Introduction to Meta-Analysis

Michael Borenstein
Biostat, Inc, New Jersey, USA.

Larry V. Hedges
Northwestern University, Evanston, USA.

Julian P.T. Higgins
MRC, Cambridge, UK.

Hannah R. Rothstein
Baruch College, New York, USA.

WILEY
A John Wiley and Sons, Ltd., Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloguing-in-Publication Data

Introduction to meta-analysis / Michael Borenstein . . . [et al.].
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-05724-7 (cloth)
1. Meta-analysis. I. Borenstein, Michael.
[DNLM: 1. Meta-Analysis as Topic. WA 950 I614 2009].
R853.M48I58 2009
610.72—dc22

2008043732

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-05724-7

Set in 10.5/13pt Times by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed in the UK by TJ International, Padstow, Cornwall

Contents

Preface

In his best-selling book *Baby and Child Care*, Dr. Benjamin Spock wrote 'I think it is preferable to accustom a baby to sleeping on his stomach from the beginning if he is willing'. This statement was included in most editions of the book, and in most of the 50 million copies sold from the 1950s into the 1990s. The advice was not unusual, in that many pediatricians made similar recommendations at the time.

During this same period, from the 1950s into the 1990s, more than 100,000 babies died of sudden infant death syndrome (SIDS), also called *crib death* in the United States and *cot death* in the United Kingdom, where a seemingly healthy baby goes to sleep and never wakes up.

In the early 1990s, researchers became aware that the risk of SIDS decreased by at least 50% when babies were put to sleep on their backs rather than face down. Governments in various countries launched educational initiatives such as the *Back to sleep* campaigns in the UK and the US, which led to an immediate and dramatic drop in the number of SIDS deaths.

While the loss of more than 100,000 children would be unspeakably sad in any event, the real tragedy lies in the fact that many of these deaths could have been prevented. Gilbert *et al*. (2005) write

> 'Advice to put infants to sleep on the front for nearly half a century was contrary to evidence available from 1970 that this was likely to be harmful. Systematic review of preventable risk factors for SIDS from 1970 would have led to earlier recognition of the risks of sleeping on the front and might have prevented over 10,000 infant deaths in the UK and at least 50,000 in the Europe, the USA and Australasia.'

AN ETHICAL IMPERATIVE

This example is one of several cited by Sir Iain Chalmers in a talk entitled *The scandalous failure of scientists to cumulate scientifically* (Chalmers, 2006). The theme of this talk was that we live in a world where the utility of almost any intervention will be tested repeatedly, and that rather than looking at any study in isolation, we need to look at the body of evidence. While not all systematic reviews carry the urgency of SIDS, the logic of looking at the body of evidence, rather than trying to understand studies in isolation, is always compelling.

Meta-analysis refers to the statistical synthesis of results from a series of studies. While the statistical procedures used in a meta-analysis can be applied to any set of data, the synthesis will be meaningful only if the studies have been collected

systematically. This could be in the context of a systematic review, the process of systematically locating, appraising, and then synthesizing data from a large number of sources. Or, it could be in the context of synthesizing data from a select group of studies, such as those conducted by a pharmaceutical company to assess the efficacy of a new drug.

If a treatment effect (or effect size) is consistent across the series of studies, these procedures enable us to report that the effect is robust across the kinds of populations sampled, and also to estimate the magnitude of the effect more precisely than we could with any of the studies alone. If the treatment effect varies across the series of studies, these procedures enable us to report on the range of effects, and may enable us to identify factors associated with the magnitude of the effect size.

FROM NARRATIVE REVIEWS TO SYSTEMATIC REVIEWS

Prior to the 1990s, the task of combining data from multiple studies had been primarily the purview of the narrative review. An expert in a given field would read the studies that addressed a question, summarize the findings, and then arrive at a conclusion – for example, that the treatment in question was, or was not, effective. However, this approach suffers from some important limitations.

One limitation is the subjectivity inherent in this approach, coupled with the lack of transparency. For example, different reviewers might use different criteria for deciding which studies to include in the review. Once a set of studies has been selected, one reviewer might give more credence to larger studies, while another gives more credence to 'quality' studies and yet another assigns a comparable weight to all studies. One reviewer may require a substantial body of evidence before concluding that a treatment is effective, while another uses a lower threshold. In fact, there are examples in the literature where two narrative reviews come to opposite conclusions, with one reporting that a treatment is effective while the other reports that it is not. As a rule, the narrative reviewer will not articulate (and may not even be fully aware of) the decision-making process used to synthesize the data and arrive at a conclusion.

A second limitation of narrative reviews is that they become *less useful as more information becomes available*. The thought process required for a synthesis requires the reviewer to capture the finding reported in each study, to assign an appropriate *weight* to that finding, and then to synthesize these findings across all studies in the synthesis. While a reviewer may be able to synthesize data from a few studies in their head, the process becomes difficult and eventually untenable as the number of studies increases. This is true even when the treatment effect (or effect size) is consistent from study to study. Often, however, the treatment effect will vary as a function of study-level covariates, such as the patient population, the dose of medication, the outcome variable, and other factors. In these cases, a proper synthesis requires that the researcher be able to understand how the treatment effect varies as a function of these variables, and the narrative review is poorly equipped to address these kinds of issues.

THE SYSTEMATIC REVIEW AND META-ANALYSIS

For these reasons, beginning in the mid 1980s and taking root in the 1990s, researchers in many fields have been moving away from the narrative review, and adopting systematic reviews and meta-analysis.

For systematic reviews, a clear set of rules is used to search for studies, and then to determine which studies will be included in or excluded from the analysis. Since there is an element of subjectivity in setting these criteria, as well as in the conclusions drawn from the meta-analysis, we cannot say that the systematic review is entirely objective. However, because all of the decisions are specified clearly, the mechanisms are transparent.

A key element in most systematic reviews is the statistical synthesis of the data, or the meta-analysis. Unlike the narrative review, where reviewers implicitly assign some level of importance to each study, in meta-analysis the weights assigned to each study are based on mathematical criteria that are specified in advance. While the reviewers and readers may still differ on the substantive meaning of the results (as they might for a primary study), the statistical analysis provides a transparent, objective, and replicable framework for this discussion.

The formulas used in meta-analysis are extensions of formulas used in primary studies, and are used to address similar kinds of questions to those addressed in primary studies. In primary studies we would typically report a mean and standard deviation for the subjects. If appropriate, we might also use analysis of variance or multiple regression to determine if (and how) subject scores were related to various factors. Similarly, in a meta-analysis, we might report a mean and standard deviation for the treatment effect. And, if appropriate, we would also use procedures analogous to analysis of variance or multiple regression to assess the relationship between the effect and study-level covariates.

Meta-analyses are conducted for a variety of reasons, not only to synthesize evidence on the effects of interventions or to support evidence-based policy or practice. The purpose of the meta-analysis, or more generally, the purpose of any research synthesis has implications for *when* it should be performed, what model should be used to analyze the data, what sensitivity analyses should be undertaken, and how the results should be interpreted. Losing sight of the fact that meta-analysis is a tool with multiple applications causes confusion and leads to pointless discussions about *what is the right way to perform a research synthesis*, when there is no single right way. It all depends on the purpose of the synthesis, and the data that are available. Much of this book will expand on this idea.

META-ANALYSIS IS USED IN MANY FIELDS OF RESEARCH

In medicine, systematic reviews and meta-analysis form the core of a movement to ensure that medical treatments are based on the best available empirical data. For example, The Cochrane Collaboration has published the results of over 3700 meta-analyses (as of January 2009) which synthesize data on treatments in all areas of

health care including headaches, cancer, allergies, cardiovascular disease, pain prevention, and depression. The reviews look at interventions relevant to neo-natal care, childbirth, infant and childhood diseases, as well as diseases common in adolescents, adults, and the elderly. The kinds of interventions assessed include surgery, drugs, acupuncture, and social interventions. BMJ publishes a series of journals on Evidence Based Medicine, built on the results from systematic reviews. Systematic reviews and meta-analyses are also used to examine the performance of diagnostic tests, and of epidemiological associations between exposure and disease prevalence, among other topics.

