CHAPTER 43

Criticisms of Meta-Analysis

INTRODUCTION

While meta-analysis has been widely embraced by large segments of the research community, this point of view is not universal and people have voiced numerous criticisms of meta-analysis.

Some of these criticisms are worth mentioning for their creative use of metaphor. The first set of Cochrane reviews dealt with studies in neonatology, and one especially creative critic, cited by Mann (1990), called the reviewers *an obstetrical Baader Meinhof gang* (*obstetrical* being a reference to the field of research, and *Baader Meinhof gang* a reference to the terrorist group that operated in Europe during the 1970s and 1980s).

Others were more circumspect in their comments. Eysenck (1978) criticized a meta-analysis as *an exercise in mega-silliness*. Shapiro (1994) published a paper entitled *Meta-Analysis / Shmeta Analysis*. Feinstein (1995) wrote an editorial in which he referred to meta-analysis as 'statistical alchemy for the 21st century'.

Introduction to Meta-Analysis. Michael Borenstein, L. V. Hedges, J. P. T. Higgins and H. R. Rothstein
 ISBN: 978-0-470-05724-7

These critics share not only an affinity for allegory and alliteration but also a common set of concerns about meta-analysis. In this chapter we address the following criticisms that have been leveled at meta-analysis, as follows.

- One number cannot summarize a research field
- The file drawer problem invalidates meta-analysis
- Mixing apples and oranges
- Garbage in, garbage out
- Important studies are ignored
- Meta-analysis can disagree with randomized trials
- Meta-analyses are performed poorly

After considering each of these questions in turn, we ask whether a traditional narrative review fares any better than a systematic review on these criticisms. And, we summarize the legitimate criticisms of meta-analysis that need to be considered whenever meta-analysis is applied.

ONE NUMBER CANNOT SUMMARIZE A RESEARCH FIELD

Criticism

A common criticism of meta-analysis is that the analysis focuses on the summary effect, and ignores the fact that the treatment effect may vary from study to study. Bailar (1997), for example, writes, 'Any attempt to reduce results to a single value, with confidence bounds, is likely to lead to conclusions that are wrong, perhaps seriously so.'

Response

In fact, the goal of a meta-analysis should be to *synthesize* the effect sizes, and not simply (or necessarily) to report a summary effect. If the effects are consistent, then the analysis shows that the effect is robust across the range of included studies. If there is modest dispersion, then this dispersion should serve to place the mean effect in context. If there is substantial dispersion, then the focus should shift from the summary effect to the dispersion itself. Researchers who report a summary effect and ignore heterogeneity are indeed missing the point of the synthesis.

THE FILE DRAWER PROBLEM INVALIDATES META-ANALYSIS

Criticism

While the meta-analysis will yield a mathematically sound synthesis of the studies included in the analysis, if these studies are a biased sample of all possible studies, then the mean effect reported by the meta-analysis will reflect this bias. Several lines of evidence show that studies finding relatively high treatment effects are more likely to be published than studies finding lower treatment effects. The latter,

unpublished, research lies dormant in the researchers' filing cabinets, and has led to the use of the term *file drawer problem* for meta-analysis.

Response

Since published studies are more likely to be included in a meta-analysis than their unpublished counterparts, there is a legitimate concern that a meta-analysis may overestimate the true effect size.

Chapter 30 (entitled *Publication Bias*) explores this question in some detail. In that chapter we discuss methods to assess the likely amount of bias in any given meta-analysis, and to distinguish between analyses that can be considered robust to the impact of publication bias from those where the results should be considered suspect.

We must remember that publication bias is a problem for any kind of literature search. The problem exists for the clinician who searches a database to locate primary studies about the utility of a treatment. It exists for persons performing a narrative review. And, it exists for persons performing a meta-analysis. Publication bias has come to be identified with meta-analysis because meta-analysis has the goal of providing a more accurate synthesis than other methods, and so we are concerned with biases that will interfere with this goal. However, it would be a mistake to conclude that this bias is not a problem for the narrative review. There, it is simply easier to ignore.

