A brief introduction to meta-analysis

Lawrence Joseph[*]

March 20, 2000

[*] Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada, and Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Quebec, H3A 1A2, Canada.

1 Introduction - why consider meta analysis?

According to Last's A Dictionary of Epidemiology, meta analysis is:

> The process of using statistical methods to combine the results of different studies. The systematic, organized and structured evaluation of a problem of interest, using information from a number of independent studies of the problem.[1]

Thus meta-analysis is helpful in synthesizing the information about a particular medical issue via statistical analysis, especially when more than one study has been carried out that relates to the question of interest. Since one should ideally consider all of the available information about any issue before making any clinical decisions, and since meta-analysis is able to summarize information from several data sets, it would seem that meta-analysis should be considered as a cornerstone of medical research and medical decision making. In some ways this is true, but unfortunately, the many problems that are associated with meta-analysis make its use controversial at best, and a poorly conducted meta-analysis can produce results that are downright misleading. Nevertheless, the use of meta-analysis is on the rise, and a carefully performed and reported meta-analysis can produce useful results. In addition, new statistical methodologies are continually being developed, many of which address specific problems that have been raised about meta-analysis.

In this chapter, we will provide an overview of the various methods that have been used in meta-analysis, including some of the most common statistical models that have appeared in the literature. We will also discuss several of the problems and biases that are difficult to avoid in carrying out a meta-analysis, and indicate how they can be minimized. In particular, Section 2 discusses the specification of the question that the meta-analysis will focus on, and the subsequent selection of studies whose data will be included in the analysis. Simple fixed effects statistical techniques for carrying out a meta-analysis will be covered in Section 3, and more complex random effects techniques will be discussed in Section 4. Throughout, we will apply all of the techniques discussed in Sections 3 and 4 to a research question first considered by Pocock et al.[2] These authors summarized the results from a collection of studies comparing coronary angioplasty (PTCA) to bypass surgery (CABG) for patients with severe angina. This example also provides us with an opportunity to emphasize the care in which meta-analytic results must be interpreted, especially when deciding upon the context to which the results can be applied. This example is summarized in Section 5, and Section 6 concludes with a summary and suggestions for further reading.

As this chapter presents a variety of statistical techniques for meta-analysis, a solid grounding in the basics of statistical analysis is an essential prerequisite. Therefore, we suggest reading through Chapter 3 before tackling the material in this chapter.

2 What is the question, and which studies should be included?

Two of the first steps in carrying out a meta-analysis are the specification of the research question that is to be investigated, and the choice of which studies to include from among all those which address the research question of interest. Neither of these tasks are as straightforward as one might first assume. Below we discuss each of these in turn.

2.1 Specifying the research question

In choosing a research topic, many factors must be considered in order to make the question as precise as possible. For example, suppose one wishes to investigate any outcome differences between PTCA and CABG in patients with severe angina. First, the definition of “severe angina” must be made precise. As for outcomes, should one consider cardiac deaths only or also include non-fatal myocardial infarction? What about the rates of subsequent revascularizations? Including composite outcomes, such as cardiac deaths or non-fatal myocardial infarction have the advantage of providing more events, and hence often provide increased statistical accuracy in estimating differences between treatments. This increase in statistical accuracy, however, comes at the steep price. If one finds a difference in a combined outcome, one cannot be sure if both or only one of the outcome rates differ between treatments, and this distinction may be important for clinical decision making. For example, a cardiac death is a more serious outcome than a non-fatal myocardial infarction. If cardiac deaths in fact occur at the same rate in both treatment groups but non-fatal MI rates differ, this may be less serious than if cardiac death rates differ. Conversely, if no differences are found between treatment arms for a combined outcome, one cannot be sure whether differences in both outcomes, but in opposite directions, have cancelled each other out. Therefore, if one will use a combined outcome, it is important to also consider each outcome alone in separate sub-analyses.

In a related issue, should one compare the rates in the two groups in terms of the difference in the absolute event rates, or use relative risks or odds ratios?

Absolute differences are often more clinically relevant, but if studies have different follow-up times, or if case-control studies are included, for example, only relative risks or odds ratios may be valid or easily available. What age range should be considered? This may be important if the success rates or adverse event rates of the different procedures can change with age. Complicating this issue further is that it is rare for two studies to consider exactly the same age range, so clinical judgement is required to decide if the age differences are such that combining results is reasonable. Other patient characteristics such as aspects of their medical histories might be of importance, as they may also differ markedly from study to study. Thus there are issues of both validity and generalizability that must be considered in narrowing down which populations to include. What about follow-up time? Should one consider events in the next year? Two years? What if there are differences in the ways that the procedures are performed from study to study? Does one need to specify the procedure, or do all of them produce similar event rates, and so can be combined in a straightforward manner? In drug studies, one must consider the dosage or the range of dosages that is acceptable.

In summary, in phrasing the research question, many issues about the procedures and populations to which the procedures are applied need to be carefully considered. Including too narrow a range of procedures or population characteristics hampers generalizability, and may exclude many important studies. On the other hand, inclusion criteria that are too wide may lead to the combining of studies with heterogeneous effects, which may make the meta-analysis results more difficult to interpret. A finding in one direction or another may not apply to all subgroups, and conversely, an overall lack of a positive effect for some intervention may still be usefully applied to some subgroups. The difficulty is that the numbers of patients in many or even all of the subgroups of interest will be small, precluding accurate estimation of differences in outcomes within these subgroups. Careful thought about these issues is therefore essential to a well planned meta-analysis.

2.2 Which studies should be included?

Once one specifies the research question, the next task is to find all studies that are related to this question, and choose from among them those that will be included in the meta-analysis. We next consider these two questions.