Pharmaceutical companies usually conduct a series of studies to assess the efficacy of a drug. They use meta-analysis to synthesize the data from these studies, yielding a more powerful test (and more precise estimate) of the drug's effect. Additionally, the meta-analysis provides a framework for evaluating the series of studies as a whole, rather than looking at each in isolation. These analyses play a role in internal research, in submissions to governmental agencies, and in marketing. Meta-analyses are also used to synthesize data on adverse events, since these events are typically rare and we need to accumulate information over a series of studies to properly assess the risk of these events.

In the field of education, meta-analysis has been applied to topics as diverse as the comparison of distance education with traditional classroom learning, assessment of the impact of schooling on developing economies, and the relationship between teacher credentials and student achievement. Results of these and similar meta-analyses have influenced practice and policy in various locations around the world.

In psychology, meta-analysis has been applied to basic science as well as in support of evidence-based practice. It has been used to assess personality change over the life span, to assess the influence of media violence on aggressive behavior, and to examine gender differences in mathematics ability, leadership, and nonverbal communication. Meta-analyses of psychological interventions have been use to compare and select treatments for psychological problems, including obsessive-compulsive disorder, impulsivity disorder, bulimia nervosa, depression, phobias, and panic disorder.

In the field of criminology, government agencies have funded meta-analyses to examine the relative effectiveness of various programs in reducing criminal behavior. These include initiatives to prevent delinquency, reduce recidivism, assess the effectiveness of different strategies for police patrols, and for the use of special courts to deal with drug-related crimes.

In business, meta-analyses of the predictive validity of tests that are used as part of the hiring process, have led to changes in the types of tests that are used to select employees in many organizations. Meta-analytic results have also been used to guide practices for the reduction of absenteeism, turnover, and counterproductive behavior, and to assess the effectiveness of programs used to train employees.

In the field of ecology, meta-analyses are being used to identify the environmental impact of wind farms, biotic resistance to exotic plant invasion, the effects of changes

in the marine food chain, plant reactions to global climate change, the effectiveness of conservation management interventions, and to guide conservation efforts.

META-ANALYSIS AS PART OF THE RESEARCH PROCESS

Systematic reviews and meta-analyses are used to synthesize the available evidence for a given question to inform policy, as in the examples cited above from medicine, social science, business, ecology, and other fields. While this is probably the most common use of the methodology, meta-analysis can also play an important role in other parts of the research process.

Systematic reviews and meta-analyses can play a role in designing new research. As a first step, they can help determine whether the planned study is necessary. It may be possible to find the required information by synthesizing data from prior studies, and in this case, the research should not be performed. Iain Chalmers (2007) made this point in an article entitled *The lethal consequences of failing to make use of all relevant evidence about the effects of medical treatments: the need for systematic reviews.*

In the event that the new study *is needed*, the meta-analysis may be useful in helping to design that study. For example, the meta-analysis may show that in the prior studies one outcome index had proven to be more sensitive than others, or that a specific mode of administration had proven to be more effective than others, and should be used in the planned study as well.

For these reasons, various government agencies, including institutes of health in various countries, have been encouraging (or requiring) researchers to conduct a meta-analysis of existing research prior to undertaking new funded studies.

The systematic review can also play a role in the publication of any new primary study. In the introductory section of the publication, a systematic review can help to place the new study in context by describing what we knew before, and what we hoped to learn from the new study. In the discussion section of the publication, a systematic review allows us to address not only the information provided by the new study, but the body of evidence as enhanced by the new study. Iain Chalmers and Michael Clarke (1998) see this approach as a way to avoid studies being reported without context, which they refer to as 'Islands in Search of Continents'. Systematic reviews would provide this context in a more rigorous and transparent manner than the narrative reviews that are typically used for this purpose.

THE INTENDED AUDIENCE FOR THIS BOOK

Since meta-analysis is a relatively new field, many people, including those who actually use meta-analysis in their work, have not had the opportunity to learn about it systematically. We hope that this volume will provide a framework that allows them to understand the logic of meta-analysis, as well as how to apply and interpret meta-analytic procedures properly.

This book is aimed at researchers, clinicians, and statisticians. Our approach is primarily conceptual. The reader will be able to skip the formulas and still understand, for example, the differences between fixed-effect and random-effects analysis, and the mechanisms used to assess the dispersion in effects from study to study. However, for those with a statistical orientation, we include all the relevant formulas, along with worked examples. Additionally, the spreadsheets and data files can be downloaded from the web at www.Meta-Analysis.com.

This book can be used as the basis for a course in meta-analysis. Supplementary materials and exercises are posted on the book's web site.

This volume is intended for readers from various substantive fields, including medicine, epidemiology, social science, business, ecology, and others. While we have included examples from many of these disciplines, the more important message is that meta-analytic methods that may have developed in any one of these fields have application to all of them.

Since our goal in using these examples is to explain the meta-analysis itself rather than to address the substantive issues, we provide only the information needed for this purpose. For example, we may present an analysis showing that a treatment reduces pain, while ignoring other analyses that show the same treatment increases the risk of adverse events. Therefore, any reader interested in the substantive issues addressed in an example should not rely on this book for that purpose.

AN OUTLINE OF THIS BOOK'S CONTENTS

Part 1 is an introduction to meta-analysis. We present a completed meta-analysis to serve as an example, and highlight the elements of this analysis – the effect size for each study, the summary effect, the dispersion of effects across studies, and so on. Our intent is to show where each element fits into the analysis, and thus provide the reader with a context as they move on to the subsequent parts of the book where each of the elements is explored in detail.

Part 2 introduces the effect sizes, such as the standardized mean difference or the risk ratio, that are computed for each study, and that serve as the unit of currency in the meta-analysis. We also discuss factors that determine the variance of an effect size and show how to compute the variance for each study, since this affects the weight assigned to that study in the meta-analysis.

Part 3 discusses the two computational models used in the vast majority of meta-analyses, the fixed-effect model and the random-effects model. We discuss the conceptual and practical differences between the two, and show how to compute a summary effect using either one.

Part 4 focuses on the issue of dispersion in effect sizes, the fact that the effect size varies from one study to the next. We discuss methods to quantify the heterogeneity, to test it, to incorporate it in the weighting scheme, and to understand it in a substantive as well as a statistical context. Then, we discuss methods to explain the heterogeneity. These include subgroup analyses to compare the effect in

different subgroups of studies (analogous to analysis of variance in primary studies), and meta-regression (analogous to multiple regression).

Part 5 shows how to work with complex data structures. These include studies that report an effect size for two or more independent subgroups, for two or more outcomes or time-points, and for two or more comparison groups (such as two treatments being compared with the same control).

Part 6 is used to address three separate issues. One chapter discusses the procedure called vote counting, common in narrative reviews, and explains the problems with this approach. One chapter discusses statistical power for a meta-analysis. We show how meta-analysis often (but not always) yields a more powerful test of the null than do any of the included studies. Another chapter addresses the question of publication bias. We explain what this is, and discuss methods that have been developed to assess its potential impact.

Part 7 focuses on the issue of why we work with effect sizes in a meta-analysis. In one chapter we explain why we work with effect sizes rather than *p*-values. In another we explain why we compute an effect size for each study, rather than summing data over all studies and then computing an effect size for the summed data. The final chapter in this part shows how the use of inverse-variance weights can be extended to other applications including Bayesian meta-analysis and analyses based on individual participant data.

Part 8 includes chapters on methods that are sometimes used in meta-analysis but that fall outside the central narrative of this volume. These include meta-analyses based on *p*-values, alternate approaches (such as the Mantel-Haenszel method) for assigning study weights, and options sometimes used in psychometric meta-analyses.

Part 9 is dedicated to a series of general issues related to meta-analysis. We address the question of when it makes sense to perform a meta-analysis. This Part is also the location for a series of chapters on separate issues such as reporting the results of a meta-analysis, and the proper use of cumulative meta-analysis. Finally, we discuss some of the criticisms of meta-analysis and try to put them in context.

