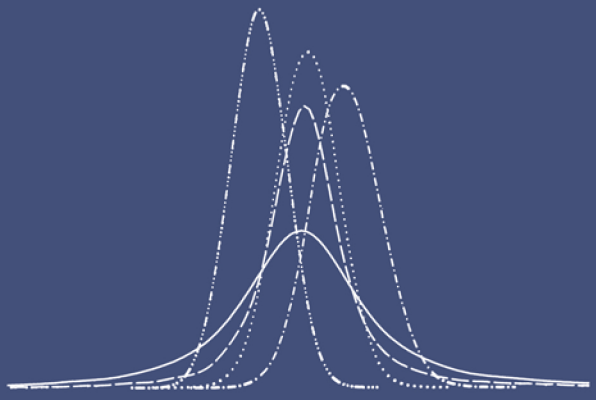
Meta-Analysis in Medicine and Health Policy



edited by
Dalene K. Stangl
Donald A. Berry

Meta-Analysis in Medicine and Health Policy

Biostatistics: A Series of References and Textbooks

Series Editor

Shein-Chung Chow

President, U.S. Operations
StatPlus, Inc.
Yardley, Pennsylvania

Adjunct Professor
Temple University
Philadelphia, Pennsylvania

1. *Design and Analysis of Animal Studies in Pharmaceutical Development*, edited by Shein-Chung Chow and Jen-pei Liu
2. *Basic Statistics and Pharmaceutical Statistical Applications*, James E. De Muth
3. *Design and Analysis of Bioavailability and Bioequivalence Studies, Second Edition, Revised and Expanded*, Shein-Chung Chow and Jen-pei Liu
4. *Meta-Analysis in Medicine and Health Policy*, edited by Dalene K. Stangl and Donald A. Berry
5. *Generalized Linear Models: A Bayesian Perspective*, edited by Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick

ADDITIONAL VOLUMES IN PREPARATION

Medical Biostatistics, Abhaya Indrayan and S.B. Sarmukaddam

Meta-Analysis in Medicine and Health Policy

edited by

Dalene K. Stangl
Duke University
Durham, North Carolina

Donald A. Berry
University of Texas M. D. Anderson Cancer Center
Houston, Texas

Marcel Dekker, Inc. New York • Basel

ISBN: 0-8247-9030-8

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fax: 41-61-261-8896

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit):
10 9 8 7 6 5 4

Series Introduction

The primary objectives of the Biostatistics series are to provide useful reference books for researchers and scientists in academia, industry, and government, and to offer textbooks for undergraduate and/or graduate courses in the area of biostatistics. This book series will provide comprehensive and unified presentations of statistical designs and analyses of important applications in biostatistics, such as those in biopharmaceuticals. A well-balanced summary will be given of current and recently developed statistical methods and interpretations for both statisticians and researchers/scientists with minimal statistical knowledge who are engaged in applied biostatistics. The series is committed to providing easy-to-understand state-of-the-art references and textbooks. In each volume, statistical concepts and methodologies will be illustrated through real examples.

Meta-analysis is a commonly employed systematic reviewing strategy for addressing research or scientific questions in health-related research. It is especially useful when results from several studies disagree with regard to direction of effect, or sample sizes are individually too small to detect an effect, or a large trial is too costly and time-consuming to perform. It has been a concern whether the results from a meta-analysis covering a number of studies, which may or may not be conducted under the same study protocol, are statistically valid. The validity of a meta-analysis depends on the selection of studies and the heterogeneity among studies. As a result, the United States Food and Drug Administration (FDA) suggests that the issues of publication bias (or selection bias) and study-by-treatment interaction should be carefully evaluated before a meta-analysis is conducted for evaluation of safety and efficacy of a pharmaceutical compound in clinical research and development. Meta-

analyses for uncombinable studies should be avoided for good clinical practice. This volume not only introduces important statistical concepts, designs, and methodologies of meta-analysis but also provides applications in clinical research through practical examples. The book will serve as a bridge among biostatisticians, health-related researchers/scientists, and regulatory agents, by providing a good understanding of key statistical concepts regarding design, analysis, and interpretation in health-related research, especially in medicine and health policy.

Shein-Chung Chow

Preface

Enhancements in research methodology and statistical computing, along with the demand for more accountable decision-making, have made the collection and analysis of data an integral component of every aspect of health (i.e., prevention, diagnosis, treatment, and policy). In the United States alone, annual spending on quantitative health-related research is in the billions of dollars. Replication of studies is often mandated and necessary to ensure validity and generality of results. However, there is also a great need for improved methods of meta-analysis to integrate research findings. This book reviews current methods of meta-analysis and introduces cutting-edge methods that improve quantitative meta-analysis and enable better decision-making.

This book is written for applied statisticians, students of statistics and biostatistics, and others who use statistical methods in their professional life. Our objective is to teach, so we rely heavily on examples. The authors present a common problem, develop a methodology to address the problem, and follow up with one or more examples. The level of chapters ranges from elementary to advanced; however, each chapter starts from first principles and proceeds to state-of-the-art techniques about which there are many open research questions suitable for graduate projects and dissertations.

Several chapters address controversies and make appropriate and practical recommendations. These recommendations are derived with the notion that statisticians must be able to persuade nonstatisticians as to the appropriateness of their conclusions. To this end, we accent pictorial presentations that are backed up by mathematical analyses.

This book is primarily a reference. However, it is ideal as a supplemental text for master's- and Ph.D.-level statistics and biostatistics

courses. An undergraduate course in statistical theory and methods will provide the necessary background for most of the chapters. Each chapter could serve as a basis for a student project in which the student can present the analysis, think through the pros and cons of the methods, investigate other evidence that bears on the medical or policy question, and suggest improvements or further steps that could be carried out.

Dalene K. Stangl
Donald A. Berry

Contents

Contributors

Keith R. Abrams, Ph.D. Department of Epidemiology and Public Health, University of Leicester, Leicester, England

Donald A. Berry, Ph.D. Department of Biostatistics, University of Texas M.D. Anderson Cancer Center, Houston, Texas

Scott M. Berry, Ph.D. Department of Statistics, Texas A & M University, College Station, Texas

James Brophy, M.D., F.R.C.P.(C), Ph.D. Department of Medicine, University of Montreal, Montreal, Quebec, Canada

Francesca Dominici, Ph.D. Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland

William DuMouchel, Ph.D. Statistics Research, AT&T Shannon Labs, Florham Park, New Jersey

Guido Francini, M.D. Division of Medical Oncology, Siena University, Siena, Italy

Geof H. Givens, Ph.D. Department of Statistics, Colorado State University, Fort Collins, Colorado

Leon J. Gleser, Ph.D. Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania

Martin Hellmich, Dr.rer.medic. Department of Medical Statistics, Informatics, and Epidemiology, University of Cologne, Cologne, Germany

David R. Jones, Ph.D. Department of Epidemiology and Public Health, University of Leicester, Leicester, England

Lawrence Joseph, Ph.D. Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada

Paul C. Lambert, M.Sc. Department of Epidemiology and Public Health, University of Leicester, Leicester, England

Daniel T. Larose, Ph.D. Department of Mathematical Sciences, Central Connecticut State University, New Britain, Connecticut

Chantal Milan, Ph.D. Fondation Française de Cancérologie Digestive (F.F.C.D.), Centre d'Epidémiologie et de Population de Bourgogne, Dijon, France

