9 Why and how sources of heterogeneity should be investigated

SIMON G THOMPSON

Summary points

- Clinical heterogeneity across the studies included in a meta-analysis is likely to lead to some degree of statistical heterogeneity in their results.
- Investigating potential sources of heterogeneity is an important component of carrying out a meta-analysis.
- Appropriate statistical methods for trial characteristics involve weighted regression and should allow for residual heterogeneity.
- Individual patient data give the greatest scope for useful analyses of heterogeneity.
- Caution is required in interpreting results, especially when analyses have been inspired by looking at the available data.
- Careful investigations of heterogeneity in meta-analysis should increase the scientific and clinical relevance of their results.

The purpose of a meta-analysis of a set of clinical trials is rather different from the specific aims of an individual trial. For example a particular clinical trial investigating the effect of serum cholesterol reduction on the risk of ischaemic heart disease tests a single treatment regimen, given for a specified duration to participants fulfilling certain eligibility criteria, using a particular definition of outcome measures. The purpose of a meta-analysis of cholesterol lowering trials is broader – that is, to estimate the extent to which serum cholesterol reduction, achieved by a variety of means, generally influences the risk of ischaemic heart disease. A meta-analysis also attempts to gain greater objectivity, applicability and precision by including all the available evidence from randomised trials that pertain to the issue.[1] Because of the broader aims of a meta-analysis, the trials included usually encompass a substantial variety of specific treatment regimens, types of patients, and outcomes. In this chapter, it is argued that the influence of these clinical differences between trials, or clinical heterogeneity, on the overall results needs to be explored carefully.

The chapter starts by clarifying the relation between clinical heterogeneity and statistical heterogeneity. Examples follow of meta-analyses of observational epidemiological studies of serum cholesterol concentration, and clinical trials of its reduction, in which exploration of heterogeneity was important in the overall conclusions reached. The statistical methods appropriate for investigating sources of heterogeneity are then described in more detail. The dangers of *post hoc* exploration of results and consequent over-interpretation are addressed at the end of the chapter.

Clinical and statistical heterogeneity

To make the concepts clear, it is useful to focus on a meta-analysis where heterogeneity posed a problem in interpretation. Figure 9.1 shows the results of 19 randomised trials investigating the use of endoscopic sclerotherapy for reducing mortality in the primary treatment of cirrhotic patients with oesophageal varices.[2] The results of each trial are shown as odds ratios and 95% confidence intervals, with odds ratios less than unity representing a beneficial effect of sclerotherapy. The trials differed considerably in patient selection, baseline disease severity, endoscopic technique, management of intermediate outcomes such as variceal bleeding, and duration of follow-up.[2] So in this meta-analysis, as in many, there is extensive clinical heterogeneity. There were also methodological differences in the mechanism of randomisation, the extent of withdrawals, and the handling of losses to follow-up.

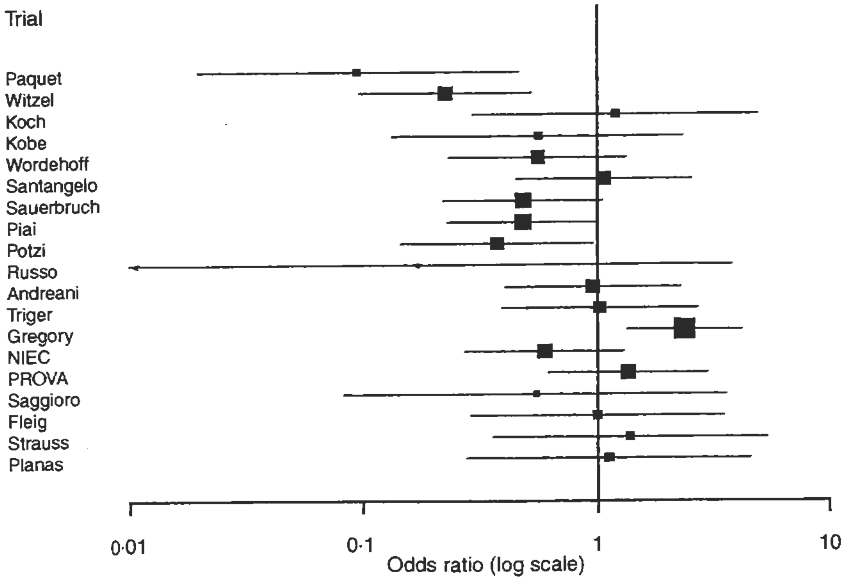

Figure 9.1 Forest plot of odds ratios of death (and 95% confidence intervals) from 19 trials of sclerotherapy. Odds ratios less than unity represent beneficial effects of sclerotherapy. Trials identified by principal author, referenced by Pagliaro *et al.*[2]

It would not be surprising, therefore, to find that the results of these trials were to some degree incompatible with one another. Such incompatibility in quantitative results is termed statistical heterogeneity. It may be caused by known clinical or methodological differences between trials, or may be related to unknown or unrecorded trial characteristics. In assessing the direct evidence of statistical heterogeneity, the imprecision in the estimate of the odds ratio from each trial, as expressed by the confidence intervals in Figure 9.1, has to be taken into account. The statistical question is then whether there is greater variation between the results of the trials than is compatible with the play of chance. As might be surmised from inspection of Figure 9.1, the statistical test (test of homogeneity, see Chapter 15) yielded a highly significant result ($\chi^2_{18} = 43$, $P < 0{\cdot}001$).