Part 10 is a discussion of resources for meta-analysis and systematic reviews. This includes an overview of several computer programs for meta-analysis. It also includes a discussion of organizations that promote the use of systematic reviews and meta-analyses in specific fields, and a list of useful web sites.

WHAT THIS BOOK DOES NOT COVER

Other elements of a systematic review

This book deals only with meta-analysis, the statistical formulas and methods used to synthesize data from a set of studies. A meta-analysis can be applied to any data, but if the goal of the analysis is to provide a synthesis of a body of data from various sources, then it is usually imperative that the data be compiled as part of a systematic review.

A systematic review incorporates many components, such as specification of the question to be addressed, determination of methods to be used for searching the literature and for including or excluding studies, specification of mechanisms to appraise the validity of the included studies, specification of methods to be used for performing the statistical analysis, and a mechanism for disseminating the results.

If the entire review is performed properly, so that the search strategy matches the research question, and yields a reasonably complete and unbiased collection of the relevant studies, then (providing that the included studies are themselves valid) the meta-analysis will also be addressing the intended question. On the other hand, if the search strategy is flawed in concept or execution, or if the studies are providing biased results, then problems exist in the review that the meta-analysis cannot correct.

In Part 10 we include an annotated listing of suggested readings for the other components in the systematic review, but these components are not otherwise addressed in this volume.

Other meta-analytic methods

In this volume we focus primarily on meta-analyses of effect sizes. That is, analyses where each study yields an estimate of some statistic (a standardized mean difference, a risk ratio, a prevalence, and so on) and our goal is to assess the dispersion in these effects and (if appropriate) compute a summary effect. The vast majority of meta-analyses performed use this approach. We deal only briefly (see Part 8) with other approaches, such as meta-analyses that combine *p*-values rather than effect sizes. We do not address meta-analysis of diagnostic tests.

Further Reading

Chalmers, I. (2007). The lethal consequences of failing to make use of all relevant evidence about the effects of medical treatments: the need for systematic reviews. In P. Rothwell(ed.), *Treating Individuals*, ed. London: Lancet: 37–58.

Chalmers, I., Hedges, L.V. & Cooper, H. (2002). A brief history of research synthesis. *Evaluation in the Health Professions*. 25(1): 12–37.

Clarke, M, Hopewell, S. & Chalmers, I. (2007). Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: a status report. *Journal of the Royal Society of Medicine*. 100: 187–190.

Hunt, M. (1999). *How Science Takes Stock: The Story of Meta-analysis*. New York: Russell Sage Foundation.

Sutton, A.J. & Higgins, J.P.T. (2008). Recent developments in meta-analysis. *Statistics in Medicine* 27: 625–650.

PART 3

Fixed-Effect Versus Random-Effects Models

CHAPTER 10

Overview

Introduction
Nomenclature

INTRODUCTION

Most meta-analyses are based on one of two statistical models, the fixed-effect model or the random-effects model.

Under the fixed-effect model we assume that there is one *true effect size* (hence the term *fixed effect*) which underlies all the studies in the analysis, and that all differences in observed effects are due to sampling error. While we follow the practice of calling this a fixed-effect model, a more descriptive term would be a *common-effect* model. In either case, we use the singular (*effect*) since there is only one true effect.

By contrast, under the random-effects model we allow that the true effect could vary from study to study. For example, the effect size might be higher (or lower) in studies where the participants are older, or more educated, or healthier than in others, or when a more intensive variant of an intervention is used, and so on. Because studies will differ in the mixes of participants and in the implementations of interventions, among other reasons, there may be *different effect sizes* underlying different studies. If it were possible to perform an infinite number of studies (based on the inclusion criteria for our analysis), the true effect sizes for these studies would be distributed about some mean. The effect sizes in the studies that actually *were performed* are assumed to represent a random sample of these effect sizes (hence the term *random effects*). Here, we use the plural (*effects*) since there is an array of true effects.

In the chapters that follow we discuss the two models and show how to compute a summary effect using each one. Because the computations for a summary effect are not always intuitive, it helps to keep in mind that the summary effect is nothing more than the mean of the effect sizes, with more weight assigned to the more precise studies. We need to consider what we mean by the *more precise* studies and

Introduction to Meta-Analysis M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein

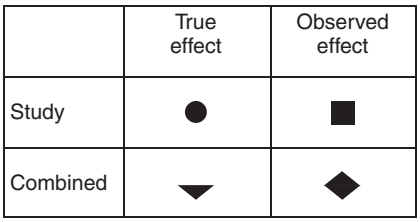

Figure 10.1 Symbols for true and observed effects.

how this translates into a study weight (this depends on the model), but not lose track of the fact that we are simply computing a weighted mean.

NOMENCLATURE

Throughout this Part we distinguish between a true effect size and an observed effect size. A study's *true effect size* is the effect size in the underlying population, and is the effect size that we would observe if the study had an infinitely large sample size (and therefore no sampling error). A study's *observed effect size* is the effect size that is actually observed.

In the schematics we use different symbols to distinguish between true effects and observed effects. For individual studies we use a circle for the former and a square for the latter (see Figure 10.1). For summary effects we use a triangle for the former and a diamond for the latter.

Worked examples

In meta-analysis the same formulas apply regardless of the effect size being used. To allow the reader to work with an effect size of their choosing, we have separated the formulas (which are presented in the following chapters) from the worked examples (which are presented in Chapter 14). There, we provide a worked example for the standardized mean difference, one for the odds ratio, and one for correlations.

The reader is encouraged to select one of the worked examples and follow the details of the computations while studying the formulas. The three datasets and all computations are available as Excel spreadsheets on the book's web site.

CHAPTER 11

Fixed-Effect Model

INTRODUCTION

In this chapter we introduce the fixed-effect model. We discuss the assumptions of this model, and show how these are reflected in the formulas used to compute a summary effect, and in the meaning of the summary effect.

THE TRUE EFFECT SIZE

Under the fixed-effect model we assume that all studies in the meta-analysis share a common (true) effect size. Put another way, all factors that could influence the effect size are the same in all the studies, and therefore the true effect size is the same (hence the label *fixed*) in all the studies. We denote the true (unknown) effect size by theta (θ)

In Figure 11.1 the true overall effect size is 0.60 and this effect (represented by a triangle) is shown at the bottom. The true effect for each study is represented by a circle. Under the definition of a fixed-effect model the true effect size for each study must also be 0.60, and so these circles are aligned directly above the triangle.

IMPACT OF SAMPLING ERROR

Since all studies share the same true effect, it follows that the observed effect size varies from one study to the next only because of the random error inherent in each study. If each study had an infinite sample size the sampling error would be zero and the observed effect for each study would be the same as the true effect. If we were to plot the observed effects rather than the true effects, the observed effects would exactly coincide with the true effects.

Introduction to Meta-Analysis M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein

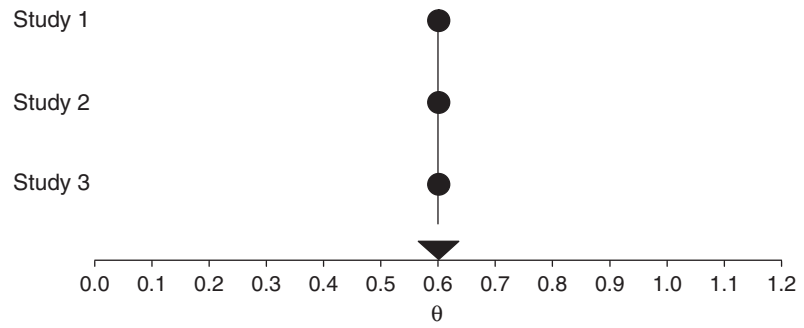

Figure 11.1 Fixed-effect model – true effects.