Sharon-Lise Normand, Ph.D. Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts

Ingram Olkin, Ph.D. Department of Statistics, Stanford University, Stanford, California

Giovanni Parmigiani, Ph.D. Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina

Donna K. Pauler, Ph.D. Biostatistics Center, Massachusetts General Hospital, Boston, Massachusetts

Nargis Rahman, Ph.D. Department of Mathematics, Imperial College of Science, Technology and Medicine, London, England

Bruno Sansó, Ph.D. Centro de Estadísticas y Software Matemático, Universidad Símon Bolívar, Caracas, Venezuela

Daniel J. Sargent, Ph.D. Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota

Chris Shaw, Ph.D. Department of Epidemiology and Public Health, University of Leicester, Leicester, England

Richard Simon, D.Sc. Biometric Research Branch, National Cancer Institute, Bethesda, Maryland

David D. Smith, Ph.D. Center for Drug Evaluation and Research, Office of Biostatistics, U.S. Food and Drug Administration, Rockville, Maryland

Dalene K. Stangl, Ph.D. Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina

Michael A. Stoto, Ph.D. Department of Epidemiology and Biostatistics, George Washington University School of Public Health and Health Services, Washington, D.C.

Alexander J. Sutton, M.Sc. Department of Epidemiology and Public Health, University of Leicester, Leicester, England

Valter Torri, M.D. Department of Oncology, Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy

Richard L. Tweedie, Ph.D., D.Sc. Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota

Jon Wakefield, Ph.D. Department of Statistics and Biostatistics, University of Washington, Seattle, Washington

Benny C. Zee, Ph.D. Department of Community Health and Epidemiology, Queen's University, Kingston, Ontario, Canada

1

Meta-analysis: Past and Present Challenges*

Dalene K. Stangl
Duke University, Durham, North Carolina

Donald A. Berry
University of Texas M.D. Anderson Cancer Center, Houston, Texas

I. META-ANALYSIS: A PARADIGM SHIFT

Most broadly defined, meta-analysis includes any methodology for combining information across sources. Of most interest to statisticians are quantitative approaches to summarize all relevant information pertaining to a research question. Introductory texts on the subject include Refs 1–8. Each of these books focuses on methods for deriving a common estimate of effect. In contrast, Ref. 9 emphasizes the need to move this focus to one of quantifying and reporting the heterogeneity between studies. This message reflects the methodological development seen in the statistics field during the last decade. Statistical methodology for meta-analysis has been moving away from approaches focused only on fitting a common estimate toward approaches that include estimating the extent and sources of heterogeneity among studies. This chapter will review this progression, highlight complexities encountered, summarize how the chapters in the present volume address heterogeneity between studies, and suggest future directions for methodological development.

*This work was partially supported by grant SBR-9809267 from the National Science Foundation.

A. Progression from Estimation of Effect to Estimation of Heterogeneity

The earliest standardized methodologies for meta-analysis include combining p-values and combining effect-size estimates. Examples of the former methodology are summarized in Refs 4 and 10. They include Tippett's minimum-p method (11) and Fisher's product of p-values (12). Tippett's minimum-p test rejects the null hypothesis, that each of the k individual-study effects is equal to zero, if any of the p-values is less than α, where $\alpha = 1 - (1 - \alpha^*)^{1/k}$, and α^* is the apriori determined significance level for the combined significance test. Fisher's product of p-values compares

$$-2\sum_{i=1}^{k} \log(p_i)$$

with the $100(1 - \alpha^*)\%$ critical value of the chi-squared distribution with $2k$ degrees of freedom. These methods require little input, but cannot give estimates of effect sizes or correct the publication biases toward statistically significant effects, and suffer the problems inherent in multiple testing.

While requiring more thorough journal reporting, combining estimates of effect size across studies overcomes some of the problems encountered in combining p-values. Examples of combined effect size (13) include combined standardized mean differences, correlations, differences between proportions, and odds ratios. These procedures are restricted in that they are based on large-sample theory, and while standard errors and tests of homogeneity are possible, these tests have low power. Hence, in practice, the focus remains on the pooled estimates of effect rather than giving due emphasis to the degree of heterogeneity between studies. Reference 14 suggests that: ". . . preparing and presenting a single estimate as the distillation of all that is known has drawn the most criticism." This distillation is responsible for much of the misunderstanding and controversy in meta-analysis today. Some of these misunderstandings are summarized in this volume by Simon (Chap. 12), and one particular controversy is addressed by Berry (Chap. 3). Berry describes how the single-estimate distillation has inappropriately resulted in discrediting of meta-analysis because of perceived disagreements between meta-analyses and the results of large clinical trials. He shows

how these perceived disagreements are errors in reasoning that are caused by failure to explicitly account for study heterogeneity.

In both the frequentist and Bayesian paradigms, the recent shift from presenting single-estimate summaries parallels the transition in popularity from fixed-effects models to random-effects models. While a brief review and comparison of these models will be presented here, nearly every chapter in this volume recognizes and promotes this shift. Further explanation and applications are given by Abrams et al. (Chap. 2), Berry (Chap. 3), Brophy and Joseph (Chap. 4), Dominici and Parmigiani (Chap. 5), DuMouchel and Normand (Chap. 6), Larose (Chap. 8), Pauler and Wakefield (Chap. 9), Rahman and Wakefield (Chap. 10), Sargent et al. (Chap. 11), and Smith et al. (Chap. 13).

1. *Fixed-effects Models*

Fixed-effects models for meta-analysis assume that the studies being modeled are homogeneous. There are no differences in underlying study populations, no differences in patient-selection criteria that might affect response to therapy, and the therapies are applied in the same way. Technically, patients who are assigned the same treatment but in different studies are taken to be exchangeable (or partially exchangeable in the case of covariates).

Let Y_i be a sufficient statistic for the effect of interest. For large individual-study sample sizes, the response, the individual-study effect, is approximately normal:

$$Y_i \sim \mathrm{N}(\mu, \sigma_i^2),$$

where σ_i is the standard deviation of the response. If we assume that the σ_i^2 are known, the minimum-square-error estimate of all linear estimators for μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{I} \frac{1}{\sigma_i^2} Y_i}{\sum_{i=1}^{I} \frac{1}{\sigma_i^2}},$$

which has distribution

$$\hat{\mu} \sim N\left(\mu, \left(\sum_{i=1}^{I} \frac{1}{\sigma_i^2}\right)^{-1}\right).$$

The sole source of variability is assumed to be the within-study variation. A test for homogeneity of study effects compares

$$Q = \sum_{i=1}^{I} \frac{1}{\sigma_i^2}(Y_i - \hat{\mu})^2$$

to the chi-squared distribution with $I - 1$ degrees of freedom.

Reference 4 presents a thorough overview of fixed-effects models. References 9 and 15 provide shorter, but more recent overviews. There are many articles in the literature that compare and contrast the main methods for fixed-effects models. Reference 16 summarizes and compares the main methods for binary-response fixed-effects models, including those of Mantel and Haenszel (17), Woolf (18), Mantel, Haenszel, and Peto (19), and logistic regression. The Mantel–Haenszel and Woolf methods use weighted averages of the maximum-likelihood estimates of the log-odds ratios and odds ratios in each study, respectively. The Mantel–Haenszel–Peto method uses a score and Fisher information statistics from the conditional likelihoods for study-specific effects to estimate pooled effects, and logistic regression derives maximum-likelihood estimates from a full binomial likelihood. Analogous methods are available for continuous outcomes.

