. We shall focus on meta-analysis of randomised trials evaluating therapies, but much the same principles apply to other comparative studies, notably case-control and cohort studies.

## Meta-analysis

### General principles

Meta-analysis is a two-stage process. In the first stage a summary statistic is calculated for each study. For controlled trials, these values describe the treatment effect observed in each individual trial. The summary statistics are usually risk ratios, odds ratios or risk differences for event data, differences in means for continuous data, or hazard ratios for survival time data. In the second stage the overall treatment effect is calculated as a weighted average of these summary statistics. The weights are chosen to reflect the amount of information that each trial contains. In practice the weights are often the inverse of the variance (the square of the standard error) of the treatment effect, which relates closely to sample size. The precision (confidence interval) and statistical significance of the overall estimate are also calculated. It is also possible to weight additionally by study quality, although this is not generally recommended (see Chapter 5). All commonly used methods of meta-analysis follow these basic principles. There are, however, some other aspects that vary between alternative methods, as described below.

In a meta-analysis we do not combine the data from all of the trials as if they were from a single large trial. Such an approach is inappropriate for several reasons and can give misleading results, especially when the number of participants in each group is not balanced within trials.[1]

### Assessing heterogeneity

An important component of a systematic review is the investigation of the consistency of the treatment effect across the primary studies. As the trials will not have been conducted according to a common protocol, there will usually be variations in patient groups, clinical settings, concomitant care and the methods of delivery of the intervention. Whilst some





divergence of trial results from the overall estimate is always expected purely by chance, the effectiveness of the treatment may also vary according to individual trial characteristics, which will increase the variability of results. The possibility of excess variability between the results of the different trials is examined by the test of homogeneity (occasionally described as a test for heterogeneity).

Consistency of trial results with a common effect despite variation in trial characteristics provides important and powerful corroboration of the generalisation of the treatment effect, so that a greater degree of certainty can be placed on its application to wider clinical practice.[2] However, the test of homogeneity has low power to detect excess variation, especially when there are not many studies, so the possibility of a type II (false negative) error must always be considered. By contrast, if the test of homogeneity is statistically significant, the between trial variability is more than expected by chance alone. In these situations it is still possible for a treatment to be shown to have a real, if not constant, benefit. In particular, the extra variation can be incorporated into the analysis using a random effects model (see below).

Where the heterogeneity is considerable, the reviewer ought to consider an investigation of reasons for the differences between trial results (see Chapters 8–11)[3] or not reporting a pooled estimate. Stratified meta-analysis (described below) and special statistical methods of meta-regression (see Chapters 9 and 11, and STATA command `metareg` in Chapter 18) can be used to test and examine potential associations between study factors and the estimated treatment effect.

## Formulae for estimates of effect from individual studies

We assume here that the meta-analysis is being carried out on summary information obtained from published papers. The case of individual patient data (see Chapter 6) is considered briefly.

### Individual study estimates of treatment effect: binary outcomes

For studies with a binary outcome the results can be presented in a 2 × 2 table (Table 15.1) giving the numbers of people who do or do not experience the event in each of the two groups (here called intervention and control).

Table 15.1 Summary information when outcome is binary.

| Study $i$ | Event | No event | Group size |
|---|---|---|---|
| Intervention | $a_i$ | $b_i$ | $n_{1i}$ |
| Control | $c_i$ | $d_i$ | $n_{2i}$ |

For the $i^{th}$ study we denote the cell counts as in Table 15.1, with $N_i = n_{1i} + n_{2i}$. Zero cells cause problems with computation of the standard errors so 0.5 is usually added to each cell ($a_i$, $b_i$, $c_i$, $d_i$) for such studies.[4]

The treatment effect can be expressed as either a relative or absolute effect. Measures of relative effect (odds ratios and risk ratios) are usually combined on the log scale. Hence we give the standard error for the log ratio measure.

The *odds ratio*[5] for each study is given by

$$OR_i = \frac{a_i d_i}{b_i c_i},$$

the standard error of the log odds ratio being

$$\mathrm{SE}[\ln(OR_i)] = \sqrt{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}},$$

where ln denotes logarithms to base e (natural logarithms).

The *risk ratio*[5] for each study is given by

$$RR_i = \frac{a_i / n_{1i}}{c_i / n_{2i}},$$

the standard error of the log risk ratio being

$$\mathrm{SE}[\ln(RR_i)] = \sqrt{\frac{1}{a_i} + \frac{1}{c_i} - \frac{1}{n_{1i}} - \frac{1}{n_{2i}}}.$$

The *risk difference*[6] for each study is given by

$$RD_i = \frac{a_i}{n_{1i}} - \frac{c_i}{n_{2i}},$$

with standard error

$$\mathrm{SE}(RD_i) = \sqrt{\frac{a_i b_i}{n_{1i}^3} + \frac{c_i d_i}{n_{2i}^3}}.$$





For the *Peto odds ratio* method[7] (see below) the individual odds ratios are given by

$$OR_i = \exp\left(\frac{a_i - E[a_i]}{v_i}\right),$$

with standard error

$$\mathrm{SE}[\ln(OR_i)] = \sqrt{1/v_i},$$

where $E[a_i] = n_{1i}(a_i + c_i) / N_i$ (the expected number of events in the intervention group under the null hypothesis of no treatment effect) and

$$v_i = \frac{n_{1i} n_{2i} (a_i + c_i)(b_i + d_i)}{N_i^2 (N_i - 1)},$$

the hypergeometric variance of $a_i$.

*Individual study estimates of treatment effect: continuous outcomes*

If the outcome is a continuous measure, we require the number of participants, the mean response and its standard deviation, for intervention and control groups (Table 15.2).

Table 15.2 Summary information when outcome is continuous.

| Study $i$ | Mean response | Standard deviation | Group size |
|---|---|---|---|
| Intervention | $m_{1i}$ | $SD_{1i}$ | $n_{1i}$ |
| Control | $m_{2i}$ | $SD_{2i}$ | $n_{2i}$ |

We let $N_i = n_{1i} + n_{2i}$ be the total number of participants in study $i$, and

$$s_i = \sqrt{\frac{(n_{1i} - 1)\mathrm{SD}_{1i}^2 + (n_{2i} - 1)\mathrm{SD}_{2i}^2}{N_i - 2}}$$

be the pooled standard deviation of the two groups.

There are two summary statistics used for meta-analysis of continuous data. The *difference in means* can be used when outcome measurements in all trials are made on the same scale. The meta-analysis computes a weighted average of these differences in means, but is confusingly termed the *weighted mean difference* (WMD) method.