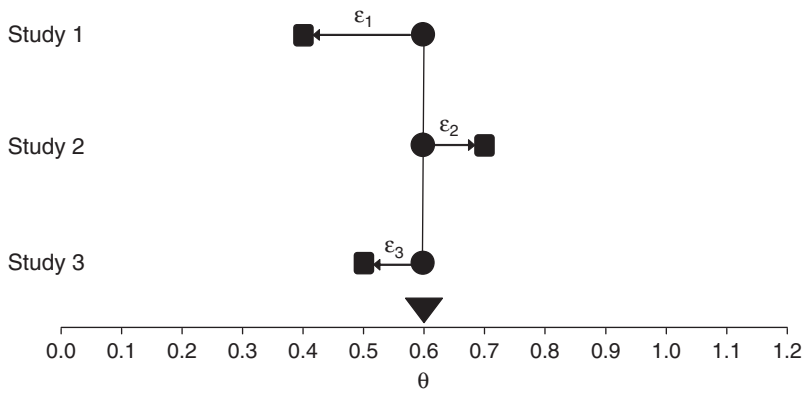

Figure 11.2 Fixed-effect model – true effects and sampling error.

In practice, of course, the sample size in each study is not infinite, and so there *is* sampling error and the effect observed in the study is not the same as the true effect. In Figure 11.2 the true effect for each study is still 0.60 (as depicted by the circles) but the observed effect (depicted by the squares) differs from one study to the next.

In Study 1 the sampling error (ε_1) is −0.20, which yields an observed effect (Y_1) of

$$Y_1 = 0.60 - 0.20 = 0.40.$$

In Study 2 the sampling error (ε_2) is 0.10, which yields an observed effect (Y_2) of

$$Y_2 = 0.60 + 0.10 = 0.70.$$

In Study 3 the sampling error (ε_3) is −0.10, which yields an observed effect (Y_3) of

$$Y_3 = 0.60 - 0.10 = 0.50.$$

More generally, the observed effect Y_i for any study is given by the population mean plus the sampling error in that study. That is,

$$Y_i = \theta + \varepsilon_i. \tag{11.1}$$

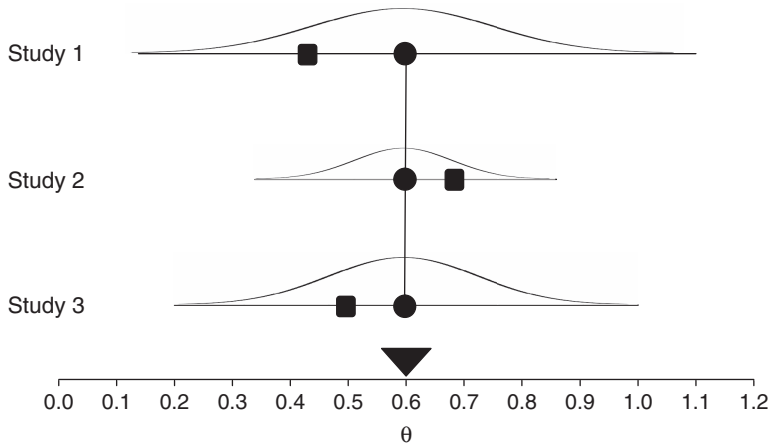

Figure 11.3 Fixed-effect model – distribution of sampling error.

While the error in any given study is random, we *can* estimate the sampling distribution of the errors. In Figure 11.3 we have placed a normal curve about the true effect size for each study, with the width of the curve being based on the variance in that study. In Study 1 the sample size was small, the variance large, and the observed effect is likely to fall anywhere in the relatively wide range of 0.20 to 1.00. By contrast, in Study 2 the sample size was relatively large, the variance is small, and the observed effect is likely to fall in the relatively narrow range of 0.40 to 0.80. (The width of the normal curve is based on the square root of the variance, or standard error).

PERFORMING A FIXED-EFFECT META-ANALYSIS

In an actual meta-analysis, of course, rather than starting with the population effect and making projections about the observed effects, we work backwards, starting with the observed effects and trying to estimate the population effect. In order to obtain the most precise estimate of the population effect (to minimize the variance) we compute a weighted mean, where the weight assigned to each study is the inverse of that study's variance. Concretely, the weight assigned to each study in a fixed-effect meta-analysis is

$$W_i = \frac{1}{V_{Y_i}}, \tag{11.2}$$

where V_{Y_i} is the within-study variance for study (i). The weighted mean (M) is then computed as

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}, \tag{11.3}$$

that is, the sum of the products $W_i Y_i$ (effect size multiplied by weight) divided by the sum of the weights.

The variance of the summary effect is estimated as the reciprocal of the sum of the weights, or

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i} \tag{11.4}$$

and the estimated standard error of the summary effect is then the square root of the variance,

$$SE_M = \sqrt{V_M}. \tag{11.5}$$

Then, 95% lower and upper limits for the summary effect are estimated as

$$LL_M = M - 1.96 \times SE_M \tag{11.6}$$

and

$$UL_M = M + 1.96 \times SE_M. \tag{11.7}$$

Finally, a Z-value to test the null hypothesis that the common true effect θ is zero can be computed using

$$Z = \frac{M}{SE_M}. \tag{11.8}$$

For a one-tailed test the p-value is given by

$$p = 1 - \Phi(\pm|Z|), \tag{11.9}$$

where we choose '+' if the difference is in the expected direction and '–' otherwise, and for a two-tailed test by

$$p = 2\Big[1 - \big(\Phi(|Z|)\big)\Big], \tag{11.10}$$

where $\Phi(Z)$ is the standard normal cumulative distribution. This function is tabled in many introductory statistics books, and is implemented in Excel as the function =NORMSDIST(Z).

Illustrative example

We suggest that you turn to a worked example for the fixed-effect model before proceeding to the random-effects model. A worked example for the standardized

mean difference (Hedges' *g*) is on page 87, a worked example for the odds ratio is on page 92, and a worked example for correlations is on page 97.

SUMMARY POINTS

- Under the fixed-effect model all studies in the analysis share a common true effect.
- The summary effect is our estimate of this common effect size, and the null hypothesis is that this common effect is zero (for a difference) or one (for a ratio).
- All observed dispersion reflects sampling error, and study weights are assigned with the goal of minimizing this within-study error.

CHAPTER 12

Random-Effects Model

INTRODUCTION

In this chapter we introduce the random-effects model. We discuss the assumptions of this model, and show how these are reflected in the formulas used to compute a summary effect, and in the meaning of the summary effect.

THE TRUE EFFECT SIZES

The fixed-effect model, discussed above, starts with the assumption that the true effect size is the same in all studies. However, in many systematic reviews this assumption is implausible. When we decide to incorporate a group of studies in a meta-analysis, we assume that the studies have enough in common that it makes sense to synthesize the information, but there is generally no reason to assume that they are *identical* in the sense that the true effect size is *exactly the same* in all the studies.

For example, suppose that we are working with studies that compare the proportion of patients developing a disease in two groups (vaccinated versus placebo). If the treatment works we would expect the effect size (say, the risk ratio) to be *similar but not identical* across studies. The effect size might be higher (or lower) when the participants are older, or more educated, or healthier than others, or when a more intensive variant of an intervention is used, and so on. Because studies will differ in the mixes of participants and in the implementations of interventions, among other reasons, there may be *different effect sizes* underlying different studies.

Introduction to Meta-Analysis M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein

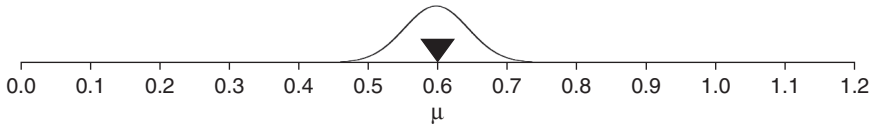

Figure 12.1 Random-effects model – distribution of true effects.

Or, suppose that we are working with studies that assess the impact of an educational intervention. The magnitude of the impact might vary depending on the other resources available to the children, the class size, the age, and other factors, which are likely to vary from study to study.

