CHAPTER 2

Why Perform a Meta-Analysis

INTRODUCTION

Why perform a meta-analysis? What are the advantages of using statistical methods to synthesize data rather than taking the results that had been reported for each study and then having these collated and synthesized by an expert?

In this chapter we start at the point where we have already selected the studies to be included in the review, and are planning the synthesis itself. We do not address the differences between systematic reviews and narrative reviews in the process of locating and selecting studies. These differences can be critically important, but (as always) our focus is on the data analysis rather than the full process of the review.

The goal of a synthesis is to understand the results of any study in the context of all the other studies. First, we need to know whether or not the effect size is consistent across the body of data. If it *is* consistent, then we want to estimate the effect size as accurately as possible and to report that it is robust across the kinds of studies included in the synthesis. On the other hand, if it varies substantially from study to study, we want to quantify the extent of the variance and consider the implications.

Meta-analysis is able to address these issues whereas the narrative review is not. We start with an example to show how meta-analysis and narrative review would approach the same question, and then use this example to highlight the key differences between the two.

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

THE STREPTOKINASE META-ANALYSIS

During the time period beginning in 1959 and ending in 1988 (a span of nearly 30 years) there were a total of 33 randomized trials performed to assess the ability of streptokinase to prevent death following a heart attack. Streptokinase, a so-called *clot buster* which is administered intravenously, was hypothesized to dissolve the clot causing the heart attack, and thus increase the likelihood of survival. The trials all followed similar protocols, with patients assigned at random to either treatment or placebo. The outcome, whether or not the patient died, was the same in all the studies.

The trials varied substantially in size. The median sample size was slightly over 100 but there was one trial with a sample size in the range of 20 patients, and two large scale trials which enrolled some 12,000 and 17,000 patients, respectively. Of the 33 studies, six were statistically significant while the other 27 were not, leading to the perception that the studies yielded conflicting results.

In 1992 Lau *et al*. published a meta-analysis that synthesized the results from the 33 studies. The presentation that follows is based on the Lau paper (though we use a risk ratio where Lau used an odds ratio).

The forest plot (Figure 2.1) provides context for the analysis. An effect size to the left of center indicates that treated patients were more likely to survive, while an

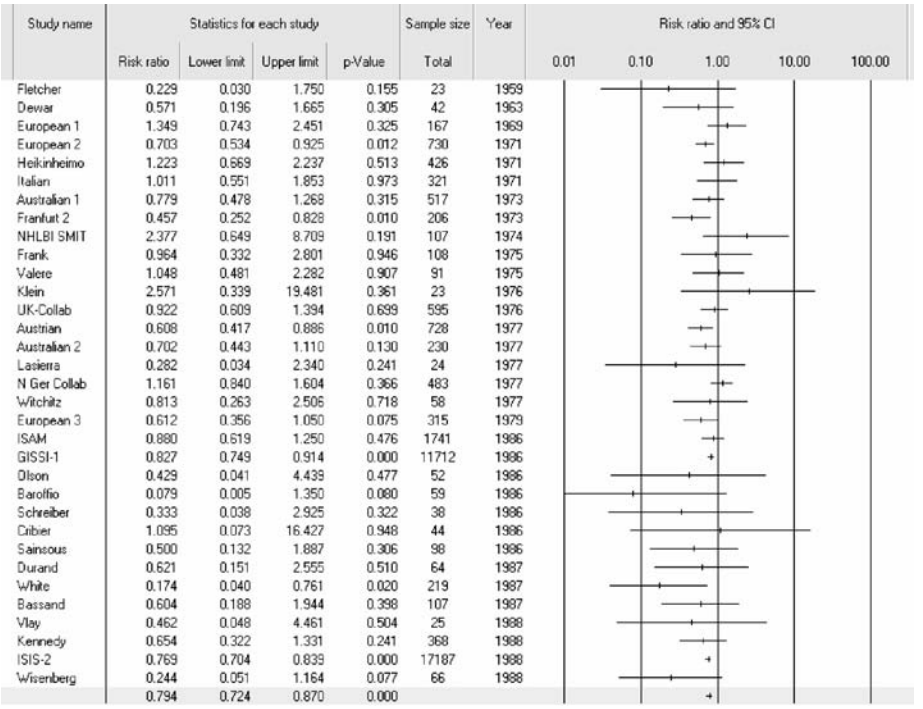

Figure 2.1 Impact of streptokinase on mortality (adapted from Lau *et al.*, 1992).

effect size to the right of center indicates that control patients were more likely to survive.