2.2.1 Finding all potentially relevant studies

Most hunts for relevant studies will start with a formal keyword and topic search in one or more databases of the medical literature. Of course, putting in a key word like "myocardial infarction" will bring up literally thousands of articles, so that search strategies usually need to be more refined. For example, one might begin with the keywords of coronary angioplasty, coronary bypass surgery, and angina. If we were considering only randomized clinical trials, we might add that term as an additional keyword.

Once the relevant articles are gathered, the references from each of these articles can be perused for any missing items of interest. For example, there may be review articles or even previous meta-analyses. Other sources of information can include government agencies, conference presentations, or conversations with experts in the field, who, for example, may be aware of ongoing or unpublished studies. It is important to find all relevant unpublished studies in order to avoid publication bias. This bias can occur if several studies of the issue of interest have been carried out, but only those with "positive" findings are submitted or accepted for publication. As a simple (and perhaps somewhat unrealistic) example, suppose that there are truly no differences between the outcomes of patients with a certain medical condition given two drugs. If 20 trials of these two drugs are carried out, by chance alone, we would expect that one of these 20 trials would be “statistically significant” at the 5% significance level (see Chapter 3). Surely it would give a completely false impression if this “positive” trial were to be the only one published! While this example is somewhat exaggerated, there has certainly been a tendency for researchers to prefer submitting articles with “positive” findings, and a tendency for most medical journal editors to accept such articles in preference to “negative” studies. This so-called “file drawer problem” is often raised as an objection to the validity of meta-analyses.

To summarize, the goal of the search for previous relevant research is not necessarily to find all articles ever written on the topic of interest, but rather to find all important sources of data that address the question of interest.

2.2.2 Which studies to include?

Once all potentially relevant articles have been gathered, the next task is to decide which will be formally included in the meta-analysis. First of all, only studies with original data (or corrections to these data!) which directly impact on the question of interest need to be included. Should one only include randomized trials, or also consider observational studies? Randomized trials may be freer from bias, but in some cases, there may be no randomized trial data available. This could occur, for example, when it is unethical to randomize patients, such as in a study of the effects of smoking or exposure to second hand smoke. Because of the specialized protocols and inclusion/exclusion criteria, randomized trials are often more difficult to generalize to all potential patients. Even more difficult is to judge the quality of each study from the article or other source where it is reported. Mixing poor studies with better conducted studies can create biases and imprecisions that can serve to either dilute or exaggerate the sizes of effects of interventions. We have already discussed publication bias, which can increase the apparent effect of an intervention. Conversely, including a study where an intervention has not been correctly implemented can dilute the true effect of the intervention when it is properly administered.

By looking closely at the features of each study, one can determine what sources of potential biases may be present (see Chapter 2), and what degree of confidence one can place in the results. In investigating bias, both internal bias (for example, non-comparability between patients in the two treatment arms in a study) and external bias (for example, strict inclusion/exclusion criteria that may limit generalizability) are of importance. One also needs to consider the degree to which the study design and main study questions of each article match those of the primary research question. What impact, for example, is expected if a drug dosage is slightly different from that which has become standard today? What if the patient population in the study is more or less severely diseased than those to whom we wish to apply the results of the meta-analysis? These are not easy questions to answer, especially if there is a paucity of results that do exactly match the desired research question. The question of whether a given study design is "close enough" to what would be considered ideal is central to the debate about the usefulness of meta-analysis. If small deviations in protocols can potentially change the results in a clinically meaningful way, one must proceed with a meta-analysis with great caution. This is an especially difficult point, since the degrees to which results can be changed by differences in designs are rarely clear.

Aside from the above concerns about bias, there are design specific issues that affect study quality. One of the most obvious design issues is sample size. As discussed in Chapter 3, all other things being equal, larger studies provide more precise estimates of effect sizes than smaller studies. Aside from size, many other design features, including the choice of randomization technique, accuracy of measures and in determination of outcomes, blinding, and length of follow-up (see Chapter 5 for a discussion of the issues pertaining to clinical trials) can all have an impact.

There is a large literature on bias and quality of studies in meta-analysis. A full discussion of this topic is beyond the scope of this chapter. The interested reader is referred to the books mentioned at the end of this chapter and the references therein for further details.

3 Fixed effects meta-analytic methods

Once one has specified the research question, found all relevant articles, judged each article for quality, and made a decision about which articles to include in the meta-analysis, the next step is to analyse the data. Recall that the major goal of this analysis is to combine the results from the different studies so that overall conclusions can be drawn. This section will consider simple methods of meta-analyses, while the following section will consider more complex models.

Before discussing the analysis techniques, we will first present the data from the previous meta-analysis concerning two different revascularization procedures for patients with severe angina.[2] Table 1 presents a summary of the results from this article. These authors considered 8 randomized clinical trials comparing coronary angioplasty (PTCA) with bypass surgery (CABG), including a total of 3371 patients with an average follow-up time of 2.7 years. While here we will focus on the combined outcome of cardiac death and non-fatal myocardial infarction, these authors also looked at rates of other cardiac interventions following the initial treatment, and the prevalence of angina in subsequent years. In addition, they performed separate analyses for multivessel and single vessel disease patients. Of course, all of these outcomes and patient groups should be considered when comparing the interventions, although the data in Table 1 will suffice for our illustrative purposes.

Looking at the results in Table 1, several preliminary observations can be immediately made:

1. There is a range of sample sizes among the 8 clinical trials, from relatively small trials (127 subjects) to relatively large trials (1054 subjects). This implies that the precision of the relative risk (RR) estimates also vary, as reflected in the widely differing lengths of the confidence intervals.

2. The mean follow-up times vary greatly, from 1 to almost 5 years. Thus the proportions of events in each trial can also expect to vary, as, all else being equal, longer trials can be expected to have more events. Nevertheless, the trial with the highest rates (up to 17% in the CABG group), King et al[5], is almost two years shorter than the longest trial, the RITA trial[4], which has a much lower rate (about 6 or 7%). Thus follow-up time alone does not fully explain the observed rate variations in the trials; one or more factors such as different trial settings, populations, and variations in procedures, as well as random variation play a role in determining these rates. One must return

to the original individual study details and consider all of these factors when drawing conclusions from this meta-analysis.