We might not have assessed these covariates in each study. Indeed, we might not even know what covariates actually are related to the size of the effect. Nevertheless, logic dictates that such factors do exist and will lead to variations in the magnitude of the effect.

One way to address this variation across studies is to perform a *random-effects* meta-analysis. In a random-effects meta-analysis we usually assume that the true effects are normally distributed. For example, in Figure 12.1 the mean of all true effect sizes is 0.60 but the individual effect sizes are distributed about this mean, as indicated by the normal curve. The width of the curve suggests that most of the true effects fall in the range of 0.50 to 0.70.

IMPACT OF SAMPLING ERROR

Suppose that our meta-analysis includes three studies drawn from the distribution of studies depicted by the normal curve, and that the true effects (denoted θ_1, θ_2, and θ_3) in these studies happen to be 0.50, 0.55 and 0.65 (see Figure 12.2).

If each study had an infinite sample size the sampling error would be zero and the observed effect for each study would be the same as the true effect for that study.

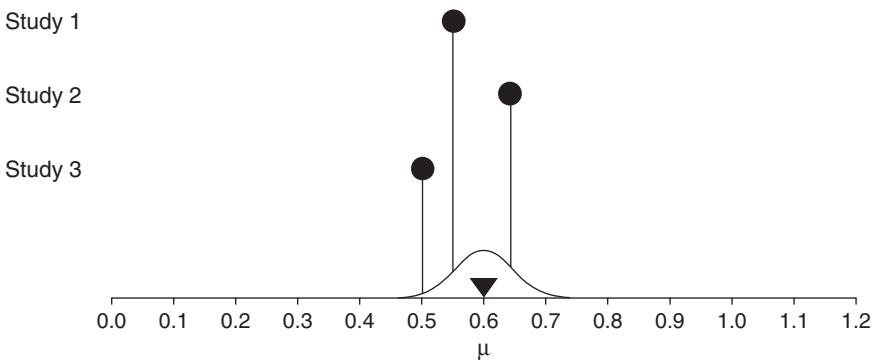

Figure 12.2 Random-effects model – true effects.

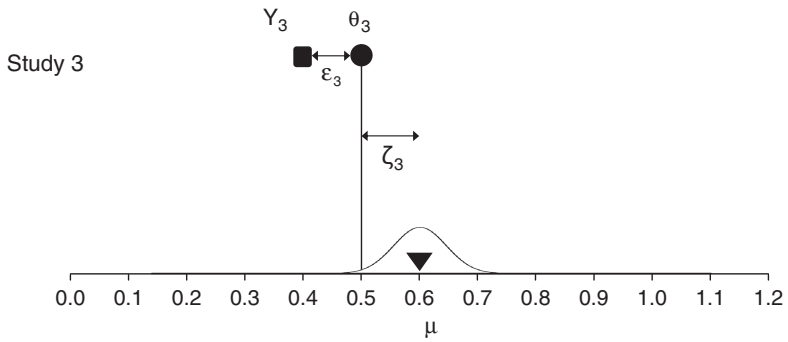

Figure 12.3 Random-effects model – true and observed effect in one study.

If we were to plot the observed effects rather than the true effects, the observed effects would exactly coincide with the true effects.

Of course, the sample size in any study is not infinite and therefore the sampling error is not zero. If the true effect size for a study is θ_i, then the observed effect for that study will be less than or greater than θ_i because of sampling error. For example, consider Study 3 in Figure 12.2. This study is the subject of Figure 12.3, where we consider the factors that control the observed effect. The true effect for Study 3 is 0.50 but the sampling error for this study is –0.10, and the observed effect for this study is 0.40.

This figure also highlights the fact that the distance between the overall mean and the observed effect in any given study consists of two distinct parts: true variation in effect sizes (ζ_i) and sampling error (ε_i). In Study 3 the total distance from μ to Y_3 is -0.20. The distance from μ to θ_3 (0.60 to 0.50) reflects the fact that the true effect size actually varies from one study to the next, while the distance from θ_3 to Y_3 (0.5 to 0.4) is sampling error.

More generally, the observed effect Y_i for any study is given by the grand mean, the deviation of the study's true effect from the grand mean, and the deviation of the study's observed effect from the study's true effect. That is,

$$Y_i = \mu + \zeta_i + \varepsilon_i. \tag{12.1}$$

Therefore, to predict how far the observed effect Y_i is likely to fall from μ in any given study we need to consider both the variance of ζ_i and the variance of ε_i.

The distance from μ (the triangle) to each θ_i (the circles) depends on the standard deviation of the distribution of the true effects across studies, called τ (tau) (or τ^2 for its variance). The same value of τ^2 applies to all studies in the meta-analysis, and in Figure 12.4 is represented by the normal curve at the bottom, which extends roughly from 0.50 to 0.70.

The distance from θ_i to Y_i depends on the sampling distribution of the sample effects about θ_i. This depends on the variance of the observed effect size from each study, V_{Y_i}, and so will vary from one study to the next. In Figure 12.4 the curve for Study 1 is relatively wide while the curve for Study 2 is relatively narrow.

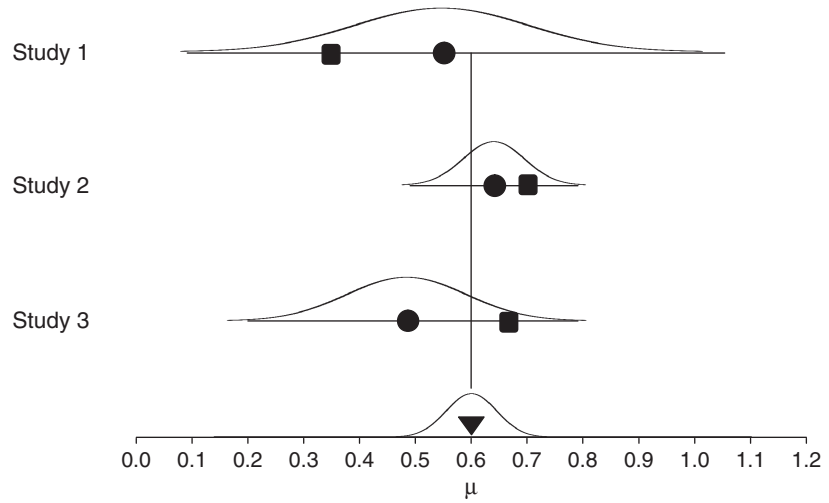

Figure 12.4 Random-effects model – between-study and within-study variance.

PERFORMING A RANDOM-EFFECTS META-ANALYSIS

In an actual meta-analysis, of course, rather than start with the population effect and make projections about the observed effects, we start with the observed effects and try to estimate the population effect. In other words our goal is to use the collection of Y_i to estimate the overall mean, μ. In order to obtain the most precise estimate of the overall mean (to minimize the variance) we compute a weighted mean, where the weight assigned to each study is the inverse of that study's variance.

To compute a study's variance under the random-effects model, we need to know both the within-study variance and τ^2, since the study's total variance is the sum of these two values. Formulas for computing the within-study variance were presented in Part 3. A method for estimating the between-studies variance is given here so that we can proceed with the worked example, but a full discussion of this method is deferred to Part 4, where we shall pursue the issue of heterogeneity in some detail.

Estimating tau-squared

The parameter τ^2 (tau-squared) is the between-studies variance (the variance of the effect size parameters across the population of studies). In other words, if we somehow knew the *true* effect size for each study, and computed the variance of these effects sizes (across an infinite number of studies), this variance would be τ^2. One method for estimating τ^2 is the method of moments (or the DerSimonian and Laird) method, as follows. We compute

$$T^2 = \frac{Q - df}{C} \quad , \tag{12.2}$$

where

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i}, \tag{12.3}$$

$$df = k - 1, \tag{12.4}$$

where k is the number of studies, and

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}. \tag{12.5}$$

Estimating the mean effect size

In the fixed-effect analysis each study was weighted by the inverse of its variance. In the random-effects analysis, too, each study will be weighted by the inverse of its variance. The difference is that the variance now includes the original (within-studies) variance plus the estimate of the between-studies variance, T^2. In keeping with the book's convention, we use τ^2 to refer to the parameter and T^2 to refer to the sample estimate of that parameter.