Fixed-effects models continue to be the most common method of meta-analysis. However, the assumption of homogeneity is usually unrealistic, given variability among studies and/or research and evaluation protocols. Indeed it is this assumption that underlies the controversy between the relative value of large clinical trials versus meta-analysis, which is addressed by Berry (Chap. 3). Equally important, the fixed-effects model underestimates variability and hence may lead to erroneous conclusions of statistical significance.

2. *Random-effects Models*

The random-effects formulation avoids the homogeneity assumption by modeling a random effect, θ_i for study i. Each θ_i is assumed to be selected from a distribution of study effects. The response at study i is

$$Y_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$$

and the individual-study effects are exchangeable with a normal distribution:

$$\theta_i \sim \mathrm{N}(\mu_\theta, \tau^2).$$

Here μ_θ represents the mean of the study effects, and τ^2 represents the between-study variability. If τ^2 is known, then μ_θ is estimated by

$$\hat{\mu}_\theta = \frac{\sum_{i=1}^{I} (\sigma_i^2 + \tau^2)^{-1} y_i}{\sum_{i=1}^{I} (\sigma_i^2 + \tau^2)^{-1}}.$$

Reference 20 provides methods of estimation if τ^2 is unknown. If there are study covariates, then one can model the study-specific effects as

$$\theta_i \sim \mathrm{N}(x_i'\beta, \tau^2),$$

where study-specific effects are assumed to be exchangeable for the partition defined by x_i.

In the random-effects model, the study effects represent samples from a population. These models "borrow strength" across studies in making estimates of both study-specific effects, θ_i, as well as an estimate of the population effect, μ. The estimate for θ_i is a weighted average of the study-specific effect estimate and the estimate of the population average. DuMouchel and Normand (Chap. 6) provide formulas for these estimates in their appendix. References 16 and 21–24 compare fixed-effects, random-effects, empirical-Bayes, and fully Bayesian models.

The primary difference in how we will define empirical-Bayes models and fully Bayesian models lies in estimating parameters of the distribution of study effects. In an empirical-Bayes model, these parameters are estimated from the "current" data, while in a fully Bayesian model another level is added, and the parameters of the distribution of study effects are given a prior distribution. In the empirical-Bayes model, the only prior information incorporated is choice of a parametric family for the distribution of study effects. The relevant experimental unit is the study; therefore, when the number of studies is small, this estimation is imprecise. Also, the uncertainty embedded in estimation of these parameters is not included in the uncertainty estimates for study effects (25).

In the fully Bayesian model, prior distributions are placed on the parameters of the distribution of the study effects, rather than estimating

these parameters from the "current" data. While experts and previous studies may provide information about these hyper-parameters, typically as we move upward in the hierarchy, less and less information is known about parameters, so priors become more diffuse. In fully Bayesian models, this uncertainty is incorporated into estimates of study effects.

Authors tend to vary in their use of the label "random-effects model." For example, DuMouchel and Normand (Chap. 6) prefer to reserve the label for empirical-Bayes models while other authors use the label to refer to both empirical-Bayes and fully Bayesian models.

a. A Frequentist Perspective

Interpreting a random-effects model from a frequentist perspective is problematic unless there are many studies. First, relying on maximum-likelihood analysis and asymptotic theory requires large samples. Second, the classical test for homogeneity between studies has low statistical power (26). Third, negative estimates of variability are possible; see, e.g., Ref. 20. And fourth, a single estimate for the variability between studies may be unsatisfactory.

Some researchers object to random-effects models for philosophical reasons. For example, in a comprehensive study of the treatment of early breast cancer (27), random-effects models were not used, because:

> The statistical assumptions needed for such statistical methods to be of direct medical evidence are unlikely to be met. In particular, the different trial designs that were adopted would have to have been randomly selected from some underlying set of possibilities that includes the populations about which predictions are to be made. This is unlikely to be the case, since trial designs are adopted for a variety of reasons, many of which depend in a complex way on the apparent results of earlier trials. Moreover, selective factors that are difficult to define may affect the types of patients in trials, and therapeutic factors that are also difficult to define may differ between trials, or between past trials and future medical practice.

These authors failed to understand that using a fixed-effects model makes a more rigid key assumption that is even more unlikely to be met. Viewing the random-effects formulation from a Bayesian perspective avoids many of these problems.

b. A Bayesian Perspective

From a Bayesian perspective, all parameters are random in that they have probability distributions. These distributions depend on all available

information and not just on the data at hand. The Bayesian paradigm is synonymous with meta-analysis. For both, the goal is to incorporate all information to predict with as much accuracy as possible some future event, and the uncertainty associated with it, and to present this prediction in a manner that leads to coherent decisions.

According to DV Lindley, "Meta-analysis is a natural for the Bayesian . . . " (28). Suppose that the meta-analyst is interested in some effect θ, common to all studies. After updating with observations from the first study, x_1, the posterior distribution is

$$f(\theta|x_1) \propto f(x_1|\theta)f(\theta),$$

where $f(\theta)$ is the prior distribution of the effect θ, and $f(x_1|\theta)$ is the likelihood of the data given the effect. Now the meta-analyst can use $f(\theta|x_1)$ as a prior for the next analysis, producing

$$f(\theta|x_1, x_2) \propto f(x_2|\theta)f(\theta|x_1)$$

and so on. This updating scheme is common in all applications of Bayesian methods and demonstrates how all Bayesian analysis can be seen as meta-analysis. A particularly clear example of such updating is presented by Brophy and Joseph (Chap. 4).

Testing for heterogeneity of effects between studies requires estimates of between-study variability. So, the greater the number of studies included in the meta-analysis the better, as long as the quality of included studies is high. Many meta-analyses include less than a dozen studies, which means that the estimates are not very precise. In fact, many authors in this volume are concerned about precisely this issue. While some Bayesians advocate the use of Bayes factors to test for heterogeneity, if the number of studies is small, some of the same problems arise as with frequentist tests of significance. Reference 29 uses Bayes factors to test for two types of heterogeneity, additive and interactive, in meta-analysis of 2×2 contingency tables. Chapters by Pauler and Wakefield (Chap. 9) and Abrams et al. (Chap. 2) both demonstrate the use of Bayes factors to test for heterogeneity. Some limitations of Bayes factors are discussed in Refs 30–32.

Using a fully Bayesian perspective for meta-analysis has several advantages. First, the Bayesian paradigm provides a method for synthesizing all available information in a formal, consistent, and coherent manner. Second, it explicitly incorporates model and parameter uncertainty. A third and very important aspect of the Bayesian approach is that one can average over the current distribution of unknown para-

meters to find a predictive distribution for future observations. For a future study outcome, x^*, the predictive distribution is

$$f(x^*|x_1, x_2) = \int f(x^*|\theta) f(\theta|x_1, x_2) d\theta.$$

This predictive distribution is a quantity of central interest to the decision maker.