The *standardised difference* is used when the trials all assess the same outcome, but measure it in a variety of ways (for example, all trials measure depression but they use different psychometric scales). In this circumstance it is necessary to standardise the results of the trials to a uniform scale before they can be combined. The *standardised mean difference* method expresses the size of the treatment effect in each trial (again in reality a difference in means and not a mean difference) relative to the variability observed in that trial. The method assumes that the differences in standard deviations between trials reflect differences in measurement scales and not real differences in variability between trial populations. This assumption may be problematic in some circumstances where pragmatic and explanatory trials (which may differ in the risk of poor outcomes) are combined in the same review. The overall treatment effect can also be difficult to interpret as it is reported in units of standard deviation rather than in units of any of the measurement scales used in the review.

For a particular study the *difference in means* (denoted MD)[8] is given by

$$MD_i = m_{1i} - m_{2i},$$

with standard error

$$\mathrm{SE}(MD_i) = \sqrt{\frac{\mathrm{SD}_{1i}^2}{n_{1i}} + \frac{\mathrm{SD}_{2i}^2}{n_{2i}}}.$$

There are three popular formulations of effect size used in the standardised mean difference method. These formulations differ with respect to the standard deviation used in calculations and whether or not a correction for *small sample bias* is included. In statistics small sample bias is defined as the difference between the expected value of an estimate given a small sample and the expected value if the sample is infinite. Simulations show that the standardised mean difference tends to be overestimated with finite samples but the bias is substantial only if total sample size is very small (less than 10).[9]

**Cohen's *d*** [10] is given by

$$d_i = \frac{m_{1i} - m_{2i}}{s_i},$$

with standard error

$$\mathrm{SE}(d_i) = \sqrt{\frac{N_i}{n_{1i}n_{2i}} + \frac{d_i^2}{2(N_i - 2)}}.$$





**Hedges' adjusted *g*** [10] is very similar to Cohen's *d* but includes an adjustment to correct for the small sample bias mentioned above. It is defined as

$$g_i = \frac{m_{1i} - m_{2i}}{s_i}\left(1 - \frac{3}{4N_i - 9}\right),$$

with standard error

$$\mathrm{SE}(g_i) = \sqrt{\frac{N_i}{n_{1i}n_{2i}} + \frac{g_i^2}{2(N_i - 3.94)}}.$$

Finally, **Glass's** $\Delta$[11] takes the standard deviation from the control group as the scaling factor, giving

$$\Delta_i = \frac{m_{1i} - m_{2i}}{\mathrm{SD}_{2i}},$$

with standard error

$$\mathrm{SE}(\Delta_i) = \sqrt{\frac{N_i}{n_{1i}n_{2i}} + \frac{\Delta_i^2}{2(n_{2i} - 1)}}.$$

This method is preferable when the intervention alters the observed variability as well as potentially changing the mean value.

Both the weighted mean difference and standardised mean difference methods assume that the outcome measurements within each trial have a Normal distribution. When these distributions are skewed or severely non-Normal, the results of these methods may be misleading.

## Formulae for deriving a summary (pooled) estimate of the treatment effect by combining trial results (meta-analysis)

The methods of meta-analysis described below all combine the individual study summary statistics described above, denoted generically by $\theta_i$, each given a weight $w_i$ which is usually related to $\mathrm{SE}(\theta_i)$. All the methods described are available in the Stata routines described in Chapter 18. The summation notation indicates summation of the $i$ trials included in the analysis.

### Fixed effect and random effects methods

In fixed effect meta-analysis it is assumed that the true effect of treatment is the same value in each study, or *fixed*, the differences between study results being due solely to the play of chance. The assumption of a fixed effect can be tested using a test of homogeneity (see below).

In a random effects meta-analysis the treatment effects for the individual studies are assumed to vary around some overall average treatment effect. Usually the effect sizes $\theta_i$ are assumed have a Normal distribution with mean $\theta$ and variance $\tau^2$. In essence the test of homogeneity described below tests whether $\tau^2$ is zero. The smaller the value of $\tau^2$ the more similar are the fixed and random effects analyses.

Peto describes his method for obtaining a summary odds ratio as assumption free,[7] arguing that it does not assume that all the studies are estimating the same treatment effect, but it is generally considered to be most similar to a fixed effect method.

There is no consensus about whether to use fixed or random effects models.[12] All of the methods given below are fixed effect approaches except the DerSimonian and Laird method.

### Inverse variance method

Inverse variance methods may be used to pool either binary or continuous data. In the general formula below, the effect size, denoted $\theta_i$, could be the log odds ratio, log relative risk, risk difference, difference in means or standardised mean difference from the $i$th trial.

The effect sizes are combined to give a pooled estimate by calculating a *weighted average* of the treatment effects from the individual trials:

$$\theta_{IV} = \frac{\sum w_i \theta_i}{\sum w_i}.$$

The weights are the reciprocals of the squared standard errors:

$$w_i = \frac{1}{\mathrm{SE}(\theta_i)^2}.$$

Thus larger studies, which have smaller standard errors, are given more weight than smaller studies, which have larger standard errors. This choice of weight minimises the variability of the pooled treatment effect $\theta_{IV}$.

The standard error of $\theta_{IV}$ is given by

$$\mathrm{SE}(\theta_{IV}) = \frac{1}{\sqrt{\sum w_i}}.$$





The heterogeneity statistic is given by

$$Q = \sum w_i (\theta_i - \theta_{IV})^2.$$

The strength of this approach is its wide applicability. It can be used to combine any estimates that have standard errors available. Thus it can be used for estimates from many types of study, including standardised mortality ratios, diagnostic test indices (Chapter 14), hazard ratios (Chapter 6), and estimates from cross-over trials and cluster-randomised trials. It is also possible to use this method when crude 2 × 2 tables cannot be obtained for each study, but treatment effects and confidence intervals are available (see Stata commands `meta` and `metan` in Chapter 18).

### Mantel–Haenszel methods

When data are sparse, both in terms of event rates being low and trials being small, the estimates of the standard errors of the treatment effects that are used in the inverse variance methods may be poor. Mantel–Haenszel methods use an alternative weighting scheme, and have been shown to be more robust when data are sparse, and may therefore be preferable to the inverse variance method. In other situations they give similar estimates to the inverse variance method. They are available only for binary outcomes (see Stata command `metan` in Chapter 18).

For each study, the effect size from each trial $\theta_i$ is given weight $w_i$ in the analysis. The overall estimate of the pooled effect, $\theta_{MH}$ is given by:

$$\theta_{MH} = \frac{\sum w_i \theta_i}{\sum w_i}.$$

Unlike with inverse variance methods, relative effect measures are combined in their natural scale, although their standard errors (and confidence intervals) are still computed on the log scale.

