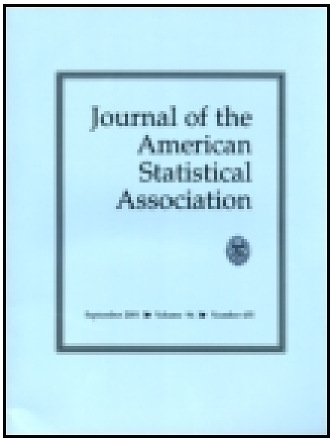


Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Meta-Analysis of Rare Binary Adverse Event Data

Dulal K. Bhaumik [a] [b] , Anup Amatya [c] , Sharon-Lise T. Normand [d] , Joel Greenhouse [e] , Eloise Kaizar [f] , Brian Neelon [g] & Robert D. Gibbons [h]

[a] Department of Psychiatry, Division of Epidemiology and Biostatistics , University of Illinois at Chicago , Chicago , IL , 60612

[b] Veteran Affairs Hospital , Hines , IL

[c] Department of Health Sciences , New Mexico State University , Las Cruces , NM , 88003

[d] Department of Health Care Policy , Harvard Medical School , Boston , MA , 02115-5899

[e] Department of Statistics , Carnegie Mellon University , Pittsburgh , PA , 15213

[f] Department of Statistics , , Ohio State University , Columbus , OH , 43210

[g] Department of Biostatistics and Bioinformatics , Duke University Medical Center , Durham , NC , 27708-0328

[h] Departments of Health Studies and Medicine, and Director of Center for Health Statistics , University of Chicago , Chicago , IL , 60637

Accepted author version posted online: 20 Mar 2012.Published online: 24 Jul 2012.

Meta-Analysis of Rare Binary Adverse Event Data

Dulal K. BHAUMIK, Anup AMATYA, Sharon-Lise T. NORMAND, Joel GREENHOUSE, Eloise KAIZAR, Brian NEELON, and Robert D. GIBBONS

We examine the use of fixed-effects and random-effects moment-based meta-analytic methods for analysis of binary adverse-event data. Special attention is paid to the case of rare adverse events that are commonly encountered in routine practice. We study estimation of model parameters and between-study heterogeneity. In addition, we examine traditional approaches to hypothesis testing of the average treatment effect and detection of the heterogeneity of treatment effect across studies. We derive three new methods, a simple (unweighted) average treatment effect estimator, a new heterogeneity estimator, and a parametric bootstrapping test for heterogeneity. We then study the statistical properties of both the traditional and the new methods via simulation. We find that in general, moment-based estimators of combined treatment effects and heterogeneity are biased and the degree of bias is proportional to the rarity of the event under study. The new methods eliminate much, but not all, of this bias. The various estimators and hypothesis testing methods are then compared and contrasted using an example dataset on treatment of stable coronary artery disease.

KEY WORDS: Background rate; Bootstrap; DerSimonian and Laird; Heterogeneity; Mantel-Haenszel; Random-effects; Sparse data.

1. INTRODUCTION

The use of meta-analysis for research synthesis has become routine in medical research. Unlike early developments for effect sizes based on continuous and normally distributed outcomes (Hedges and Olkin 1985), applications of meta-analysis in medical research often focus on the odds ratio (Engels et al. 2000; Deeks 2002) between treated and control conditions in terms of a binary indicator of efficacy and/or the presence or absence of an adverse drug reaction (ADR). The two most widely used statistical methods for meta-analysis of a binary outcome are the fixed-effects model [Mantel and Haenszel (MH) 1959] and the random-effect model [DerSimonian and Laird (DSL) 1986]. A special statistical problem arises when the focus of research synthesis is on a rare binary event, such as a rare ADR.

The literature of fixed-effects meta-analysis for sparse data provides a solid guideline for both continuity correction and methods to use. The standard use of a continuity correction for binary data may not be appropriate for sparse data as the number of zero cells for such data becomes large. Sweeting, Sutton, and Lambert (2004) showed via simulation that for sparse data with homogeneous treatment effect, the "empirical correction," which incorporates information on odds ratios from other studies, and the "treatment arm correction," which uses the reciprocal of the size from the other arm, perform better than the constant 0.5 correction for both the MH and the inverse-variance weighted methods. Their investigation reveals that for fixed-effects models, the MH method performs consistently better than the inverse-variance weighted method for imbalanced group sizes and all continuity corrections. They found that the Peto method is almost unbiased for balanced group sizes and the bias increases with respect to the group imbalance.

Bradburn et al. (2007) performed an extensive simulation study to compare a number of fixed-effects methods of pooling odds ratios for sparse data meta-analysis. They considered balanced as well as highly imbalanced group sizes and used a constant 0.5 zero-cell correction only when required. Their investigation revealed that most of the well-known meta-analysis methods are biased for sparse data. They found that the Peto method is the least biased and the most powerful for within-study balanced sparse data, which matches with the findings of Sweeting, Sutton, and Lambert (2004); whereas for unbalanced cases, the MH without zero-cell correction, logistic regression, and the exact method have similar performance and are less biased than the Peto method. They concluded that the method of analysis should be chosen based on the expected treatment effect size, imbalance of the study arms, and the underlying event rates. The general recommendation is to use the MH method with an appropriate continuity correction and avoid the inverse-variance weighted average and DSL methods when dealing with sparse data with homogeneous treatment effect.

Relatively less attention has been paid to heterogeneous treatment effects or moment-based meta-analysis with random effects for sparse data. Sweeting, Sutton, and Lambert (2004) performed a limited simulation study using random-effects models to combine odds ratios for sparse data. In 95% of the cases, they did not get valid estimates (i.e., positive estimates) of the between-study variance. As a consequence, their results for random-effects models were close to those for the fixed-effects model. For random-effects meta-analysis, Shuster (2010) showed via simulation that inverse-variance weighted average estimates, including the DSL method, are highly biased. Based

Dulal K. Bhaumik is Professor of Biostatistics, Department of Psychiatry, Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, IL 60612; Senior Biostatistician at the Veteran Affairs Hospital, Hines, IL (E-mail: *dbhaumik@psych.uic.edu*). Anup Amatya is Assistant Professor of Biostatistics, Department of Health Sciences, New Mexico State University, Las Cruces, NM 88003 (E-mail: *aamatya@nmsu.edu*). Sharon-Lise T. Normand is Professor, Department of Health Care Policy, Harvard Medical School, Boston, MA 02115-5899 (E-mail: *sharon@hcp.med.harvard.edu*). Joel Greenhouse is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *joel@stat.cmu.edu*). Eloise Kaizar is Assistant Professor, Department of Statistics, Ohio State University, Columbus, OH 43210 (E-mail: *ekaizar@stat.osu.edu*). Brian Neelon is Assistant Professor, Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27708-0328 (E-mail: *neelo003@duke.edu*). Robert D. Gibbons is Professor of Biostatistics in the Departments of Health Studies and Medicine, and Director of Center for Health Statistics, University of Chicago, Chicago, IL 60637 (E-mail: *rdg@uchicago.edu*). This work was supported by NIMH (National Institute of Mental Health) grants MH8012201 (RDG, DKB, and AA), RO1-MH7862 (JG) and MH054693 (SLTN). The authors are grateful to an associate editor and two referees for suggestions that considerably improved the article.

Journal of the American Statistical Association
June 2012, Vol. 107, No. 498, Applications and Case Studies
DOI: 10.1080/01621459.2012.664484

on his findings, he strongly advocated for the simple (unweighted) average estimate for random-effects meta-analysis.

Available random-effect methods consistently underestimate the heterogeneity parameter (DerSimonian and Kacker 2007). The random-effects meta-analysis also requires an appropriate continuity correction to estimate the treatment effect. Although Sweeting, Sutton, and Lambert (2004) showed that the empirical and treatment arm corrections performed better than 0.5 cell correction for fixed-effects models, they cautioned against the applicability of the empirical continuity correction for the random-effects model, as for such models, the underlying treatment effect varies between studies.

The focus of this article is on random-effects meta-analysis for sparse data. We first look for a continuity correction to make our moment-based estimate of the treatment effect asymptotically unbiased for a single study. Next, we extend this concept of bias correction for multiple studies and propose an asymptotically unbiased estimate that matches with the finding of Shuster (2010). We organize the article as follows. In Section 2, we discuss various meta-analytic methods for estimating relevant model parameters. In this section, we propose two new methods: one for estimating the treatment effect and the other for estimating the heterogeneity parameter. In Section 3, we investigate hypothesis-testing problems for these parameters. For the heterogeneity parameter, we show that standard testing procedures have very poor power. Using the concept of parametric bootstrapping (PB), we propose a testing procedure for the heterogeneity parameter that provides better power. In Section 4, we compare performances of several methods via simulation and show that our proposed methods provide very satisfactory results. In Section 5, we illustrate our results with an example of percutaneous coronary intervention (PCI) versus medical treatment alone (MED) in the treatment of patients with stable coronary artery disease. We conclude with a discussion of our results in Section 6.

2. ESTIMATION OF MODEL PARAMETERS

Consider a meta-analysis consisting of k randomized studies. In the ith study, n_{it} subjects are randomly assigned to the treatment group, and the remaining n_{ic} subjects are assigned to the control group. The outcome variable is characterized as a success or failure and accordingly assigned a value of 1 or 0. Let x_{it} and x_{ic} be, respectively, the numbers of observed events of interest in the treatment and control groups of the ith study. One general approach to model the between-study variation is to use a binomial-normal hierarchical model. Let p_{it} and p_{ic} be, respectively, the probabilities of observing an event in the treatment and control groups. The model can be expressed as

$$\begin{gathered} x_{ic} \sim B(p_{ic}, n_{ic}), \quad x_{it} \sim B(p_{it}, n_{it}), \\ \text{logit}(p_{ic}) = \mu_i, \quad \text{logit}(p_{it}) = \mu_i + \theta + \epsilon_i, \\ \ln\left\{\frac{p_{it}/q_{it}}{p_{ic}/q_{ic}}\right\} = \theta + \epsilon_i, \qquad \text{and} \qquad \epsilon_i \sim N(0, \tau^2), \end{gathered} \tag{1}$$

where $q_{it} = 1 - p_{it}$ and $q_{ic} = 1 - p_{ic}$. The primary focus of this article is on estimation and testing of the treatment effect (θ) and the heterogeneity parameter (τ^2).

We start by reviewing the moment-based estimators that form the basis of current routine practice in this area (e.g., the Cochrane Reviews). As we will show, the moment-based estimators can result in quite biased estimates of the overall treatment effect in the presence of heterogeneity of the treatment effect across studies. We study the bias of these estimators and propose new alternative moment-based methods that improve overall performance and testing.

The MH and empirical logit (EL) methods (also known as the inverse-variance method) used for estimating the odds ratios are moment-based approaches that ignore heterogeneity among studies. The DSL method incorporates the between-study variability in the weighted average estimate of θ, assuming τ^2 is known. When the binary outcome is rare, and one or more cells in a study are zero, traditional moment-based approaches such as EL and DSL break down as it becomes impossible to compute the odds ratio. Several numeric adjustments to correct this problem have been suggested. Haldane (1955) added 1/2 to the observed frequencies to estimate the treatment effect for a single study, whereas for multiple studies, Cox (1970) added −1/2 to estimate the same parameter. Both of the authors proved that for fixed-effects models, their estimates are optimal in terms of reducing the bias to the first-order approximation. For studies of rare events, when a large number of observed frequencies are zero, the empirical and treatment arm corrections proposed by Sweeting, Sutton, and Lambert (2004) provide better results than the constant 0.5 correction for fixed-effect meta-analysis. In what follows, we discuss our approach to select the continuity correction for random-effects meta-analysis with sparse data.