References that cover the topic of decision making from an applied perspective include Refs 33–36, while Refs 37 and 38 take more theoretical perspectives. Applications using a decision-theoretic perspective which could be easily extended to the meta-analysis context may be found in Refs 39–44. These applications cover the development of clinical recommendations, determining the effectiveness of vaccines, analyzing multi-center clinical trial data, and determining who should remediate contaminated geographic areas. Rahman & Wakefield (Chap. 10) demonstrate the value of predictive distributions for decision making in a meta-analysis of pharmacokinetic data. Data from four phase I studies are combined via a hierarchical model that is nonlinear at the first stage. The authors derive overall and study-specific predictive distributions for drug clearance, half life, and volume parameters as well as for drug concentration as a function of time. These distributions are useful for decisions such as determining optimal dosages and designing future studies. Simon (Chap. 12) and Stoto (Chap. 14) review the role of meta-analysis in medical and health-policy decision making respectively, and both authors discuss controversies that ensue.

Bayesian methods offer flexible modeling schemes. One incorporates information contained in the data at hand with available prior information. The posterior distribution of the variance components as well as posterior and predictive distributions for study-specific effects are used to investigate heterogeneity in effects across studies. A typical Bayesian hierarchical model for meta-analysis is (45, 21, 16):

Level I: $Y_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$

Level II: $\theta_i \sim \mathrm{N}(\mu, \tau^2)$

Level III: $\mu \sim \pi(\mu), \tau^2 \sim \pi(\tau^2)$.

Here Y_i is the effect from the ith study, and σ_i represents the standard deviation for the effect. Although σ_i is often assumed known, a prior for the σ_i may be incorporated as well. As with the empirical-Bayes models these models "borrow strength" across studies in estimating both study-

specific effects, θ_i, and the population effect, μ. The posterior mean for θ_i is a weighted average of the study-specific effect estimate and the posterior mean of the population effect. Study-specific effects are shrunk toward the overall population mean, μ, while more accurate estimates of uncertainty are derived for the study-specific effects and the population effect. Formulas for posterior means and variances are provided in the appendix of DuMouchel and Normand (Chap. 6).

c. *A Brief Literature Review of Empirical-Bayesian and Fully Bayesian Approaches*

Early development and application of random-effect (empirical-Bayes and fully Bayesian) meta-analyses are included in the following publications. Bayesian hierarchical models for meta-analysis were introduced in Ref. 46. Empirical-Bayesian approaches are used in Ref. 47 (and elsewhere). Reference 49 considers general parametric approaches for meta-analysis of clinical trials. Sampling-based methods to hierarchical-Bayesian models for normally distributed data are applied in Ref 21. Reference 49 compares Bayesian and empirical-Bayes methods for 2×2 tables and demonstrates that empirical-Bayes methods underestimate the variance of the pooled estimate. References 3 and 50–52 introduced the Confidence Profile Method, a software package for carrying out Bayesian meta-analyses evaluating health-care interventions. Reference 53 developed a random-effects dose–response meta-analysis model incorporating correlation between observations within studies and including study-level covariates. Results from controlled and uncontrolled studies using random-effects models for meta-analysis are combined in Ref. 54. Studies involving variation in response classification are examined in Ref. 55. The use of meta-analysis in pharmaceutical studies is investigated in Ref. 56; more specifically Ref. 57 investigates meta-analysis for dose–response models. References 22 and 24 apply Bayesian meta-analytic models to lung-cancer studies.

Random-effects, asymptotic-Bayesian and exact-Bayesian methods are compared in Ref. 58. A meta-analysis for 2×2 tables for vaccine efficacy using empirical-Bayes methods is presented in Ref. 59. Reference 60 estimates and adjusts for selection bias in Bayesian meta-analysis. Bayesian approaches to model discrimination in meta-analysis are investigated in Ref. 61. Variability in the underlying population risk across studies in meta-analysis of clinical trials is also investigated.

Empirical-Bayesian and fully-Bayesian approaches applicable to meta-analysis of time-to-event data are presented in Refs 43 and 62–69.

In this volume, Sargent et al. (Chap. 11) present a meta-analysis of randomized trails of chemotherapy for colon cancer. Using individual patient survival data in a random-effects proportional-hazards model, Sargent et al. demonstrate the importance of carefully choosing a parametrization. They show how subtle model changes can change results. An example is whether the model is parametrized so that shrinkage occurs within treatment groups, or in the relative treatment effect, or in both.

II. COMPLEXITIES ADDRESSED IN THIS VOLUME

Difficulties in implementing either the Bayesian or classical paradigm include defining a method for choosing studies, incorporating inconsistency between study designs and outcomes, assessing and including measures of study quality, matching outcome scales of measurement, adjusting for publication bias, accommodating missing data, checking for model fit, incorporating study-level covariates, and finding or developing appropriate software. A thorough review of recent research on these complexities is available in Ref. 70. Solutions or partial solutions to many of these complexities are presented in this volume.

A. Inconsistent Outcomes and Result Reporting

Outcomes are rarely identical across studies. While endpoints may be named the same, different measures, different implementations, and different populations may prevent direct comparison and pooling of data. Even when measures, implementation, and populations are very close, researchers may have chosen to present results in different ways. To compare two groups on a binary outcome, the researcher may choose measures such as differences in proportions, relative risks, odds ratios, risk differences, or the number needed to treat. References 71–72 provide detailed comparisons and discussion of the advantages and disadvantages of these measures, and DuMouchel and Normand (Chap. 6) briefly do so. Reference 70 provides a concise summary and discusses transforming between measures when results are inconsistently reported. If enough information is given in individual studies, transforming between any of these measures is possible. For example, if the odds ratio and the incidence of the event in the control group are reported, then the odds ratio can be transformed to the relative risk. The amount of additional infor-

mation reported will dictate the possible transformations. The problem of combining across different measures is not restricted to binary data. Solutions for analogous problems with continuous data are presented by Abrams et al. (Chap. 2) and with mixed, binary, and continuous data, by Dominici and Parmigiani (Chap. 5).

Abrams et al. (Chap. 2) vary, between their studies, both the outcome measures used and the way results are reported. The six studies included in the meta-analysis use four different instruments to measure anxiety level in two groups of patients. For each group, four of the six studies reported anxiety levels at baseline and follow-up, while the other two studies reported only the mean change from baseline. The authors present both frequentist and Bayesian analyses that take these differences into account. In the frequentist analysis, results from each study are transformed to standardized group differences by assuming that the within-subject correlation is zero for studies that do not report the correlation. Both fixed- and random-effects models are then fitted to these group differences. In the Bayesian analysis, a three-level hierarchy is introduced which places a distribution not only on the standardized group differences and the means of these differences, but also on the variances of the standardized group differences, the overall pooled effect, and the variance in results across studies. This Bayesian model is fitted using a range of values for the within-subject correlation. The model is then extended so that the within-subject correlation is incorporated as a random parameter. Bayes factors are used to examine competing models and to average across plausible models.

A second chapter that addresses variability in reporting of results is that of Dominici and Parmigiani (Chap. 5). Their focus is conducting meta-analyses when some outcomes are continuous and others binary. Using assumptions similar to those in Ref. 73, they develop a hierarchical-Bayesian latent-variables approach. Rather than dichotomizing continuous responses, they assume an underlying latent continuous variable for the discrete outcomes. They apply their approach to a meta-analysis of efficacy of calcium-blockers for preventing migraine headaches.