3. The point estimates of the relative risks vary from study to study, in both magnitude and direction. Thus it is difficult to tell if the overall results indicate a RR above or below 1. Furthermore, all confidence intervals include the value of RR = 1.

In carrying out a meta-analysis, as in all statistical analyses, various assumptions must be made. One of the major assumptions in a meta-analysis concerns the parameter of main interest, which in our example is the relative risk. If we believe that the relative risks from all studies are in fact the same, then a simpler meta-analysis technique can be used compared to if we believe that the relative risks vary from study to study. Note that it is the true (unobserved) value for the RR, not the calculated point estimates, that must be the same. Thus one cannot necessarily decide that the parameters differ between studies from looking at the point estimates alone, since the observed values are subject to random fluctuations about their true values. No two study settings are ever exactly the same, because populations, inclusion and exclusion criteria, specifics of the interventions, and so on, always differ from study to study. Therefore, assuming that the relative risks from two or more studies are exactly the same seems unreasonable. Nevertheless, if they can be expected to be very close, either because the study settings did not differ greatly across studies or because one expects the effects not to greatly depend on situation specific variables, then this “fixed effects” model may be a reasonable approximation to reality. In most other cases, where effects can be expected to differ at least a bit from study to study, a so-called “random effects” model is preferable, since such a model explicitly accounts for between study differences.

These more complex models will be discussed in Section 4, while this section discusses three different fixed effects models. From a frequentist viewpoint, we will discuss the pooled and the variance weighted fixed effect approaches. We will also look at a fixed effects Bayesian approach. See Chapter 3 for a discussion of the philosophical and practical differences between the frequentist and Bayesian approaches to statistical analyses.

3.1 Meta-analysis using a pooled data approach

The pooled data approach is the most simple (some might say simplistic!) of all meta-analytic approaches. To apply this method one acts as if all of the data comes from a single large experiment, and analyses the data using the appropriate "single study" technique. In the meta-analyses of Pocock et al[2], one can form a numerator by summing the numbers of events across each study in each of the PTCA and CABG treatment arms, and similarly sum across studies to obtain the total numbers of patients in each study arm. From Table 1, these totals would be 127 cardiac death or non-fatal MI events in 1661 patients in the CABG group, and 136 events in 1710 patients in the PTCA group. We can then calculate a relative risk and corresponding confidence interval using the methods for single studies, as described in Chapter 3, Section 5. Thus the estimated pooled relative risk would be

$$\hat{RR} = \frac{\frac{136}{1710}}{\frac{127}{1661}} = 1.04$$

with a confidence interval of (0.83, 1.32).

Pooled meta-analytic estimates can similarly be calculated for other measures, such as risk differences or mean differences. Use of this method implies that the rates (or means) are the same within each treatment arm of each study. Clearly this is not a reasonable assumption here, since even if the rates would theoretically be the same, the length of follow-up of each study varied, so that the reported rates cannot possibly be the same. Therefore, this method is innappropriate for this study, and should at least be seriously questioned in most other situations.

3.2 Meta-analysis using a variance weighted approach

All meta-analysis techniques are, at least in a loose sense, a weighted average of the results from the individual studies. The pooled analysis of the previous section, for example, can be considered to be a weighted average where the weights are obtained directly from the sample sizes from each study. If all studies provided equally reliable estimates and all were equally free from bias, then a simple mean (or average) of the results would seem to be appropriate as an overall estimate of the effect of the intervention on the outcome of interest. It is usually the case, however, that some studies provide more reliable estimates than others. This can happen, for example, if some studies have larger sample sizes than others, and thus provide estimates with smaller standard errors (again, see Chapter 3 for definitions of these basic statistical terms). Here a simple average is sub-optimal as an overall estimate of the effect, since we would somehow like those studies with more precise estimates to "count" for more than those providing less precise estimates. In other words, a weighted average giving more weight to more precise studies would be preferred to a simple average. As the name implies, in a variance weighted approach, the weights we use in the weighted average are related to the variance of the estimates from each particular study. The smaller the variance, the more precise an estimate one obtains from an individual study, so the more weight it should receive in calculating the overall estimate of the effect of interest.

If we let the weight $w_i = \frac{1}{var_i}$ be the reciprocal of the variance from study i (so small variances imply large weight, and vice versa), then the variance weighted meta-analytic estimate, $\hat{E}$ of the overall effect E of interest is given by

$$\hat{E} = \frac{\sum_{i=1}^{k} w_i \times \hat{E}_i}{\sum_{i=1}^{k} w_i} \ . \tag{1}$$

Here k is the total number of studies included in the meta analysis, and study i has estimated effect $\hat{E}_i$.

Note that E can be any effect of interest, such as a relative risk, an odds ratio, a between treatment difference in means or in rates, and so on. In our case we are looking at the relative risk, so that outcome E is the RR, and the weights are related to the estimated variance of the RR (see Chapter 3, Section 5). Formulae for the variance of the overall estimate ($\hat{E}$) are also available, see, for example, Chapter 19 of Cooper and Hedges.[3]

Applying equation (1) to the data from Table 1 gives the results in the second line of Table 2. The variance weighted point estimate of the relative risk is 0.88. This means that there is a 12% decrease in the rate of cardiac death or non-fatal myocardial infarction following CABG versus PTCA, but note that the 95% confidence interval (0.42, 1.34) includes the null value of 1, and does not rule out an effect in the direction opposite to the point estimate. Since a risk reduction of 58% (corresponding to the lower limit of the confidence interval of 0.42) would presumably be of great clinical interest if it were the true value, one concludes that this analysis is inconclusive (see Figure 8 of Chapter 3), since the null value is included in the confidence interval, but a clinically important effect cannot be ruled out.