To highlight the parallel between the formulas here (random effects) and those in the previous chapter (fixed effect) we use the same notations but add an asterisk (*) to represent the random-effects version. Under the random-effects model the weight assigned to each study is

$$W_i^* = \frac{1}{V_{Y_i}^*} \tag{12.6}$$

where $V_{Y_i}^*$ is the within-study variance for study i plus the between-studies variance, T^2. That is,

$$V_{Y_i}^* = V_{Y_i} + T^2.$$

The weighted mean, M^*, is then computed as

$$M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*} \tag{12.7}$$

that is, the sum of the products (effect size multiplied by weight) divided by the sum of the weights.

The variance of the summary effect is estimated as the reciprocal of the sum of the weights, or

$$V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*} \tag{12.8}$$

and the estimated standard error of the summary effect is then the square root of the variance,

$$SE_{M^*} = \sqrt{V_{M^*}}. \tag{12.9}$$

The 95% lower and upper limits for the summary effect would be computed as

$$LL_{M^*} = M^* - 1.96 \times SE_{M^*}, \tag{12.10}$$

and

$$UL_{M^*} = M^* + 1.96 \times SE_{M^*}. \tag{12.11}$$

Finally, a Z-value to test the null hypothesis that the mean effect μ is zero could be computed using

$$Z^* = \frac{M^*}{SE_{M^*}}. \tag{12.12}$$

For a one-tailed test the p-value is given by

$$p^* = 1 - \Phi(\pm|Z^*|), \tag{12.13}$$

where we choose '+' if the difference is in the expected direction or '–' otherwise, and for a two-tailed test by

$$p^* = 2[1 - (\Phi(|Z^*|))], \tag{12.14}$$

where $\Phi(Z^*)$ is the standard normal cumulative distribution. This function is tabled in many introductory statistics books, and is implemented in Excel as the function =NORMSDIST(Z*).

Illustrative example

As before, we suggest that you turn to one of the worked examples in the next chapter before proceeding with this discussion.

SUMMARY POINTS

- Under the random-effects model, the true effects in the studies are assumed to have been sampled from a distribution of true effects.
- The summary effect is our estimate of the mean of all relevant true effects, and the null hypothesis is that the mean of these effects is 0.0 (equivalent to a ratio of 1.0 for ratio measures).

- Since our goal is to estimate the mean of the distribution, we need to take account of two sources of variance. First, there is within-study error in estimating the effect in each study. Second (even if we knew the true mean for each of our studies), there is variation in the true effects across studies. Study weights are assigned with the goal of minimizing both sources of variance.

CHAPTER 13

Fixed-Effect Versus Random-Effects Models

INTRODUCTION

In Chapter 11 and Chapter 12 we introduced the fixed-effect and random-effects models. Here, we highlight the conceptual and practical differences between them.

Consider the forest plots in Figures 13.1 and 13.2. They include the same six studies, but the first uses a fixed-effect analysis and the second a random-effects analysis. These plots provide a context for the discussion that follows.

DEFINITION OF A SUMMARY EFFECT

Both plots show a summary effect on the bottom line, but the meaning of this summary effect is different in the two models. In the fixed-effect analysis we assume that the true effect size is the same in all studies, and the summary effect is our estimate of this common effect size. In the random-effects analysis we assume that the true effect size varies from one study to the next, and that the studies in our analysis represent a random sample of effect sizes that could

Introduction to Meta-Analysis M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein

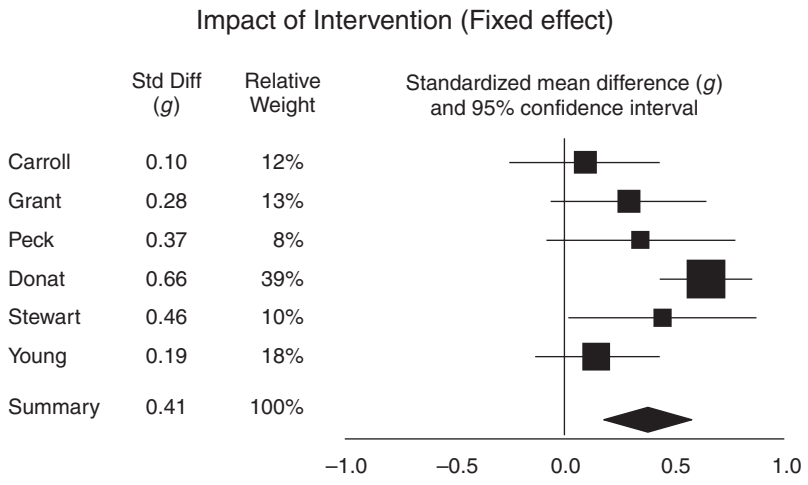

Figure 13.1 Fixed-effect model – forest plot showing relative weights.

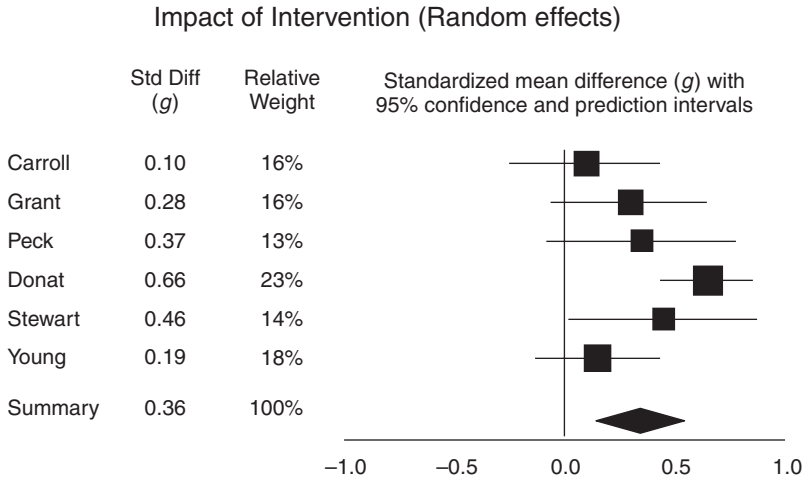

Figure 13.2 Random-effects model – forest plot showing relative weights.

have been observed. The summary effect is our estimate of the mean of these effects.

ESTIMATING THE SUMMARY EFFECT

Under the fixed-effect model we assume that the true effect size for all studies is identical, and the only reason the effect size varies between studies is sampling error (error in estimating the effect size). Therefore, when assigning

weights to the different studies we can largely ignore the information in the smaller studies since we have better information about the same effect size in the larger studies.

By contrast, under the random-effects model the goal is not to estimate one true effect, but to estimate the mean of a distribution of effects. Since each study provides information about a different effect size, we want to be sure that all these effect sizes are represented in the summary estimate. This means that we cannot discount a small study by giving it a very small weight (the way we would in a fixed-effect analysis). The estimate provided by that study may be imprecise, but it is information about an effect that no other study has estimated. By the same logic we cannot give too much weight to a very large study (the way we might in a fixed-effect analysis). Our goal is to estimate the mean effect in a range of studies, and we do not want that overall estimate to be overly influenced by any one of them.

In these graphs, the weight assigned to each study is reflected in the size of the box (specifically, the area) for that study. Under the fixed-effect model there is a wide range of weights (as reflected in the size of the boxes) whereas under the random-effects model the weights fall in a relatively narrow range. For example, compare the weight assigned to the largest study (Donat) with that assigned to the smallest study (Peck) under the two models. Under the fixed-effect model Donat is given about five times as much weight as Peck. Under the random-effects model Donat is given only 1.8 times as much weight as Peck.

EXTREME EFFECT SIZE IN A LARGE STUDY OR A SMALL STUDY

How will the selection of a model influence the overall effect size? In this example Donat is the largest study, and also happens to have the highest effect size. Under the fixed-effect model Donat was assigned a large share (39%) of the total weight and pulled the mean effect up to 0.41. By contrast, under the random-effects model Donat was assigned a relatively modest share of the weight (23%). It therefore had less pull on the mean, which was computed as 0.36.