B. Model Uncertainty

Like any sophisticated statistical analysis, meta-analysis is tricky. Getting the model correct or at least "good enough" requires careful thought and a delicate separation of the wheat from the chaff. In the hierarchical

models promoted in much of this volume, the problem is multiplied because there are several levels of the model: one for within-study variability, one for between-study variability, and often one for the priors on parameters of the between-study variability. Practical graphical techniques for model selection are described by DuMouchel and Normand (Chap. 6). For binary data, they describe plots to assess heterogeneity in risk differences, risk ratios, and odds ratios. This determination is important in choosing which parametrization to analyze. They explain how various plots can be used to help detect publication bias and the presence of study-level covariates that may be related to effect size. Extending Ref. 56, they discuss the use of cross-validated residuals to check for model fit and the presence of outliers. They suggest sensitivity analysis to the prior for the between-study variance component.

Many chapters in this volume address choosing between fixed- and random-effects models: Abrams et al. (Chap. 2), Berry (Chap. 3), Brophy and Joseph (Chap. 4), DuMouchel and Normand (Chap. 6), Larose (Chap. 8), Pauler and Wakefield (Chap. 9), Rahman and Wakefield (Chap. 10), Sargent et al. (Chap. 11), and Smith et al. (Chap. 13). Most of these chapters use residual plots and/or Bayes factors to examine the heterogeneity between studies and to compare the fit of the fixed- and random-effects models. Several chapters demonstrate the difficulty of this task in view of the typically small number of studies included in the meta-analysis. Cautionary warnings on the use of Bayes factors are presented in subsection I.A.2. Sargent et al. (Chap. 11) argue that the decision to fit a fixed-effects versus a random-effects model should be made prior to the analysis.

C. Assessment and Prior Specifications

For most problems, investigators will not have the luxury of postponing decisions until enough data are available to make the decision insensitive to all possible prior distributions. Decisions must be made under uncertainty, and effort needs to go into modeling prior information as well as assessing the impact of possible priors post data analysis. Eliciting a prior distribution is challenging in most Bayesian analyses, and meta-analysis is no exception. In meta-analysis, elicitation is not just about assessing distributions of model parameters. Choosing which studies to include, whether to weight studies depending upon quality of design and implementation, and related decisions are part of the elicitation process. These

topics are covered in Ref. 2, and a brief review of elicitation of prior distributions will be provided here.

Most applications using Bayesian models rely on prior distributions that are either estimated from data or are reference priors (eg., Jeffreys', constant, or unit-information). In many problems, resorting to reference priors is essentially the same as doing a frequentist analysis, although the interpretations differ. Reference 32 presents a review of these priors and discusses their limitations. Graphical methods as described in Refs 74–75, quantile prediction of outcomes as described in Refs 76–80, and providing a range of priors representing beliefs from "skeptical" to "sold" (81–82) are likely to become more popular as software is disseminated and made more user-friendly. Alternate methods are examining classes of priors (83–84) and partitioning priors into subspaces which either support or do not support particular decisions (85–87). Common practice mandates a sensitivity analysis that checks the robustness of results to the prior specification.

In meta-analysis, one of the most important elicitations is the prior distribution of the between-study variability. This distribution controls how much shrinkage occurs across studies, and will have an impact on the variance of the posterior distribution of the overall population effect. At times, there may be little expert opinion or previous empirical evidence to guide the choice of this prior distribution. When the number of studies included in a meta-analysis is small, as is typical, this prior distribution can be very influential on the conclusions. DuMouchel and Normand (Chap. 6) explain the impact of this prior on shrinkage, derive shrinkage factors, and make suggestions for this choice of prior. Pauler and Wakefield (Chap. 9) also address this issue and make suggestions.

D. Missing Data

Many types of missing data can occur in meta-analyses. One type is when entire studies are excluded, because they were not known to the meta-analysts. Publication bias refers to the possibility that the results of these studies may differ from those of published studies. Because publication bias is one of the most widely researched and most important topics of meta-analysis, it merits separate discussion and will be considered in a later section in this introductory chapter. At the study level, missing data may mean that the relevant summary statistics, such as effect sizes and standard errors, are not available for all studies being considered, that the

same covariate information is not available for all studies, or that different subsets of effects are available in each study. At the individual-subject level, all data are usually missing in meta-analysis. While rapidly improving data storage and transfer may change this in the future, few meta-analyses use the individual-subject data.

Missing data are the norm in meta-analysis. Several chapters in this volume address missing data of one type or another, although this may not be the primary focus. Standard approaches for dealing with missing data include complete-case analysis, use of missing-value indicator variables, single-value imputation (either an unconditional mean, conditional mean based on other covariates, or conditional means based on other covariates and the outcome variable), maximum likelihood, and multiple imputation. Bayesian methods for missing covariates include data augmentation and the use of hierarchical submodels. These methods assume the missing data to be a random variable, with an underlying stochastic mechanism. When using data augmentation, one imputes missing data from the variable's predictive distribution. When using hierarchical submodels, missing data are considered to be model parameters.

Gleser and Olkin (Chap. 7) address missing-data problems that arise in meta-analyses when multiple interventions are being compared and different subsets of interventions are included in each study. In the context of a meta-analysis, addressing the effectiveness of three antihypertension therapies for preventing heart disease, they develop and apply an approximation method to estimate effectiveness when each study examines only one or two of three possible therapies. All studies include a control group that receives none of the three therapies. While it is extremely helpful that each study includes this control group, the analysis must adjust for the fact that the comparisons of the therapies with the control are positively correlated within studies. Their approach approximates these correlations and uses weighted least squares to estimate effect sizes and simultaneous confidence intervals. They demonstrate their method both when effects are estimated as increments in proportions and as log-odds ratios. The former may use an arcsine variance-stabilizing transformation that eliminates the need for sample variance estimates.

E. Covariates

In addition to combining results across studies, meta-analyses allow for examining the impact of covariates in a way that individual studies cannot. Differences in the dosage of intervention administered, and differences in the severity of subjects' impairment are but two examples of covariates that may vary more across studies than within studies. In a carefully conducted meta-analysis, researchers can examine the impact of these covariates, and the resultant knowledge is greater than the sum of the parts. Chapters by Brophy and Joseph (Chap. 4), Larose (Chap. 8), DuMouchel and Normand (Chap. 6), Pauler and Wakefield (Chap. 9), and Sargent et al. (Chap. 11) present models for incorporating study-level covariates and provide examples. Larose examines the impact of the study-level covariate—duration of estrogen exposure—on development of endometrial cancer. Duration of estrogen exposure is incorporated through a Bayesian random-effects model. The outcome variable is the logarithm of the relative risk of contracting cancer for estrogen users relative to those who have never used estrogen. Similarly, DuMouchel and Normand (Chap. 6) use a Bayesian hierarchical model to examine the impact of two study-level covariates: route of nicotine-replacement therapy (gum versus patch) and the intensity of support (high versus low) on rates of smoking cessation. They also demonstrate how standard errors of individual-study effects can be included as a covariate to determine whether study size is correlated with outcome. This is also demonstrated by Pauler and Wakefield (Chap. 9).

In a similar vein, Brophy and Joseph (Chap. 4) adjust for two study-level covariates that threaten to bias the results of their meta-analysis. These authors look at clinical trials that compare the ability of two thrombolytic agents to reduce mortality following acute myocardial infarction. The covariates were rates of revascularization (angioplasty and/or by-pass surgery) and method of drug administration. Because this meta-analysis combines only three studies, incorporating these covariates in the regression manner of Larose (Chap. 8) or DuMouchel and Normand (Chap. 6) was not possible. Instead, the authors create prior distributions for the increased mortality rates due to decreased revascularization and less-effective administration protocols used in some studies, and then they recalculate posterior distributions using these priors.