We first add a positive constant a to the observed frequencies and estimate the treatment effect for the model defined in (1) and then determine the optimal value of a to make the estimate unbiased. Let $\hat{\theta}_{ia}$ be an estimate of the treatment effect based on the ith study:

$$\hat{\theta}_{ia} = \ln\left\{\frac{x_{it} + a}{n_{it} - x_{it} + a}\right\} - \ln\left\{\frac{x_{ic} + a}{n_{ic} - x_{ic} + a}\right\}, \tag{2}$$

where x_{ij} and n_{ij} are, respectively, the observed number of events and total sample size in the jth group, $j = t, c$, of the ith study. Using the result presented in Appendix A, we compute the following expression for the expected value of $\hat{\theta}_{ia}$:

$$\begin{aligned} E(\hat{\theta}_{ia}) = \theta + \left(\frac{\frac{1}{2} - a}{n_{it}}\right)[e^{\mu+\theta+\tau^2/2} - e^{-\mu-\theta+\tau^2/2}] \\ - \left(\frac{\frac{1}{2} - a}{n_{ic}}\right)[e^{\mu} - e^{-\mu}] + O(n^{-2}). \end{aligned} \tag{3}$$

Inspecting the right-hand expression in (3), we observe that the first-order term of bias (i.e., terms of order n^{-1}) will vanish if $a = 1/2$. The implication of this result is that the estimate $\hat{\theta}_{i1/2}$ is unbiased up to the order of n^{-1}, and hence, the simple average estimate $\hat{\theta}_{s1/2} = \sum_{i=1}^{k} \frac{\hat{\theta}_{i1/2}}{k}$ is also unbiased. We now explore the properties of the weighted average estimate. The variance of $\hat{\theta}_{i1/2}$, shown in Appendix B, is

$$\begin{aligned} V(\hat{\theta}_{i1/2}) = \frac{1}{n_{it}}\left[e^{\mu+\theta+\frac{\tau^2}{2}} + e^{-\mu-\theta+\frac{\tau^2}{2}}\right] + \frac{1}{n_{ic}}[e^{\mu} + e^{-\mu}] \\ + \frac{2}{n_{it}} + \frac{2}{n_{ic}} + \tau^2 + O(n^{-2}). \end{aligned} \tag{4}$$

On the right-hand side of (4), the quantity $\frac{1}{n_{it}}[e^{\mu+\theta+\frac{\tau^2}{2}} + e^{-\mu-\theta+\frac{\tau^2}{2}}] + \frac{1}{n_{ic}}[e^{\mu}+e^{-\mu}] + \frac{2}{n_{it}} + \frac{2}{n_{ic}}$ is the expression for the within-study variance, and τ^2 is the between-study variance. Thus, $V(\hat{\theta}_{i1/2})$ is the sum of the within- and between-study variances. Note that

$$E_{\epsilon}\left(\frac{1}{n_{it}p_{it}q_{it}} + \frac{1}{n_{ic}p_{ic}q_{ic}}\right) = \frac{1}{n_{it}}\left[e^{\mu+\theta+\frac{\tau^2}{2}} + e^{-\mu-\theta+\frac{\tau^2}{2}}\right] + \frac{1}{n_{ic}}[e^{\mu}+e^{-\mu}]+\frac{2}{n_{it}}+\frac{2}{n_{ic}}. \quad (5)$$

Hence, the quantity $\frac{1}{n_{it}p_{it}q_{it}} + \frac{1}{n_{ic}p_{ic}q_{ic}} + \hat{\tau}^2$ is unbiased for $V(\hat{\theta}_{i1/2})$, provided that $\hat{\tau}^2$ is unbiased for τ^2. A usual estimate of $V(\hat{\theta}_{i1/2})$ denoted by $\hat{V}(\hat{\theta}_{i1/2})$ is

$$\hat{V}(\hat{\theta}_{i1/2}) = \hat{\sigma}_i^2(\hat{\tau}^2) = \frac{1}{n_{it}\hat{p}_{it}\hat{q}_{it}} + \frac{1}{n_{ic}\hat{p}_{ic}\hat{q}_{ic}} + \hat{\tau}^2, \quad (6)$$

where $\hat{p}_{it} = \frac{x_{it}+1/2}{n_{it}+1}$ and $\hat{p}_{ic} = \frac{x_{ic}+1/2}{n_{ic}+1}$. The study-specific estimate of variance of the treatment effect recommended by DSL has the same expression as on the right-hand side of (6). Let $\hat{w}_i(\tau^2) = \frac{1}{\hat{\sigma}_i^2(\tau^2)}$. The weighted average estimate of θ denoted by $\hat{\theta}_{wa}$ is

$$\hat{\theta}_{wa}(\tau^2) = \frac{\sum_{i=1}^{k}\hat{w}_i(\tau^2)\hat{\theta}_{ia}}{\sum_{i=1}^{k}\hat{w}_i(\tau^2)}. \quad (7)$$

When τ^2 is assumed to be 0, the estimate in Equation (7) is known as the EL or inverse-variance estimator of common log odds ratio. In contrast, the MH estimate of the common odds ratio, which also assumes $\tau^2 = 0$, is given by $\hat{\theta}_{\text{MH}} = \sum_{i=1}^{K}\frac{x_{it}(n_{ic}-x_{ic})}{n_{it}+n_{ic}} / \sum_{i=1}^{K}\frac{x_{ic}(n_{it}-x_{it})}{n_{it}+n_{ic}}$.

Note that (a) the weights $\hat{w}_i(\tau^2)$ are biased (Malzahn, Bohning, and Holling 2000; Bohning et al. 2002), and (b) $\hat{w}_i(\tau^2)$ and $\hat{\theta}_{ia}$ are correlated (as $\hat{p}_i$ and $\frac{1}{\hat{p}_i\hat{q}_i}$ are correlated). Shuster (2010) proved that when estimated effect size and empirically derived weights are correlated, the weighted average estimate from the random-effects model provided in (7) is biased. In Appendix E, we show that $\hat{\theta}_{wa}$ has the following bias:

$$\text{Bias}(\hat{\theta}_{wa}) = E\left(\frac{E\left\{\frac{\hat{w}}{n}|\epsilon\right\}}{E\left\{\frac{\hat{w}}{n}\right\}}L_a(x) - L(p)\right) = H(p_t, q_t, p_c, q_c), \quad (8)$$

where $L_a(x) = \hat{\theta}_{wa}$, $L(p) = E_x(L_a(x))$, $H(p_{t|\epsilon}, q_{t|\epsilon}, p_c, q_c) = -\frac{(p_{t|\epsilon}-q_{t|\epsilon})}{n(q_{t|\epsilon}p_{t|\epsilon})}\{a + \frac{p_cq_c - p_{t|\epsilon}q_{t|\epsilon}}{2(p_{t|\epsilon}q_{t|\epsilon}+p_cq_c)}\} + \frac{(p_c-q_c)}{n(p_cq_c)}\{a + \frac{p_{t|\epsilon}q_{t|\epsilon}-p_cq_c}{2(p_{t|\epsilon}q_{t|\epsilon}+p_cq_c)}\}$, and $H(p_t, q_t, p_c, q_c) = E(H(p_{t|\epsilon}, q_{t|\epsilon}, p_c, q_c))$. We observe in Figure 1(b) and 1(c) that for studies of rare events, this bias is usually positive and is an increasing function of both τ^2 and a. As a result, the bias of the weighted average method does not vanish even after adding 1/2, as it does for the simple average method. It is a well-known result in linear models that the weighted average estimate (when weights are inversely proportional to variance) is the best linear unbiased estimate (BLUE) for the mean effect. However, the concept of the BLUE is not applicable in the current context as the weights are biased and correlated. Shuster (2010) strongly recommended using the simple average (unweighted) estimate for the random-effects model as the correlation between the weights and effect size produces serious bias to the weighted average estimate. The estimated variances of $\hat{\theta}_{s1/2}$ and $\hat{\theta}_{w1/2}$ are, respectively,

$$\hat{V}(\hat{\theta}_{s1/2}) = \frac{\sum_{i=1}^{k}\hat{\sigma}_i^2(\tau^2)}{k^2} \quad \text{and} \quad \hat{V}(\hat{\theta}_{w1/2}) = \frac{1}{\sum_{i=1}^{k}\hat{w}_i(\tau^2)}. \quad (9)$$

Note that $\hat{V}(\hat{\theta}_{s1/2}) \geq \hat{V}(\hat{\theta}_{w1/2})$ because the arithmetic mean is larger than the harmonic mean. However, this ordering of the variance estimates may not hold for the true variances. In our case, the study-specific estimate of θ and the estimate of its variance are statistically dependent; hence, $E(\hat{w}_i(\tau^2)\hat{\theta}_{i1/2}) \neq E(\hat{w}_i(\tau^2))E(\hat{\theta}_{i1/2})$ (as $\hat{p}_i$ and $\frac{1}{\hat{p}_i\hat{q}_i}$ are correlated). Consequently, the exact variance of the weighted average estimate is not $\frac{1}{\sum_{i=1}^{k}w_i(\tau^2)}$. Therefore, in such situations, not only is the weighted average estimate biased but also the superiority of the weighted average estimate in terms of its smaller variance compared with that of the simple average estimate is questionable.

2.1 Estimation of τ^2

In practice, τ^2 has primarily been estimated using the method of moments and likelihood-based methods. Following Cochran's (1954) Q statistics, DerSimonian and Laird (1986) proposed a moment-based estimator for τ^2 that is easy to compute and has been used extensively. Hardy and Thompson (1996) explored the DSL procedure in connection with constructing a confidence interval for θ for unknown τ^2. They concluded that the likelihood-based method outperforms the DSL method because it incorporates extra variability due to the estimation of τ^2. In the context of a mixed-effects linear meta-analysis model, Sidik and Jonkman (2007) compared seven different estimators of the heterogeneity in a simulation study. To obtain confidence intervals for the variance components, Viechtbauer (2007) compared those seven approaches with a new method and showed that the new method had the correct coverage probability. Alternatively, the I^2 statistic is used to quantify the impact of heterogeneity on the treatment effect (Higgins and Thompson 2002). From the inferential point of view, I^2 and Q have been shown to have similar performance (Huedo-Medina et al. 2006).