For combining *odds ratios*, each study's OR is given weight[13,14]

$$w_i = \frac{b_i c_i}{N_i},$$

and the logarithm of $OR_{MH}$ has standard error given by[15]

$$\mathrm{SE}[\ln(OR_{MH})] = \sqrt{\frac{1}{2}\left(\frac{E}{R^2} + \frac{F+G}{R \times S} + \frac{H}{S^2}\right)},$$

where

$$R = \sum \frac{a_i d_i}{N_i};\ S = \sum \frac{b_i c_i}{N_i};$$

$$E = \sum \frac{(a_i + d_i)a_i d_i}{N_i^2};\ F = \sum \frac{(a_i + d_i)b_i c_i}{N_i^2};$$

$$G = \sum \frac{(b_i + c_i)a_i d_i}{N_i^2};\ H = \sum \frac{(b_i + c_i)b_i c_i}{N_i^2}.$$

For combining *risk ratios*, each study's RR is given weight[16]

$$w_i = \frac{c_i n_{1i}}{N_i},$$

and the logarithm of $RR_{MH}$ has standard error given by

$$\mathrm{SE}[\ln(RR_{MH})] = \sqrt{\frac{P}{R \times S}},$$

where

$$P = \sum \frac{(n_{1i} n_{2i}(a_i + c_i) - a_i c_i N_i)}{N_i^2};\ R = \sum \frac{a_i n_{2i}}{N_i};\ S = \sum \frac{c_i n_{1i}}{N_i}.$$

For *risk differences*, each study's RD has the weight[16]

$$w_i = \frac{n_{1i} n_{2i}}{N_i},$$

and $RD_{MH}$ has standard error given by

$$\mathrm{SE}(RD_{MH}) = \sqrt{\mathcal{J} / K^2},$$

where

$$\mathcal{J} = \sum \left( \frac{a_i b_i n_{2i}^3 + c_i d_i n_{1i}^3}{n_{1i} n_{2i} N_i^2} \right);\ K = \sum \left( \frac{n_{1i} n_{2i}}{N_i} \right).$$





However, the test of homogeneity is based upon the inverse variance weights and not the Mantel–Haenszel weights. The heterogeneity statistic is given by

$$Q = \sum w_i (\theta_i - \theta_{MH})^2$$

where $\theta$ is the log odds ratio, log relative risk or risk difference.

### Peto's odds ratio method

An alternative to the Mantel–Haenszel method is a method due to Peto (sometimes attributed to Yusuf, or to Yusuf and Peto).[7] The overall odds ratio is given by

$$OR_{\text{Peto}} = \exp\left( \frac{\sum w_i \ln(OR_i)}{\sum w_i} \right),$$

where the odds ratio $OR_i$ is calculated using the approximate Peto method described in the individual trial section, and the weight $w_i$ is equal to the hypergeometric variance of the event count in the intervention group, $v_i$.

The logarithm of the odds ratio has standard error

$$\mathrm{SE}[\ln(OR_{\text{Peto}})] = \frac{1}{\sqrt{\sum v_i}}.$$

The heterogeneity statistic is given by

$$Q = \sum v_i (\ln OR_i - \ln OR_{\text{Peto}})^2.$$

The approximation upon which Peto's method relies has shown to fail when treatment effects are very large, and when the sizes of the arms of the trials are seriously unbalanced.[17] Severe imbalance, with, for example, four or more times as many participants in one group than the other, would rarely occur in randomised trials. In other circumstances, including when event rates are very low, the method performs well.[18] Corrections for zero cell counts are not necessary for this method (see Stata command `metan` in Chapter 18).

### Extending the Peto method for pooling time-to-event data

Pooling of time-to-event outcomes can be achieved either by computing hazard ratios for each trial and pooling them using the inverse variance

method (as explained above), or by exploiting a link between the log rank test statistic and the Peto method, as follows.

For each trial, the calculation of a log rank statistic involves dividing the follow-up period into a series of discrete time intervals. For each interval the number of events observed in the treated group $O_{ij}$, the number of events that would be expected in the treatment group under the null hypothesis $E_{ij}$ and its variance $v_{ij}$ are calculated (for formulae, see for example Altman[19]). The expected count and its variance are computed taking into account the number still at risk of the event within each time period. The log-rank test for the $i$th trial is computed from $\sum O_{ij}$, $\sum E_{ij}$ and $\sum v_{ij}$ summed over all the time periods, $j$.

Following the same format as the Peto odds ratio method, an estimate of the hazard ratio in each trial is given by[19]

$$HR_i = \exp\left(\frac{\sum O_{ij} - \sum E_{ij}}{\sum v_{ij}}\right),$$

with standard error

$$\mathrm{SE}[\ln(HR_i)] = \sqrt{1/\sum v_{ij}}.$$

The overall hazard ratio is given by the weighted average of the log hazard ratios

$$HR_{\text{Peto}} = \exp\left(\frac{\sum w_i \ln(HR_i)}{\sum w_i}\right),$$

where the weights $w_i$ are equal to the variances computed from the trials, $\sum v_{ij}$.

The logarithm of the overall hazard ratio has standard error

$$\mathrm{SE}[\ln(HR_{\text{Peto}})] = \frac{1}{\sqrt{\sum w_i}}.$$

Computation of the components of the log-rank statistic $\sum O_{ij}$, $\sum E_{ij}$ and $\sum v_{ij}$ is straightforward if individual patient data are available. Methods have been proposed for indirectly estimating the log hazard ratio and its variance from graphical and numerical summaries commonly published in reports of randomised controlled trials.[20]





**DerSimonian and Laird random effects models**

Under the random effects model, the assumption of a common treatment effect is relaxed, and the effect sizes $\theta_i$ are assumed have a Normal distribution with mean and variance $\tau^2$. The usual DerSimonian and Laird[21] estimate of $\tau^2$ is given by

$$\tau^2 = \frac{Q-(k-1)}{\sum w_i - \left(\frac{\sum w_i^2}{\sum w_i}\right)},$$

where $Q$ is the heterogeneity statistic, with $\tau^2$ set to zero if $Q < k-1$, and the $w_i$ are calculated as in the inverse variance method. The estimate of the combined effect for the heterogeneity may be taken as the inverse variance estimate, although the Mantel–Haenszel estimate may be preferred. Again, for odds ratios and risk ratios, the effect size is taken as the natural logarithm of the OR and RR. Each study's effect size is given weight

$$w_i' = \frac{1}{\mathrm{SE}(\theta_i)^2 + \tau^2}.$$

The pooled effect size is given by

$$\theta_{\text{DL}} = \frac{\sum w_i' \theta_i}{\sum w_i'},$$

with standard error

$$\mathrm{SE}(\theta_{\text{DL}}) = \frac{1}{\sqrt{\sum w_i'}}.$$

Note that when $\tau^2 = 0$, i.e. where the heterogeneity statistic $Q$ is as small as or smaller than its degrees of freedom $(k-1)$, the weights reduce to those given by the inverse variance method.