Similarly, Carroll is one of the smaller studies and happens to have the smallest effect size. Under the fixed-effect model Carroll was assigned a relatively small proportion of the total weight (12%), and had little influence on the summary effect. By contrast, under the random-effects model Carroll carried a somewhat higher proportion of the total weight (16%) and was able to pull the weighted mean toward the left.

The operating premise, as illustrated in these examples, is that whenever τ^2 is nonzero, the relative weights assigned under random effects will be *more balanced* than those assigned under fixed effects. As we move from fixed effect to random effects, extreme studies will lose influence if they are large, and will gain influence if they are small.

CONFIDENCE INTERVAL

Under the fixed-effect model the only source of uncertainty is the within-study (sampling or estimation) error. Under the random-effects model there is this same source of uncertainty plus an additional source (between-studies variance). It follows that the variance, standard error, and confidence interval for the summary effect will always be larger (or wider) under the random-effects model than under the fixed-effect model (unless T^2 is zero, in which case the two models are the same). In this example, the standard error is 0.064 for the fixed-effect model, and 0.105 for the random-effects model.

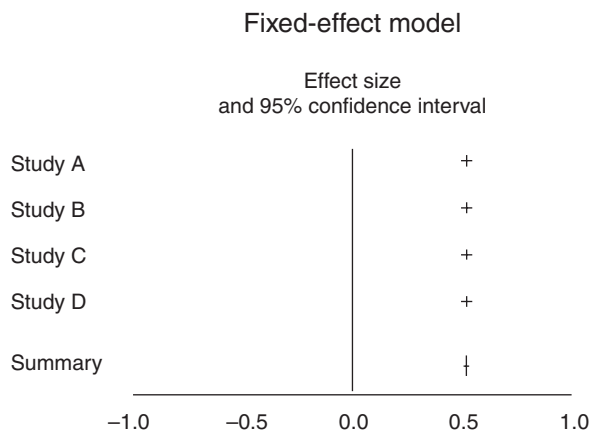

Figure 13.3 Very large studies under fixed-effect model.

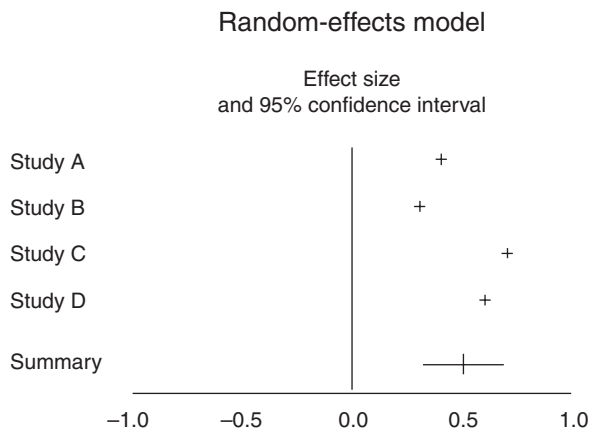

Figure 13.4 Very large studies under random-effects model.

Consider what would happen if we had five studies, and each study had an infinitely large sample size. Under either model the confidence interval for the effect size in each study would have a width approaching zero, since we know the effect size in that study with perfect precision. Under the fixed-effect model the summary effect would also have a confidence interval with a width of zero, since we know the common effect precisely (Figure 13.3). By contrast, under the random-effects model the width of the confidence interval would not approach zero (Figure 13.4). While we know the effect in each study precisely, these effects have been sampled from a universe of possible effect sizes, and provide only an estimate of the mean effect. Just as the error within a study will approach zero only as the sample size approaches infinity, so too the error of these studies as an estimate of the mean effect will approach zero only as the number of studies approaches infinity.

More generally, it is instructive to consider what factors influence the standard error of the summary effect under the two models. The following formulas are based on a meta-analysis of means from k one-group studies, but the conceptual argument applies to all meta-analyses. The within-study variance of each mean depends on the standard deviation (denoted σ) of participants' scores and the sample size of each study (n). For simplicity we assume that all of the studies have the same sample size and the same standard deviation (see Box 13.1 for details).

Under the fixed-effect model the standard error of the summary effect is given by

$$SE_M = \sqrt{\frac{\sigma^2}{k \times n}}. \tag{13.1}$$

It follows that with a large enough sample size the standard error will approach zero, and this is true whether the sample size is concentrated on one or two studies, or dispersed across any number of studies.

Under the random-effects model the standard error of the summary effect is given by

$$SE_M = \sqrt{\frac{\sigma^2}{k \times n} + \frac{\tau^2}{k}}. \tag{13.2}$$

The first term is identical to that for the fixed-effect model and, again, with a large enough sample size, this term will approach zero. By contrast, the second term (which reflects the between-studies variance) will only approach zero as the number of studies approaches infinity. These formulas do not apply exactly in practice, but the conceptual argument does. Namely, increasing the sample size within studies is not sufficient to reduce the standard error beyond a certain point (where that point is determined by τ^2 and k). If there is only a small number of studies, then the standard error could still be substantial even if the total n is in the tens of thousands or higher.

BOX 13.1 FACTORS THAT INFLUENCE THE STANDARD ERROR OF THE SUMMARY EFFECT.

To illustrate the concepts with some simple formulas, let us consider a meta-analysis of studies with the very simplest design, such that each study comprises a single sample of n observations with standard deviation σ. We combine estimates of the mean in a meta-analysis. The variance of each estimate is

$$V_{Y_i} = \frac{\sigma^2}{n}$$

so the (inverse-variance) weight in a fixed-effect meta-analysis is

$$W_i = \frac{1}{\sigma^2/n} = \frac{n}{\sigma^2}$$

and the variance of the summary effect under the fixed-effect model the standard error is given by

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i} = \frac{1}{k \times n/\sigma^2} = \frac{\sigma^2}{k \times n}.$$

Therefore under the fixed-effect model the (true) standard error of the summary mean is given by

$$SE_M = \sqrt{\frac{\sigma^2}{k \times n}}.$$

Under the random-effects model the weight awarded to each study is

$$W^*_i = \frac{1}{(\sigma^2/n) + \tau^2}$$

and the (true) standard error of the summary mean turns out to be

$$SE_{M^*} = \sqrt{\frac{\sigma^2}{k \times n} + \frac{\tau^2}{k}}.$$

THE NULL HYPOTHESIS

Often, after computing a summary effect, researchers perform a test of the null hypothesis. Under the fixed-effect model the null hypothesis being tested is that there is zero effect in *every study*. Under the random-effects model the null hypothesis being tested is that the *mean effect* is zero. Although some may treat these hypotheses as interchangeable, they are in fact different, and it is imperative to choose the test that is appropriate to the inference a researcher wishes to make.

WHICH MODEL SHOULD WE USE?

The selection of a computational model should be based on our expectation about whether or not the studies share a common effect size and on our goals in performing the analysis.

Fixed effect

It makes sense to use the fixed-effect model if two conditions are met. First, we believe that all the studies included in the analysis are functionally identical. Second, our goal is to compute the common effect size for the identified population, and not to generalize to other populations.

For example, suppose that a pharmaceutical company will use a thousand patients to compare a drug versus placebo. Because the staff can work with only 100 patients at a time, the company will run a series of ten trials with 100 patients in each. The studies are identical in the sense that any variables which can have an impact on the outcome are the same across the ten studies. Specifically, the studies draw patients from a common pool, using the same researchers, dose, measure, and so on (we assume that there is no concern about practice effects for the researchers, nor for the different starting times of the various cohorts). All the studies are expected to share a common effect and so the first condition is met. The goal of the analysis is to see if the drug works in the population from which the patients were drawn (and not to extrapolate to other populations), and so the second condition is met, as well.

In this example the fixed-effect model is a plausible fit for the data and meets the goal of the researchers. It should be clear, however, that this situation is relatively rare. The vast majority of cases will more closely resemble those discussed immediately below.