To illustrate, we use results from the previous section to find the DSL estimator of τ^2. We take the within-study estimates of variance and corresponding weights to be

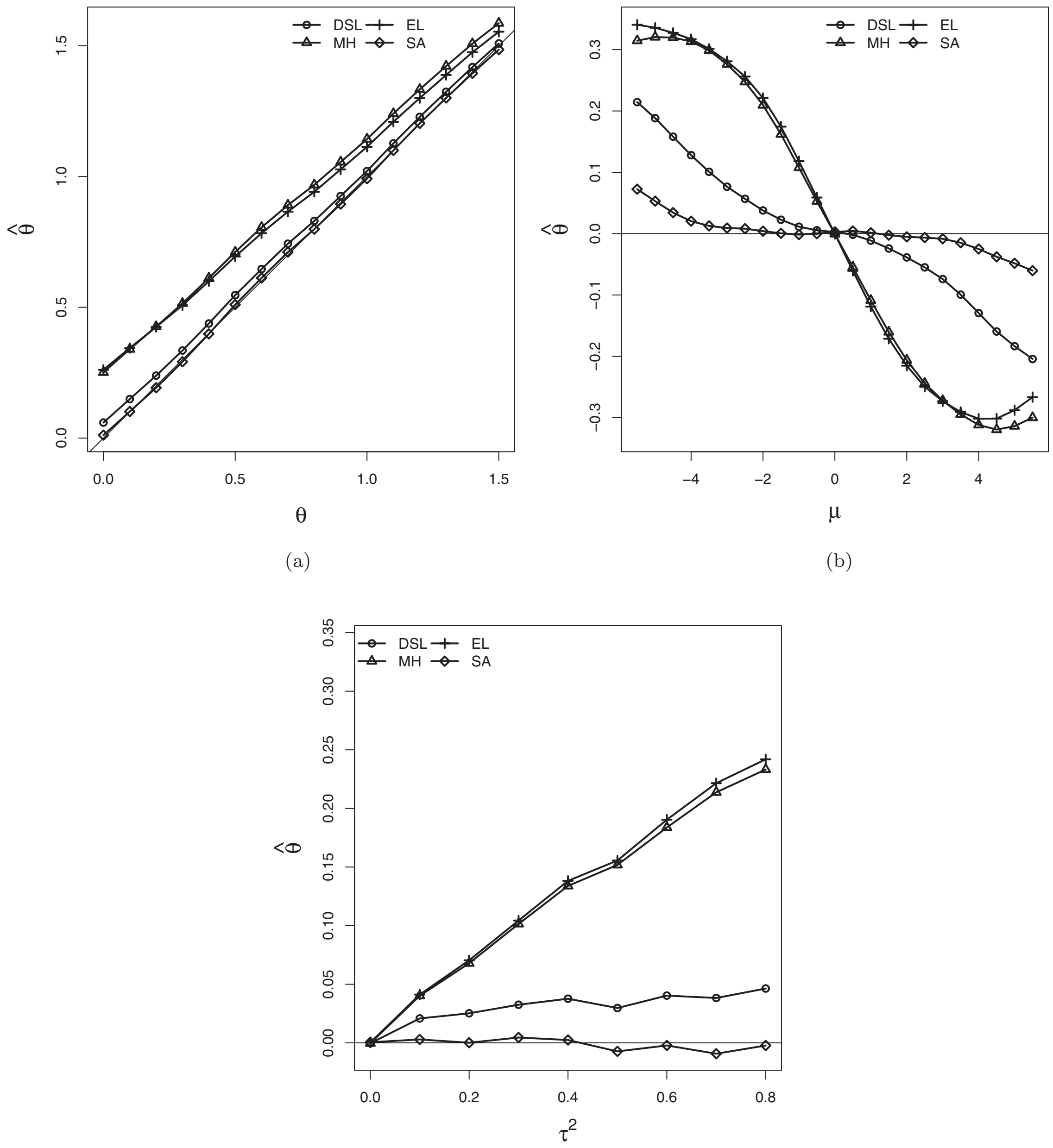
$$\hat{\sigma}_i^2(0) = \frac{1}{n_{it}\hat{p}_{it}\hat{q}_{it}} + \frac{1}{n_{ic}\hat{p}_{ic}\hat{q}_{ic}} \quad \text{and} \quad \hat{w}_i(0) = \frac{1}{\hat{\sigma}_i^2(0)}. \quad (10)$$

Let

$$Q = \sum_{i=1}^{k}\hat{w}_i(0)(\hat{\theta}_{i1/2} - \hat{\theta}_{w(0)1/2})^2, \quad (11)$$

where $\hat{\theta}_{w(0)1/2}$ is a weighted average estimate of θ defined in (7). Note that the weight function does not include the between-study variability. The DSL (1986) estimate of τ^2 (see Appendix C) is

$$\hat{\tau}^2_{\text{DSL}} = \frac{Q-(k-1)}{\sum_{i=1}^{k}\hat{w}_i(0) - \sum_{i=1}^{k}\hat{w}_i^2(0)/\sum_{i=1}^{k}\hat{w}_i(0)}. \quad (12)$$



(a)

(b)

(c)

Figure 1. Comparison of estimates of θ by DerSimonian and Laird (DSL), empirical logit (EL), Mantel and Henszel (MH), and simple average (SA) methods for (a) $\mu = -2.5$, $\tau^2 = 0.8$; (b) $\theta = 0$, $\tau^2 = 0.8$; and (c) $\mu = -2.5$, $\theta = 0$. Parameters: $n \sim U(50, 1000)$ and $k = 20$.

This estimator has two drawbacks. First, in the presence of the heterogeneity, $\hat{\tau}^2_{\text{DSL}}$ is a biased estimate (Malzahn, Bohning, and Holling 2000; Bohning et al. 2002; Sidik and Jonkman 2005). Second, the weights $\hat{w}_i(0)$ assume that $\tau^2 = 0$ and hence the between-study variability is not included in it. Incorporating estimates of the proper weights $\hat{w}_i^2(\tau^2)$, DerSimonian and Kacker (DSK) (2007) proposed a two-step procedure to estimate τ^2: (a) use the within-study weights described in Equation (10) to compute $\hat{\tau}^2_{\text{DSL}}$ as described in (12), (b) compute adjusted weights that incorporate the DSL estimate of between-study variance, $\hat{w}_i(\hat{\tau}^2_{\text{DSL}})$, and use these in Equations (10) and (12) to compute the adjusted estimate $\hat{\tau}^2_{\text{DSK}}$. Theoretical properties of this two-step method have not been well studied. We have numerically investigated the properties of τ^2_{DSK} and found that for rare events, this estimate has considerable downward bias.

Another procedure for estimating τ^2 was proposed by Paule and Mandel (PM) (1982) (τ^2_{PM}) and is based on the following estimating equation:

$$F(\tau^2) = \sum_i w_i(\tau^2)[\theta_i - \theta_{w(\tau^2)}]^2 - (k-1) = 0. \qquad (13)$$

A unique solution of Equation (13), τ^2_{PM}, can be determined by numerical iteration, starting with $\tau^2 = 0$. If $F(\tau^2)$ is negative for all positive τ^2, $\hat{\tau}^2_{\text{PM}}$ is set to 0. Note that $\hat{\tau}^2_{\text{PM}}$ is based on $\hat{\sigma}_i(0)$, which varies from study to study. The variance estimator may be improved by borrowing strength from all studies when estimating each within-study variance $\hat{\sigma}_i^2$:

$$\hat{\sigma}_i^2(*) = \frac{1}{(n_{it}+1)}\left[\exp\left(-\hat{\mu} - \hat{\theta}_{sa} + \frac{\tau^2}{2}\right) + 2 + \exp\left(\hat{\mu} + \hat{\theta}_{sa} + \frac{\tau^2}{2}\right)\right] + \frac{1}{(n_{ic}+1)} \times [\exp(-\hat{\mu}) + 2 + \exp(\hat{\mu})]. \tag{14}$$

We propose a new estimator, denoted $\hat{\tau}^2_{\mathrm{IPM}}$, that is the solution to a modified version of Equation (13), where the weights $w_i(\tau^2)$ are replaced by the shared-strength weights $w_i(*) = \frac{1}{\tau^2 + \hat{\sigma}_i^2(*)}$. We numerically compare the performance of τ^2_{PM} and τ^2_{IPM} in Section 4.

3. HYPOTHESIS TESTING

In addition to parameter estimation, an equally important problem in meta-analysis is hypothesis testing regarding θ and τ^2. A large sample test for θ using $\hat{\theta}_{w1/2}$ and test for τ^2 using a mixture of chi-square distributions are available (see Shapiro 1985; Self and Liang 1987; Hartung, Knapp, and Sinha 2008). In this section, we explore some tests for both of these parameters.

3.1 Testing the Treatment Effect

Hartung, Knapp, and Sinha (2008) constructed the following test statistic for testing the null hypothesis $H_0:\ \theta = 0$.

$$T_1 = \frac{\hat{\theta}_{\hat{w}_i(\hat{\tau}^2)} - 0}{\sqrt{\frac{1}{\sum_{i=1}^k \hat{w}_i(\hat{\tau}^2)}}}. \tag{15}$$

Even though asymptotically T_1 follows a t distribution with $k-1$ degrees of freedom (df), that is, $T_1 \sim t_{k-1}$, its small-sample property has not been well studied, particularly for moderate values of τ^2 (say $0.5 \leq \tau^2 \leq 1.5$). As the bias of $\hat{\theta}_{w1/2}$ is significant even for moderate τ^2, a natural question to ask is how good the performance of T_1 will be in terms of controlling Type I error rates for small-to-large within-study sample sizes ($n = 50, \ldots, 1000$). Based on the unbiased simple average estimate $\theta_{s1/2}$, we construct the following test statistic:

$$T_2 = \frac{k(\hat{\theta}_{s1/2} - 0)}{\sqrt{\sum_{i=1}^k \hat{\sigma}_i^2(\tau^2)}}. \tag{16}$$

In Section 4.2, we investigate the performance of T_1 and T_2 numerically and provide guidelines for application.

3.2 Testing the Heterogeneity Parameter

In practice, we use the likelihood ratio test or Wald's test to determine whether the heterogeneity parameter is zero, that is, $H_0 : \tau^2 = 0$. This puts the variance component on the boundary of the parametric space defined by the alternative hypothesis. Under this scenario, the limiting distribution of the likelihood ratio test statistic $-2\ln(\mathrm{LR})$ under the null hypothesis does not follow a χ^2 distribution. Shapiro (1985) derived the asymptotic distribution of $-2\ln(\mathrm{LR})$ as a mixture of χ^2 distributions when τ^2 falls on the boundary. Self and Liang (1987) generalized these results. The general conclusion is that standard tests for $H_0 : \tau^2 = 0$ are too conservative in terms of controlling Type I error rates (α) and exhibit inadequate power in the neighborhood of the null hypothesis. We propose two tests for τ^2. The first test is based on the Q statistic:

$$Q(\hat{\tau}^2) = \sum_{i=1}^k \hat{w}_i(\tau^2)[\hat{\theta}_{i1/2} - \hat{\theta}_{w(\tau^2)1/2}]^2.$$

As shown in Appendix D, an estimate of the variance of the Q statistic is

$$\widehat{\mathrm{var}}(Q(\hat{\tau}^2)) = 2\sum_{i=1}^k \hat{w}_i^2(\hat{\tau}^2)\left[\hat{\tau}^2 + \hat{\sigma}_i^2(0) + \frac{1}{\sum_{i=1}^k \hat{w}_i(\hat{\tau}^2)} - \frac{2\hat{w}_i(\hat{\tau}^2)(\hat{\tau}^2 + \hat{\sigma}_i^2(0))}{\sum_{i=1}^k \hat{w}_i(\hat{\tau}^2)}\right]^2. \tag{17}$$

Cochran (1950) showed that the asymptotic distribution of $Q(\hat{\tau}^2)$ is χ^2, with degrees of freedom $(k-1)$, which is a gamma distribution. Krishnamoorthy, Mathew, and Mukherjee (2008) and Bhaumik, Kapur, and Gibbons (2009) observed that the normal distribution approximation is better for a standardized log-transformed gamma distribution than for the standardized gamma distribution. Based on this result, we propose the following test for $H_0 : \tau^2 = 0$:

$$T_3 = \frac{(k-1)[\ln(Q(\hat{\tau}^2)) - \ln(k-1)]}{\sqrt{\widehat{\mathrm{var}}[\ln(\hat{Q}(0))]}}. \tag{18}$$

In Equation (18), the variance of $\ln(Q(\hat{\tau}^2))$ is computed using the delta method and $\widehat{\mathrm{var}}[\ln(Q(0))]$ is obtained by substituting $\hat{\tau}^2 = 0$ in the expression of $\widehat{\mathrm{var}}[\ln(Q(\hat{\tau}^2))]$. For many situations, T_3 performs extremely well in terms of controlling Type I error rates. Our simulation results show that for extremely rare events, this test has inflated Type I error rates. The simulated results are not reported here. Our second proposed test for $\tau^2 = 0$ is based on the simple average estimate of θ, which is specifically designed for rare events. Let $y_i = (\hat{\theta}_{i1/2} - \hat{\theta}_{s1/2})^2$. Under the null hypothesis, an estimate of the mean of y_i, denoted by B_i, is $\frac{k-2}{k}\hat{\sigma}_i^2(0) + \frac{\sum_{i=1}^k \hat{\sigma}_i^2(0)}{k^2}$. An estimate of the variance of y_i is $\widehat{\Sigma}_i$, where $\widehat{\Sigma}_i = 2B_i$. The second proposed test for $H_0 : \tau^2 = 0$ is

$$T_4 = \frac{\sum_{i=1}^k (y_i - B_i)}{\sqrt{\sum_{i=1}^k \widehat{\Sigma}_i}}. \tag{19}$$

Even though the asymptotic distribution of T_4 is standard normal, for small samples, its distribution is not known. In Section 4, we show that for finite samples this test is very conservative. To maintain the proper Type I error rate, we propose to determine the critical value using the PB technique, as follows:

(1) For a given dataset of sample size n, estimate θ and μ by the simple average method. For large n, $\hat{\theta}$ and $\hat{\mu}$ are consistent estimators for θ and μ. In the following, we replace $\hat{\theta}$ by θ and $\hat{\mu}$ by μ.