MIXING APPLES AND ORANGES

Criticism

A common criticism of meta-analysis is that researchers combine different kinds of studies (*apples and oranges*) in the same analysis. The argument is that the summary effect will ignore possibly important differences across studies.

Response

The studies that are brought together in a meta-analysis will inevitably differ in their characteristics, and the difficulty is deciding just how similar they need to be. The decision as to which studies should be included is always a judgment, and people will have different opinions on the appropriateness of combining results across studies. Some meta-analysts may make questionable judgments, and some critics may make unreasonable demands on similarity.

We need to remember that meta-analyses almost always, by their very nature, address broader questions than individual studies. Hence a meta-analysis may be thought of as asking a question about fruit, for which both apples and oranges (and indeed pears and melons) contribute valuable information. One of the strengths of meta-analysis is that the consistency, and hence generalizability, of findings from one type of study to the next can be assessed formally.

Of course, we always need to remember that we are dealing with different kinds of fruit, and to anticipate that effects may vary from one kind to the other. It is a further strength of meta-analysis that these differences, if identified, can be investigated formally. Assume, for example, that a treatment is very effective for patients with acute symptoms but has no effect for patients with chronic symptoms. If we were to combine data from studies that used both types of patients, and conclude that the treatment was modestly effective (on average), this conclusion would not be accurate for either kind of patient. If we were to restrict our attention to studies in only patients with acute symptoms, or only patients with chronic symptoms, we could report how the treatment worked with one type of patient, but could only speculate about how it would have worked with the other type. By contrast, a meta-analysis that includes data for both types of patients may allow us to address this question empirically.

GARBAGE IN, GARBAGE OUT

Criticism

The often-heard metaphor *garbage in, garbage out* refers to the notion that if a meta-analysis includes many low-quality studies, then fundamental errors in the primary studies will be carried over to the meta-analysis, where the errors may be harder to identify.

Response

Rather than thinking of meta-analysis as a process of *garbage in, garbage out* we can think of it as a process of waste management. A systematic review or meta-analysis will always have a set of inclusion criteria and these should include criteria based on the quality of the study. For trials, we may decide to limit the studies to those that use random assignment, or a placebo control. For observational studies we may decide to limit the studies to those where confounders were adequately addressed in the design or analysis. And so on. In fact, it is common in a systematic review to start with a large pool of studies and end with a much smaller set of studies after all inclusion/exclusion criteria are applied.

Nevertheless, the studies that do make it as far as a meta-analysis are unlikely to be perfect, and close attention should be paid to the possibility of bias due to study limitations. A meta-analysis of a collection of studies that is each biased in the same direction will suffer from the same bias and have higher precision. In this case, performing a meta-analysis can indeed be more dangerous than not performing one.

However, as noted in the response to the previous criticism about *apples and oranges*, a strength of meta-analysis is the ability to investigate whether variation in characteristics of studies is related to the size of the effect. Suppose that ten studies used an acceptable method to randomize patients while another ten used a questionable method. In the analysis we can compare the effect size in these two subgroups, and determine whether or not the effect size actually differs between

the two. Note that such analyses (those comparing effects in different subgroups) can have very low power so need to be interpreted carefully, especially when there are not many studies within subgroups.

IMPORTANT STUDIES ARE IGNORED

Criticism

Whereas the *garbage in, garbage out* problem relates to the inclusion of studies that perhaps should not be included, a common complementary criticism is that important studies were left out. The criticism is often leveled by people who are uncomfortable with the findings of a meta-analysis. For example, a meta-analysis to assess the effects of antioxidant supplements (beta-carotene, vitamin A, vitamin C, vitamin E, and selenium) on overall mortality was met with accusations on the web site of the Linus Pauling Institute (Oregon State University) that in this 'flawed analysis of flawed data' the authors looked at 815 human clinical trials of antioxidant supplements, but only 68 were included in the meta-analysis.