If the estimate of $\tau^2$ is greater than zero then the weights in random-effects models ($w_i' = 1/(\mathrm{SE}(\theta_i)^2 + \tau^2)$) will be smaller and more similar to each other than the weights in fixed effect models ($w_i = 1/\mathrm{SE}(\theta_i)^2$). This means than random-effects meta-analyses will be more conservative (the confidence intervals will be wider) than fixed effect analyses[22] since the variance of the pooled effect is the inverse of the sum of the weights. It also

means that random effects models give relatively more weight to smaller studies than the fixed effect model. This may not always be desirable (see Chapter 11).

The DerSimonian and Laird method has the same wide applicability as the inverse variance method, and can be used to combine any type of estimates provided standard errors are available (see Stata commands **meta** and **metan** in Chapter 18).

## Confidence interval for overall effect

The $100(1-\alpha)\%$ confidence interval for the overall estimate $\theta$ is given by

$$\theta - \left(z_{1-\alpha/2} \times \mathrm{SE}(\theta)\right) \text{ to } \theta + \left(z_{1-\alpha/2} \times \mathrm{SE}(\theta)\right),$$

where $\theta$ is the log odds ratio, log relative risk, risk difference, mean difference or standardised mean difference, and $z$ is the standard Normal deviate. For example, if $\alpha = 0.05$, then $z_{1-\alpha/2} = 1.96$ and the 95% confidence interval is given by

$$\theta - \left(1.96 \times \mathrm{SE}(\theta)\right) \text{ to } \theta + \left(1.96 \times \mathrm{SE}(\theta)\right).$$

Confidence intervals for log odds ratios and log risk ratios are exponentiated to provide confidence intervals for the pooled OR or RR.

## Test statistic for overall effect

In all cases a test statistic for the overall difference between groups is derived as

$$z = \frac{\theta}{\mathrm{SE}(\theta)}$$

(where the odds ratio or risk ratio is again considered on the log scale). Under the null hypothesis that there is no treatment effect, $z$ will follow a standard Normal distribution.

For odds ratios an alternative test statistic is given by comparing the number of observed and expected events in the treatment group given no difference is present between the groups. This test is given by

$$\chi^2 = \frac{\sum (a_i - E[a_i])^2}{\sum v_i},$$





where $E[a_i]$ and $v_i$ are as defined above. Under the null hypothesis of no treatment effect, this statistic follows a chi-squared distribution on one degree of freedom.

## Test statistics of homogeneity

For a formal test of homogeneity, the statistic $Q$ will follow a chi-squared distribution on $k-1$ degrees of freedom under the null hypothesis that the true treatment effect is the same for all trials.

Breslow and Day proposed an alternative test of the homogeneity of odds ratios,[14] based upon a comparison of the observed number of events in the intervention groups of each trial ($a_i$), with those expected when the common treatment effect $OR$ is applied (calculation of these expected values involves solving quadratic expressions). The test statistic is given by

$$Q_{\mathrm{BD}} = \sum \left( \frac{a_i - E[a_i \mid OR]}{v_i} \right),$$

where each trial's variance $v_i$ is computed using the fitted cell counts

$$v_i = \frac{1}{E[a_i \mid OR]} + \frac{1}{E[b_i \mid OR]} + \frac{1}{E[c_i \mid OR]} + \frac{1}{E[d_i \mid OR]}.$$

Under the null hypothesis of homogeneity $Q_{\mathrm{BD}}$ also has a chi-squared distribution on $k-1$ degrees of freedom.

## Use of stratified analyses for investigating sources of heterogeneity

In a stratified analysis the trials are grouped according to a particular feature or characteristic and a separate meta-analysis carried out of the trials within each subgroup. The overall summaries calculated within each subgroup can then be inspected for evidence of variation in the effect of the intervention, which would suggest that the stratifying characteristic is an important source of heterogeneity and may moderate treatment efficacy.

Stratified analysis can be used when the trials can be grouped into a small number of categories according to the study characteristic; meta-regression (see Chapter 9) can be used when the characteristic is a continuous measure.

An inference that the treatment effect differs between two or more subsets of the trials should be based on a formal test of statistical significance. There are three methods to assess statistical significance.

Consider first a stratified analysis with the trials grouped into $k$ subgroups. By performing separate meta-analyses within each subgroup, we obtain for the $k$th subgroup:

$\theta_k$, an estimate of the overall effect within each group,

$SE(\theta_k)$, the standard error of these estimates,

$Q_k$, the heterogeneity observed within each group.

If there are only 2 groups, the significance of the difference between the two groups can be examined by comparing the $z$ statistic

$$z = \frac{\theta_1 - \theta_2}{\sqrt{[SE(\theta_1)]^2 + [SE(\theta_2)]^2}},$$

with critical values of the Normal distribution.

An alternative test, which can be used regardless of the number of subgroups, involves explicitly partitioning the overall heterogeneity into that which can be explained by differences between subgroups, and that which remains unexplained within the subgroups. If the heterogeneity of the overall unstratified analysis is $Q_T$, the heterogeneity explained by differences between subgroups, $Q_B$, is given by:

$$Q_B = Q_T - \sum_k Q_k,$$

which can be compared with critical values of the chi-squared distribution with $k$–1 degrees of freedom.

The problem can also be formulated as a meta-regression (see Chapter 9), using $k$–1 dummy variables to indicate membership of the $k$ subgroups, in the standard manner used in multiple regression. The meta-regression will also produce estimates of the differences between a baseline reference subgroup and each of the other subgroups. If the categories are ordered, meta-regression should be used to perform a test for trend by denoting group membership by a single variable indicating the ranked order of each subgroup.

The interpretation of comparisons between subgroups should be undertaken cautiously, as significant differences can easily arise by chance (a type I error), or are explicable by other factors. Even when the studies in the meta-analysis are randomised controlled trials, the investigation of differences between subgroups is a non-randomised comparison, and is





prone to all of the difficulties in inferring causality in observational studies (see Chapter 12). Where multiple possible sources of heterogeneity are investigated, the chance of one of them being found to be statistically significant increases, so the number of factors considered should be restricted. Pre-specification (in a protocol) of possible sources of heterogeneity increases the credibility of any statistically significant findings, as there is evidence that the findings are not data-derived. Examples of stratified meta-analyses are shown in the Case studies 1 and 3 below.