Like the pooled method, implicit in variance weighting is the assumption that the effects are identical across studies, so that there are no internal or external biases in any of the studies that might make their effects different. In other words, in variance weighting, one still retains the usual fixed effects assumption of equal relative risks across studies, an assumption that may or may not be reasonable. While in theory one can devise a test for homogeneity of relative risks across studies, these tests are subject to all of the problems associated with p-values (see Section 4 of Chapter 3), especially low power. Thus these tests should be used with great caution, or avoided completely. Unlike the pooled method, however, in variance weighting one does not need to assume that the rates themselves are identical across studies. Assuming that the rate of events remains stable over time, longer studies will on average have more events. Since this increase in numbers of events applies to both the PTCA and CABG groups, however, the relative risk should not be affected, and so RR's from studies with different follow-up times can still be combined. Note that the weaker assumptions of the variance weighted method compared to the pooled method results in a wider confidence interval for the variance weighted method. This is a common phenomenon in statistics, where the relaxing of assumptions leads to less certainty in the estimated results. Of course, if the assumption does not hold in the first place (as is certainly the case in the pooled analyses in our example), the entire analyses that depends on that assumption is not valid.

3.3 Meta-analysis using a fixed effects Bayesian approach

Recall from Chapter 3 that in a Bayesian approach, one combines prior information (in the prior distribution) with the information in the data (through the likelihood function) to derive the posterior distribution. The posterior distribution summarizes all of the information that was known about the effect of interest *a priori* together with the additional information contained in the data being analyzed.

A very simple Bayesian model can be formed by following the idea from the pooled analysis discussed above. In each treatment arm, the total numbers of events can be considered to follow independent binomial distributions. If the prior information about each event rate can be summarized in by beta densities, the posterior distributions also follow a beta densities (see Chapter 3, Section 9.1). Roughly speaking, the ratio of these two beta densities provides the posterior distribution for the relative risk, from which a 95% credible interval can be formed. In our example from Table 1, the binomial likelihood functions would be

$$\theta_1^{136}(1-\theta_1)^{(1710-136)}$$

and

$$\theta_2^{127}(1-\theta_2)^{(1661-127)}$$

where θ_1 and θ_2 are the event rates in the PTCA and CABG groups, respectively. If a diffuse or noninformative beta(0.5, 0.5) prior density is used for each group, the posterior densities are the same as the likelihood functions above, except that 0.5 is added to each exponent (four times). Thus the posterior density for θ_1 is a beta(136.5, 1574.5) density, and the posterior density for θ_2 is a beta(127.5, 1534.5) density. These densities are depicted in Figure 1, where one can see that the two event rates are very similar, and both are almost surely between 6% and 10%.

Technically, using Bayes Theorem to derive the posterior distribution for the RR involves methods from calculus, including a change of variable transformation and integration, so the details will be omitted here. A simple way to conceptualize the analysis is to imagine that first a large random sample is taken from each of the two beta densities, representing the rates from each treatment arm, and a new random sample is created from these by forming the ratios of the randomly selected rates in the two groups. This random sample can be considered as having arisen from the posterior distribution of the relative risk, so that a histogram of the variables generated in this way approximates the posterior density curve. Either way, the results should be very similar to those presented in Table 2 for the fixed effects Bayesian approach, with a point estimate for the relative risk of 1.04 and a 95% credible interval of (0.83, 1.31). Note that this interval is very similar the previous pooled estimate. Figure 2 displays the posterior density for the RR. One of the advantages of the Bayesian approach is the ability to directly calculate and report probabilities relating to the RR. For example, if the region of clinical equivalence is determined to be (0.9, 1.1), representing a 10% difference in rates in either direction, then the area under the curve (Figure 2) between these two points provides the probability of clinical equivalence of these two treatments. Here we find that $Pr\{0.9 < RR < 1.1\} = 0.57$. Therefore, we are 57% certain that these two treatments are clinically equivalent.

Of course, this very simple Bayesian model suffers from the same drawbacks as the frequentist pooled estimate discussed above, and thus cannot be realistically applied to our cardiology example. Nevertheless, the Bayesian approach does retain all of its usual advantages, including the possibility of formally including prior information into the analysis, including a range of prior distributions (see Chapter 3, Section 9), and ease of interpretation of credible intervals compared to confidence intervals. In addition, the Bayesian approach offers a way of evaluating the possible effects of publication bias as a partial solution to the "file drawer problem". After having analysed the data with a "flat" prior distribution, to see the information that the data themselves provide, one can ask the following question: How strong would my prior distribution need to be in order to change the conclusions that I have formed from this meta-analysis? For example, if we add a study with 200 subjects and with an observed relative risk of 2 in favour of PTCA, say, does this change our conclusions? Such a study might have an event rate of 10% in the PTCA group, and a 20% rate in the CABG group. In other words, if there were an unpublished or undiscovered study with 20 events in 100 patients in the CABG group and 10 events in 100 patients in the PTCA group, would this be enough to change our main conclusions from the meta-analysis? Redoing the analysis with this added study as our "prior information" (thus using prior densities of beta(20, 80) and beta(10, 90) for the CABG and PTCA groups, respectively) finds the results moved to RR = 0.97 (95% CI (0.78, 1.21)), which is not a large change from the previous estimate. However, if it is possible that a similar "undiscovered" study could have had a size of 1000 rather than 200, we find RR = 0.71 (95% CI (0.60, 0.83)), which would be a strong result in favour of PTCA. Therefore, by this "backwards Bayes" technique (so-called because the prior distribution is not formed before, but rather after the data are collected), we can claim robustness to small undiscovered studies, but not to large ones. If it is considered very unlikely that such large studies would go unreported, then this analysis could show that our results are "robust" to publication bias or the file drawer problem.