In the example of the sclerotherapy trials, the evidence for statistical heterogeneity is substantial. In many meta-analyses, however, such statistical evidence is lacking and the test of homogeneity is non-significant. Yet this cannot be interpreted as evidence of homogeneity (that is, total consistency) of the results of all the trials included. This is not only because a non-significant test can never be interpreted as direct evidence in favour of the null hypothesis of homogeneity,[3] but in particular because tests of homogeneity have low power and may fail to detect as statistically significant even a moderate degree of genuine heterogeneity.[4,5]

We might be somewhat happier to ignore the problems of clinical heterogeneity in the interpretation of the results if direct evidence of statistical heterogeneity is lacking, and more inclined to try to understand the reasons for any heterogeneity for which the evidence is more convincing. However, the extent of statistical heterogeneity, which can be quantified,[6] is more important than the evidence of its existence. Indeed it is reasonable to argue that testing for heterogeneity is largely irrelevant, because the studies in any meta-analysis will necessarily be clinically heterogeneous.[7] The guiding principle should be to investigate the influences of the specific clinical differences between studies rather than rely on an overall statistical test for heterogeneity. This focuses attention on particular contrasts among the trials included, which will be more likely to detect genuine differences – and more relevant to the overall conclusions. For example, in the sclerotherapy trials, the underlying disease severity was identified as being potentially related to the benefits of sclerotherapy observed (see also Chapter 10).[2]

The quantitative summary of the results, for example in terms of an overall odds ratio and 95% confidence interval, is generally considered the most important conclusion from a meta-analysis. For the sclerotherapy trials, the overall odds ratio for death was given as 0·76 with 95% confidence interval 0·61 to 0·94,[2] calculated under the "fixed effect" assumption of homogeneity.[5] A naive interpretation of this would be that sclerotherapy convincingly decreased the risk of death with an odds reduction of around 25%. However, what are the implications of clinical and statistical heterogeneity in the interpretation of this result? Given the clinical heterogeneity, we do not know to which endoscopic technique, to which selection of patients, or in conjunction with what ancillary clinical management such a conclusion is supposed to refer. It is some sort of "average" statement that is not easy to interpret quantitatively in relation to the benefits that might accrue from the use of a specific clinical protocol. In this particular case the evidence for statistical heterogeneity is also overwhelming and this introduces even more doubt about the interpretation of any single overall estimate of effect. Even if we accept that some sort of average or typical[8] effect is being estimated, the confidence interval given is too narrow in terms of extrapolating the results to future trials or patients, since the extra variability between the results of the different trials is ignored.[5]

The answer to such problems is that meta-analyses should incorporate a careful investigation of potential sources of heterogeneity. Meta-analysis can go further than simply producing a single estimate of effect.[9] For example, in a meta-analysis of trials of thrombolysis in the acute phase of myocardial infarction, the survival benefit has been shown to be greater when there is less delay between onset of symptoms and treatment.[10] Quantifying this relation is important in drawing up policy recommendations for the use of thrombolysis in routine clinical practice. More generally, the benefits of trying to understand why differences in treatment effects occur across trials often outweigh the potential disadvantages.[11] The same is true for differences in exposure-disease associations across epidemiological studies.[12] Such analyses, often called meta-regressions,[13] can in principle be extended, for example in a meta-analysis of clinical trials, to investigate how a number of trial or patient characteristics act together to influence treatment effects (see also Chapters 8, 10 and 11 for more discussion of the use of regression models in meta-analysis). Two examples of the benefits of applying such an approach in published meta-analyses follow.

Serum cholesterol concentration and risk of ischaemic heart disease

An extreme example of heterogeneity was evident in a 1994 review[14] of the 10 largest prospective cohort studies of serum cholesterol concentration and the risk of ischaemic heart disease in men, which included data on 19 000 myocardial infarctions or deaths from ischaemic heart disease. The purpose was to summarise the magnitude of the relation between serum cholesterol and risk of ischaemic heart disease in order to estimate the long term benefit that might be expected to accrue from reduction in serum cholesterol concentrations.

The results from the 10 prospective studies are shown in Figure 9.2. These are expressed as proportionate reductions in risk associated with a reduction in serum cholesterol of 0·6 mmol/l (about 10% of average levels in Western countries), having been derived from the apparently log-linear associations of risk of ischaemic heart disease with serum cholesterol concentration in individual studies. They also take into account the underestimation that results from the fact that a single measurement of serum cholesterol is an imprecise estimate of long term level, sometimes termed

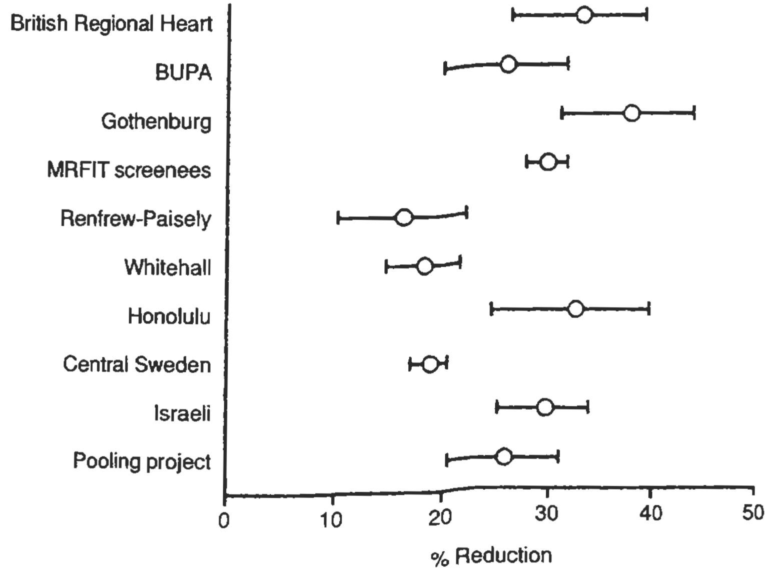

Figure 9.2 Percentage reduction in risk of ischaemic heart disease (and 95% confidence intervals) associated with 0·6 mmol/l serum cholesterol reduction in 10 prospective studies of men. Studies referenced by Law *et al.*[14]