(2) Using $\hat{\theta}, \hat{\mu}$ (obtained from Step 1), and τ_0, generate k studies, each of size (n_t, n_c), from $B(n_t, p_t)$ and $B(n_c, p_c)$ for the treatment and control groups, respectively, under the null hypothesis.

(3) Compute y_i and B_i.

(4) Compute T_4 using Equation (19).

(5) Repeat Steps 2–4 for 10,000 times.

(6) Find the $100(1-\alpha)$th percentile point $T_4(\alpha)$ from the generated T_4's.
(7) Reject H_0 if $T_4 > T_4(\alpha)$, where T_4 is obtained from (19) for the original data.

We investigate the performance of T_4 in Section 4.

4. SIMULATION STUDY

In this section, we describe the results of a simulation study designed to compare several methods for evaluating (a) performance of moment-based estimates of θ, (b) the Type I error rates and power functions for testing θ, (c) performance of estimates of τ^2, and (d) the Type I error rates and power functions for testing τ^2. All the simulation studies were performed using the R software.

4.1 Performance of Estimates of θ

We compare the performance of the overall treatment effect estimate θ via simulation for four different estimation procedures: (a) simple average (SA), (b) DSL with estimated τ^2_{DSL}, (c) EL, and (d) MH. For this comparison, we set $\theta = 0, 0.5, 1.0, \ldots, 2.5$; $k = 20$; $\mu = -5.5, -5, \ldots, 0, \ldots, 5, 5.5$; $\sigma^2_\mu = 0.5$, and $\tau^2 = 0, 0.2, 0.4, 0.6, 0.8$. The number of subjects in each arm was chosen independently by rounding random draws from a uniform distribution with $\min(n) = 50$ and $\max(n) = 1000$, that is, n_t and $n_c \sim U(50, 1000)$. Next, we generated the responses x_{ic} for the control group based on a binomial distribution $B(n_{ic}, p_{ic})$ for $i = 1, \ldots, k$, where p_{ic} is computed as

$$\frac{\exp(\mu + \epsilon_1)}{1 + \exp(\mu + \epsilon_1)}.$$

The responses x_{it} for the treatment group were drawn from a binomial distribution $B(n_{it}, p_{it})$ for $i = 1, \ldots, k$, with

$$p_{it} = \frac{\exp(\mu + \epsilon_1 + \theta + \epsilon_2)}{1 + \exp(\mu + \epsilon_1 + \theta + \epsilon_2)},$$

$\epsilon_1 \sim N(0, 0.5)$, and $\epsilon_2 \sim N(0, \tau^2)$. Inclusion of ϵ_1 in p_{ic} and ϵ_2 in p_{it} implies that both the control and the treatment groups have varying rates of events. We have added 0.5 correction when necessary for EL, DSL, and MH, whereas for SA estimates, we have added 0.5 correction for all studies. We simulated 1000 replications for each combination of k, n, θ, μ, and τ^2.

Figure 1 compares various estimates of the overall treatment effect. Panel (a) reveals that the MH and EL methods overestimate the overall treatment effect. This figure also suggests that SA estimates are less biased for all values of the treatment effect under consideration. For moderate overall treatment effects, DSL and SA estimates are comparable. Panel (b) exhibits the effect of the background incidence rate in estimating the overall treatment effect. It is evident from this figure that MH, EL, and DSL estimates are biased when $\mu \neq 0$ (i.e., mean $(p_c) \neq 0.5$). The SA estimate, on the other hand, is almost unbiased for moderate background incidence rates (i.e., $-4 \leq \mu \leq 4$). Also, the bias of the SA estimate for rare-event cases is comparatively smaller. Panel (c) shows the effect of heterogeneity of treatment effects across studies on the estimate of the overall treatment effect. It is apparent that bias of the SA estimate in the presence of even significant heterogeneity is minimal compared with the other estimators (EL, DSL, and MH) under consideration.

In addition, we have used the "empirical continuity correction" suggested by Sweeting, Sutton, and Lambert (2004) for all moment-based estimators. We observe that moment-based estimates of θ with this continuity correction are more biased than 0.5 correction for rare events in the presence of noticeable heterogeneity. This is not surprising as Sweeting, Sutton, and Lambert (2004) have pointed out the nonapplicability of empirical continuity correction for random-effects models.

In this context, it is important to mention that theoretical properties of the simple average estimate of θ are derived under the fixed response rate in the control group but its properties are studied via simulation under both fixed (not reported here) and random response rates in the control group. In both cases, we observe that the simple average estimate performs better than MH, EL, and DSL.

4.2 Performance of Tests for Treatment Effect

In Section 3, we discussed the general testing procedure for $H_0 : \theta = 0$ against the alternative $H_1 : \theta \neq 0$. In this section, we use a simulation study to compare the performance of three tests: (a) z-test based on the EL estimate of θ in Equation (15), that is, T_1, with $\hat{\tau}^2 = 0$; (b) z-test based on the DSL estimate of θ, including τ^2 in (15), that is, T_1, with $\hat{\tau}^2 = \hat{\tau}^2_{\text{DSK}}$; and (c) z-test based on the simple average estimate of θ, that is, T_2, with $\hat{\tau}^2 = \hat{\tau}^2_{\text{IPM}}$. Our Monte Carlo simulation is based on 10,000 replications, with $\tau^2 = 0.8$, $k = 20$, and a nominal significance rate of $\alpha = 0.05$. Figure 2 graphically presents the results of our simulation study.

Figure 2(a) displays the estimated significance level of each test when the underlying incidence rate is varied, and when the null hypothesis (i.e., $\theta = 0$) is true. Although DSL performs quite well when μ is in the neighborhood of zero, the Type I error of DSL is inflated (to almost 20%) for extreme values of μ. Type I error rates of EL and MH are highly inflated for all values of μ. On the other hand, the SA-based test has Type I error rates close to the nominal level regardless of the background incidence rates.

Next, we numerically study the power functions of these tests by varying $\theta = 0, 0.1, \ldots, 1.5$, $\mu = -2.5$ (i.e., mean $(p_c) = 0.08$), and $n_c, n_t \sim U(50, 1000)$. We present these results in Figure 2(b). The SA test demonstrates a desirable power curve that controls the Type I error rate at the nominal level and grows to the power close to 1 for an effect size close to 0.6. On the contrary, power curves of EL and MH are overlapping in the figure. Both EL and MH have highly inflated Type I error rates (close to 1) and hence naturally have deceptively high power.

4.3 Performance of Estimates of the Heterogeneity Parameter τ^2

To study the performance of estimates of τ^2, we simulate datasets following the same procedure described in the first paragraph of Section 4.1. For this simulation, values of τ^2 are set between 0 and 1.2. To demonstrate the effect of rare events as well as prevalent events on the estimates of τ^2, the values of μ are varied from –5.5 (0.4%) to 5.5 (99%). For this comparison, we set θ at 0 and the number of studies were set to 20. Sample sizes in each treatment arm were drawn from $U(50, 1000)$. Figure 3(a) shows that for extremely rare (or prevalent) cases, moment-based estimates by PM and DSK fail to

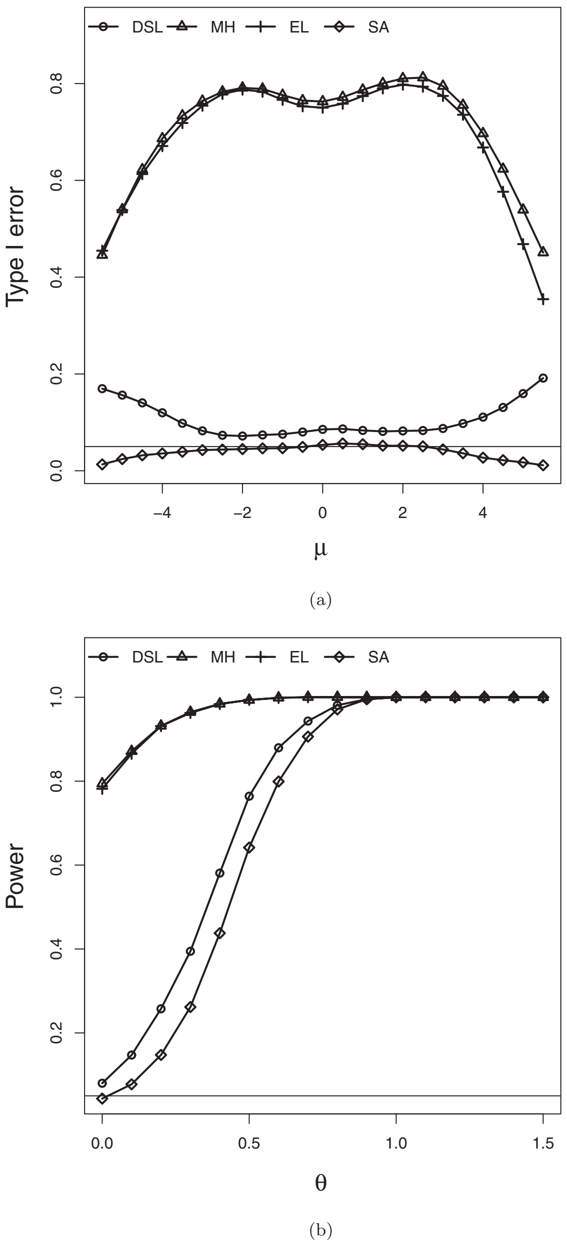


Figure 2. Comparison of Type I error rates of DerSimonian and Laird (DSL), empirical logit (EL), Mantel and Henszel (MH), and simple average (SA) methods for testing (a) $\theta = 0$, and (b) $\mu = -2.5$ at various background rates. Parameters: $\tau^2 = 0.8$, $n \sim U(50, 1000)$, and $k = 20$.