In yet another example, Sargent et al. (Chap. 11) incorporate two study-level covariates, treatment duration and drug dosage, within a random-effects Cox proportional-hazards regression model. They use

stratified Cox models within a hierarchical random-effects structure, grouping treatment effects by level of covariate—that is, studies with covariates in common were given a single prior. Results showed that dosage, but not duration, had an impact on parameter estimates.

F. Publication Bias

A serious threat to the validity of any meta-analysis is publication bias. Publication bias can occur at pre-publication (called the "file-drawer problem"), journal review, and/or post-publication (88). This bias arises when submission, review, or publication of meta-analyses is restricted to studies having a particular type of result. Investigators, reviewers, and editors often base decisions regarding submission or acceptance of manuscripts for publication on whether the study shows a "statistically significant" effect. Published studies are the most accessible. Careless meta-analysis compiles a set of biased studies, overestimates effects, and underestimates standard errors. Unwary users of such analyses are presented with highly convincing arguments that can lead to inappropriate decision making. For example, Ref. 89 examined a particular treatment for ovarian cancer. This demonstrated that a meta-analysis of only the published trials led to a conclusion of significant improvement, while inclusion of all studies registered with the International Cancer Research Data Bank did not. The scientific literature in most fields includes documentation that publication bias exists. Discussions and literature reviews on publication bias can be found in Refs 90–91. Both cite studies in the natural sciences, social sciences, and medicine demonstrating the pervasiveness of this problem. Other references include Refs 92–96.

Reference 97 reviews methods for identifying publication bias. It explains why sample sizes provide an important clue for detecting publication bias, describes the "funnel graph" (5), which plots sample size against effect size across studies to determine the potential of publication bias, and presents statistical tests that help formally assess the presence of publication bias. Reference 98 proposes methods to quantify the funnel graph, structuring the association between bias and sample size through a model which assumes no correlation between the effects and sample size. Graphical methods for detecting publication bias are also reviewed by DuMouchel and Normand (Chap. 6).

The impact of publication bias can be mediated in several ways. Meta-analyses can be restricted to sampling studies contained in certain

pre-defined sampling frames, such as "complete" registers. However, this method is restrictive, and more appropriate methods have been developed. In Ref. 99, a pooled z-score from published studies is calculated. Then the number of zero-effect studies that would be required to deem this pooled z-score no longer significant is compared with an estimate of the number of studies that have gone unpublished. This "file-drawer" method is easy to implement, but missing results are centered at the null hypothesis of no effect, it does not adjust for degree of publication bias, and does not provide a corrected estimate. DuMouchel and Normand (Chap. 6) adjust for publication bias by incorporating standard errors of study-specific effects as a covariate. Their example on assessing the impact of nicotine-replacement therapy on smoking cessation shows large differences in effect estimates after adjusting for possible publication bias. Pauler and Wakefield (Chap. 9) also use this approach in the context of a meta-analysis of 13 randomized trials comparing drugs to reduce hypertension.

Early efforts to model publication bias include Refs. 100–102. References 100–101 examine publication bias where a study is published only if it yields significant results. Their methods use weight functions that assume observations in certain parts of a distribution are more likely to be observed. Reference 102 uses an approach to include specific fixed monotonic families of weight functions, and Ref. 103 further extends publication-bias models to account for heterogeneity as well as bias. More sophisticated work with weight functions appear in Refs 104–105. Both consider publication-bias models in which the weight function is estimated by maximum likelihood. Using Bayesian selection models, Refs 106–107 model the selection mechanisms of published results. Data augmentation is used to estimate and adjust for publication bias in Ref. 60. Synthesizing and building on these works, Refs 108–109 show that statistical models describing publication bias can be constructed quite naturally using weighted distributions. This method models bias by adjusting the probabilities of actual event occurrence to arrive at the probabilities that events are observed and recorded. Reference 110 uses hierarchical selection models, with parametric and step weight functions, to address selection bias along with heterogeneity between study effects and sensitivity of results to any unobserved study effects. Most of these authors adjust for publication bias based on a single factor, such as significance level. Smith et al. (Chap. 13) extend the data-augmentation methods of Ref. 60 by developing a strategy that adjusts for significance level and also for the number of and outcomes of studies which may be

missing, within strata defined by research quality. They apply their method to a meta-analysis of studies of cervical cancer rates associated with use of oral contraceptives.

Another type of publication bias, called first-report bias, is discussed by Simon (Chap. 12). Within the context of clinical trials, he argues that the first study should often be interpreted differently than confirmatory studies. He presents a rationale based on calculating the posterior odds of an effective treatment given the design (detectable treatment effect, power, and desired significance) and the prior probability that the treatment represents a medically important improvement.

G. Software

Several computer-software packages and collections of program macros are available for doing meta-analysis. A comparative review of three packages—DSTAT, TRUE EPISTAT, and FAST*PRO—is available in Ref. 111, where a review covers data requirements, model assumptions, and data input/output. As noted in the article, two of the three companies had released new versions which addressed all criticisms made in the article by the time the article was published. This, along with the review of Sutton et al. (Chap. 15), demonstrates the speed with which meta-analysis software is becoming available and adapting to user needs. Sutton et al. briefly highlight others' reviews of DSTAT, TRUE EPISTAT, and FAST*PRO and present a thorough review of seven newer packages: four stand-alone packages (Review Manager, EasyMA, MetaGraphs, and Descartes), two collections of macros that run on SAS and Stata, and one general statistical package with meta-analysis options (Arcus). The authors point out the improvements and versatility of the available software, but also note that these resources are still deficient in providing model checking, incorporating publication bias, and performing sensitivity analysis.

III. OTHER CHALLENGES ADDRESSED IN THIS VOLUME

Most chapters in this volume propose improving meta-analysis methods by increasing the level of quantitative sophistication. A chapter of a different sort is Stoto (Chap. 14). Based upon first-hand experience

with two Institute of Medicine projects, he addresses the role of quantification versus professional-group judgement in meta-analysis, and proposes pragmatic guidelines. While not dismissing the need for improved methods, his development contrasts with the increasingly technical advancements proposed in most of the other chapters. Stoto is in accord with most chapters on the importance of the review of heterogeneity and systematic variation between studies. He also argues that (a) group judgements require both quantitative and qualitative aspects, but reporting of results within a few standard qualitative categories (e.g., no association, inadequate/insufficient, limited/suggestive, and sufficient) was most useful, most reliable, and does not make finer distinctions about the quality of evidence than the data support; (b) professional groups should be comprised of "non-biased experts," naïve on the topic in question, that adopt a neutral starting point in evaluating evidence; and (c) formal quantitative meta-analysis approaches are often impractical because of the heterogeneity in studies, the difficulty in extracting data from published literature, and the focus on causation rather than statistical association. These conclusions challenge the increasingly sophisticated methods proposed in most of the volume concerning the role of the statistician in conducting meta-analyses.

IV. THE FUTURE OF META-ANALYSIS

So what lies ahead? What will be the fate of quantitative methods for meta-analysis? The key may lie in our ability to develop and use increasingly sophisticated statistical models to produce more reasoned conclusions.