regression dilution bias.[15] Although all of the 10 studies showed that cholesterol reduction was associated with a reduction in the risk of ischaemic heart disease, they differed substantially in the estimated magnitude of this effect. This is clear from Figure 9.2, and the extreme value that is obtained from an overall test of homogeneity ($\chi^2_9 = 127$, $P < 0{\cdot}001$). This shows that simply combining the results of these studies into one overall estimate is misleading; an understanding of the reasons for the heterogeneity is necessary.

The most obvious cause of the heterogeneity relates to the ages of the participants, or more particularly the average age of experiencing coronary events during follow-up, since it is well known that the relative risk association of ischaemic heart disease with a given serum cholesterol increment declines with advancing age.[16,17] The data from the 10 studies were therefore divided, as far as was possible from published and unpublished information, into groups according to age at entry.[14] This yielded 26 substudies, the results of which are plotted against the average age of experiencing a coronary event in Figure 9.3. The percentage reduction in risk of ischaemic heart disease clearly decreases markedly with age. This relation can be summarised using a quadratic regression of log relative risk reduction on age, appropriately weighted to take account of the different precisions of each estimate. It was concluded that a decrease in cholesterol concentration of 0.6 mmol/l was associated with a decrease in risk of ischaemic heart disease of 54% at age 40, 39% at age 50, 27% at age 60, 20% at age 70, and 19% at age 80. In fact, there remains considerable evidence of heterogeneity in Figure 9.3 even from this summary of results ($\chi^2_{23} = 45$, $P = 0{\cdot}005$), but it is far less extreme than the original heterogeneity evident before considering age (Figure 9.2).

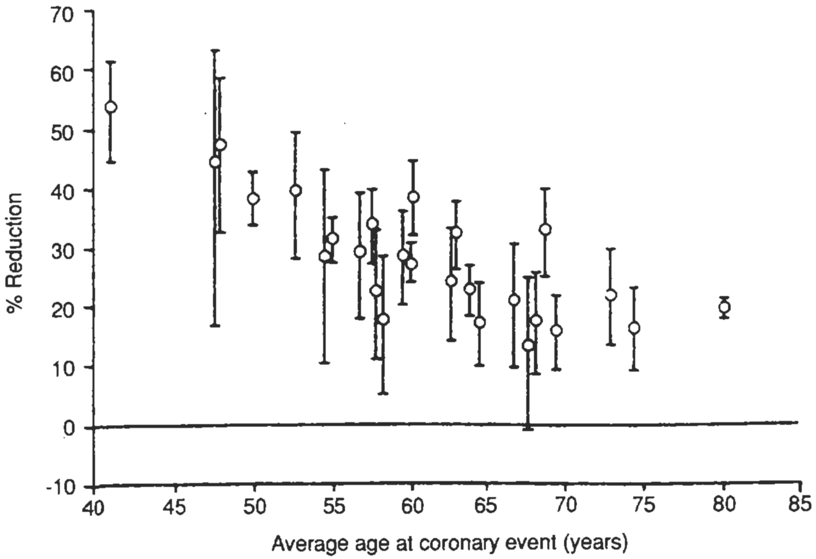

Figure 9.3 Percentage reduction in risk of ischaemic heart disease (and 95% confidence intervals) associated with 0·6 mmol/l serum cholesterol reduction, according to average age of experiencing a coronary event.

The effect on the conclusions brought about by considering age are crucial, for example in considering the impact of cholesterol reduction in the population. The proportionate reductions in the risk of ischaemic heart disease associated with reduction in serum cholesterol are strongly age-related. The large proportionate reductions in early middle age cannot be extrapolated to old ages, at which more modest proportionate reductions are evident. In meta-analyses of observational epidemiological studies, such investigation of sources of heterogeneity may often be a principal rather than subsidiary aim.[18] Systematic reviews of observational studies are discussed in detail in Chapters 12–14.

Serum cholesterol reduction and risk of ischaemic heart disease

The randomised controlled trials of serum cholesterol reduction have been the subject of a number of meta-analyses[14,19,20] and much controversy. In conjunction with the review of the 10 prospective studies just described, the results of 28 randomised trials available in 1994 were summarised;[14] this omits the results of trials of serum cholesterol reduction, notably those using statins, that have become available more recently. The aim was to quantify the effect of serum cholesterol reduction on the risk of ischaemic heart disease in the short term, the trials having an average duration of about five years. There was considerable clinical heterogeneity between the trials in the interventions tested (different drugs, different diets, and in one case surgical intervention using partial ileal bypass grafting), in the duration of the trials (0·3–10 years), in the average extent of serum cholesterol reduction achieved (0·3–1·5 mmol/l), and in the selection criteria for the patients such as pre-existing disease (for example, primary or secondary prevention trials) and level of serum cholesterol concentration at entry. As before it would seem likely that these substantial clinical differences would lead to some heterogeneity in the observed results.