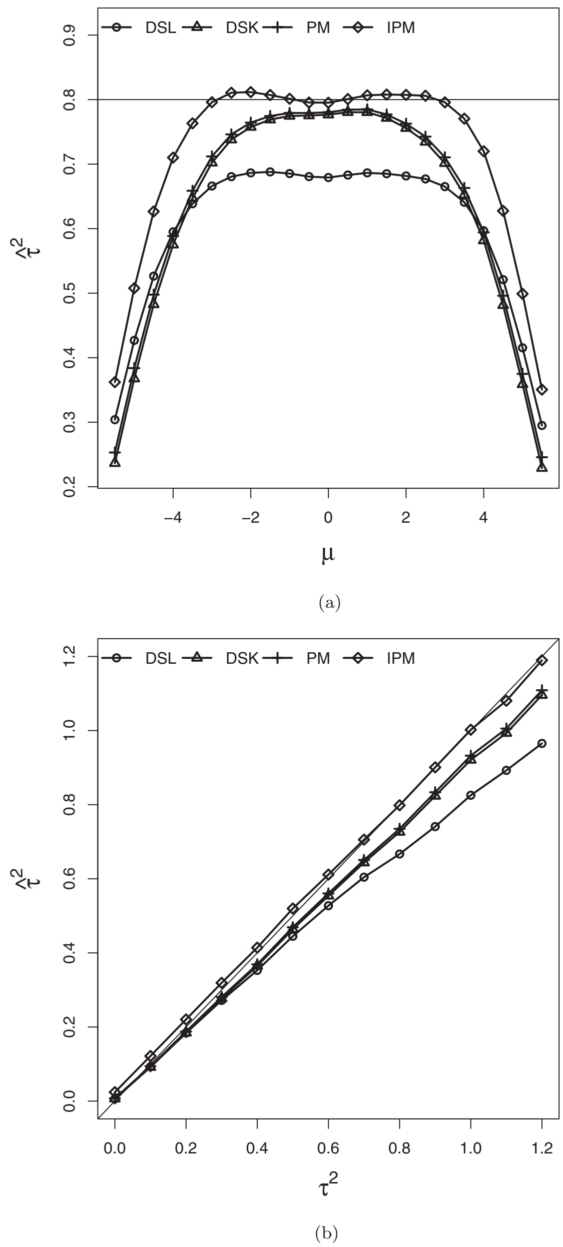


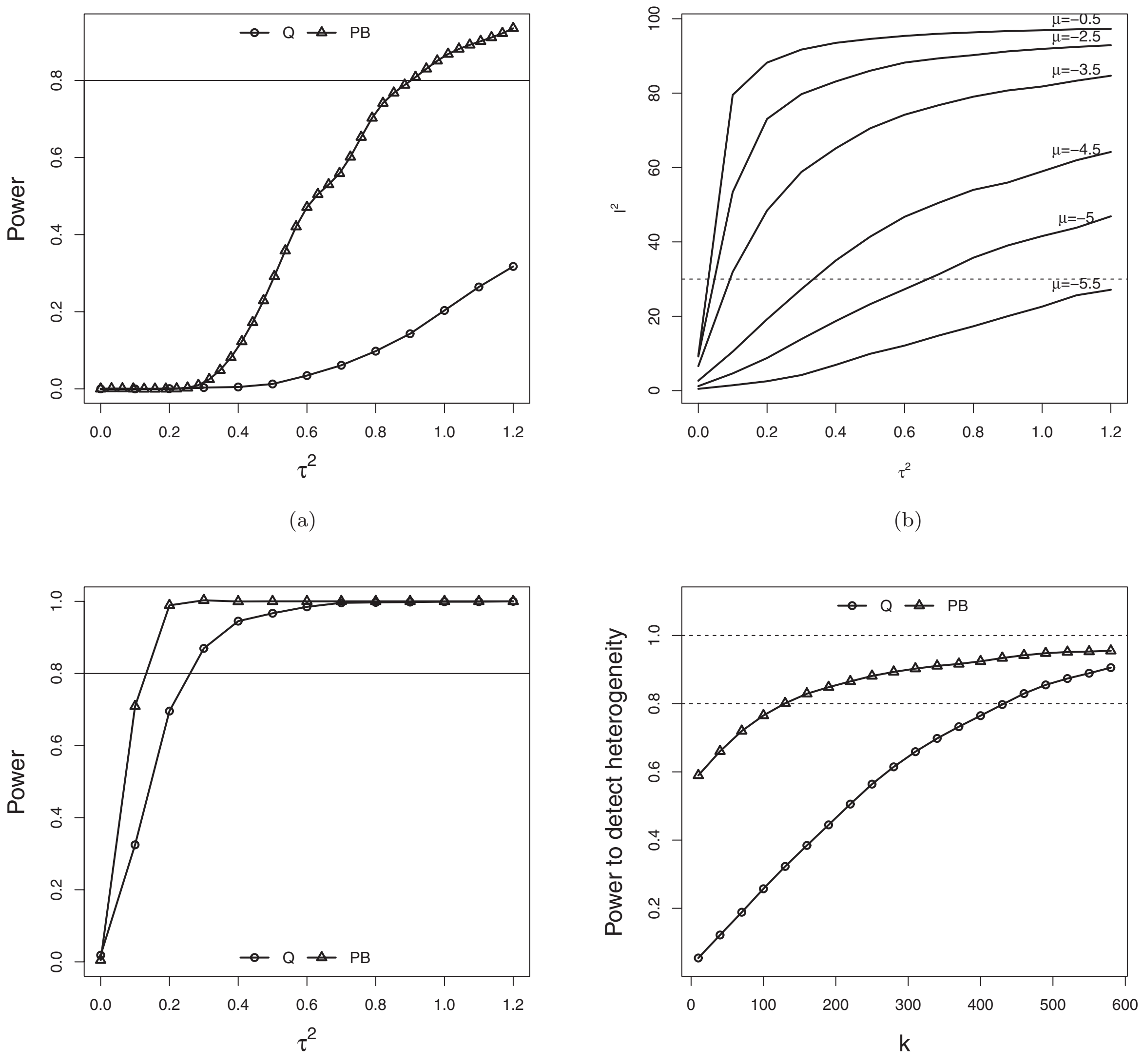
Figure 3. Comparison of estimates of τ^2 by DerSimonian and Laird (DSL), DerSimonian and Kacker (DSK), Paul and Mandel (PM), and Improved Paul and Mandel (IPM) methods for various values of (a) background rates $\tau^2 = 0.8$; and (b) treatment heterogeneity $\mu = -2.5$. Parameters: $\theta = 0.8$, $n \sim U(50, 1000)$, and $k = 20$.

detect the presence of heterogeneity ($\tau^2 = 0.8$) in the treatment effect across studies. This figure also illustrates the reduction of bias achieved by the IPM procedure compared with DSL, DSK, and PM for $-4 \leq \mu \leq 4$. Figure 3(b) shows that the IPM approach performs the best among all four methods for all values of $0 < \tau^2 \leq 1.2$ when μ is set at -2.5 (i.e., a moderate background rate of 7.5%).

4.4 Performance of Tests of the Heterogeneity Parameter

In this section, we compare the performance of Cochran's Q and PB statistics for testing $H_0 : \tau^2 = 0$ against $H_1 : \tau^2 > 0$. For the current simulation, we follow the same parametric conditions as described in Section 4.1. For this study, we set $\mu = -4.5$ and -2.5, along with $\theta = 0$, $k = 20$, and $n \sim U(50, 1000)$. We present the power curves in Figure 4. Panels (a) and (c) of Figure 4 reveal that Cochran's Q has lower power compared with PB for both event rates, -4.5 and -2.5. It is clear from the comparison of the figures in the two panels that for moderate event rates (i.e., $\mu = -2.5$), the performances of both the tests have improved significantly. In Figure 4(a), we see that for a rare-event case ($\mu = -4.5$), the power of$Q$ is extremely poor. The new test, PB, also performs poorly for smaller values of τ^2, but for $\tau^2 > 0.6$, it performs far better than Q.

For $\mu = -4.5$, $\tau^2 = 0.6$ plays an important role, as for this value of τ^2, the proportion of between-study variance (I^2) exceeds 40% of the total variance [see Figure 4(b)]. Higgins and Thompson (2002) suggested that any value of I^2 less than 30% indicates only for a mild heterogeneity. Following their tentative rule, we can infer that $\tau^2 \leq 0.6$ indicates only mild



(a)

(b)

(c)

(d)

Figure 4. Comparison of power curves of Q-test (Q) and parametric bootstrapping (PB) for testing heterogeneity at (a) $\mu = -4.5$; (c) $\mu = -2.5$; (d) $\tau^2 = 0.6$, $\mu = -4.5$; and (b) comparison of I^2 statistics for different background rates at different values of treatment heterogeneity. Parameters: $\theta = 0$, $n \sim U(50, 1000)$, and $k = 20$.

heterogeneity [see Figure 4(b)]. For a significant heterogeneity (when $I^2 \geq 50\%$, or in the current context, $\tau^2 \geq 0.6$), PB outperforms Q in terms of power [see Figure 4(a)]. In Figure 4(d), we see that the power of each test depends on the number of studies included in the meta-analysis. The power curve of Q grows with a slower rate compared with that of PB. To attain 80% power, PB requires about 125 studies, whereas for the same power, Q needs almost 450 studies.

4.5 Correlated Arms

In Figure 1, we observed that all methods overestimated the treatment effect. To give a definite answer to the question when moment-based estimates will underestimate the treatment effect, we perform a simulation study with correlated arms (i.e., we assume that ϵ_1 and ϵ_2 are correlated and the data were generated accordingly for this simulation study). Figure 5 shows that usual moment-based methods run into even more problems when the event rates of the control and treatment are correlated. For larger negative correlations (i.e., $\rho < -0.5$), treatment shows a protective effect, and for larger positive correlations (i.e., $\rho < -0.5$), it shows a harmful effect when the generating parameters are under the null. The performance of the simple average is better compared with its counterparts DSL, MH, and EL even for correlated arms. We have extended our simulation study for a wide range of θ ($-.4 \leq \theta \leq 1.0$) and observed the same pattern. For negative correlations, θ is underestimated, and for positive correlations, it is overestimated, *on average*. The SA method always has better performance.

4.6 Some Results for Extremely Rare Events

As we have mentioned earlier in Section 2, a special statistical problem arises when the focus of research synthesis is on a rare binary event. Our simulation studies show that moment-based estimates have undesirable statistical properties when the event is very rare or very frequent. We noted in Figure 1(b) that the bias of the overall treatment effect estimate is attenuated for very rare cases. One explanation for this undesirable behavior is that when events are rare, estimates and inferences are unduly influenced

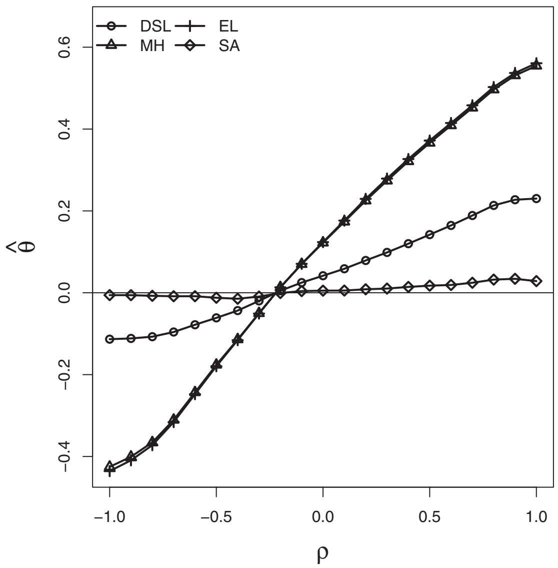


Figure 5. Estimate of θ by DerSimonian and Laird (DSL), empirical logit (EL), Mantel and Henszel (MH), and simple average (SA) methods for correlated background rates and treatment effects. Parameters: $\mu = -2.77$, $\theta = 0$, $\sigma_{\mu}^2 = 1.2$, $\tau^2 = 0.45$, $n \sim U(50, 1000)$, and $k = 17$.

by the factor used for continuity correction. One way to avoid this is to conduct larger studies, each with more samples. Our simulation results conducted for rare events (mean $(p_c) = 2/1000$) show that the moment-based estimates produce significant biases for sample sizes smaller than 400, and only the SA estimate converges asymptotically to the true value. The DSL estimate shows very weak convergence, and MH and EL estimates fail to converge to the true value even for a sample size as large as 2400 in each arm. Our simulation also shows that for rare-event studies, all available estimating methods fail to adequately estimate the heterogeneity parameter for random treatment effects. However, as we have seen in Figure 1(c) and 2(a) for fixed-effects models, biases and Type I error rates can be severely affected, depending on the magnitude of the true heterogeneity. Our simulation shows that for rare events, the PB method outperforms the Q statistic. When the number of studies are fairly large ($k \geq 40$), our proposed tests for the treatment effect maintain nominal Type 1 error rates and provide adequate power.

