CHAPTER 40

When Does it Make Sense to Perform a Meta-Analysis?

Introduction
Are the studies similar enough to combine?
Can I combine studies with different designs?
How many studies are enough to carry out a meta-analysis?

INTRODUCTION

In the early days of meta-analysis (at least in its current incarnation) Robert Rosenthal was asked if it makes sense to perform a meta-analysis, given that the studies differ in various ways, and the analysis amounts to *combining apples and oranges*. Rosenthal answered that combining apples and oranges makes sense if your goal is to produce a fruit salad.

The goal of a meta-analysis is only rarely to synthesize data from a set of identical studies. Almost invariably, the goal is to broaden the base of studies in some way, expand the question, and study the pattern of answers. The question of whether it makes sense to perform a meta-analysis, and the question of what kinds of studies to include, must be asked and answered in the context of specific goals.

The ability to combine data from different studies to estimate the common effect (or mean effect), continues to be an important function of meta-analysis. However, it is not the only function. The goal of some syntheses will be to report the summary effect, but the goal of other syntheses will be to assess the dispersion as well as the mean effect, and the goal of others will be to focus on the dispersion exclusively.

For example, suppose that we are looking at the impact of a teaching intervention on student performance. Does it make sense to include studies that measured verbal skills and also studies that measured math skills? If our goal is to assess the impact on performance in general, then the answer is *Yes*. If our goal is to assess the impact on verbal skills alone, then the answer is *No*. Does it make sense to include studies

Introduction to Meta-Analysis. Michael Borenstein, L. V. Hedges, J. P. T. Higgins and H. R. Rothstein
© 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-05724-7

that enrolled middle-school students and also studies that enrolled high-school students? Again, the answer depends on the question being asked.

In some cases, however, the decisions are less clear cut. For example, does it make sense to include both randomized trials and observational studies in the same analysis? What about quasi-experimental studies? Is it acceptable to include studies that used independent groups and also studies that used matched designs? The answers to these and other questions will need to be decided in the context of the research question being addressed. Our goal here is to outline the kinds of issues that may arise and provide a context for making these kinds of decisions.

ARE THE STUDIES SIMILAR ENOUGH TO COMBINE?

From a statistical perspective, there is no restriction on the similarity of studies based on the types of participants, interventions, or exposures. However, for the analysis to be meaningful, we need to pay careful consideration to the diversity of studies in these respects. For example, the research question might be *Does Drug A reduce the risk of heart attack as compared with a placebo; when used in a population of males, age 40–60 years, with no prior history of heart attack, and a cholesterol level of 250–300 on initial screening; where the dose was between 10–15 mg per day; in studies that followed patients for at least a year, with a drop-out rate no higher than five percent, where the randomization and blinding met specific criteria; and where patients were under the care of a primary care physicians for the duration of the study?* In this case, the criteria for including studies in the analysis would be very narrow, and the goal of the analysis would be to yield a more precise estimate (and more powerful test) of the effect than would be possible with any single study. This kind of meta-analysis might be planned by a pharmaceutical company as part of the approval process, and this approach is entirely legitimate.

In most meta-analyses, however, the inclusion criteria will be broader than this. It is an important feature of a meta-analysis that it may (and usually must) address a broader question than those addressed by the primary studies it includes. Thus a certain amount of diversity among the studies is not only inevitable but also desirable. A good meta-analysis will anticipate this diversity and will interpret the findings with attention to the dispersion of results across studies. To modify the prior example by relaxing some of the criteria, a *pragmatic* review of the effects of Drug A versus placebo on the risk of heart attack might include both sexes, adults of any age with no prior history of heart attack, any cholesterol level; any dose of drug; in studies that followed patients for at least a year, with a drop-out rate no higher than twenty percent, where the randomization and blinding met specific criteria. The diversity of studies meeting these broader criteria may lead to heterogeneous results, and this heterogeneity needs to be recognized in the analysis and interpretation.

One approach to diversity is to apply the random-effects model and then address the diversity by reporting the expected range of true effects over the populations and interventions sampled. This could take the form of a prediction interval as explained

in Chapter 17. This is appropriate if the effects fall over a small range, so that the substantive implications of the finding are the same across the range.