Often the stratifying factor is the type of intervention. For example, a systematic review may include placebo controlled trials of several drugs, all for the same condition. The meta-analysis will be stratified by drug, and will provide estimates of treatment effect for each drug. Here a test of differences between subgroups is effectively an indirect comparison of the effects of the drugs. Although such a test can provide indirect evidence of relative treatment effects, it is much less reliable than evidence from randomised controlled trials which compare the drugs directly (head-to-head comparisons). Similar situations also arise with non-pharmacological interventions. Such indirect comparisons are considered by Bucher *et al.*[23] and Song *et al.*[24]

## Meta-analysis with individual patient data

The same basic approaches and meta-analysis methods are used for meta-analyses of individual patient data (IPD)[25] (see Chapter 6). However, there are two principal differences between IPD analyses and those based on published summary statistics. Firstly, the IPD meta-analyst calculates the summary tables or statistics for each study, and therefore can ensure all data are complete and up-to-date, and that the same method of analysis is used for all trials. Secondly, summary statistics can be calculated for specific groups of participants enabling full intention-to-treat (see below) and subgroup analyses to be produced. Additionally, it is worth noting that IPD meta-analyses often combine time-to-event data rather than binary or continuous outcomes, the meta-analyst calculating the required components of the log rank statistic in the same manner for each of the trials.

## Additional analyses

Additional analyses undertaken after the main meta-analysis investigate *influence*, *robustness* and *bias*. Influence and robustness can be assessed in sensitivity analyses by repeating the meta-analysis on subsets of the original dataset (see Chapter 2 for an example). The influence of each study can be estimated by deleting each in turn from the analysis and noting the degree

to which the size and significance of the treatment effect changes (see Stata command `metainf` in Chapter 18). Other sensitivity analyses can assess robustness to uncertainties and assumptions by removing or adding sets of trials, or by changing the data for individual trials. Situations where these may be considered include when some of the trials are of poorer quality (Chapter 5), when it is unclear whether some trials meet the inclusion criteria, or when the results of trials in the published reports are ambiguous and assumptions are made when extracting data. Methods for investigating bias, including publication bias, are described in detail in Chapter 11 (see also Stata command `metabias` in Chapter 18).

## Some practical issues

Although it is desirable to include trial results from intention to treat analyses, this is not always possible given the data provided in published reports. Reports commonly omit participants who do not comply, receive the wrong treatment, or who drop out of the study. All of these individuals can easily be included in intention to treat analyses if follow-up data are available, and it is most important that they are included if the reasons for exclusion relate to the treatment that they received (such as drop-outs due to side-effects and poor tolerability of treatment). Occasionally full details of the outcomes of those excluded during the trial may be mentioned in the text of the report, but in many situations assumptions must be made regarding their fate. By inventive use of sensitivity analysis (using *worst case*, *best case* and *most likely case* scenarios for every trial) it is possible to assess the influence of these excluded cases on the final results. The issue is more problematic for continuous outcomes, where there is a continuum of possible scenarios for every excluded participant.

Other problems can occur when trials have no events in one or both arms. In these situations inverse variance, Mantel–Haenszel and DerSimonian and Laird methods require the addition of a small quantity (usually 0.5) to the cell counts to avoid division by zero errors. (Many software implementations of these methods automatically add this correction to all cell counts regardless of whether it is strictly needed.) When both groups have event rates of zero (there being no events in either arm) odds ratios and relative risks are undefined, and such trials must be excluded from the analysis. The risk difference in such situations is zero, so the trials will still contribute to the analysis. However, both inverse variance and Mantel–Haenszel methods perform poorly when event rates are very low, underestimating both treatment effects and statistical significance.[18] Peto's odds ratio method gives more accurate estimates of the treatment effects and their confidence intervals providing the sample sizes of the arms in the trials are not severely unbalanced.



## Other methods of meta-analysis

The meta-analytical methods described above are straightforward and easy to implement in most statistical software and spreadsheet packages. Other more complex methods exist, and are implemented in specialist statistical software packages, such as Stata (see Chapter 18), SAS, and StatXact (see Box 17.1 in Chapter 17). Maximum likelihood logistic regression can also be used to perform fixed effect meta-analysis, and will give similar answers to the Mantel–Haenszel and inverse variance methods provided sample sizes are large. Maximum likelihood (ML) and restricted maximum likelihood (REML) estimation techniques also enable better estimation of the between trial variance $\tau^2$,[26] and can estimate additional parameters, such as the standard error of $\tau^2$.[27] Bayesian methods (see Chapter 2) can incorporate prior information from other sources, such as is available from qualitative research,[28] whilst exact methods[29] use challenging permutation algorithms to compute treatment effects and P values.

### Case study 1 : support from caregivers during childbirth

Descriptive studies of women's childbirth experiences have suggested that women appreciate advice and information from their caregivers, comfort measures and other forms of tangible assistance to cope with labour, and the continuous presence of a sympathetic person. A systematic review included studies that evaluated the effects of intrapartum support from caregivers on a variety of childbirth outcomes, medical as well as psychosocial.[31] One outcome included in the review was the use of epidural anaesthesia during delivery. Six trials reported this outcome, four from America and two from Europe. In four of the six trials husbands, partners or other family members were also usually present. The person providing the support intervention was variously described in the trials as a midwife, nurse, *monitrice* and a *doula*. The results of the six studies are given in Table 15.3.

Ten alternative methods have been described in this chapter which can be used to perform a meta-analysis of these data. The results are shown in Table 15.4.