Zero studies: For extremely rare events, a large proportion of studies tends to have zero events in both arms. It has been argued by various authors that the zero studies do not contribute to the odds ratio estimation and hence should not be included in the analysis. The contrasting argument in support of inclusion of those studies to take an advantage is also found in the literature (Whitehead and Whitehead 1991). We have performed an extensive simulation study to examine the effects of inclusion (with 0.5 correction) and exclusion of zero studies in the estimation of both the treatment effect parameter (θ) and the heterogeneity parameter (τ^2). We find that as the proportion of zero studies increases, the discrepancy in the estimation of θ also increases. It reveals that estimates with the inclusion of zero cells have smaller bias compared with estimates without the inclusion of zero cells. The SA estimator is least affected by the proportion of zero studies. It also reveals that the exclusion of zero studies improves our ability to estimate τ^2. In summary, we have two contradictory results. First, the inclusion of zero studies with 0.5 continuity correction helps us in estimating θ more accurately. Second, an exclusion of zero studies is helpful in estimating τ^2. Therefore, two separate strategies are needed to be implemented when estimating these two parameters.

5. ILLUSTRATION

Coronary artery disease is the single-largest killer of American men and women (Rosamond et al. 2007). In 2004, in the U.S., there were 840,000 cases discharged with the diagnosis of acute coronary syndrome, most of them with acute myocardial infarction (MI). Percutaneous coronary intervention or PCI (commonly known as angioplasty) is increasingly being used in patients with various manifestations of coronary artery disease. PCI is an established treatment strategy that improves overall survival and survival time free of recurrent MI for patients with acute coronary disease; however, less is known about the effects of PCI in the treatment of patients with stable coronary artery disease. Research in this area has been limited for two reasons. First, patients with stable coronary artery disease have a very good prognosis and large-sample size studies are required to assess potential differences in treatments regarding rare events (Rihal et al. 2003; Timmis, Feder, and Hemingway 2007). Second, there is a period of early risk associated with PCI, which requires longer follow-up periods when compared with MED to offset this early excess risk.

In an effort to better study the efficacy of PCI versus MED, Schomig et al. (2008) conducted a meta-analysis of 17 randomized clinical trials (RCTs) that compared PCI to MED in patients with stable coronary artery disease. They studied a total of 7513 patients (3675 PCI and 3838 MED). Overall, the average age of the patients was 60 years, 18% of them were women, 54% had incurred MI, and the average length of follow-up was 51 months. Ninety-two percent of the patients in the PCI-based strategy group received revascularization (43% balloon angioplasty, 41% stents, and 8% coronary artery bypass graft [CABG]). In the PCI group, 271 patients died, and in the MED group, 335 died. Among the 13 studies that reported cardiac mortality, the PCI and MED arms had a combined 115/2814 and 151/2805 cardiac deaths, respectively. Finally, MI rates were provided in all 17 studies, reporting 319 in the PCI group and 357 in the MED group. The original authors used Cochran's Q test to assess heterogeneity and performed MH and DSL methods to estimate the overall treatment effect. The test of heterogeneity (Cochran's Q statistic) was not significant for total mortality ($p = 0.263$) or cardiac mortality ($p = 0.161$), but was significant for MI ($p = 0.003$). The fixed-effects (MH) model showed a significant protective effect of PCI on total mortality [odds ratio (OR) = 0.80, confidence interval (CI) = 0.68–0.95] and cardiac mortality (OR = 0.74, CI = 0.57–0.96), and approached significance for MI (OR = 0.91, CI = 0.77–1.06). The random-effect (DSL) model showed a significant protective effect of PCI on total mortality (OR = 0.80, CI = 0.64–0.99), an effect that approached significance for cardiac mortality (OR = 0.74, CI = 0.51–1.06), and a nonsignificant effect for MI (OR = 0.90, CI = 0.66–1.23), which was in the same protective direction as the other effects. On the basis of these results, the authors concluded that a PCI-based invasive strategy may improve long-term survival compared with a medical treatment in patients with stable coronary artery disease.

Table 1. Analysis of PCI versus MED for myocardial infarction data

Estimate	$\hat{\theta}$	Std. error	p-value
SA	0.284	0.482	0.221[a]
MH	–0.096	0.080	0.230[b]
DSL	–0.106	0.155	0.493[b]

[a](T_2) was used to perform the test.
[b](T_1) was used to perform the test.

To illustrate the performance of the moment-based approaches, we reanalyzed these data for all three outcomes (total mortality, cardiac mortality, and MI). For MI, there are 17 studies, with an average of 200 subjects per arm. Table 1 presents the estimates of the treatment effects, and Table 2 presents the corresponding heterogeneity parameter estimates. Inspection of Table 2 reveals the significance of the heterogeneity. The IPM provides a significantly larger estimate of the heterogeneity parameter. The SA estimate of the treatment effect is positive in contrast with the other estimators. All estimators (SA, MH, and DSL) indicate that PCI has a nonsignificant effect for MI. We note the reversal of bias in this example when compared with the figure presented earlier (Figure 1). Our simulation shows (in Figure 5) that such reversal occurs when there is a strong negative correlation between background event rates and treatment effects across studies.

For the cardiac death data, none of the methods (DSL, DSK, and IPM) found significant heterogeneity, consistent with the findings of Schomig et al. (2008). Estimates of heterogeneity by DSL, DSK, and IPM for the all-cause mortality data were also nonsignificant. Hence, to estimate the treatment effect for the cardiac death and all-cause mortality data, our recommendation is to use the MH method using the continuity correction proposed by Sweeting, Sutton, and Lambert (2004). Our analysis by MH method with empirical continuity correction shows nonsignificant protective effects of PCI on both total mortality and cardiac mortality.

6. DISCUSSION

It is with some trepidation that we present our findings on the limitations of the most commonly used methods for research synthesis of rare events. These methods (MH and DSL) have routinely been used for decades to advise physicians of the best evidence-based practice (e.g., Cochrane Reviews: *http://www.cochrane.org/*), and to identify potential adverse reactions of pharmaceuticals. For example, the U.S. Food and Drug Administration (FDA) used the MH test to perform an analysis of the risk of suicidal thoughts and behaviors associated with antidepressant medications in children, which led to a black box warning that is now present on every antidepressant medication and was further extended to young adults (*http://www.fda.gov/Drugs/DrugSafety/InformationbyDrugClass/ucm096352*). Our findings reveal that these methods, in particular, and moment-based methods, in general, can be quite limited in their ability to detect heterogeneity in treatment effect, and in the presence of such heterogeneity, can yield biased estimates of overall treatment effects and corresponding tests of hypotheses. In some cases, these biases can be large enough to even change the direction of the overall treatment effect. Furthermore, our simulations indicate that sample size requirements for very rare outcomes are enormous and generally require hundreds of studies, each with hundreds of patients per treatment arm. In practice, such studies are rarely of sufficient size or number to provide anything close to these requirements. Finally, the need to discard studies with zero events in both arms and/or impute a constant for studies with zero events in a single arm further limits our ability to estimate and test heterogeneity, which in turn biases estimation and testing of the overall treatment effect.

Table 2. Analysis of PCI versus MED for myocardial infarction data

Estimate	$\hat{\tau}^2$	p-value
DSL	0.168	0.003[a]
DSK	0.175	0.003[a]
IPM	0.358	< 0.05[b]

[a]Cochran's Q was used to perform the test.
[b]Parametric bootstrapping (T_4) was used to perform the test.

In summary, research synthesis of rare binary event data appears to be more complicated than traditional meta-analysis for continuous outcomes where more traditional effect size estimates are available from a series of studies. The nonlinear form of the models produces more complicated relationships between the overall average treatment effect and its variance than for the case of meta-analysis based on linear models. Bias of moment-based estimates is a complicated function of the degree to which the treatment effect varies across studies and the volume of data analyzed. Furthermore, there are a number of different moment-based approaches for estimating the combined treatment effect and heterogeneity, and depending on the combination of the above factors, some work better than others.

In the absence of heterogeneity, the MH method with the empirical continuity correction performs well, and is to be recommended for moment-based fixed-effects meta-analysis. The three new methods developed in this article (SA for estimating the overall treatment effect, IPM for estimating heterogeneity, and PB for testing heterogeneity) perform reasonably well. We recommend the SA with the 0.5 continuity correction for sparse data with heterogeneity. To estimate the heterogeneity parameter, our recommendation is to use the IPM. We recommend the PB for testing the heterogeneity parameter. Finally, it should be noted that we have not considered full-likelihood approaches to the problem of parameter estimation and hypothesis testing in connection with random-effects meta-analysis. Research along these lines is currently underway.