With sufficient data, we can also explore the diversity across studies. For example, we could investigate how the effect of a drug (as compared with a placebo) depends on the sex of a patient. Assume that these studies reported outcomes for males and females separately. We now have the ability to compute a summary effect in each of these subgroups and to determine whether (and how) the effect is related to sex. If the effect is similar in both groups, then we can report that the effect is robust, something that was not possible with the more narrow criteria. If the effect varies (say, the drug is effective for males but not for females) then the meta-analysis may have yielded important information that, again, was not possible when all studies adhered to the more narrow criteria. Note, however, that for many meta-analyses there is insufficient power to do this reliably. There may also be problems of confounding.

This basic approach, that we can define the inclusion criteria narrowly and focus on the summary effect, or define the inclusion criteria more broadly and explore the dispersion, holds true for any meta-analysis. This idea can play itself out in various ways, and we explore some of them here.

CAN I COMBINE STUDIES WITH DIFFERENT DESIGNS?

The appropriate types of study to include in a meta-analysis depend primarily on the type of question being addressed. For example, meta-analyses to evaluate the effect of an intervention will tend to seek randomized trials, in which interventions are assigned in an experimental fashion so that there are no important differences between those receiving and not receiving the intervention of interest. Meta-analyses to investigate the cause of a rare disease will tend to seek case-control studies, in which the past exposures of a collection of people with the disease are compared with those of a collection of people without the disease. Meta-analyses to examine the prevalence of a condition or a belief will tend to seek cross-sectional studies or surveys, in which a single group is examined and no within-study comparisons are made. And so on. Nevertheless, for any particular question there are typically several types of study that could yield a meaningful answer. A frequent question is whether studies with different designs can be combined in a meta-analysis.

Randomized trials versus observational studies

Some have argued that systematic reviews on the effects of interventions should be limited to randomized controlled trials, since these are protected from internal bias by design, and should exclude observational studies, since the effect sizes in these are almost invariably affected by confounders (and the confounders may vary from one study to the next). In our opinion, this distinction is somewhat arbitrary. It suggests that we would be better off with a set of poor quality randomized trials

than with a set of high-quality observational studies (and leaves open the question of quasi-experimental studies). The key distinction should not be the design of the studies but the extent to which the studies are able to yield an unbiased estimate of the effect size in question.

For example, suppose we wish to evaluate the effects of going to a support group to give up smoking. We might locate five studies in which smokers were recruited and then randomly assigned to either of two conditions (invitation to a support group, or a control intervention). Because the trials use random assignment, differences between groups are attributed to differences in the effects of the interventions. If we include these trials in a meta-analysis, we are able to obtain a more precise estimate of the effect than we could from any single trial, and this effect can be attributed to the treatment. However, since trials cannot *impose* an intervention on people, the effect is of being *invited* to the support group rather than, necessarily, of attending the support group. Furthermore, the types of smokers who volunteer to be randomized into a trial may not be the types of smokers who might volunteer to join a support group.

Alternatively, suppose that we locate five studies that compared the outcomes of smokers who had voluntarily joined a support group with others who had not. Because these studies are observational, any differences allow us to draw conclusions about what proportions of people are likely to be smoking after joining a support group or not joining a support group, but do not allow us to attribute these differences to the treatment itself. For instance, those who enrolled for treatment are likely to have been more motivated to stop smoking. If we include these observational studies in a meta-analysis we are able to obtain a more precise estimate of the difference than we could from any single study, but the interpretation of this difference is subject to the same limitations as that of the primary studies.

Does it make sense to include both these randomized trials and these observational studies in the same meta-analysis? The two kinds of studies are asking different questions. The randomized trial asks if there is a relationship between treatment and outcome when we control for all other factors, while the observational study asks if there is a relationship when we do not control for these factors. Furthermore, the *treatment* is different, in that the randomized trial evaluates the effect of the invitation, and the observational study collects information based on actual participation in the support group.