Forest plots such as in Figure 9.1, are not very useful for investigating heterogeneity. A better diagram for this purpose was proposed by Galbraith,[21] and is shown for the cholesterol lowering trials in Figure 9.4. For each trial the ratio of the log odds ratio of ischaemic heart disease to its

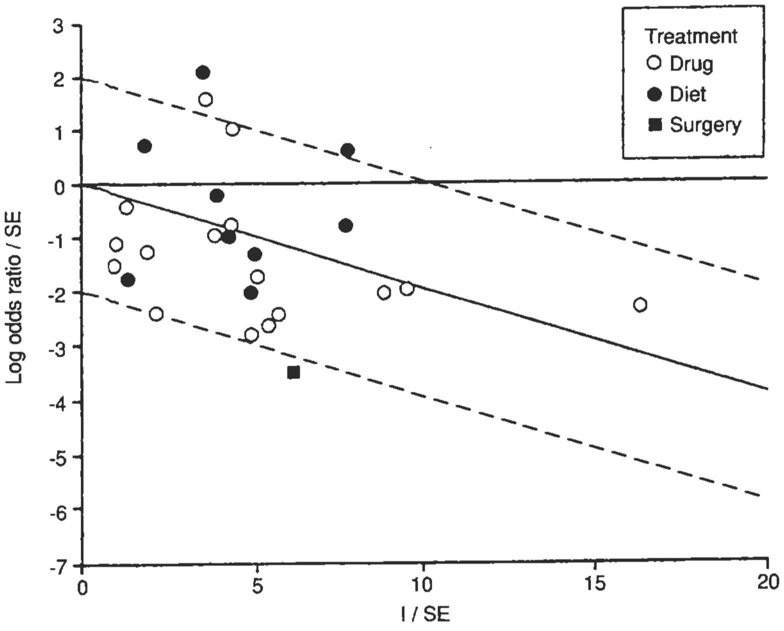

Figure 9.4 Galbraith plot of odds ratios of ischaemic heart disease in 28 trials of serum cholesterol reduction (see text for explanation). Two trials were omitted because of no events in one group.

standard error (the *z*-statistic) is plotted against the reciprocal of the standard error. Hence the least precise results from small trials appear towards the left of the figure and results from the largest trials towards the right. An overall log odds ratio is represented by the slope of the solid line in the figure; this is an unweighted regression line constrained to pass through the origin. The dotted lines are positioned two units above and below the solid line and delimit an area within which, in the absence of statistical heterogeneity, the great majority (that is, about 95%) of the trial results would be expected to lie. It is thus interesting to note the characteristics of those trials which lie near or outside these dotted lines. For example, in Figure 9.4, there are two dietary trials that lie above the upper line and showed apparently adverse effects of serum cholesterol reduction on the risk of ischaemic heart disease. One of these trials achieved only a very small cholesterol reduction while the other had a particularly short duration.[22] Conversely the surgical trial, below the bottom dotted line and showing a large reduction in the risk of ischaemic heart disease, was both the longest trial and the one that achieved the greatest cholesterol reduction.[22] These observations add weight to the need to investigate heterogeneity of results according to extent and duration of cholesterol reduction.

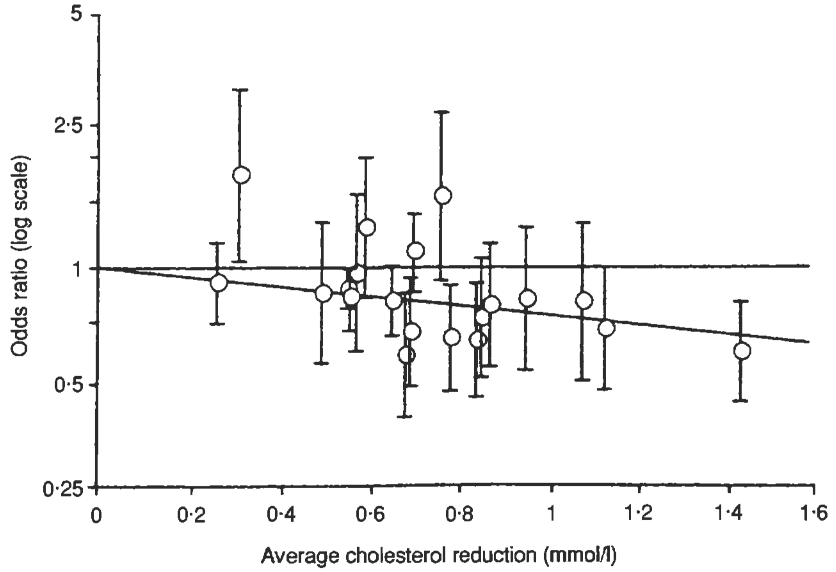

Figure 9.5 Odds ratios of ischaemic heart disease (and 95% confidence intervals) according to the average extent of serum cholesterol reduction achieved in each of 28 trials. Overall summary of results indicated by sloping line. Results of the nine smallest trials have been combined.

Figure 9.5 shows the results according to average extent of cholesterol reduction achieved. There is very strong evidence (P = 0·002) that the proportionate reduction in the risk of ischaemic heart disease increases with the extent of average cholesterol reduction; the appropriate methods for this analysis are explained in the next section. A suitable summary of the trial results, represented by the sloping line in Figure 9.5, is that the risk of ischaemic heart disease is reduced by an estimated 18% (95% confidence interval 13 to 22%) for each 0.6 mmol/l reduction in serum cholesterol concentration.[22] Obtaining data subdivided by time since randomisation[14] to investigate the effect of duration was also informative (Figure 9.6). Whereas the reduction in the risk of ischaemic heart disease in the first two years was rather limited, the reductions thereafter were around 25% per 0·6 mmol/l reduction. After extent and duration of cholesterol reduction were allowed for in this way, the evidence for further heterogeneity of the results from the different trials was limited (P = 0·11). In particular there was no evidence of further differences in the results between the drug and the dietary trials, or between the primary prevention and the secondary prevention trials.[14,22]