APPENDIX A: EXPECTATION OF $\hat{\theta}_{ia}$

$$\ln\left(\frac{p_{it|\epsilon_i}}{1-p_{it|\epsilon_i}}\right) = \mu + \theta + \epsilon_i,$$

$$\mu = \ln\left(\frac{p_{ic}}{1-p_{ic}}\right),$$

$$\epsilon_i \sim N(0, \tau^2).$$

From Gart, Hugh, and Thomas (1985),

$$E_{x|\epsilon}\left\{\ln\left(\frac{x_{it}+a}{n_{it}-x_{it}+a}\right)\right\} = \ln\left(\frac{p_{it|\epsilon_i}}{q_{it|\epsilon_i}}\right) + \frac{(p_{it|\epsilon_i}-q_{it|\epsilon_i})\left(\frac{1}{2}-a\right)}{n_{it}\,p_{it|\epsilon_i}\,q_{it|\epsilon_i}} + O(n^{-2}).$$

Then,

$$E_x\left\{\ln\left(\frac{x_{ic}+a}{n_{ic}-x_{ic}+a}\right)\right\}=\ln\left(\frac{p_{ic}}{q_{ic}}\right)+\frac{(p_{ic}-q_{ic})\left(\frac{1}{2}-a\right)}{n_{ic}p_{ic}q_{ic}}+O(n^{-2}).$$

Thus,

$$\begin{aligned}
E[\hat{\theta}_{ia}] &= E_\epsilon[E_{x|\epsilon}[\hat{\theta}_i]]\\
&= E_\epsilon\left\{\ln\left(\frac{p_{it|\epsilon_i}}{q_{it|\epsilon_i}}\right)\right\}+\frac{\left(\frac{1}{2}-a\right)}{n_{it}}E_\epsilon\left\{\frac{1}{q_{it|\epsilon_i}}-\frac{1}{p_{it|\epsilon_i}}\right\}+O(n^{-2})\\
&\quad -E_\epsilon\left\{\ln\left(\frac{p_{ic}}{q_{ic}}\right)\right\}-\frac{\left(\frac{1}{2}-a\right)}{n_{ic}}E_\epsilon\left\{\frac{1}{q_{ic}}-\frac{1}{p_{ic}}\right\}+O(n^{-2})\\
&= E_\epsilon\left\{\mu+\theta+\epsilon_i\right\}+\frac{\left(\frac{1}{2}-a\right)}{n_{it}}E_\epsilon\left\{\frac{1}{q_{it|\epsilon_i}}-\frac{1}{p_{it|\epsilon_i}}\right\}-\mu\\
&\quad -\frac{\left(\frac{1}{2}-a\right)}{n_{ic}}\left\{\frac{1}{q_{ic}}-\frac{1}{p_{ic}}\right\}+O(n^{-2})\\
&= \theta+\frac{\left(\frac{1}{2}-a\right)}{n_{it}}E_\epsilon\left\{1+e^{\mu+\theta+\epsilon_i}-1-e^{-\mu-\theta-\epsilon_i}\right\}\\
&\quad -\frac{\left(\frac{1}{2}-a\right)}{n_{ic}}\{e^{\mu}-e^{-\mu}\}+O(n^{-2})\\
&= \theta+\frac{\left(\frac{1}{2}-a\right)}{n_{it}}\left\{e^{\mu+\theta+\tau^2/2}-e^{-\mu-\theta+\tau^2/2}\right\}\\
&\quad -\frac{\left(\frac{1}{2}-a\right)}{n_{ic}}\{e^{\mu}-e^{-\mu}\}+O(n^{-2}).
\end{aligned}$$

$$\begin{aligned}
\text{Bias of }\hat{\theta}_{ia} &= \frac{\left(\frac{1}{2}-a\right)}{n_{it}}\left\{e^{\mu+\theta+\tau^2/2}-e^{-\mu-\theta+\tau^2/2}\right\}\\
&\quad -\frac{\left(\frac{1}{2}-a\right)}{n_{ic}}\{e^{\mu}-e^{-\mu}\}+O(n^{-2}).
\end{aligned}$$

APPENDIX B: VARIANCE OF $\hat{\theta}_{i\frac{1}{2}}$

$$\begin{aligned}
V(\hat{\theta}_{ia}) &= E_\epsilon V(\hat{\theta}_{ia}|\epsilon)+V_\epsilon E(\hat{\theta}_{ia}|\epsilon)\\
&= E_\epsilon\left\{\frac{1}{n_{it}p_{it|\epsilon_i}q_{it|\epsilon_i}}+\frac{1}{n_{ic}p_{ic}q_{ic}}\right\}+V_\epsilon\left\{E_{x|\epsilon}\left(\ln\frac{\hat{p}_{it|\epsilon_i}}{\hat{q}_{it|\epsilon_i}}-\ln\frac{\hat{p}_{ic|\epsilon_i}}{\hat{q}_{ic|\epsilon_i}}\right)\right\}\\
&= E_\epsilon\left\{\frac{1}{n_{it}p_{it|\epsilon_i}q_{it|\epsilon_i}}+\frac{1}{n_{ic}p_{ic}q_{ic}}\right\}+V_\epsilon\left\{\ln\frac{p_{it|\epsilon_i}}{q_{it|\epsilon_i}}-\ln\frac{p_{ic}}{q_{ic}}+\text{const}\right\},
\end{aligned}$$

where, const$=\frac{(\frac{1}{2}-a)}{n_{it}}\{\frac{1}{q_{it|\epsilon_i}}-\frac{1}{p_{it|\epsilon_i}}\}-\mu-\frac{(\frac{1}{2}-a)}{n_{ic}}\{\frac{1}{q_{ic}}-\frac{1}{p_{ic}}\}+O(n^{-2})$. Then,

$$\begin{aligned}
V(\hat{\theta}_{i\frac{1}{2}}) &= \frac{1}{n_{it}}E_\epsilon\left\{e^{\mu+\theta+\epsilon_i}+e^{-\mu-\theta-\epsilon_i}+2\right\}+\frac{1}{n_{ic}}\{e^{\mu}+e^{-\mu}+2\}\\
&\quad +V_\epsilon\left\{\ln\frac{p_{it|\epsilon_i}}{q_{it|\epsilon_i}}-\ln\frac{p_{ic}}{q_{ic}}\right\}\\
&= \frac{1}{n_{it}}\left\{e^{\mu+\theta+\tau^2/2}+e^{-\mu-\theta+\tau^2/2}\right\}+\frac{1}{n_{ic}}\{e^{\mu}+e^{-\mu}\}+\frac{2}{n_{it}}\\
&\quad +\frac{1}{n_{ic}}+V_\epsilon(\theta+\epsilon_i)\\
&= \frac{1}{n_{it}}\left\{e^{\mu+\theta+\tau^2/2}+e^{-\mu-\theta+\tau^2/2}\right\}+\frac{1}{n_{ic}}\{e^{\mu}+e^{-\mu}\}\\
&\quad +\frac{2}{n_{it}}+\frac{2}{n_{ic}}+\tau^2.
\end{aligned}$$

APPENDIX C: EXPRESSION OF τ^2

Let, $\hat{\theta}_{ia}\sim\mathcal{N}\left(\theta,\tau^2+\sigma_i^2\right)$,

$$\tilde{\theta}=\frac{\sum_{i=1}^k\hat{w}_i\hat{\theta}_{ia}}{\sum_{i=1}^k\hat{w}_i},\quad \hat{w}_i=\frac{1}{\hat{\sigma}_i^2};\text{ note that }\hat{w}_i\xrightarrow{p}w_i.$$

Then, $V(\tilde{\theta})=\dfrac{\sum_{i=1}^k w_i^2(\tau^2+\sigma_i^2)}{\left(\sum_{i=1}^k w_i\right)^2}=\tau^2\dfrac{\sum_{i=1}^k w_i^2}{\left(\sum_{i=1}^k w_i\right)^2}+\dfrac{1}{\sum_{i=1}^k w_i}$.

$$Q=\sum_{i=1}^k w_i(\hat{\theta}_{ia}-\tilde{\theta})^2.$$

$$\begin{aligned}
E[Q] &= \sum_{i=1}^k w_iE(\hat{\theta}_{ia}-\tilde{\theta})^2\\
&= \sum_{i=1}^k w_iV(\hat{\theta}_{ia}-\tilde{\theta})\\
&= \sum_{i=1}^k w_i[V(\hat{\theta}_{ia})+V(\tilde{\theta})-2\text{Cov}(\hat{\theta}_{ia},\tilde{\theta})]\\
&= \sum_{i=1}^k w_i\left[\tau^2+\sigma_i^2+\tau^2\frac{\sum_{i=1}^k w_i^2}{\left(\sum_{i=1}^k w_i\right)^2}+\frac{1}{\sum_{i=1}^k w_i}-2\text{cov}(\hat{\theta}_{ia},\tilde{\theta})\right]\\
&= \sum_{i=1}^k w_i\left[\tau^2+\sigma_i^2+\tau^2\frac{\sum_{i=1}^k w_i^2}{\left(\sum_{i=1}^k w_i\right)^2}+\frac{1}{\sum_{i=1}^k w_i}-2\frac{w_i(\tau^2+\sigma_i^2)}{\sum_{i=1}^k w_i}\right]\\
&= \tau^2\sum_{i=1}^k w_i-\tau^2\frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i}+(k-1).
\end{aligned}$$

Thus,

$$\tau^2=\frac{E[Q]-(k-1)}{\sum_{i=1}^k w_i-\frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i}}.$$

The DSL estimate of τ^2, denoted by τ^2_{DSL}, has the following expression:

$$\tau^2_{\text{DSL}}=\frac{\hat{Q}-(k-1)}{\sum_{i=1}^k\hat{w}_i-\frac{\sum_{i=1}^k\hat{w}_i^2}{\sum_{i=1}^k\hat{w}_i}}.$$

APPENDIX D: ESTIMATE OF $V[\widehat{Q(\tau^2)}]$

$$Q(\tau^2)=\sum_{i=1}^k w_i(\tau^2)(\hat{\theta}_{ia}-\tilde{\theta}_{w(\tau^2)})^2,\quad w_i(\tau^2)=\frac{1}{\tau^2+\sigma_i^2},$$

$$V[Q(\tau^2)]=\sum_{i=1}^k[w_i(\tau^2)]^2V(\hat{\theta}_{ia}-\tilde{\theta}_{w(\tau^2)})^2.$$

Since $(\hat{\theta}_{ia}-\tilde{\theta}_{wa(\tau^2)})\sim\mathcal{N}(0,A_i^2)$, where $A_i^2=E[\hat{\theta}_{ia}-\tilde{\theta}_{w(\tau^2)}]^2=(\tau^2+\sigma_i^2)(1-2\frac{w_i(\tau^2)}{[\sum_{i=1}^k w_i(\tau^2)]^2})+\frac{1}{\sum_{i=1}^k w_i(\tau^2)}$ and $E[\hat{\theta}_{ia}-\tilde{\theta}_{w(\tau^2)}]^4=3A^4$ (using the recursive relation of moment of a normal distribution), then,

$$\begin{aligned}
V[Q(\tau^2)] &= \sum_{i=1}^k[w_i(\tau^2)]^2\left\{E(\hat{\theta}_{ia}-\tilde{\theta}_{w(\tau^2)})^4-[E(\hat{\theta}_{ia}-\tilde{\theta}_{w(\tau^2)})^2]^2\right\}\\
&= \sum_{i=1}^k[w_i(\tau^2)]^2\left\{3A_i^4-\left[A_i^2\right]^2\right\}\\
&= \sum_{i=1}^k[w_i(\tau^2)]^2\ \left(2A_i^4\right).
\end{aligned}$$

Hence, $V[\widehat{Q(\tau^2)}]=\sum_{i=1}^k[\hat{w}(\tau^2)_i]^2\ \left(2\hat{A}_i^4\right)$.