The plot serves to highlight the following points.

- The effect sizes are reasonably consistent from study to study. Most fall in the range of 0.50 to 0.90, which suggests that it would be appropriate to compute a summary effect size.
- The summary effect is a risk ratio of 0.79 with a 95% confidence interval of 0.72 to 0.87 (that is, a 21% decrease in risk of death, with 95% confidence interval of 13% to 28%). The *p*-value for the summary effect is 0.0000008.
- The confidence interval that bounds each effect size indicates the precision in that study. If the interval excludes 1.0, the *p*-value is less than 0.05 and the study is statistically significant. Six of the studies were statistically significant while 27 were not.

In sum, the treatment reduces the risk of death by some 21%. And, this effect was reasonably consistent across all studies in the analysis.

Over the course of this volume we explain the statistical procedures that led to these conclusions. Our goal in the present chapter is simply to explain that meta-analysis does offer these mechanisms, whereas the narrative review does not. The key differences are as follows.

STATISTICAL SIGNIFICANCE

One of the first questions asked of a study is the statistical significance of the results. The narrative review has no mechanism for synthesizing the *p*-values from the different studies, and must deal with them as discrete pieces of data. In this example six of the studies were statistically significant while the other 27 were not, which led some to conclude that there was evidence against an effect, or that the results were inconsistent (see vote counting in Chapter 28). By contrast, the meta-analysis allows us to combine the effects and evaluate the statistical significance of the summary effect. The *p*-value for the summary effect is $p = 0.0000008$.

While one might assume that 27 studies failed to reach statistical significance because they reported small effects, it is clear from the forest plot that this is not the case. In fact, the treatment effect in many of these studies was actually *larger* than the treatment effect in the six studies that *were* statistically significant. Rather, the reason that 82% of the studies were not statistically significant is that these studies had small sample sizes and low statistical power. In fact, as discussed in Chapter 29, most had power of less than 20%. By contrast, power for the meta-analysis exceeded 99.9% (see Chapter 29).

As in this example, if the goal of a synthesis is to test the null hypothesis, then meta-analysis provides a mathematically rigorous mechanism for this purpose. However, meta-analysis also allows us to move beyond the question of

statistical significance, and address questions that are more interesting and also more relevant.

CLINICAL IMPORTANCE OF THE EFFECT

Since the point of departure for a narrative review is usually the *p*-values reported by the various studies, the review will often focus on the question of whether or not the body of evidence allows us to reject the null hypothesis. There is no good mechanism for discussing the magnitude of the effect. By contrast, the meta-analytic approaches discussed in this volume allow us to compute an estimate of the effect size for each study, and these effect sizes fall at the core of the analysis.

This is important because the effect size is what we care about. If a clinician or patient needs to make a decision about whether or not to employ a treatment, they want to know if the treatment reduces the risk of death by 5% or 10% or 20%, and this is the information carried by the effect size. Similarly, if we are thinking of implementing an intervention to increase the test scores of students, or to reduce the number of incarcerations among at-risk juveniles, or to increase the survival time for patients with pancreatic cancer, the question we ask is about the magnitude of the effect. The *p*-value can tell us only that the effect is not zero, and to report simply that the effect is not zero is to miss the point.

CONSISTENCY OF EFFECTS

When we are working with a collection of studies, it is critically important to ask whether or not the effect size is consistent across studies. The implications are quite different for a drug that consistently reduces the risk of death by 20%, as compared with a drug that reduces the risk of death by 20% on average, but that increases the risk by 20% in some populations while reducing it by 60% in others.