It would probably not make sense to compute a summary value across both kinds of studies. The meta-analyst should first decide which question is of greater interest. Unfortunately neither would seem to address the fundamental question of whether participating in the support group increases the likelihood of stopping smoking. As is often the case, the researcher must decide between asking the sub-optimal question (about invitations) with minimal bias (through randomization) or the right question (about participation) with likely bias (using observational studies). Most would argue that randomized trials do ask highly relevant questions, allowing important conclusions to be drawn about causality even if they do not fully reflect the way intervention would be applied on a day to day basis. Thus the majority of

meta-analyses of interventions are restricted to randomized trials, at least in health care, where randomized trials have long been the established method of evaluation. Of course, some important effects of interventions, such as long-term or rare outcomes (especially harms) often cannot be studied in randomized trials, so may need to be addressed using observational studies. We would generally recommend that randomized trials and observational studies be analyzed separately, though they might be put together if they do not disagree with each other and are believed to address a common question.

Studies that used independent groups, paired groups, clustered groups

Suppose that some of the studies compared means for treatment versus control using two independent groups, others compared means using paired groups and others used cluster-randomized trials. There is no technical problem with combining data from the three kinds of studies, but we need to assume that the studies are functionally similar in all other important respects. On the one hand, studies that used different designs may differ from each other in substantive ways as well. On the other hand, these differences may be no more important than the difference between (say) studies that enrolled subjects in cities and others that enrolled subjects in rural areas. If we are looking at the impact of a vaccination, then the biological function is probably the same in all three kinds of studies. If we are looking at the impact of an educational intervention, then we would probably want to test this assumption rather than take it on faith.

Can I combine studies that report results in different ways?

Meta-analysts frequently have to deal with results reported in different ways. Suppose we are looking at ways to increase the yield of grain, and are interested in whether a high dose of fertilizer works better than the standard dose. We might find studies that measure the impact of dose by randomizing different plots to receive one of the two doses, but which measure the outcome in different ways. Some studies might measure the average growth rate for the plants while others measure the yield after a certain number of weeks (and the timings might vary across studies). Some studies might measure the proportion of plants achieving a specific growth rate while others measure the time from application to production of a certain volume of grain. We might find further studies that apply a range of doses and examine the correlation between the dose and, for example, yield.

Even within studies investigating the same outcome, results can be reported in different ways. There are two types of variation here. First, different approaches to analysis could be used. For example, two studies might focus on the proportion of plants that fail under each dose of fertilizer, but one reports this as a ratio while another reports this as a difference in proportions. Second, even the same analysis can be reported using different statistics. For example, if several studies compare

the mean yields between the two doses, some may report means with standard deviations, others means with a *p*-value, others differences in means with confidence intervals, and others *F* statistics from analysis of variance.

To what extent can all of these variations be combined in a meta-analysis? We address here only the statistical considerations, and assume that there is sound rationale for combining the different outcome measures in the analysis. Note that we have described binary outcomes (proportion of failing plants), continuous outcomes using different measurement scales (growth rate, yield), survival outcomes (time to fruit) and correlational data (dose-yield). The list of possibilities is longer, and we do not attempt a comprehensive summary of all options.

When studies are addressing the same outcome, measured in the same way, using the same approach to analysis, but presenting results in different ways, then the only obstacles to meta-analysis are practical. If sufficient information is available to estimate the effect size of interest, then a meta-analysis is possible. For instance, means with standard deviations, means with a *p*-value, and differences in means with a confidence interval can all be used to estimate the difference in mean yield (providing, in the first two situations, that the sample sizes are known). These three also allow calculation of a standardized difference in means, as does a suitable *F* statistic in combination with sample size. Detailed discussions of such conversions are provided in Borenstein *et al*. (2009).

When studies are addressing the same outcome, measured in the same way, but using different approaches to analysis, then the possibility of a meta-analysis depends on both statistical and practical considerations. One important point is that all studies in a meta-analysis must use essentially the same index of treatment effect. For example, we cannot combine a risk difference with a risk ratio. Rather, we would need to use the summary data to compute the same index for all studies.

There are some indices that are similar, if not exactly the same, and judgments are required as to whether it is acceptable to combine them. One example is odds ratios and risk ratios. When the event is rare, then these are approximately equal and can readily be combined. As the event gets more common the two diverge and should not be combined. Other indices that are similar to risk ratios are hazard ratios and rate ratios. Some people decide these are similar enough to combine; others do not. The judgment of the meta-analyst in the context of the aims of the meta-analysis will be required to make such decisions on a case by case basis.