Response

We have explained that systematic reviews and meta-analyses require explicit mechanisms for deciding which studies to include and which ones to exclude. These eligibility criteria are determined by a combination of considerations of relevance and considerations of bias, and are typically decided before the search for studies is implemented. Studies should be sufficiently similar to yield results that can be interpreted, and sufficiently free of bias to yield results that can be believed. For both purposes, judgments are required, and not all meta-analysts or readers would reach the same judgments on each occasion. Importantly, in meta-analysis the criteria are transparent and are described as part of the report.

META-ANALYSIS CAN DISAGREE WITH RANDOMIZED TRIALS

Criticism

LeLorier *et al*. (1997) published a paper in which they pointed out that meta-analyses sometimes yield different results than large scale randomized trials. Specifically, they located cases in the medical literature where someone had performed a meta-analysis, and someone else subsequently performed a large scale randomized trial that addressed the same question (e.g. *Does the treatment work*?). The authors reported that the results of the meta-analysis and the randomized trial *matched* (both were statistically significant, or neither was statistically significant) in about 66% of cases, but did not match (one was statistically significant but the other was not) in the remaining 34%. Since randomized trials are generally accepted as the gold standard they conclude that some 34% of these meta-analyses were wrong, and that meta-analyses in general cannot be trusted.

Response

There are both technical and conceptual flaws in this criticism. The technical flaws relate to the question of what we mean by *matching*, and the authors' decision to define *matching* as both studies being (or not being) statistically significant. The discussion that follows draws in part on comments by Ioannidis *et al*. (1998), Lelorier *et al*. (1997, 536–543) and others (see further readings at the end of this chapter).

Consider Figure 43.1, which shows a meta-analysis of five randomized controlled trials (RCTs) at the top, and a subsequent large-scale randomized trial at the bottom.

In this fictional example the five studies in the meta-analysis each showed precisely the same effect, an odds ratio of 0.80. The summary effect in the meta-analysis is (it follows) an odds ratio of 0.80. And, the subsequent study showed the same effect, an odds ratio of 0.80.

The only difference between the summary effect in the meta-analysis and the effect in the subsequent study is that the former is reported with greater precision (since it is based on more data) and therefore yields a *p*-value under 0.05. By the LeLorier criterion these two conclusions would be seen as conflicting, when in fact they have the identical effect size.

Additionally, LeLorier concludes that in the face of this conflict the single randomized trial is correct and the meta-analysis is wrong. In fact, though, it is the meta-analysis, which incorporates data from five randomized trials rather than one, that has the more powerful position. (What would happen if we performed a new meta-analysis which incorporated the most recent randomized trial? Would

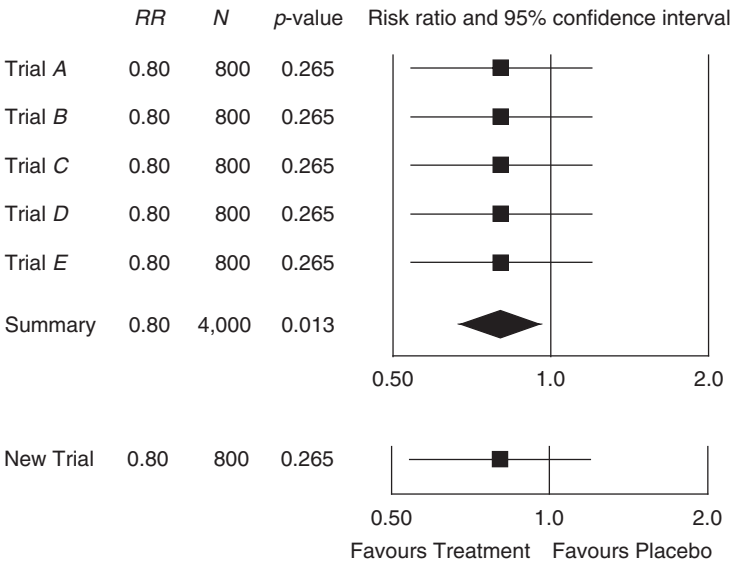

Figure 43.1 Forest plot of five fictional studies and a new trail (consistent effects).