4 Random effects meta-analysis models

As discussed in the previous section, fixed effects models are, at best, very rough approximations to reality. Again, this is because we rarely expect two studies to have exactly the same true effect. Variations in population characteristics, interventions applied, study protocols, and other factors mean that the effects being estimated in each study usually vary from study to study. Random effects models are specifically designed to accomodate variations in effects between studies. The basic idea is to include two variance terms in the model. The first variance term represents the usual random variability about the observed results within each study. The second variance term represents the between study variations in true effect rates (or other outcome). By considering both of these terms, one models both within and between study sources of variation. In this section we will examine two random effects models, one from each of the Bayesian and frequentist schools. As our goal is to present the basic ideas behind random effects models, technical details are left to a minimum.

4.1 Bayesian hierarchical random effects models

First consider a single study, for example the CABRI trial, whose data are given in Table 1. Within this single trial, we can apply simple methods to estimate the RR from this study alone. From the Bayesian viewpoint, this means we can start with two beta prior densities on the binomial probabilities for the event rates in the CABG and PTCA groups. Adding in the information provided by the data from this trial and using the methods briefly hinted at in Section 3.3, a posterior distribution for the RR for this study can be derived. Similarly, we can create posterior densities for each of the RR's from each of the eight studies considered by Pocock et al[2] listed in Table 1. Thus we consider the set of relative risks from each study, which we can lable as $RR_1, RR_2, \ldots, RR_8$.

Now, suppose that, for the various reasons mentioned above, these RR's are not all identical to each other, but rather are distributed according to

$$RR_i \sim N(\mu, \sigma^2),\ i = 1, 2, \ldots, 8. \tag{2}$$

Thus we assume that the relative risks from the 8 studies follow a normal distribution. The parameter μ represents the overall mean value, or the true average effect among all the effects in such studies. The variance parameter, σ^2 represents the study to study variability in RR's, due to the various settings of each study. Now, if $\sigma^2 = 0$, then all RR_i's are assumed to have the same true value, given by μ, and we are back to a fixed effects model. Thus we can see that a fixed effeects model is really just a special case of the random effects model, when $\sigma^2 = 0$. If σ^2 is greater than 0, however, the studies do vary in their true effects, and a hierarchical random effects model is indicated.

Hierarchical models such as these are often called two stage models. At the first stage, the relative risks (or other outcome of interest) have variabilities that are largely determined by the study specific data, while at the second stage these study specific parameters themselves follow a distribution that applies between studies. Thus equation (2) represents the second stage of this hierarchical model. Hierarchical models allow for the "borrowing of strength" between studies, in the following sense. Studies with large sample sizes tend to have more stable estimates compared to studies with smaller sizes. Because equation (2) acts as a sort of "meta-prior" over the collection of relative risks among the studies, μ will depend more on, and tend to be closer to, the relative risks from the larger studies. Thus small studies will tend to have posterior estimates for their relative risks that are slightly "pulled towards" the overall estimate μ. This is similar to the effect of the prior density when the data set is small in Chapter 3, Section 9.2.

In a hierarchical analysis, one can present two different types of results. First, one can present the posterior distribution for the overall mean, μ. This posterior distribution is interpreted as describing the uncertainty in the overall mean effect among the studies. Looking at Table 2, we see that the overall estimate of the mean RR for the 8 studies from Table 1 is 1.05, with a 95% credible interval of (0.74, 1.44). Notice that this is slightly wider than the previous confidence interval from the simple Bayesian estimate, because we no longer assume that the RR is constant across all studies, as in the simple Bayesian pooled approach (in fact, that approach went further, and assumed that the event rates in each trials arm were identical across all studies). Considering the entire distribution $N(\mu, \sigma^2)$ allows one to include the study to study variability due to the different settings. Thus we are able to derive the posterior density for the "next similar study" that might be performed, assuming that the true parameter value is a random draw from the normal distribution presented in equation (2). Here we try to capture what might occur in a randomly selected clinical setting that is subject to similar sources of variation in effects as those studies included in the meta analysis. Assuming that the study settings included in the meta analysis represent a reasonble range of all possible settings, this represents a more realistic assessment of what would happen in real clinical practice. Again referring to Table 2, the estimate of 1.09 (95% CI 0.55, 1.97) for the "next study" using a Bayesian hierarchical approach is even wider than that for the mean effect. This result recognizes that not all study settings would have relative risks near the overall mean μ. The difference between the results for the mean effect versus the results for the "next study" are analogous to the difference between reporting a standard error or a standard deviation for a mean (see Chapter 3).

Figure 3 presents the posterior densities for the three Bayesian approaches. As can be expected, the fixed effects Bayesian estimate provides the narrowest posterior density, although it is almost surely making unrealistic assumptions. The posterior distribution for the mean effect shows that we are quite certain that the overall mean effect is between about 0.7 and 1.5, but the final density shows that we have not ruled out substantial variation in relative risks from setting to setting.

More sophisticated Bayesian models may try to explain these differences by forming another level in the hierarchy where the mean μ is not considered as fixed, but may vary in a regression model depending on study specific covariates. See Brophy and Joseph[4] for an example of such a hierarchical model. It is important to emphasize that random effects models are appropriate when the sources of variations in the effects from study to study are unknown or uncertain. If the variations arise from well identified sources (for example, studies with different drug dosages that produce dose-effects responses), then these sources should be identified and incorporated into a more complex model, rather than considered as random effects.

Finally, as always in a Bayesian approach, prior densities are needed for the parameters μ and σ. If there is substantial prior information and the studies are either few in number or small in size, then prior information can be very useful, especially if care is taken in their elicitation and the results are presented across a reasonable range of prior densities. Most meta-analyses use "non-informative" priors, however, so that subjective input is kept to a bare minumum.