Table 15.3 Rates of use of epidural anaesthesia in trials of caregiver support.

| Trial | Caregiver present Epidurals / *N* | Standard Care Epidurals / *N* |
|---|---|---|
| Bréart 1992 (France) | 55/133 | 62/131 |
| Bréart 1992 (Belgium) | 281/656 | 319/664 |
| Gagnon 1997 (Canada) | 139/209 | 142/204 |
| Hodnett 1989 (Canada) | 30/72 | 43/73 |
| Kennell 1991 (USA) | 24/212 | 55/200 |
| Langer 1998 (Mexico) | 205/361 | 303/363 |

Table 15.4 Results of meta-analyses of epidural rates from trials of caregiver support.

| Method | Estimate of effect (95% CI) | Significance of effect | Test for heterogeneity |
|---|---|---|---|
| Odds ratio | | | |
| Peto | 0·59 (0·51 to 0·69) | $z = 7·05$, $P < 0·0001$ | $\chi^2_5 = 38·5$, $P < 0·001$ |
| Mantel–Haenszel | 0·59 (0·51 to 0·69) | $z = 6·98$, $P < 0·0001$ | $\chi^2_5 = 38·9$, $P < 0·001$ |
| Inverse variance | 0·60 (0·52 to 0·70) | $z = 6·70$, $P < 0·0001$ | $\chi^2_5 = 38·8$, $P < 0·001$ |
| DerSimonian and Laird | 0·54 (0·34 to 0·85) | $z = 2·64$, $P = 0·008$ | |
| Risk ratio | | | |
| Mantel–Haenszel | 0·79 (0·74 to 0·85) | $z = 6·95$, $P < 0·0001$ | $\chi^2_5 = 29·8$, $P < 0·001$ |
| Inverse variance | 0·80 (0·75 to 0·85) | $z = 7·14$, $P < 0·0001$ | $\chi^2_5 = 29·7$, $P < 0·001$ |
| DerSimonian and Laird | 0·77 (0·64 to 0·92) | $z = 2·93$, $P = 0·003$ | |
| Risk difference | | | |
| Mantel–Haenszel | −0·117 (−0·149 to −0·085) | $z = 7·13$, $P < 0·0001$ | $\chi^2_5 = 33·1$, $P < 0·001$ |
| Inverse variance | −0·127 (−0·158 to −0·095) | $z = 7·86$, $P < 0·0001$ | $\chi^2_5 = 32·7$, $P < 0·001$ |
| DerSimonian and Laird | −0·124 (−0·211 to −0·038) | $z = 2·81$, $P = 0·005$ | |

There are some notable patterns in the results in Table 15.4. First, there is substantial agreement between Peto, Mantel–Haenszel and inverse variance methods for odds ratios and for risk ratios, indicating that in this instance, where trials are large and event rates reasonably high, the choice of the fixed effect weighting method makes little difference to the results. Secondly, there are substantial differences between treatment effects expressed as odds ratios and risk ratios. Considering the Mantel–Haenszel results, the reduction in the odds of having an epidural with additional caregiver support is 41% ($100 \times (1 - 0·59)$), whilst the relative risk reduction is 21% ($100 \times (1 - 0·79)$), only around half the size. Where events are common (around half the women in the standard care groups received epidurals) odds and risks are very different, and care must be taken to ensure that a reader of the review is not misled into believing that benefits of intervention are larger than is truly the case.[32]

The tests of homogeneity were also statistically significant for odds ratios, risk ratios and risk differences. As a result the confidence intervals for the DerSimonian and Laird random effects estimates are wider than those calculated from fixed effect models. The estimates of the benefit of treatment expressed as relative risks and odds ratios also increase as the random effects model attributes proportionally greater weight to the smallest trials, which in this example report larger relative benefits of treatment.

The report mentions that the benefit of the intervention may be expected to be greater when partners or other family members are absent at the birth,





which could explain the significant heterogeneity. Stratifying the analysis into 'accompanied' and 'unaccompanied' trials (partners were absent in the Kennell and Langer trials) does explain a large proportion of the heterogeneity. The relative risk reduction in the four trials where partners were also present is 11% (95% CI: 3 to 18%; heterogeneity test $\chi^2_3 = 2·92$, $P = 0·4$), whilst in the two trials where partners were absent it is 36% (95% CI: 29 to 43%; heterogeneity test $\chi^2_1 = 5·39$, $P = 0·02$). The differences between the subgroups is highly statistically significant (heterogeneity explained by the subgroups $\chi^2_1 = 29·8 - (2·92 + 5·39) = 21·5$: $P < 0·0001$).

The conclusion of the analysis is that the presence of a caregiver is of benefit in reducing the use of epidural analgesia in all situations, but that the benefit seems much greater in situations were partners are usually absent.

### Case study 2 : Assertive community treatment for severe mental disorders

Assertive community treatment (ACT) is a multidisciplinary team based approach to care for the severely mentally ill in the community. It is assertive in that it continues to offer services to uncooperative and reluctant people, and places emphasis on treatment compliance with the aim of improving mental state. A systematic review comparing ACT to standard care (which consists of outpatient appointments and assistance from community mental health teams) found three trials that assessed mental state at around 12 months.[33] The results are shown in Table 15.5.

Table 15.5 Trials comparing mental state at 12 months between ACT and standard care.

| Trial | ACT | | Standard care | | Assessment scale |
|---|---|---|---|---|---|
| | *N* | Mean (SD) | *N* | Mean (SD) | |
| Audini (London) | 30 | 41·4 (14·0) | 28 | 42·3 (12·4) | Brief psychiatric rating scale |
| Morse (St Louis) | 37 | 0·95 (0·76) | 35 | 0·89 (0·65) | Brief symptom inventory |
| Lehman (Baltimore) | 67 | 4·10 (0·83) | 58 | 3·80 (0·87) | Colorado symptom index |

All three trials have used different scoring systems so the trial results require standardisation to a common scale before they can be combined. In addition, high scores on the Colorado symptom index indicate good outcomes, whilst high scores on the other two scales are poor outcomes, so the direction of the results for Lehman must be reversed before the data can be combined (this is easily accomplished by multiplying the means by −1). Six alternative models for combining the data were described above, and their results are given in Table 15.6.

In this situation, the differences between the analyses are minimal.

Table 15.6 Results of meta-analyses of mental status from trials of ACT.

| Method | Estimate of effect (95% CI) | Significance of effect | Test for heterogeneity |
|---|---|---|---|
| Fixed effect models | | | |
| Cohen's *d* | −0·16 (−0·41 to 0·08) | $z = 1·29$, P = 0·20 | $\chi^2_2 = 2·34$, P = 0·31 |
| Hedges' adjusted *g* | −0·16 (−0·41 to 0·08) | $z = 1·29$, P = 0·20 | $\chi^2_2 = 2·31$, P = 0·32 |
| Glass's Δ | −0·16 (−0·40 to 0·09) | $z = 1·24$, P = 0·21 | $\chi^2_2 = 2·28$, P = 0·32 |
| Random effects models | | | |
| Cohen's *d* | −0·15 (−0·42 to 0·12) | $z = 1·12$, P = 0·26 | |
| Hedges' adjusted *g* | −0·15 (−0·42 to 0·11) | $z = 1·12$, P = 0·26 | |
| Glass's Δ | −0·15 (−0·42 to 0·12) | $z = 1·10$, P = 0·27 | |

Cohen's *d* and Hedges' adjusted *g* will only differ in very small samples. Glass's Δ will differ when the standard deviations vary substantially between treatment and control groups, which was not the case here. Very little heterogeneity was observed, so random and fixed effects analyses are very similar. The analysis can conclude that although all trials favoured ACT no significant change in mental status at 12 months was found with ACT. Also benefits of ACT larger than 0·5 standard deviations or more can probably be excluded as they are outside the lower limit of the confidence interval. To express the findings in a more accessible way consider the standard deviations from each of the trials. A change of 0·5 standard deviations can be estimated to be 6–7 points on the brief psychiatric rating scale, 0·45–0·5 points on the brief symptom inventory and 0·4–0·45 points on the Colorado symptom index.