LeLorier now see this new meta-analysis as flawed?) In fact, the real issue is not that a meta-analysis disagrees with a randomized trial, but that randomized trials disagree with each other.

At a meeting of The Cochrane Collaboration in Baltimore (1996), a plenary speaker made the same argument being made by LeLorier *et al*. (that meta-analyses sometimes yield different results than randomized trials) and, like the paper, cited the statistic that roughly a third of meta-analyses fail to match the *comparable* randomized trial. A distinguished member of the audience, Harris Cooper, asked the speaker if he knew what percentage of randomized trials fail to match the next randomized trial on the same topic. It turns out that the percentage is roughly a third.

However, to move on to a more interesting question, let's assume that the results from a meta-analysis and a randomized trial really do differ. Suppose that the meta-analysis yields a risk ratio of 0.67 (with a 95% confidence interval of 0.84 to 0.77) while the new trial yields a risk ratio of 0.91 (0.82 to 1.0). According to the meta-analysis the treatment reduces the risk by at least 23%, while the new trial says that its impact is no more than 18%.

In this case the effect *is different* in the two analyses, but that does not mean that one is wrong and the other is right. Rather, it behooves us to ask why the two results should differ, much as we would if we had two large scale randomized trials with significantly different results. Often, it will turn out that the different analyses either were asking different questions or differed in some important way. A careful examination of the differences in method, patient population, and so on, may help to uncover the source of the difference.

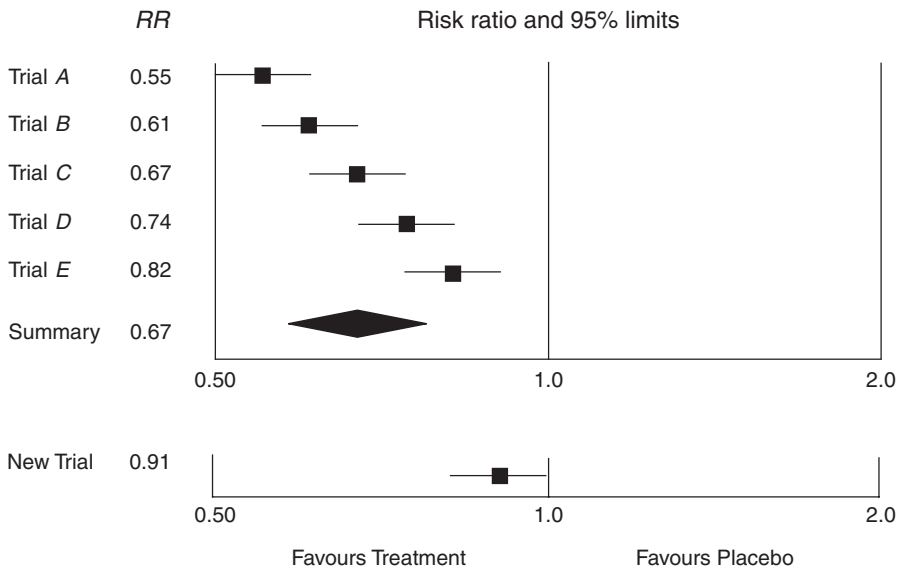

Figure 43.2 Forest plot of five fictional studies and a new trial (heterogeneous effects).

Consider the following scenario, depicted in Figure 43.2. A new compound is introduced, which is meant to minimize neurological damage in stroke patients. In 1990, the compound is tested in a randomized trial involving patients with a very poor prognosis, and yields a risk ratio of 0.55. Based on these encouraging results, in 1994 it is tested in patients with a somewhat better prognosis. Since the patients in this group are more likely to recover without treatment, the impact of the drug is less pronounced, and the risk ratio is 0.61. By 1998 the drug is being tested with all patients, and the risk ratio is 0.82. These are the studies included in the meta-analysis. The new trial is performed using a relatively healthy population and (following the trend seen in the meta-analysis) yields a risk ratio of 0.91.