Random effects

By contrast, when the researcher is accumulating data from a series of studies that had been performed by researchers operating independently, it would be unlikely that all the studies were functionally equivalent. Typically, the subjects or interventions in these studies would have differed in ways that would have impacted on

the results, and therefore we should not assume a common effect size. Therefore, in these cases the random-effects model is more easily justified than the fixed-effect model.

Additionally, the goal of this analysis is usually to generalize to a range of scenarios. Therefore, if one did make the argument that all the studies used an identical, narrowly defined population, then it would not be possible to extrapolate from this population to others, and the utility of the analysis would be severely limited.

A caveat

There is one caveat to the above. If the number of studies is very small, then the estimate of the between-studies variance (τ^2) will have poor precision. While the random-effects model is still the appropriate model, we lack the information needed to apply it correctly. In this case the reviewer may choose among several options, each of them problematic.

One option is to report the separate effects and *not* report a summary effect. The hope is that the reader will understand that we cannot draw conclusions about the effect size and its confidence interval. The problem is that some readers will revert to vote counting (see Chapter 28) and possibly reach an erroneous conclusion.

Another option is to perform a fixed-effect analysis. This approach would yield a descriptive analysis of the included studies, but would not allow us to make inferences about a wider population. The problem with this approach is that (a) we do want to make inferences about a wider population and (b) readers will make these inferences even if they are not warranted.

A third option is to take a Bayesian approach, where the estimate of τ^2 is based on data from outside of the current set of studies. This is probably the best option, but the problem is that relatively few researchers have expertise in Bayesian meta-analysis. Additionally, some researchers have a philosophical objection to this approach.

For a more general discussion of this issue see *When does it make sense to perform a meta-analysis* in Chapter 40.

MODEL SHOULD NOT BE BASED ON THE TEST FOR HETEROGENEITY

In the next chapter we will introduce a test of the null hypothesis that the between-studies variance is zero. This test is based on the amount of between-studies variance observed, relative to the amount we would expect if the studies actually shared a common effect size.

Some have adopted the practice of starting with a fixed-effect model and then switching to a random-effects model if the test of homogeneity is statistically significant. This practice should be strongly discouraged because the decision to use the random-effects model should be based on our understanding of whether or not all studies share a common effect size, and not on the outcome of a statistical test (especially since the test for heterogeneity often suffers from low power).

If the study effect sizes are seen as having been sampled from a *distribution* of effect sizes, then the random-effects model, which reflects this idea, is the logical one to use. If the between-studies variance is substantial (and statistically significant) then the fixed-effect model is inappropriate. However, even if the between-studies variance does not meet the criterion for statistical significance (which may be due simply to low power) we should still take account of this variance when assigning weights. If T^2 turns out to be zero, then the random-effects analysis reduces to the fixed-effect analysis, and so there is no cost to using this model.

On the other hand, if one has elected to use the fixed-effect model *a priori* but the test of homogeneity is statistically significant, then it would be important to revisit the assumptions that led to the selection of a fixed-effect model.

CONCLUDING REMARKS

Our discussion of differences between the fixed-model and the random-effects model focused largely on the computation of a summary effect and the confidence intervals for the summary effect. We did not address the implications of the dispersion itself. Under the fixed-effect model we assume that all dispersion in observed effects is due to sampling error, but under the random-effects model we allow that some of that dispersion reflects real differences in effect size across studies. In the chapters that follow we discuss methods to quantify that dispersion and to consider its substantive implications.

Although throughout this book we define a fixed-effect meta-analysis as assuming that every study has a common true effect size, some have argued that the fixed-effect method is valid without making this assumption. The point estimate of the effect in a fixed-effect meta-analysis is simply a weighted average and does not strictly require the assumption that all studies estimate the same thing. For simplicity and clarity we adopt a definition of a fixed-effect meta-analysis that does assume homogeneity of effect.

SUMMARY POINTS

- A fixed-effect meta-analysis estimates a single effect that is assumed to be common to every study, while a random-effects meta-analysis estimates the mean of a distribution of effects.
- Study weights are more balanced under the random-effects model than under the fixed-effect model. Large studies are assigned less relative weight and small studies are assigned more relative weight as compared with the fixed-effect model.
- The standard error of the summary effect and (it follows) the confidence intervals for the summary effect are wider under the random-effects model than under the fixed-effect model.

- The selection of a model must be based solely on the question of which model fits the distribution of effect sizes, and takes account of the relevant source(s) of error. When studies are gathered from the published literature, the random-effects model is generally a more plausible match.
- The strategy of starting with a fixed-effect model and then moving to a random-effects model if the test for heterogeneity is significant is a mistake, and should be strongly discouraged.

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 M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein

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

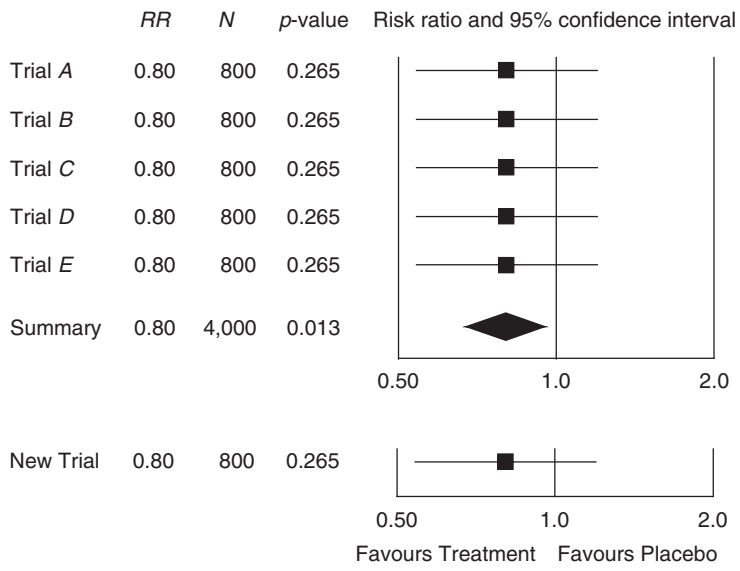

Figure 43.1 Forest plot of five fictional studies and a new trail (consistent effects).

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.

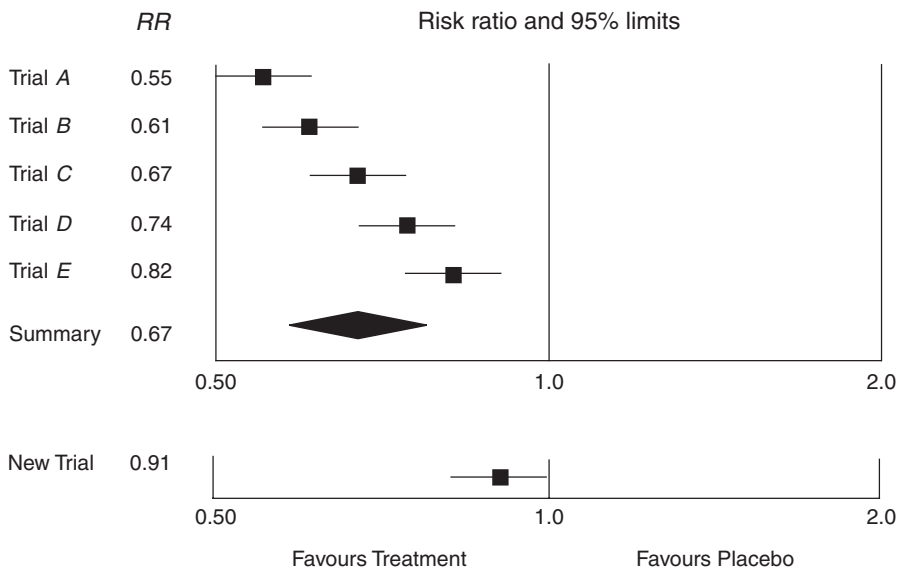

Figure 43.2 Forest plot of five fictional studies and a new trial (heterogeneous effects).

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.