### Case study 3 : effect of reduced dietary sodium on blood pressure

Restricting the intake of salt in diet has been proposed as a method of lowering blood pressure, both in hypertensives and people with normal blood pressure. A systematic review of randomised studies of dietary sodium restrictions compared to control included 56 trials comparing salt lowering diets with control diets.[34] Only trials which assessed salt reduction through measurement of sodium excretion were included. Twenty-eight of the studies recruited hypertensive participants, and 28 recruited normotensive participants; 41 studies used a cross-over design, whilst 15 used a parallel group design.

The focus of interest in these trials is the difference in mean blood pressure (both diastolic and systolic) between the salt reducing diet and the control diet. As all measurements are in the same units (mmHg) the difference in means can be used directly as a summary statistic in the meta-analysis. The trials estimated this difference in mean blood pressure in four different ways:





(i) in a parallel group trial, as the difference in mean final blood pressure between those receiving the salt lowering diet and the control diet
(ii) in a parallel group trial, as the difference in mean change in blood pressure whilst on the diets, between those on the salt lowering diet and those on the control diet
(iii) in a cross-over trial, as the mean within person difference between final blood pressure at the end of the salt lowering diet and at the end of the control diet
(iv) in a cross-over trial, as the mean within person difference in the change in blood pressure whilst on the salt lowering diet compared to the control diet.

Results from these four different designs all estimate the same summary measure. However, it is likely that trials that use within person changes are more efficient than those that use final values, and that those which use cross-over designs are more efficient than those recruiting parallel groups. These differences are encapsulated in the standard errors of the estimates in differences in mean blood pressure between the two diets, provided appropriate consideration is given to the within person pairing of the data for change scores and cross-over trials in the analysis of those trials. As the standard inverse variance approach to combining trials uses weights inversely proportional to the square of these standard errors, it copes naturally with data of these different formats, so that the trials are given appropriate weightings according to the relative efficiency of their designs.

The authors of the review reported that they had had to use a variety of techniques to estimate these standard errors, as they were not always available in the original reports. If necessary standard errors can be derived directly from standard deviations, confidence intervals, *t* values and exact P values. However, when paired data (both for change scores and cross-over trials) are used it is occasionally necessary to make an assumption about the within participant correlation between two time-points if the analysis presented mistakenly ignores the pairings. Similarly, when results are reported simply either as significant or non-significant, particular P values must be assumed from which the standard errors can be derived. Such problems are common in meta-analyses of continuous data due to the use of inappropriate analyses and the poor standard of presentation commonly encountered in published trial reports.

Meta-analyses were undertaken separately for the trials in normotensive and hypertensive groups, and for systolic and diastolic blood pressure. The results are given in the Table 15.7.

The analysis shows statistically significant reductions of around 5–6 mmHg in systolic blood pressure in hypertensive participants, with a

Table 15.7 Impact of salt lowering diets on systolic and diastolic blood pressure.

| Method | Estimated difference in blood pressure reduction (95% CI) (diet–control) (mmHg) | Test of overall effect | Test for heterogeneity |
|---|---|---|---|
| *Normotensive trials* | | | |
| Systolic | | | |
| Inverse variance | −1·2 (−1·6 to −0·8) | $z = 6·4$, $P < 0·001$ | $\chi^2_{27} = 75·1$, $P < 0·001$ |
| DerSimonian and Laird | −1·7 (−2·4 to −0·9) | $z = 4·2$, $P < 0·001$ | |
| Diastolic | | | |
| Inverse variance | −0·7 (−1·0 to −0·3) | $z = 3·4$, $P = 0·001$ | $\chi^2_{27} = 56·1$, $P = 0·001$ |
| DerSimonian and Laird | −0·5 (−1·2 to 0·1) | $z = 1·63$, $P = 0·10$ | |
| *Hypertensive trials* | | | |
| Systolic | | | |
| Inverse variance | −5·4 (−6·3 to −4·5) | $z = 12·0$, $P < 0·001$ | $\chi^2_{27} = 99·2$, $P < 0·001$ |
| DerSimonian and Laird | −5·9 (−7·8 to −4·1) | $z = 6·4$, $P < 0·001$ | |
| Diastolic | | | |
| Inverse variance | −3·5 (−4·0 to −2·9) | $z = 11·6$, $P < 0·001$ | $\chi^2_{27} = 57·3$, $P = 0·001$ |
| DerSimonian and Laird | −3·8 (−4·8 to −2·9) | $z = 8·0$, $P < 0·001$ | |

smaller reduction in diastolic blood pressure. The size of the reductions observed in normotensive participants was much smaller, the differences between the hypertensive and normotensive subgroups being statistically significant for both systolic ($z = 4·12$: $P < 0·0001$) and diastolic ($z = 5·61$: $P < 0·0001$) measurements. The confidence intervals for the DerSimonian and Laird random effects analyses for all reductions are much wider than those of the inverse variance fixed effect analyses, reflecting the significant heterogeneity detected in all analyses. The authors investigated this further using methods of meta-regression (see Chapters 9 and 11 and Stata command `metareg` in Chapter 18) and showed that the heterogeneity between trials could in part be explained by a relationship between the reduction in blood pressure and the reduction in salt intake achieved in each trial. This regression analysis, and the possible presence of bias, is discussed in Chapter 11.

On the basis of these analyses the authors concluded that salt-lowering diets may have some worthwhile impact on blood pressure for hypertensive people but not for normotensive people, contrary to current recommendations for universal dietary salt reduction.





## Discussion

We have outlined a variety of methods for combining results from several studies in a systematic review. There are three aspects of choosing the right method for a particular meta-analysis: identifying the data type (binary, continuous, time to event), choosing an appropriate summary statistic, and selecting a weighting method for combining the studies, as summarised below and in Box 15.1.