The narrative review has no good mechanism for assessing the consistency of effects. The narrative review starts with *p*-values, and because the *p*-value is driven by the size of a study as well as the effect in that study, the fact that one study reported a *p*-value of 0.001 and another reported a *p*-value of 0.50 does not mean that the effect was larger in the former. The *p*-value of 0.001 *could* reflect a large effect size but it could also reflect a moderate or small effect in a large study (see the GISSI-1 study in Figure 2.1, for example). The *p*-value of 0.50 *could* reflect a small (or nil) effect size but could also reflect a large effect in a small study (see the Fletcher study, for example).

This point is often missed in narrative reviews. Often, researchers interpret a nonsignificant result to mean that there is no effect. If some studies are statistically significant while others are not, the reviewers see the results as conflicting. This problem runs through many fields of research. To borrow a phrase from Cary Grant's character in *Arsenic and Old Lace*, we might say that it practically gallops.

Schmidt (1996) outlines the impact of this practice on research and policy. Suppose an idea is proposed that will improve test scores for African-American children. A number of studies are performed to test the intervention. The effect size is positive and consistent across studies but power is around 50%, and only around 50% of the studies yield statistically significant results. Researchers report that the evidence is 'conflicting' and launch a series of studies to determine why the intervention had a positive effect in some studies but not others (Is it the teacher's attitude? Is it the students' socioeconomic status?), entirely missing the point that the effect was actually consistent from one study to the next. No pattern can be found (since none exists). Eventually, researchers decide that the issue cannot be understood. A promising idea is lost, and a perception builds that research is not to be trusted. A similar point is made by Meehl (1978, 1990).

Rossi (1997) gives an example from the field of memory research that shows what can happen to a field of research when reviewers work with discrete *p*-values. The issue of whether or not researchers could demonstrate the spontaneous recovery of previously extinguished associations had a bearing on a number of important learning theories, and some 40 studies on the topic were published between 1948 and 1969. Evidence of the effect (that is, statistically significant findings) was obtained in only about half the studies, which led most texts and reviews to conclude that the effect was ephemeral and 'the issue was not so much resolved as it was abandoned' (p. 179). Later, Rossi returned to these studies and found that the average effect size (*d*) was 0.39. If we assume that this is the population effect size, the mean power for these studies would have been slightly under 50%. On this basis we would expect about half the studies to yield a significant effect, which is exactly what happened.

Even worse, when the significant study was performed in one type of sample and the nonsignificant study was performed in another type of sample, researchers would sometimes interpret this difference as meaning that the effect existed in one population but not the other. Abelson (1997) notes that if a treatment effect yields a *p*-value of 0.07 for wombats and 0.05 for dingbats we are likely to see a discussion explaining why the treatment is effective only in the latter group—completely missing the point that the treatment effect may have been virtually identical in the two. The treatment effect may have even been *larger* for the wombats if the sample size was smaller.

By contrast, meta-analysis completely changes the landscape. First, we work with effect sizes (not *p*-values) to determine whether or not the effect size is consistent across studies. Additionally, we apply methods based on statistical theory to allow that some (or all) of the observed dispersion is due to random sampling variation rather than differences in the true effect sizes. Then, we apply formulas to partition the variance into random error versus real variance, to quantify the true differences among studies, and to consider the implications of this variance. In the Schmidt and the Rossi examples, a meta-analysis might have found that the effect size was

consistent across studies, and that all of the observed variation in effects could be attributed to random sampling error.

SUMMARY POINTS

- Since the narrative review is based on discrete reports from a series of studies, it provides no real mechanism for synthesizing the data. To borrow a phrase from Abelson, it involves *doing arithmetic with words*. And, when the words are based on *p*-values *the words are the wrong words*.
- By contrast, in a meta-analysis we introduce two fundamental changes. First, we work directly with the effect size from each study rather than the *p*-value. Second, we include all of the effects in a single statistical synthesis. This is critically important for the goal of computing (and testing) a summary effect. Meta-analysis also allows us to assess the dispersion of effects, and distinguish between real dispersion and spurious dispersion.