When studies are addressing the same outcome measured in different ways, or different outcomes altogether, then the suitability of a meta-analysis depends mainly on substantive considerations. The researcher will have to decide whether a combined analysis would have a meaningful interpretation. If so, then the above statistical and practical considerations apply. A further consideration is how different scales used for different outcomes are to be dealt with. The standard approach for continuous outcome measures is to analyze each study as a standardized mean difference, so that all studies share a common metric.

There is a useful class of indices that are, perhaps surprisingly, combinable under some simple transformations. In particular, formulas are available to convert standardized mean differences, odds ratios and correlations to a common metric (see Chapter 7). These kinds of conversions require some assumptions about the underlying nature of the data, and violations of these assumptions can have an impact on the validity of the process. Also, we must remember that studies which used dichotomous data may be different in some substantive ways than studies which used continuous data, and studies measuring correlations may be different from those that compared two groups. As before, these are questions of degree rather than of qualitative differences among the studies.

HOW MANY STUDIES ARE ENOUGH TO CARRY OUT A META-ANALYSIS?

If we are working with a fixed-effect model, then it makes sense to perform a meta-analysis as soon as we have two studies, since a summary based on two or more studies yields a more precise estimate of the true effect than either study alone. Importantly, we are not concerned with dispersion in the observed effects because this is assumed to reflect nothing more than sampling error. There might be a concern that by reporting a summary effect we are implying a level of certainty that is not warranted. In fact, though, the summary effect is qualified by a confidence interval that describes the uncertainty of the estimate. Additionally, research shows that if we fail to provide this information researchers will impose their own synthesis on the data, which will invariably be less accurate and more idiosyncratic than the value than we compute using known formulas.

In most cases however, we should be working with the random-effects model, where the dispersion in effects is assumed to be real (at least in part). Unlike the fixed-effect analysis, where the estimate of the error is based on sampling theory (and therefore reliable), in a random-effects analysis, our estimate of the error may itself be unreliable. Specifically, when based on a small number of studies, the estimate of the between-studies variance (T^2), may be substantially in error. The standard error of the summary effect is based (in part) on this value, and therefore, if we present a summary effect with confidence interval, not only is the point estimate likely to be wrong but the confidence interval may provide a false sense of assurance.

A separate problem is that in a random-effects analysis, our understanding of the dispersion affects not only our estimate of the summary effect but also the thrust of the analysis. In other words, if the effect is consistent across studies we would report that the effect is robust. By contrast, if the effect varies substantially from study to study we would want to consider the impact of the dispersion. The problem is that when we have only a few studies to work with, we may not know what the dispersion actually looks like.

This suggests that if the number of studies is small enough it might be better not to summarize them statistically. However many statisticians would argue that, when

faced with a series of studies, people have an almost irresistible tendency to draw some summary conclusions from them. Experience has shown that seemingly intuitive *ad hoc* summaries (such as vote counting, Chapter 28) are also often highly misleading. This suggests that a statistical summary with known, but perhaps poor, properties (such as high uncertainty) may be superior to inviting an *ad hoc* summary with unknown properties.

In sum, when the number of studies is small, there are no really good options. As a starting point we would suggest reporting the usual statistics and then explaining the limitations as clearly as possible. This helps preclude the kinds of ad hoc analyses mentioned in the previous paragraph, and is an accurate representation of what we can do with limited data.

SUMMARY POINTS

- The question of whether or not it makes sense to perform a meta-analysis is a question of matching the synthesis to the research question.
- If our goal is to report a summary effect, then the populations and interventions (and other variables) in the studies should match those in our target population. If our goal is to report on the dispersion of effects as a function of a covariate, then the synthesis must include the relevant studies and the analysis should focus on the differences in effects.
- Generally, we need to be aware of substantive differences among studies, but technical differences can be addressed in the analysis.
- A potentially serious problem exists when the synthesis is based on a small number of studies. Without sufficient numbers of studies we will have a problem estimating the between-studies variance, which has important implications for many aspects of the analysis.

Further Reading

Ioannidis, J.P.A., Patsopoulos, N.A., Rothstein, H.R. (2008). Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ* 336: 1413–1415.