What is clearly required from a summary statistic is that it is as stable as possible over the trials in the meta-analysis and subdivisions of the population to which the treatment will be applied. The more consistent it is, the greater is the justification for expressing the effect of treatment in a single summary number.[30] A second consideration is that the summary statistic should be easily understood and applied by those using the review. For binary data the choice is not straightforward, and no measure is best in all circumstances. These issues are considered in detail in Chapter 16.

Selection of summary statistics for continuous data is principally determined by whether trials all report the outcome using the same scale. If

### Box 15.1 Considerations in choosing a method of meta-analysis

**Choice of summary statistic depends upon:**

(a) the type of data being analysed (binary, continuous, time-to-event)
(b) the consistency of estimates of the treatment effect across trials and subgroups
(c) the ease of interpretation of the summary statistic.

**Choice of weighting method depends upon:**

(a) the reliability of the method when sample sizes are small
(b) the reliability of the method if events are very rare
(c) the degree of imbalance in allocation ratios in the trials.

**Consideration of heterogeneity can affect:**

(a) whether a meta-analysis should be considered, depending on the similarity of trial characteristics
(b) whether an overall summary can have a sensible meaning, depending on the degree of disagreement observed between the trial results
(c) whether a random effects method is used to account for extra between-trial variation and to modify the significance and precision of the estimate of overall effect
(d) whether the impact of other factors on the treatment effect can be investigated using stratified analyses and methods of meta-regression.

this is not the case use of a weighted mean difference method would be erroneous. However, the standardised mean difference method can be used for either circumstance. Differences in results between these two methods can reflect differences in both the treatment effects calculated for each study, and the study weights. Interpretation of a weighted mean difference is easier than that of a standardised mean difference as it is expressed in natural units of measurement rather than standard deviations.

For all types of outcome, the choice of weighting scheme involves deciding between random and fixed effect models, and for fixed effect analyses of binary outcome measures, between inverse variance, Mantel–Haenszel and Peto methods. There is no consensus regarding the choice of fixed or random effects models, although they differ only in the presence of heterogeneity, when the random effects result will usually be more conservative. It is important to be aware of circumstances in which Mantel–Haenszel, inverse variance and Peto methods give erroneous results when deciding between them. Inverse variance methods are poor when trials are small and are rarely preferable to Mantel–Haenszel methods. Both Mantel–Haenszel and inverse variance methods are poor when event rates are very low, and Peto's method can be misleading when treatment effects are large, and when there are severely unequal numbers of participants in treatment and control groups in some or all of the trials.[17] Some of these points are illustrated in the case studies discussed above.

It is important to note that none of the analyses described can compensate for any publication bias (see Chapter 11), nor can they account for bias introduced through poor trial design and execution.

## Acknowledgements

We are grateful to Julian Midgley for allowing us access to the data on which the third case study is based.

1 Deeks JJ. Systematic reviews of published evidence: miracles or minefields? *Ann Oncol* 1998;**9**:703–9.
2 Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1992;**102**:305S–311S.
3 Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;**309**:1351–5.
4 Haldane JBS. The estimation and significance of the logarithm of a ratio of frequencies. *Ann Hum Genet* 1955;**20**:309–14.
5 Morris JA, Gardner MJ. Epidemiological studies. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with confidence,* 2nd edn. London: BMJ Books, 2000:57–72.
6 Newcombe RG, Altman DG. Proportions and their differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with confidence,* 2nd edn. London: BMJ Books, 2000:45–56.
7 Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis* 1985;**27**:335–71.





8 Sinclair JC, Bracken MB. *Effective care of the newborn infant.* Oxford: Oxford University Press, 1992: chapter 2.
9 Hedges LV, Olkin I. *Statistical methods for meta-analysis.* San Diego: Academic Press 1985: chapter 5.
10 Rosenthal R. Parametric measures of effect size. In: Cooper H, Hedges LV, eds. *The Handbook of research synthesis.* New York: Russell Sage Foundation, 1994.
11 Glass GV. Primary, secondary, and meta-analysis of research. *Educat Res* 1976;5:3–8.
12 Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet* 1991;**338**:1127–30.
13 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;**22**:719–48.
14 Breslow NE, Day NE. Combination of results from a series of 2 × 2 tables; control of confounding. In: *Statistical methods in cancer research, Vol. 1: The analysis of case-control data.* IARC Scientific Publications No.32. Lyon: International Agency for Health Research on Cancer, 1980.
15 Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. *Am J Epidemiol* 1986;**124**:719–23.
16 Greenland S, Robins J. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 1985;**41**:55–68.
17 Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Stat Med* 1990;**9**:247–52.
18 Deeks JJ, Bradburn MJ, Localio R, Berlin J. Much ado about nothing: meta-analysis for rare events [abstract]. *6th Cochrane Colloquium.* Baltimore, MD, 1998.
19 Altman DG. *Practical statistics for medical research.* London: Chapman and Hall, 1991: 379.
20 Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998;**17**:2815–34.
21 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;7:177–88.
22 Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining events rates from clinical trials. *Stat Med* 1989;**8**:141–51.
23 Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analyses of randomized controlled trials. *J Clin Epidemiol* 1997;**50**:683–91.
24 Song F, Glenny A-M, Altman DG. Indirect comparison in evaluating relative efficacy of antimicrobial prophylaxis in colorectal surgery. *Controlled Clin Trials* 2000;**21**:488–97.
25 Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Stat Med* 1995;**14**:2057–79.
26 Normand ST. Meta-analysis: formulating, evaluating, combining and reporting. *Stat Med* 1999;**18**:321–60.
27 Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med* 1996;**15**:619–29.
28 Louis TA, Zelterman D. Bayesian approaches to research synthesis. In: Cooper H, Hedges LV, eds. *The handbook of research synthesis.* New York: Russell Sage Foundation, 1994.
26 Normand ST. Meta-analysis: formulating, evaluating, combining and reporting. *Stat Med* 1999;**18**:321–60.
27 Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med* 1996;**15**:619–29.
28 Louis TA, Zelterman D. Bayesian approaches to research synthesis. In: Cooper H, Hedges LV, eds. *The handbook of research synthesis.* New York: Russell Sage Foundation, 1994.
29 Gart J. Point and interval estimation of the common odds ratio in the combination of 2 × 2 tables with fixed marginals. *Biometrika* 1970;**38**:141–9.
30 Hodnett ED. Caregiver support for women during childbirth (Cochrane Review). In: *The Cochrane Library,* Issue 4. Oxford: Update Software, 1999.
31 Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! *Evidence-Based Med* 1996;1:164–6.
32 Marshall M, Lockwood A. Assertive community treatment for people with severe mental disorders. (Cochrane Review) In: *The Cochrane Library,* Issue 4. Oxford: Update Software, 1999.