4.2 The random effects method of DerSimonian and Laird

DerSimonian and Laird[5] presented a frequentist method for random effects in meta-analysis. Similar to the Bayesian approach, the basic idea is to hypothesize both within and between study variances, such that

$$\text{total variance} = \text{within study variance} + \text{between study variance}. \tag{3}$$

One then estimates the overall effect and the two variances on the right hand side of equation (3). We omit the lengthy details of the estimation procedure here, see DerSimonian and Laird[5] for the full details.

In practice, the frequentist random effect model often provides similar inferences to those from the hierarchical Bayesian approach, as can be seen from Table 2, where the estimated RR's and the 95% CI's are very similar between these two methods.

Random effects models, both frequentist and Bayesian, can be criticized for making unverifiable assumptions. It is typically assumed that the between study effects follow a normal distribution. Unless there is a large number of studies, the distribution of study effects is difficult to verify. While one can conceptualize using hierarchical distributions other than the Normal, this is rarely done in practice, and any choice would still suffer from the same difficulties of model verification. Furthermore, when making inferences about the "next study" setting, one assumes that the "next study" is similar to those included in the meta-analysis, which may also be either unverifiable or even unlikely, depending on the particular circumstances. Thus while random effects models are often more plausible than fixed effects models, they too suffer from great uncertainty about their applicability to many situations.

5 Summary of the application

Considering the totality of the results presented in Table 2, one finds at least slightly different inferences from the different methods, even within classes of fixed and random effects models. Furthermore, Pocock et al[2] used a fixed effects model on the logarithm of the relative risk, finding estimating RR = 1.10, with 95% CI = (0.89, 1.37), which is again slightly different from any result in Table 2. Because they are formed from the ratios of two probabilities, the logarithms of the relative risks may be closer to normality than the relative risks themselves. Looking at the death rate alone, Pocock et al[2] report a similar result as for the combined endpoint, with RR = 1.08 and 95% CI = (0.79, 1.50). They further reported potentially important differences in the rates of revascularization and relief of angina. Thus a further difficulty with meta-analysis is that the results can depend on the specific method used to combine the data, and, of course, on the choice of endpoint.

Nevertheless, all six methods agree that the null value of RR = 1 cannot be ruled out. Furthermore, the random effects models agree that effects of up to almost 2 fold differences in event rates in *either* direction cannot be ruled out, at least in some settings. Thus, while no strong evidence is found for differences in the rates of cardiac deaths or non-fatal myocardial infarction, the data also do not support a strong conclusion of no difference in event rates either. Despite combining data on a total of over 3,000 patients, this meta analysis must be considered as inconclusive, and further evidence should be gathered. Both between study and within study variability contributes to our uncertainty. Pocock et al[2] discuss many other limitations to their study.

6 Conclusion

All meta-analyses require a variety of expertises, which must be assembled before embarking on the analysis. Clinicians very familiar with the substantive area are clearly important, but so are epidemiologists who must carefully consider each study for possible biases, design flaws and other differences that may create special problems for combining study results. Statisticians should be available to provide advice on selection of the appropriate techniques to use, how to adjust for biases, if necessary, and so on.

This chapter has presented the basic issues and simple analytic techniques behind meta-analysis. Clearly meta-analysis is a complex topic, both clinically and methodologically, and this chapter should be considered as only a very brief introduction. Several textbooks on meta analysis have been written which should be consulted for other methods of meta-analysis and for further details about issues related to performing and interpreting meta-analyses. Hedges and Olkin[6] is a classic text on the subject, although many new analytic techniques have appeared since its' publication. Eddy et al.[7] take a Bayesian approach to meta-analysis, and include discussions of bias adjustments, combining randomized with non-randomized studies, adjusting for differing lengths of follow-up, and other more complex topics. Cooper and Hedges[3] is a comprehensive modern textbook on the subject that includes not only the relevant statistical techniques, but also extensive chapters on selecting research questions, searching the literature, judging the quality of the research, and reporting the results of meta-analyses. The recent book edited by Berry and Stangl[8] contains many examples of complex meta-analyses, and discusses the bridge between meta-analyses and health policy decisions. A variety of software packages for meta-analysis are also available, such as the FastPro software of Eddy et al.[7] Some standard statistical packages also include meta-analytic techniques.

Some authors[9,10] have raised the question of whether meta-analysis has anything to offer over and above what can be concluded from a non-quantitative critical review of the literature. Clearly meta-analysis would have something important to offer IF all of the assumptions of either a fixed or a random effects model were perfectly correct, so that the answer to this question hinges on whether these assumptions are reasonable or not, and if not, on the the robustness of the conclusions to deviations from the assumptions. The fact that these assumptions can be very difficult or even impossible to verify in most cases is at the root of the controversy that surrounds the usefulness of meta-analysis. Does meta-analysis provide the best possible summary of the available evidence, or does it provide an overly simplistic estimate of the uncertainty surrounding a given medical question, therefore providing a false sense of security? The jury is still out, but the answer probably lies somewhere in between these extremes for most problems.

References

1. Last, J. A dictionary of Epidemiology (Second Edition). Oxford University Press, Oxford, 1988.

2. Pocock S, Henderson R, Rickards A, Hampton J, King S, Hamm C, Puel J, Hueb W, Goy J, Rodriguez A. Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. Lancet 1995;346(8984):1184-1189.

3. Cooper H, Hedges L (eds). The handbook of research synthesis. Sage, New York, 1994.

4. Brophy J, Joseph L, Theroux P, on behalf of the Quebec Acute Coronary Care Working Group. Medical decision making about the choice of thrombolytic agent for Acute Myocardial Infarction. Medical Decision Making 1999;19(4):411-418.

5. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986;7:177-188.