Improved technologies will allow increasing sophistication. We are likely to see much more emphasis on analyses that make use of the original raw data rather than the few summary statistics that fit on journal pages. Computing and communication advances will make data collection more complete, data transfer much simpler, and data analysis across studies much more sophisticated. Research will continue to develop methods for combining studies with different types of outcome and different sampling and data collection designs. We are also likely to see increasing use of observational data, available via international registries, to supplement clinical trials.

The models presented in this volume will be able to capitalize on these improved technologies. However, hurdles remain. Three such hurdles are small samples, unclear definition of the unit of analysis, and choice of parametrization. To derive good estimates of between-study heterogeneity and to be able to check model fit, a sufficient number of studies must be included. If the number of studies is small—say less than 10—but the studies themselves are large, our estimate of variability may be satisfactory. But assessing model fit will still be problematic for a small number of large studies. We need to continue searching for adequate ways of checking that a sample of studies represents the population of studies, or for ways to adjust if it does not.

Related to the problem of estimating variability is the unit-of-analysis question: what constitutes a trial or study in a meta-analysis? The meta-analysis by Brophy and Joseph (Chap. 4) included three clinical trials each sampling more than 20,000 patients with several outcome events each. Within each of these trials, patients were recruited from many treatment centers. The authors explored the heterogeneity across the three clinical trials, but the heterogeneity between treatment centers within trials, and between clinicians within treatment centers, is not explored. The question of concern is: at what level of disaggregation can we stop worrying about heterogeneity?

Several chapters demonstrate that the choice of parametrization makes a difference. This has implications for decision making. For example, in the Cox proportional-hazards models of Sargent et al. (Chap. 11), parametrization determined whether shrinkage occurred within each treatment group or in the relative treatment effect. Shrinkage of survival proportions within each treatment group will not give the same answer as shrinking the difference between these two proportions. Most authors chose decision models analogous to hypothesis testing, and used standard significance criteria. Given that the choice of parametrization can affect whether confidence and posterior intervals cover the region of the null hypothesis, it is clear that we must do a better job of educating consumers and decision makers about this problem and its implications. This brings us to the biggest challenge for the new millennium.

Will improved but more sophisticated methods have any more impact than previous methods? Stoto argues that simple is better. How do we employ increasingly sophisticated statistical models to produce increasingly relevant output? We need to focus on decision making. For example, consider the choice between two treatments for cancer. Which is more useful: a predictive survival curve for each treatment or

a log hazards ratio? If we want to test a point null hypothesis, then the latter is more useful. If we want to make a decision about the relative utility of the two treatments, the former is more useful. The former is in units which the decision maker can directly and easily consider, while the latter is not. Using a predictive distribution, the decision maker can assess the various decisions under a variety of utility functions, while the units of a log hazards ratio may be irrelevant for making decisions. The shift from fixed- to random-effects models has been a desirable one. But, do we need to take it a step further by providing predictive distributions for the outcomes of interest?

References

1. TD Cook, H Cooper, DS Cordray, H Hartman, LV Hedges, RJ Light, TA Louis, F Mosteller. Meta-Analysis for Explanation. New York: Russell Sage Foundation, 1992.
2. H Cooper, LV Hedges. The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1994.
3. DM Eddy. The confidence profile method: a Bayesian method for assessing health technologies. Operations Research 37:210–228, 1989.
4. LV Hedges, I Olkin. Statistical Methods for Metaanalysis. Orlando, FL: Academic Press, 1985.
5. RJ Light, DB Pillemer. Summing Up: The Science of Reviewing Research. Cambridge, MA: Harvard University Press, 1984.
6. DB Petitti. Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine. Oxford University Press, 1994.
7. R Rosenthal. Meta-Analytic Procedures for Social Research. Beverly Hills: Sage Publications, 1984.
8. KW Wachter, ML Straf. The Future of Meta-Analysis. New York: Russell Sage Foundation, 1990.
9. D Draper, D Gaver, P Goel, J Greenhouse, L Hedges, C Morris, C Waternaux. On combining information: statistical issues and opportunities for research. A report of a panel convened by the National Academy of Sciences, 1992.
10. BJ Becker. Combining significance levels. In: H Cooper and LV Hedges, eds. The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1994, pp 215–230.
11. LHC Tippett. The Method of Statistics. London: Williams and Norgate, 1931.

12. RA Fisher. Statistical Methods for Research Workers. London: Oliver and Boyd, 1932.
13. WR Shadish, CK Haddock. Combining estimates of effect size. In: H Cooper and LV Hedges, eds. The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1994, pp 261–282.
14. JC Bailar III. The promise and problems of meta-analysis. New England Journal of Medicine 337:559–561, 1997.
15. LV Hedges. Fixed effects models. In: H Cooper and LV Hedges, eds. The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1994, pp 285–300.
16. TC Smith, DJ Spiegelhalter, A Thomas. Bayesian approaches to random-effects meta-analysis: A comparative study. Statistics in Medicine 14:2685–99, 1995.
17. N Mantel, W Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 22:719–748, 1959.
18. B Wolfe. On estimating the relation between blood group and disease. Annals of Human Genetics 19:251–253, 1955.
19. S Yusef, R Peto, J Lewis, R Collins, P Sleight. Beta blockade during and after myocardial infarction: An overview of the randomized trials. Progress in Cardiovascular Diseases 27:335–371, 1985.
20. R DerSimonian, N Laird. Meta-analysis in clinical trials. Controlled Clinical Trials 7:177–188, 1986.
21. Morris, C., Normand, S.L. Hierarchical models for combining information and for meta-analyses with discussion. In: J. Bernardo, J. Berger, A. Dawid, and A. Smith, eds. Bayesian Statistics 4. London: Oxford University Press, 1992, pp 321–344.
22. BJ Biggerstaff, RL Tweedie, KL Mengersen. Passive smoking in the work-place: classical and Bayesian meta-analyses. Int. Arch Occup. Environ. Health 66:260-77, 1994.
23. XY Su, ALW Po. Combining event rates from clinical trials: Comparison of Bayesian and classical methods. Annals of Pharmacotherapy 30:460–465, 1996.
24. RL Tweedie, DJ Scott, BJ Biggerstaff, KL Mengersen. Bayesian meta-analysis, with application to studies of ETS and lung cancer. Lung Cancer 14:171–94, 1996.
25. R Kass, D Steffey. Approximate Bayesian inference in conditionally independent hierarchical models. Journal of the American Statistical Association, 84(407):717–726, 1989.
26. JL Fleiss. Measures of effect size for categorical data. In: H Cooper and LV Hedges, eds. The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1994, pp 245–260.