This investigation of heterogeneity was again crucial to the conclusions

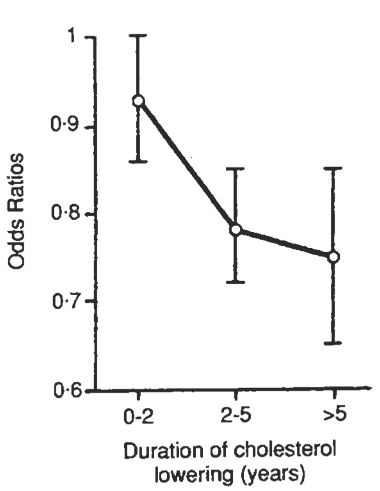

Figure 9.6 Odds ratios of ischaemic heart disease (and 95% confidence intervals) per 0·6 mmol/l serum cholesterol reduction in 28 trials, according to the duration of cholesterol lowering.

reached. The analysis quantified how the percentage reduction in the risk of ischaemic heart disease depends both on the extent and the duration of cholesterol reduction. Meta-analyses ignoring these factors may well be misleading. It also seems that these factors are more important determinants of the proportionate reduction in ischaemic heart disease than the mode of intervention or the underlying risk of the patient.

Statistical methods for investigating sources of heterogeneity

How should analyses such as those described above be carried out? To simplify terminology we consider treatment effects in trials, but the same methods are appropriate for investigating heterogeneity of effects in observational epidemiological studies. We focus on meta-regression, where the aim is to investigate whether a particular covariate or characteristic, with a value defined for each trial in the meta-analysis, is related to the extent of treatment benefit. Figures 9.3 and 9.5 in the meta-analyses discussed above are examples of such analyses. The statistical methods described below are discussed in more detail elsewhere,[23] and can be extended to consider simultaneously the effects of more than one covariate (see also Chapter 11).

The simplest form of analysis assumes that the observed treatment effects in each trial, say log odds ratios, are normally distributed. In the same way that calculating a single overall summary of effect in a meta-analysis takes into account the precision of the estimate in each study,[24] an analysis of a particular covariate as a source of heterogeneity in meta-analysis should be based on weighted regression. The weight that applies to each study is equal to the inverse of the variance of the estimate for that study. This variance has two components: the within-trial variance and the between-trial variance. For example, in the case of log odds ratios, the within-trial variance is simply estimated as the sum of the reciprocal cell counts in the 2 × 2 table[25] (see Chapter 15). The between-trial variance represents the residual heterogeneity in treatment effects, that is the variability between trial results which is not explained by the covariate. Analyses which assume that the between-trial variance is zero, where weighting is therefore simply according to the within-trial variance, correspond to a "fixed effect" analysis. In general, it is an unwarranted assumption that all the heterogeneity is explained by the covariate, and the between-trial variance should be included as well, corresponding to a "random effects" analysis[26] (see Chapter 15). The same arguments apply here as when estimating a single overall treatment effect, ignoring sources of heterogeneity.[24]

To be explicit, consider the analysis presented in Figure 9.5. Here there are 28 trials, which we index by $i = 1...28$. For the ith trial, we denote the observed log odds ratio of ischaemic heart disease by y_i, its estimated within-trial variance by v_i, and the extent of serum cholesterol reduction in mmol/l by x_i. The linear regression of log odds ratios on extent of cholesterol reduction can be expressed as $y_i = \alpha + \beta x_i$; here we are not forcing the regression through the origin as in Figure 9.5, and α represents the intercept of the regression line. The purpose of the analysis is to provide estimates of α and β, together with their standard errors. An additional point of interest is the extent to which the heterogeneity between results is reduced by including the covariate. The weights for the regression are equal to $1/(v_i + \tau^2)$, where τ^2 is the residual heterogeneity variance. There are a number of ways of estimating τ^2, amongst which a restricted maximum likelihood estimate is generally recommended.[23] Programs to carry out such weighted regression analyses are available in the statistical package STATA[27] (see Chapter 18). Note that these analyses are not the same as usual weighted regression where weights are inversely proportional (rather than equal) to the variances.

Table 9.1 presents results from two weighted regressions, the first assuming that there is no residual heterogeneity ($\tau^2 = 0$) and the second allowing the extent of residual heterogeneity to be estimated. The first analysis provides no evidence that the intercept α is non-zero, and convincing evidence that the slope β is negative (as in Figure 9.5). However the estimate of τ^2 in the second analysis is positive, indicating at least some residual heterogeneity. In fact, about 85% of the heterogeneity variance of results in a simple meta-analysis is explained by considering the extent of

Table 9.1 Estimates of the linear regression relationship between log odds ratio of ischaemic heart disease and extent of serum cholesterol reduction (mmol/l) in 28 randomised trials, obtained by different methods (from Thompson and Sharp[23]).

Method	Residual heterogeneity	Estimates (SEs)		Residual heterogeneity variance (τ^2)
		Intercept (α)	Slope (β)	
Weighted regression:				
	None	0·121 (0·097)	–0·475 (0·138)	0
	Additive[a]	0·135 (0·112)	–0·492 (0·153)	0·005
Logistic regression:				
	None	0·121 (0·097)	–0·476 (0·137)	0
	Additive[b]	0·148 (0·126)	–0·509 (0·167)	0·011

[a] Estimated using restricted maximum likelihood.