6. Hedges L, Olkin I. Statistical methods for meta-analysis. Academic Press, San Diego, 1985.

7. Eddy D, Hasselblad, V, Shachter, R. Meta-Analysis by the Confidence Profile Method, Academic Press, Boston, 1991.

8. Berry D, Stangl D (eds). Meta-Analysis in Medicine and Health Policy. Marcel Dekker, New York, 2000

9. Bailar J 3rd. The practice of meta-analysis. Journal of Clinical Epidemiology 1995;48(1):149-57.

10. Bailar J 3rd. The promise and problems of meta-analysis. New England Journal of Medicine 1997;337:559-561.

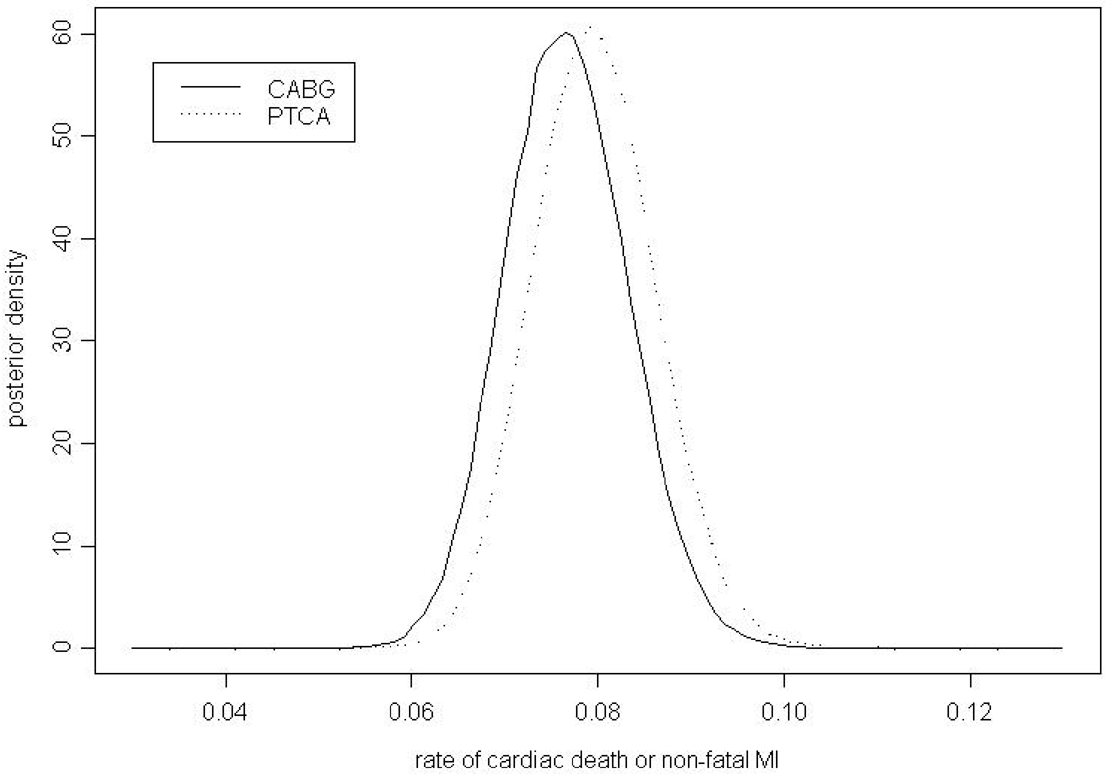

Figure 1: Posterior densities for the rate of the combined endpoint of cardiac death and non-fatal myocardial infarction for the CABG and PTCA groups, for all data combined via a fixed effects Bayesian model.

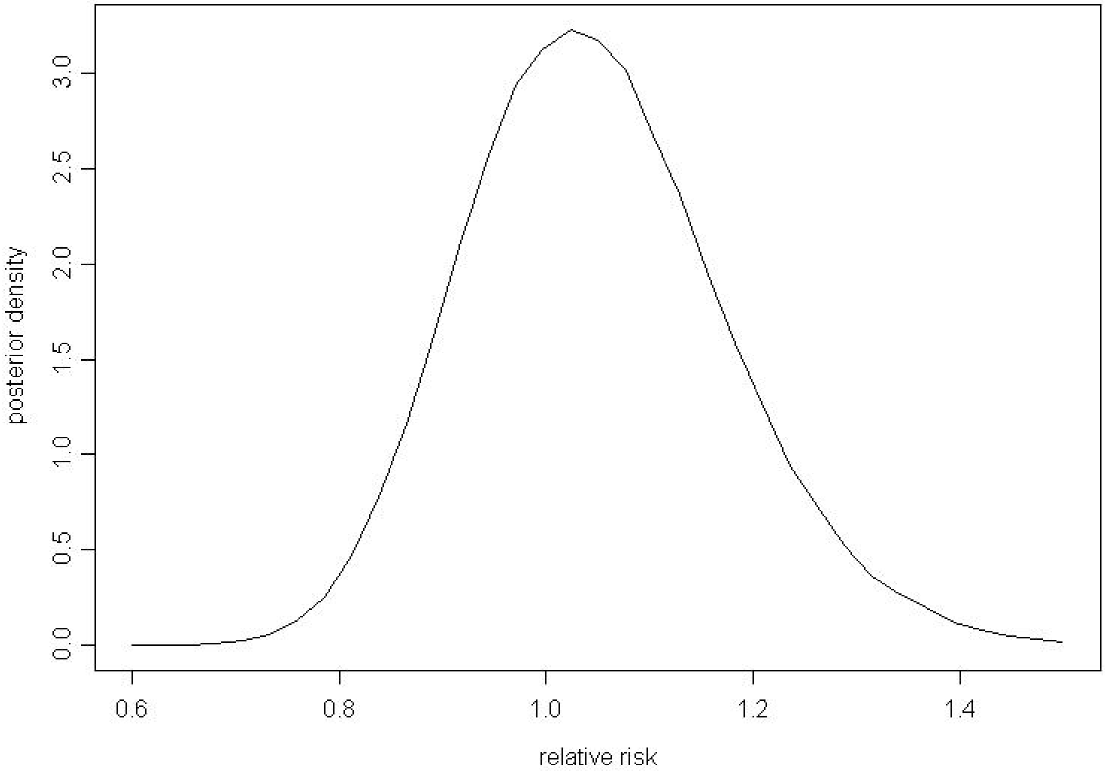

Figure 2: Posterior density for the relative risk of PTCA versus CABG for the combined endpoint of cardiac death and non-fatal myocardial infarction for the CABG and PTCA groups, for all data combined via a simple Bayesian model.