If one were to report a mean effect of 0.67 for the meta-analysis versus 0.91 for the new trial there would indeed be a problem. But, as we have emphasized throughout this volume, the meta-analysis should focus on the dispersion in effects and try to identify the reason for the dispersion. In this example, using either health status or study year as a covariate we can explain the pattern of the effects, and would have predicted that the effect size in the new study would fall where it did.

META-ANALYSES ARE PERFORMED POORLY

Criticism

John C. Bailar, in an editorial for the *New England Journal of Medicine* (Bailar, 1997), writes that mistakes such as those outlined in the prior criticisms are common in meta-analysis. He argues that a meta-analysis is inherently so complicated that mistakes by the persons performing the analysis are all but inevitable. He also argues that journal editors are unlikely to uncover all of these mistakes.

Response

The specific points made by Bailar about problems with meta-analysis are entirely reasonable. He is correct that many meta-analyses contain errors, some of them important ones. His list of potential (and common) problems can serve as a bullet list of mistakes to avoid when performing a meta-analysis.

However, the mistakes cited by Bailar are flaws in the application of the method, rather than problems with the method itself. Many primary studies suffer from flaws in the design, analyses, and conclusions. In fact, some serious kinds of problems are endemic in the literature. The response of the research community is to locate these flaws, consider their impact for the study in question, and (hopefully) take steps to avoid similar mistakes in the future. In the case of meta-analysis, as in the case of primary studies, we cannot condemn a method because some people have used that method improperly. As Bob Abelson once remarked in a related context, 'Think of all the things that people abuse. There are college educations. And oboes.'

IS A NARRATIVE REVIEW BETTER?

In his editorial Bailar concludes that, until such time as the quality of meta-analyses is improved, he would prefer to work with the traditional narrative reviews: 'I still prefer conventional narrative reviews of the literature, a type of summary familiar to readers of the countless review articles on important medical issues.'

We disagree with the conclusion that narrative reviews are preferable to systematic reviews, and that meta-analyses should be avoided. The narrative review suffers from every one of the problems cited for the systematic review. The only difference is that, in the narrative review, these problems are less obvious. For example:

- The process of determining which studies to include in the systematic review or meta-analysis is difficult and prone to error. But at least there is a set of criteria for determining which studies to include. If the narrative review also has such criteria, then it is subject to the same kinds of error. If not, then we have no way of knowing how studies are being selected, which only compounds the problem.
- Meta-analyses can be affected by publication bias. But the same biases exist in the material upon which narrative reviews are based. Meta-analysis offers a means to investigate the likelihood of these biases and their potential impact on the results.
- Meta-analyses may be based on low quality primary research. But a good systematic review includes a careful assessment of the included studies with regard to their quality or risk of bias, and meta-analytic methods enable formal examination of the potential impact of these biases. A narrative reviewer may discount a study because of a belief that the results are suspect for some reason. However, a limitation can be found for virtually any study, so in the absence of a systematic quality assessment of every study, a narrative reviewer is free to be suspect about any study's results and to lay the blame on one or more of its limitations.
- The weighting scheme in a meta-analysis may give a lot (or little) weight to specific studies in ways that may appear inappropriate. But in a meta-analysis the weights reflect specific goals (to minimize the variance, or to reflect the range of effects) and the weighting scheme is detailed as part of the report, so a reader is able to agree or disagree with it. By contrast, in the case of a narrative review, the reviewer assigns *weights* to studies based on criteria that he or she does not communicate, and may not even be able to fully articulate. Here, the problem involves not only the relative weights assigned to small or large studies. It extends also to the propensity of one reviewer to focus on effect sizes, and of another to focus on (and possibly be misled by) significance tests.
- Some meta-analyses focus on the summary effect and ignore the pattern of dispersion in the results. To ignore the dispersion is clearly a mistake both in a narrative review and in a meta-analysis. However, meta-analysis provides a full complement of tools to assess the pattern of dispersion, and possibly to explain it as a function of study-level covariates. By contrast, it would be an almost

impossible task for a narrative reviewer to accurately assess the pattern of dispersion, or to understand its relationship to other variables.