[b] Estimated using a random effects logistic regression with second order predictive quasi-likelihood[28] in the software MLwiN[29].

cholesterol reduction as a covariate.[23] The standard errors of the estimates of α and β are quite markedly increased in this second analysis, even though the estimate of τ^2 is small. This exemplifies the point that it is important to allow for residual heterogeneity, otherwise the precision of estimated regression coefficients may be misleadingly overstated and sources of heterogeneity mistakenly claimed. Indeed in examples where the residual heterogeneity is substantial, the effects of making allowance for it will be much more marked than in Table 9.1.

Intuitive interpretation of the estimate of τ^2 is not straightforward. However, consider the predicted odds ratio of ischaemic heart disease if serum cholesterol were reduced, for example, by 1 mmol/l, that is $\exp(0{\cdot}135 - 0{\cdot}492) = 0{\cdot}70$. Given the heterogeneity between studies expressed by τ^2, and for a 1 mmol/l cholesterol reduction, the 95% range of true odds ratios for different studies is estimated as $\exp(0{\cdot}135 - 0{\cdot}492 \pm 2 \times \sqrt{0{\cdot}005})$, that is 0·61–0·81. The estimated value of τ, $\sqrt{0{\cdot}005} = 0{\cdot}07$ or 7%, can thus be interpreted approximately as the coefficient of variation on the overall odds ratio caused by heterogeneity between studies. This coefficient of variation would apply to the predicted odds ratio for any given reduction in serum cholesterol.

The assumption that estimated log odds ratios can be considered normally distributed and that the variances v_i are known may be inadequate for small trials or when the number of events is small. It is possible to frame the analyses presented above as logistic regressions for binary outcome data to overcome these problems.[23] The results assuming no residual heterogeneity were almost identical to the weighted regression results; the estimates from the second analysis were slightly different because a larger estimate of τ^2 was obtained (Table 9.1). Another extension to the analysis, which is appropriate in principle, is to allow for the imprecision in estimating τ^2. This can be achieved in a fully Bayesian analysis, but again results for the cholesterol trials were similar.[23] In general, the use of logistic regression or fully Bayesian analyses, rather than weighted regression, will probably make very little difference to the results. Only when all the trials are small (when the normality assumption will fail, and the results will not be dominated by other larger trials) or the number of trials is limited (when τ^2 is particularly imprecise) might different results be anticipated. Indeed one advantage of the weighted regression approach is that it can easily be used for treatment effects on scales other than log odds ratios, such as log relative risks or absolute risk differences, which are more interpretable for clinical practice.[30]

The relationship between underlying risk and treatment benefit

It is reasonable to ask whether the extent of treatment benefit relates to the underlying risk of the patients in the different trials included in a meta-analysis.[31,32] Underlying risk is a convenient summary of a number of characteristics which may be measurable risk factors but for which individual patient data are not available from some or all of the trials. Here it is atural to plot the treatment effect in each trial against the risk of events observed in the control group. Returning to the sclerotherapy meta-analysis introduced at the beginning of the chapter (Figure 9.1), such a plot is shown in Figure 9.7. Each trial is represented by a circle, the area of which represents the trial precision, so trials which contribute more information are represented by larger circles. A weighted regression line, according to the methods of the previous section, is superimposed and gives strong evidence of a negative association ($P < 0{\cdot}001$). A naive interpretation of the line would claim that the treatment effect increases (lower odds ratio) with increasing proportion of events in the control group, and that underlying risk is a significant source of heterogeneity. Furthermore, there is a temptation to use the point T in Figure 9.7 to define a cut-off value of risk in the control group and conclude that treatment is effective (odds ratio below 1) only in patients with an underlying risk higher than this value. As discussed in Chapters 8 and 10, these conclusions are flawed and seriously misleading. The reason for this stems from regression to the mean, since the outcome in the control group is being related to the treatment effect, a quantity which itself includes the control group outcome.[31,33–35] Statistical approaches that overcome this problem are described in Chapter 10.

To a clinician, the "underlying risk" of a patient is only known through certain measured characteristics. So a clinically more useful, and statistically less problematic, alternative to these analyses is to relate treatment

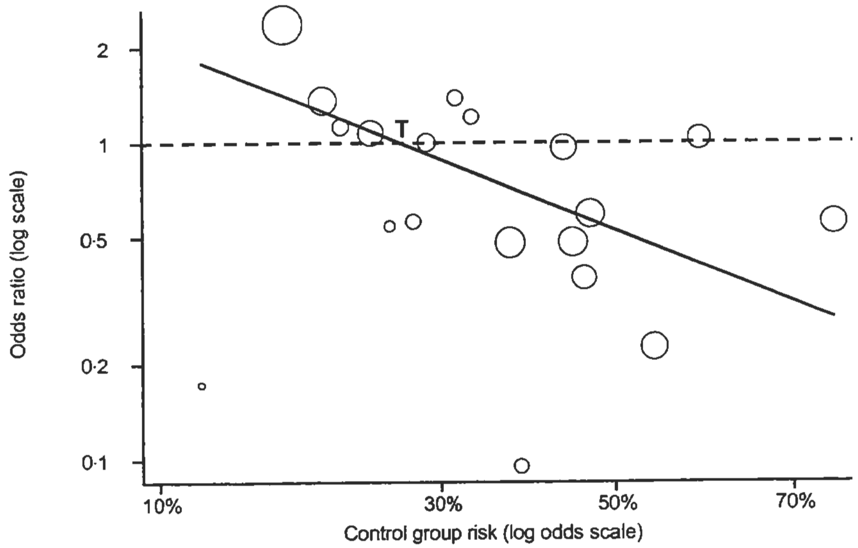

Figure 9.7 Treatment effect versus percentage of events in control group for 19 trials of sclerotherapy. The area of each circle is inversely proportional to the variance of the estimated treatment effect in the trial.

benefit to measurable patient characteristics. This is one of the advantages of individual patient data, as discussed in Chapter 6 and below.