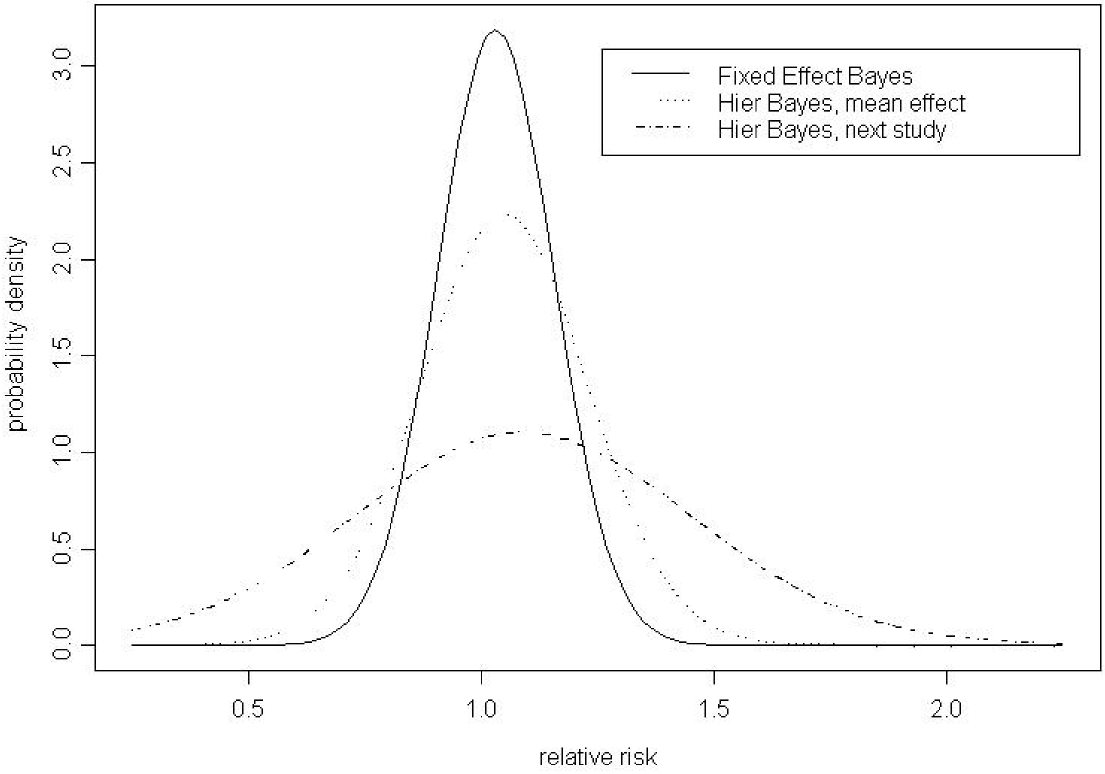

Figure 3: Posterior density for the relative risk of PTCA versus CABG for the combined endpoint of cardiac death and non-fatal myocardial infarction for the CABG and PTCA groups, for all data combined from two different Bayesian models.

Trial	# Patients		# Deaths or MI (%)		Mean follow-up	RR	(95% CI)
	CABG	PTCA	CABG	PTCA			
CABRI Trial[3]	513	541	29 (5.7)	43 (7.9)	1.0	1.44	(0.89, 2.22)
RITA Trial[4]	501	510	31 (6.2)	34 (6.7)	4.7	1.11	(0.67, 1.73)
King et al.[5]	194	198	33 (17.0)	24 (12.1)	3.0	0.74	(0.44, 1.16)
Hamm et al.[6]	177	182	18 (10.2)	10 (5.5)	1.0	0.58	(0.26, 1.14)
Puel et al.[7]	76	76	6 (7.9)	6 (7.9)	2.8	1.17	(0.34, 2.96)
Hueb et al.[8]	70	72	1 (1.4)	5 (6.9)	3.2	8.16	(0.60, 36.89)
Goy et al.[9]	66	68	2 (3.0)	6 (8.8)	3.2	3.95	(0.61, 13.63)
Rodriguez et al[10]	64	63	7 (10.9)	8 (12.7)	3.8	1.31	(0.45, 3.01)

Table 1: Summary of the results of the 8 trials of PTCA versus CABG as given by Pocock et al.[2]

Method	Results for the mean effect		Results for the "next study"	
	RR	95% CI	RR	95% CI
Pooled	1.04	(0.83, 1.32)	NA	NA
Variance weighted	0.88	(0.42, 1.34)	NA	NA
Simple Bayesian	1.03	(0.81, 1.30)	NA	NA
Hierarchical Bayesian	1.05	(0.74, 1.44)	1.09	(0.55, 1.97)
DerSimonian and Laird	1.02	(0.74, 1.41)	1.02	(0.56, 1.88)

Table 2: Summary of the results of the meta analysis of the data from in Table 1, as given by five different meta-analytic techniques. Descriptions of each technique are as given in the text. NA indicates non-applicable, as only random effects models have results for the "next study".