- In support of the narrative review, Bailer cites the role of the expert with substantive knowledge of the field, who can identify flaws in specific studies, or the presence of potentially important moderator variables. However, this is not an advantage of the narrative review, since the expert is expected to play the same role in a meta-analysis. Steve Goodman (1991) wrote, 'The best meta-analyses knit clinical insight with quantitative results in a way that enhances both. They should combine the careful thought and synthesis of a good review with the scientific rigor of a good experiment.'

CONCLUDING REMARKS

Most of the criticisms raised in this chapter point to problems with meta-analysis, and make the implicit argument that the problem would go away if we dispensed with the meta-analysis and performed a narrative review. We have argued that these problems exist also for the narrative review, and that the key advantage of the systematic approach of a meta-analysis is that all steps are clearly described so that the process is transparent.

Is meta-analysis so difficult that the method should be abandoned, as some have suggested? Our answer is obviously that it is not. Most of the criticisms raised deal with the application of the method, rather than with the method itself. What we should do is take the valid criticisms seriously and protect against them in planned analyses and by thoughtful interpretation of results.

Steven Goodman, in his editorial for *Annals of Internal Medicine* (1991) writes,

> Regardless of the summary number, meta-analysis should shed light on why trial results differ; raise research and editorial standards by calling attention to the strengths and weaknesses of the body of research in an area; and give the practitioner an objective view of the research literature, unaffected by the sometimes distorting lens of individual experience and personal preference that can affect a less structured review.

SUMMARY POINTS

- Meta-analyses are sometimes criticized for a number of flaws, and critics have argued that narrative reviews provide a better solution.
- Some of these flaws, such as the idea that we cannot summarize a body of data in a single number, are based on misunderstandings of meta-analysis.
- Many of the flaws (such as ignoring dispersion in effect sizes) reflect problems in the way that meta-analysis is used, rather than problems in the method itself.

- Other flaws (such as publication bias) are a problem for meta-analysis. However, the suggestion that these problems do not exist in narrative reviews is wrong. These problems exist for narrative reviews as well, but are simply easier to ignore since those reviews lack a clear structure.

Further Reading

Bailar, J.C. (1995). The practice of meta-analysis. *J Clin Epidemiol* 48: 149–157.

Bailar, J.C. (1997). The promise and problems of meta-analysis. *New Engl J Med* 337: 559–561.

Boden, W.E. (1992). Meta-analysis in clinical trials reporting: has a tool become a weapon? *Am J Cardiol* 69: 681–686.

Egger, M, & Davey Smith, G. (1998). Bias in location and selection of studies. *BMJ* 316: 61–66.

Eysenck, H.J. (1978). An exercise in mega-silliness. *Am Psychol* 33: 517.

Lau, J., Ioannidis, J.P., Terrin, N., Schmid, C.H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ* 333: 597–600.

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med* 337: 536–543.

Responses to Lelorier *et al*.

- Bent, S., Kerlikowske, K., & Grady, D. (1998). *NEJM*, 338(1), 60.
- Imperiale, T.F. (1998). *NEJM*, 338(1), 61.
- Ioannidis, J.P., Cappelleri, J.C., & Lau, J. (1998). *NEJM*, 338(1), 59.
- Khan, S., Williamson, P., & Sutton, R. (1998). *NEJM*, 338(1), 60–61
- LeLorier, J., & Gregoire, G. (1998). *NEJM*, 338(1), 61–62.
- Song, F. J., & Sheldon, T. A. (1998). *NEJM*, 338(1), 60.
- Stewart, L. A., Parmar, M. K., & Tierney, J. F. (1998). *NEJM*, 338(1), 61

Sharpe, D. (1997) Of apples and oranges, file drawers and garbage: why validity issues in meta-analysis will not go away. *Clin Psychol Rev* 17: 881–901.

Thompson, S.G & Pocock, S. J. (1991). Can meta-analysis be trusted? *Lancet* 338: 1127–1130.