The virtues of individual patient data

Meta-regression using trial-level characteristics can only partially address issues of heterogeneity. The aspects that can be investigated as sources of heterogeneity in such analyses are limited to characteristics of each trial as a whole, for example relating to treatment regimens. Furthermore, analyses using averages of patient characteristics in each trial (such as the mean age of all the patients) can give a misleading impression of the relation for individual patients. This is as a result of the so-called ecological fallacy, whereby the relation with treatment benefit may be different across trials as compared to within trials.[13,36] Clinically more useful information comes from analyses which relate the extent of treatment benefit to individual patient characteristics. As discussed in Chapter 6, meta-analysis based on individual patient data, rather than summary data obtained from publications, has many advantages.[37] Amongst these is the ability to carry out a more thorough and extensive investigation of sources of heterogeneity, since subdivisions according to patients' characteristics can be made within trials and these results combined across trials.

Large individual patient data meta-analyses, undertaken under the auspices of collaborative groups of researchers, have this potential. Even such analyses should allow for the possibility of residual heterogeneity of treatment effects not explained by the patient characteristics available. In practice, however, there may be no great difference between those who advocate a fixed effect approach[8] and those who would be more cautious[5,38,39] when it comes to undertaking particular meta-analyses. For example, a large-scale overview of early breast cancer treatment,[40] carried out ostensibly with a fixed effect approach, included an extensive investigation of heterogeneity according to type and duration of therapy, dose of drug, use of concomitant therapy, age, nodal status, oestrogen receptor status, and outcome (recurrence or mortality).

Exactly how such analyses should be carried out needs further development. For example, assumptions of linearity of covariate effects or normality of the residual variation between trial results can be difficult to assess in practice.[7] The analysis can be viewed as an example of a multilevel model,[41,42] in which information is available at both the trial level and on individuals within trials. Using this structure a general framework for meta-analysis can be proposed, incorporating both trial-level and patient-level covariates, from either a classical or Bayesian viewpoint.[43,44] Some patient characteristics may vary more between trials than within trials; for example, gender would be a within-trial covariate if all the trials in a meta-analysis included both men and women, and a between-trial covariate if trials were either of men alone or of women alone. The strength of inference about how a covariate affects treatment benefit depends on the extent to which it varies within trials. Covariates that vary only between trials have relations with treatment benefit that may be confounded by other trial characteristics. These associations are observational in nature, and do not necessarily have the same interpretation that can be ascribed to treatment comparisons within randomised clinical trials. Covariates that vary within trials are less prone to such biases.

Conclusions

As meta-analysis becomes widely used as a technique for synthesising the results of separate primary studies, an overly simplistic approach to its implementation needs to be avoided. A failure to investigate potential sources of heterogeneity is one aspect of this. As shown in the examples in this chapter, such investigation can importantly affect the overall conclusions drawn, as well as the clinical implications of the review. Therefore the issues of clinical and statistical heterogeneity and how to approach them need emphasis in guidelines and in computer software being developed for conducting meta-analyses, for example by the Cochrane Collaboration.[45]

Although a simple random effects method of analysis[6] may be useful when statistical heterogeneity is present but cannot be obviously explained, the main focus should be on trying to understand any sources of heterogeneity that are present.

There are, however, dangers of over-interpretation induced by attempting to explore possible reasons for heterogeneity, since such investigations are usually inspired, at least to some extent, by looking at the results to hand.[11] Moreover apparent, even statistically significant, heterogeneity may always be due to chance and searching for its causes would then be misleading. The problem is akin to that of subgroup analyses within an individual clinical trial.[46] However the degree of clinical heterogeneity across different trials is greater than that within individual trials, and represents a more serious problem. Guidelines for deciding whether to believe results that stem from investigating heterogeneity depend on, for example, the magnitude and statistical significance of the differences identified, the extent to which the potential sources of heterogeneity have been specified in advance, and indirect evidence and biological considerations which support the investigation.[47] These problems in meta-analysis are greatest when there are many clinical differences but only a small number of trials available. In such situations there may be several alternative explanations for statistical heterogeneity, and ideas about sources of heterogeneity can be considered only as hypotheses for evaluation in future studies.

Although clinical causes of heterogeneity have been focused on here, it is important to recognise that there are other potential causes. For example, statistical heterogeneity may be caused by publication bias[48] whereby, amongst small trials, those with dramatic results may more often be published (see Chapter 3). Statistical heterogeneity can also be caused by defects of methodological quality,[49] as discussed in detail in Chapter 5. For example, poor methodological quality was of concern in the meta-analysis of sclerotherapy trials[2] discussed at the beginning of this chapter. The evidence for publication bias, or other small study biases, can be explored by regression on a Galbraith plot (such as Figure 9.4) without constraining the intercept through the origin.[50] An equivalent analysis can be undertaken using meta-regression of treatment effects against their standard errors, using the methods of this chapter, which also then allow for possible residual heterogeneity.[23] These and other methods are discussed in detail in Chapter 11. Statistical heterogeneity may also be induced by employing an inappropriate scale for measuring treatment effects, for example using absolute rather than relative differences, or even by early termination of clinical trials for ethical or other reasons.[51]

Despite the laudable attempts to achieve objectivity in reviewing scientific data, considerable areas of subjectivity remain in carrying out systematic reviews. These judgments include decisions about which studies are "relevant", which studies are methodologically sound enough to be included in a statistical synthesis, as well as the issue of whether and how to investigate sources of heterogeneity. Such scientific judgements are as necessary in meta-analysis as they are in other forms of medical research, and skills in recognising appropriate analyses and dismissing overly speculative interpretations are required. In many meta-analyses, however, heterogeneity can and should be investigated so as to increase the scientific understanding of the studies reviewed and the clinical relevance of the conclusions drawn.