27. Early Breast Cancer Trialists' Collaborative Group. Treatment of Early Breast Cancer. Volume 1: Worldwide Evidence, 1985–1990. Oxford: Oxford University Press, 1990.
28. DV Lindley. A discussion of: Hierarchical models for combining information by C.N. Morris and S.L. Normand. In: JM Bernardo, JO Berger, AP Dawid, AFM Smith, eds. Bayesian Statistics IV. London: Oxford Science Publications, 1992, pp 341–342.
29. S Berry. Understanding and testing across 2×2 tables: Application to meta-analysis. Statistics in Medicine 17:2353–2369, 1998.
30. M Lavine, M Schervish. Bayes factors: What they are and what they are not. American Statistician 53(2):119–122, 1999.
31. JB Kadane, JM Dickey. Bayesian decision theory and the simplification of models. In: J Kmenta, J. Ramsey, eds. Evaluation of Econometric Models, Academic Press, 1980, pp 245–268.
32. D Stangl. The use of reference priors and Bayes factors in the analysis of clinical trials. *Proceedings of the IMA Conference on Statistical Methods in the Health Sciences*, Minneapolis, 1997, pp 237–250.
33. HC Sox, MA Blatt, MS Higgins, KI Marton. Medical Decision Making. Boston: Butterworth and Heinemann, 1988.
34. DA Berry. Decision analysis and Bayesian methods. In: PF Thall, ed. Recent Advances in Clinical Trial Design and Analysis. Boston: Kluwer Academic Publishers, 1995, pp 125–154.
35. V Hasselblad, D McCrory. Meta-analysis Tools for Medical Decision Making: A Practical Guide. Medical Decision Making 15:81–96, 1995.
36. R Lewis, DA Berry. Decision theory. In: P Armitage and T Colton, eds. Encyclopedia of Biostatistics. New York: John Wiley & Sons, 1998.
37. MH Degroot. Optimal Statistical Decisions. McGraw Hill Book Company, 1970.
38. JO Berger. Statistical Decision Theory and Bayesian Analysis. 2nd ed. New York: Springer-Verlag, 1985.
39. G Parmigiani, MS Kamlet. A cost–utility analysis of alternative strategies in screening for breast cancer. In: C Gatsonis, J Hodges, R Kass, N Singpurwalla, eds. Case Studies in Bayesian Statistics, New York: Springer-Verlag, 1993, pp 390–402.
40. G Parmigiani, M Ancukiewicz, D Matchar. Decision models in clinical recommendations development: The stroke prevention policy model. In: DA Berry, DK Stangl, eds. Bayesian Biostatistics. New York: Marcel Dekker, 1996, pp 207–233.
41. DA Berry, MC Wolff, D Sack. Public health decision making: A sequential vaccine trial with discussion. In: JM Bernardo, JO Berger, AP Dawid, AFM Smith, eds. Bayesian Statistics. Oxford, England: Oxford University Press, 1992, pp 79–96.

42. DA Berry, MC Wolff, D Sack. Decision making during a phase III randomized controlled trial. Controlled Clinical Trials 15:360–379, 1994.
43. D Stangl. Prediction and decision making using Bayesian hierarchical models. Statistics in Medicine 14:2173–2190, 1995.
44. L Wolfson, J Kadane, M Small. Expected utility as a policy-making tool: An environmental health example. In: D Berry and D Stangl, eds. Bayesian Biostatistics. New York: Marcel Dekker, 1996.
45. W DuMouchel. Bayesian meta-analysis, In DA Berry, ed. Statistical Methodology in the Pharmaceutical Sciences. New York: Marcel Dekker, 1990, pp 509–529.
46. WH DuMouchel, JE Harris. Bayes methods for combining the results of cancer studies in humans and other species with discussion. Journal of the American Statistical Association 78:293–315, 1983.
47. SW Raudenbush, AS Bryk. Empirical Bayes metaanalysis. Journal of Educational Statistics 10:75–98, 1985.
48. JB Carlin. Meta-analysis for 2 × 2 tables: A Bayesian approach. Statistics in Medicine 11:141–158, 1992.
49. A Whitehead, J Whitehead. A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine 10:1665–1677, 1991.
50. DM Eddy, V Hasselblad, R Schachter. A Bayesian method for synthesizing evidence. International Journal of Technical Assistance in Health Care 6:31–55, 1990.
51. DM Eddy, V Hasselblad, R Shachter. An introduction to a Bayesian method for meta-analysis: the confidence profile method. Medical Decision Making 10:15–23, 1990.
52. DM Eddy, V Hasselblad, R Shachter. Meta-analysis by the Confidence Profile Method. The Statistical Synthesis of evidence. Academic Press Inc., 1992.
53. SL Normand. Random effects methods for dose-response meta-analyses. Technical Series in Statistics, Report 1–1994, Dept. of Health Care Policy, Harvard Medical School, 1994.
54. Z Li, CB Begg. Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. Journal of the American Statistical Association 89:1523–1527, 1994.
55. A Whitehead, NMB Jones. A meta-analysis of clinical trials involving different classifications of response into ordered categories. Statistics in Medicine 13:2503–2515, 1994.
56. WH DuMouchel. Predictive cross-validation in Bayesian meta-analysis. In: JM Bernardo, JO Berger, AP Dawid, AFM Smith, eds. Bayesian Statistics 5. Oxford: Clarendon Press, 1996, pp 107–124.
57. W DuMouchel. Meta-analysis for dose-response models. Statistics in Medicine 14:679–685, 1995.

58. A Rogatko. Bayesian Approach for meta-analysis of controlled clinical trials. Communications in Statistics, Part A (Theory and Methods) 21:1441–1462, 1992.
59. CS Berkey, DC Hoaglin, F Mosteller, DA Colditz. A random-effects model for meta-analysis. Statistics in Medicine 14:395–411, 1995.
60. GH Givens, DD Smith, RL Tweedie. Bayesian data-augmented meta-analysis that account for publication bias issues exemplified in the passive smoking debate, Statistical Science 12:221–250, 1997.
61. K Abrams, B Sanso. Discrimination in meta-analysis—a Bayesian perspective. Technical Report 95-03. Department of Epidemiology and Public Health, University of Leicester, U.K., 1995.
62. DG Clayton. A Monte Carlo method for Bayesian Inference in frailty models. Biometrics 47:467–485, 1991.
63. CB Begg, L Pilote. A model for incorporating historical controls into a meta-analysis. Biometrics 47:899–906, 1991.
64. RT Clemen. Making Hard Decisions. Boston: PWS-Kent Publishing Company, 1991.
65. R Gray. A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. Biometrics 50:244–253, 1994.
66. C Morris, C Christiansen. Fitting Weibull duration models with random effects. Lifetime Data Analysis 1:347–359, 1995.
67. SK Sahu, DK Dey, H Aslanidou, D Sinha. A Weibull regression model with gamma frailties for multivariate survival data. Lifetime Data Analysis 3:123–137, 1997.
68. D Stangl. Modeling heterogeneity in multicenter clinical trials using Bayesian hierarchical models. Ph.D. Dissertation, Carnegie Mellon University, 1991.
69. D Stangl. Hierarchical analysis of continuous-time survival models. In: D Berry, D Stangl, eds. Bayesian Biostatistics. New York: Marcel Dekker, 1996, pp 429–450.
70. AJ Sutton, KR Abrams, DR Jones, TA Sheldon, F Song. Systematic reviews of trials and other studies. Health Technology Assessment 2(19), 1999.
71. JC Sinclair, MB Bracken. Clinically useful measures of effect in binary analyses of randomized trials. Journal of Clinical Epidemiology 47:881–889, 1994.
72. JL Fleiss. The statistical basis of meta-analysis. Statistical Methods in Medical Research 2:147–160, 1993.
73. A Whitehead, AJ Bailey, D Elbourne. Combining summaries of binary outcomes with those of continuous outcomes in a meta-analysis. Journal of Biopharmaceutical Statistics 9:1–16, 1999.
74. K Chaloner. Elicitation of Prior Distributions. In: D Berry, D Stangl, eds. Bayesian Biostatistics. New York: Marcel Dekker, 1996, pp 141–156.