Acknowledgements

This chapter draws on material published earlier in the *BMJ*.[52]

1 Dickersin K, Berlin J. Meta-analysis: state-of-the-science. *Epidemiol Rev* 1992;**14**:154–76.
2 Pagliaro L, D'Amico G, Sorensen TIA, *et al.* Prevention of first bleeding in cirrhosis: a meta-analysis of randomised trials of non-surgical treatment. *Ann Intern Med* 1992;**117**:59–70.
3 Altman DG. *Practical statistics for medical research.* London: Chapman and Hall, 1991:167–70.
4 Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. *Stat Med* 1991;**10**:1665–77.
5 Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet* 1991;**338**:1127–30.
6 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177–88.
7 Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998; **17**:841–56.
8 Peto R. Why do we need systematic overviews of randomised trials? *Stat Med* 1987;**6**:233–40.
9 Chalmers TC. Problems induced by meta-analysis. *Stat Med* 1991;**10**:971–80.
10 Boersma E, Maas ACP, Deckers JW, Simoons ML. Early thrombolytic treatment in acute myocardial infarction: reappraisal of the golden hour. *Lancet* 1996;**348**:771–5.
11 Davey Smith G, Egger M, Phillips AN. Meta-analysis: Beyond the grand mean? *BMJ* 1997;**315**:1610–14.
12 Berlin JA. Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *Am J Epidemiol* 1995;**142**:383–7.
13 Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998;**351**:123–7.
14 Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ* 1994;**308**:367–73.
15 MacMahon S, Peto R, Cutler J, *et al.* Blood pressure, stroke, and coronary heart disease. Part I, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 1990;**335**:765–74.
16 Manolio TA, Pearson TA, Wenger NK, Barrett-Connor E, Payne GH, Harlan WR. Cholesterol and heart disease in older persons and women: review of an NHLBI workshop. *Ann Epidemiol* 1992;**2**:161–76.
17 Shipley MJ, Pocock SJ, Marmot MG. Does plasma cholesterol concentration predict mortality from coronary heart disease in elderly people? 18 year follow-up in Whitehall study. *BMJ* 1991;**303**:89–92.
18 Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ* 1998;**316**:140–4.

19 Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *BMJ* 1992;**305**:15–19.

20 Davey Smith G, Song F, Sheldon T. Cholesterol lowering and mortality: the importance of considering initial level of risk. *BMJ* 1993;**306**:1367–73.

21 Galbraith RF. A note on the graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;**7**:889–94.

22 Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Stat Meth Med Res* 1993;**2**:173–92.

23 Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999;**18**:2693–708.

24 Thompson SG. Meta-analysis of clinical trials. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. New York: Wiley, 1998:2570–9.

25 Cox DR, Snell EJ. *Analysis of binary data*, 2nd edn. London: Chapman and Hall, 1989.

26 Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med* 1995; **14**: 395–411.

27 Sharp SJ. Meta-analysis regression. *Stata Tech Bull* 1998;**42**:16–22.

28 Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. *J Roy Statist Soc A* 1996;**159**:505–13.

29 Goldstein H, Rasbash J, Plewis I, *et al. A user's guide to MLwiN*. London: Institute of Education, 1998.

30 McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. *Ann Intern Med* 1997;**126**:712–20.

31 Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *BMJ* 1996;**313**:735–8.

32 Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med* 1998;**17**:1923–42.

33 Egger M, Davey Smith G. Risks and benefits of treating mild hypertension: a misleading meta-analysis? *J Hypertens* 1995;**13**:813–15.

34 McIntosh M. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med* 1996;**15**:1713–28.

35 Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* 1997;**16**:2741–58.

36 Morgenstern H. Uses of ecological analysis in epidemiologic research. *Am J Public Health* 1982;**72**:127–30.

37 Stewart LA, Clarke MJ. Practical methodology of meta-analysis (overviews) using updated individual patient data. *Stat Med* 1995;**14**:2057–79.

38 Meier P. Meta-analysis of clinical trials as a scientific discipline [commentary]. *Stat Med* 1987;**6**:329–31.

39 Bailey KR. Inter-study differences: how should they influence the interpretation and analysis of results? *Stat Med* 1987;**6**:351–8.

40 Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. *Lancet* 1992;**339**:1–15, 71–85.

41 Goldstein H. *Multilevel statistical models*, 2nd edn. London: Edward Arnold, 1995.

42 Stram DO. Meta-analysis of published data using a linear mixed-effects model. *Biometrics* 1996;**52**:536–44.

43 Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2000, in press.

44 Pauler DK, Wakefield J. Modeling and implementation issues in Bayesian meta-analysis. In: Stangl DK, Berry DA, eds. *Meta-analysis in medicine and policy health*. New York: Marcel Dekker 2000, 205–30.

45 Oxman A. Preparing and maintaining systematic reviews. In: Sackett D, ed. *Cochrane Collaboration Handbook, Sect. VI*. Oxford: The Cochrane Collaboration, 1998.

46 Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials. *JAMA* 1991;**266**:93–8.

47 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;**116**:78–84.

48 Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;**337**:867–72

49 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodologic quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273**:408–12.

50 Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**:629–34.

51 Hughes MD, Freedman LS, Pocock SJ. The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* 1992;**48**:41–53.

52 Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;**309**:1351–5.