APPENDIX E: BIAS OF $\hat{\theta}_{wi}$

Recall that $\hat{\upsilon}_t = \frac{(x_t+a)(n_t-x_t+a)}{n_t+2a}$ and $\hat{\upsilon}_c = \frac{(x_c+a)(n_c-x_c+a)}{n_c+2a}$, where $\hat{\upsilon}_t$ and $\hat{\upsilon}_c$ are the estimates variance of log odds of the treatment and the control groups, respectively. Then, $\hat{w} = \frac{1}{\hat{\upsilon}_t+\hat{\upsilon}_c}$, and for $A_1 = a/n_t + p_{t|\epsilon}q_{t|\epsilon}$, $A_2 = a/n_c + p_c q_c$, $B_1 = q_{t|\epsilon} - p_{t|\epsilon}$, $B_2 = q_c - p_c$, $e_1 = \hat{p}_t - p_{t|\epsilon}$, and $e_2 = \hat{p}_c - p_c$, we have

$$f(\hat{p}_{t|\epsilon}, \hat{p}_c) = \frac{\hat{w}}{n} = \frac{n}{n+2a}\left\{\frac{\left(A_1 + B_1 e_1 - e_1^2\right)\left(A_2 + B_2 e_2 - e_2^2\right)}{(2A + B(e_1+e_2) - \left(e_1^2 + e_2^2\right)}\right\}.$$

Expanding $f(\hat{p}_{t|\epsilon}, \hat{p}_c)$ in a Taylor series about $\hat{p}_{t|\epsilon}, \hat{p}_C = (p_{t|\epsilon}, p_c)$, we can write

$$\begin{aligned}
\frac{\hat{w}}{n} = {} & \frac{a}{n} + \frac{(p_{t|\epsilon}q_{t|\epsilon})(p_c q_c)}{p_{t|\epsilon}q_{t|\epsilon} + p_c q_c} + \frac{(q_{t|\epsilon} - p_{t|\epsilon})(p_c^2 q_c^2)}{(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^2} e_1 + \frac{(q_c - p_c)(p_{t|\epsilon}^2 q_{t|\epsilon}^2)}{(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^2} e_2 \\
& - \left\{\frac{2(p_c^2 q_c^2)}{(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^2} + \frac{2(p_{t|\epsilon} - q_{t|\epsilon})^2 (p_c^2 q_c^2)}{(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^3}\right\} e_1^2 \\
& - \left\{\frac{2(p_{t|\epsilon}^2 q_{t|\epsilon}^2)}{(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^2} + \frac{2(p_c - q_c)^2 (p_{t|\epsilon}^2 q_{t|\epsilon}^2)}{(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^3}\right\} e_2^2 \\
& + \left\{\frac{2(q_{t|\epsilon} - p_{t|\epsilon})(q_c - p_c)(q_{t|\epsilon} p_{t|\epsilon} q_c p_c)}{(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^3}\right\} e_1 e_2.
\end{aligned}$$

$$\begin{aligned}
E\left(\frac{\hat{w}}{n}|\epsilon\right) = {} & \frac{(p_{t|\epsilon}q_{t|\epsilon})(p_c q_c)}{p_{t|\epsilon}q_{t|\epsilon} + p_c q_c} + \frac{1}{n} \\
& \times \left\{\frac{a - 2(q_{t|\epsilon} p_{t|\epsilon} q_c p_c)[(p_{t|\epsilon} - q_{t|\epsilon})^2 (p_c q_c) + (p_c - q_c)^2 p_{t|\epsilon} q_{t|\epsilon}]}{(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^3}\right. \\
& \left. \times \frac{2(q_{t|\epsilon} p_{t|\epsilon} q_c p_c)}{p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)}\right\}.
\end{aligned}$$

Let, $L_1(x) = \ln\left(\frac{x_{it} + a}{n_{it} - x_{it} + a}\right)$ and $L_2(x) = \ln\left(\frac{x_{ic} + a}{n_{ic} - x_{ic} + a}\right)$. And $l(p_t|\epsilon) = E_x(L_1(x))$ and $l(p_c) = E_x(L_2(x))$, then

$$\begin{aligned}
L_a(x) &= L_1(x) - L_2(x) \\
&= l(p_{t|\epsilon}) + B_{(t)0} + \sum_{i=1}^{\infty} B_{(t)i} e_1^i - \left[l(p_c) + B_{(c)0} + \sum_{i=1}^{\infty} B_{(c)i} e_2^i\right] \\
&= l(p_{t|\epsilon}) - l(p_c) + B_{(t)0} - B_{(c)0} + \sum_{i=1}^{\infty} B_{(t)i} e_1^i - \sum_{i=1}^{\infty} B_{(c)i} e_2^i.
\end{aligned}$$

Thus, if $L(p) = E_x(L_a(x))$, then

$$\begin{aligned}
E\left\{\frac{\hat{w}}{n} L_a(x)|\epsilon\right\} &= E\left\{\frac{\hat{w}}{n}|\epsilon\right\} L(p) + E\left\{\frac{\hat{w}}{n}|\epsilon\right\}[B_{(t)0} - B_{(c)0}] \\
&\quad + \frac{(q_{t|\epsilon} - p_{t|\epsilon})(p_c^2 q_c^2)}{n(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^2} \\
&\quad - \frac{(p_{t|\epsilon} - q_{t|\epsilon})(p_c q_c)}{2n(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)} - \frac{(q_c - p_c)(p_{t|\epsilon}^2 q_{t|\epsilon}^2)}{n(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)^2} \\
&\quad + \frac{(q_c - p_c)(p_{t|\epsilon} q_{t|\epsilon})}{2n(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)} \\
&= E\left\{\frac{\hat{w}}{n}|\epsilon\right\}[L(p) + (B_{(t)0} - B_{(c)0})] + H(p_{t|\epsilon} q_{t|\epsilon} p_c q_c),
\end{aligned}$$

where

$$B_{(.)i} = \sum_{j=0}^{\infty} B_{ij}/n^j.$$

All the B_{ij}'s are constant with respect to n. In particular,

$$\begin{aligned}
B_{(c)0} &= -a(p_c - q_c)/(n_c p_c q_c) + O(n^{-2}), \\
B_{(c)1} &= 1/(p_c q_c) - a\left(p_c^2 + q_c^2\right)/\{n_c(p_c q_c)^2\} + O(n^{-2}), \\
B_{(c)2} &= \frac{1}{2}\left(p_c^2 - q_c^2\right)/(p_c q_c)^2 - a\left(p_c^3 - q_c^3\right)/\{n_c(p_c q_c)^3\} + O(n^{-2}), \\
B_{(t)0} &= -a(p_{t|\epsilon} - q_{t|\epsilon})/(n_t p_{t|\epsilon} q_{t|\epsilon}) + O(n^{-2}), \\
B_{(t)1} &= 1/(p_{t|\epsilon} q_{t|\epsilon}) - a\left(p_{t|\epsilon}^2 + q_{t|\epsilon}^2\right)/\{n_t(p_{t|\epsilon} q_{t|\epsilon})^2\} + O(n^{-2}), \\
B_{(t)2} &= \frac{1}{2}\left(p_{t|\epsilon}^2 - q_{t|\epsilon}^2\right)/(p_{t|\epsilon} q_{t|\epsilon})^2 \\
&\quad - a\left(p_{t|\epsilon}^3 - q_{t|\epsilon}^3\right)/\{n_t(p_{t|\epsilon} q_{t|\epsilon})^3\} + O(n^{-2}).
\end{aligned}$$

Then,

$$\begin{aligned}
& E\left(\frac{E\left\{\frac{\hat{w}}{n}|\epsilon\right\}}{E\left\{\frac{\hat{w}}{n}\right\}} L_a(x) - L(p)\right) \\
&= E\left(-\frac{(p_{t|\epsilon} - q_{t|\epsilon})}{n(q_{t|\epsilon} p_{t|\epsilon})}\left\{a + \frac{p_c q_c - p_{t|\epsilon} q_{t|\epsilon}}{2(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)}\right\}\right. \\
&\quad \left. + \frac{(p_c - q_c)}{n(p_c q_c)}\left\{a + \frac{p_{t|\epsilon} q_{t|\epsilon} - p_c q_c}{2(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)}\right\}\right) \\
&= E\left\{H(p_{t|\epsilon}, q_{t|\epsilon}, p_c, q_c)\right\},
\end{aligned}$$

where $H(p_{t|\epsilon}, q_{t|\epsilon}, p_c, q_c) = -\frac{(p_{t|\epsilon} - q_{t|\epsilon})}{n(q_{t|\epsilon} p_{t|\epsilon})}\{a + \frac{p_c q_c - p_{t|\epsilon} q_{t|\epsilon}}{2(p_{t|\epsilon} q_{t|\epsilon} + p_c q_c)}\} + \frac{(p_c - q_c)}{n(p_c q_c)}\{a + \frac{p_{t|\epsilon} q_{t|\epsilon} - p_c q_c}{2(p_{t|\epsilon} q_{t|\epsilon} + p_c q_c)}\}$. Note that H is a continuous function of $\frac{p_{t|\epsilon}}{n}$, and $\frac{p_{t|\epsilon}}{n} \xrightarrow{p} \frac{e^{\mu+\theta+\tau^2/2}}{1+e^{\mu+\theta+\tau^2/2}}/n$. Let $p_t = \frac{e^{\mu+\theta+\tau^2/2}}{1+e^{\mu+\theta+\tau^2/2}}$. Then,

$$E\left(\frac{E\left\{\frac{\hat{w}}{n}|\epsilon\right\}}{E\left\{\frac{\hat{w}}{n}\right\}} - L(p)\right) = H(p_t, q_t, p_c, q_c).$$

[Received July 2010. Revised December 2011.]

REFERENCES

Bhaumik, D., Kapur, K., and Gibbons, R. (2009), "Testing Parameters of a Gamma Distribution for Small Samples," *Technometrics*, 51, 326–334. [559]

Bohning, D., Malzahn, U., Dietz, E., Schlattmann, P., Viwatwongkasem, C., and Biggeri, A. (2002), "Some General Points in Estimating Heterogeneity Variance With the DerSimonian Laird Estimator," *Biostatistics*, 3, 445–457. [557]

Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Russel, L. A. (2007), "Much Ado About Nothing: A Comparison of the Performance of Meta-analytical Methods With Rare Events," *Statistics in Medicine*, 26, 53–77. [555]

Cochran, W. (1950), "The Comparison of Percentages in Matched Samples," *Biometrika*, 37, 256–266. [559]

——— (1954), "The Contribution of Estimates From Different Experiments," *Biometrics*, 10, 101–129. [557]

Cox, D. (1970), *The Analysis of Binary Data*, London: Methuen. [556]

Deeks, J. (2002), "Issues in the Selection of a Summary Statistic in Meta-Analysis of Clinical Trials With Binary Outcomes," *Statistics in Medicine*, 21, 1575–1600. [555]

DerSimonian, R., and Kacker, R. (2007), "Random-Effects Model for Meta-analysis of Clinical Trials: An Update," *Contemporary Clinical Trial*, 28, 105–114. [556]

DerSimonian, R., and Laird, N. (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical Trial*, 7, 177–188. [557]

Engels, E., Schmid, C., Terrin, N., Olkin, I., and Lau, J. (2000), "Heterogeneity and Statistical Significance in Meta-Analysis: An Empirical Study of 125 Meta-Analyses," *Statistics in Medicine*, 19, 1707–1728. [555]

Gart, J., Hugh, M., and Thomas, D. G. I. (1985), "The Effect of Bias, Variance Estimation, Skewness and Kurtosis of the Empirical Logit on Weighted Least Squares Analyses," *Biometrika*, 72, 179–190. [564]

Haldane, J. B. S. (1955), "The Estimation and Significance of the Logarithm of a Ratio Frequencies," *Annals of Human Genetics*, 20, 309–311. [556]

Hardy, R., and Thompson, S. (1996), "A Likelihood Approach to Meta-Analysis With Random Effects," *Statistics in Medicine*, 15, 619–629. [557]

Hartung, J., Knapp, G., and Sinha, B. (2008), *Statistical Methods With Applications*, New York: Wiley. [559]

Hedges, L., and Olkin, I. (1985), *Statistical Methods for Meta-Analysis*, Orlando, FL: Academic Press. [555]

Higgins, J., and Thompson, S. (2002), "Quantifying Heterogeneity in a Meta-Analysis," *Statistics in Medicine*, 21, 1539–1558. [557,561]

Huedo-Medina, T. B., Snchez-Meca, J., Marn-Martnez, F., and Botella, J. (2006), "Assessing Heterogeneity in Meta-Analysis: Q Statistic or I^2 Index," *Psychological Methods*, 11, 193–206. [557]

Krishnamoorthy, K., Mathew, T., and Mukherjee, S. (2008), "Normal Based Methods for a Gamma Distribution: Prediction and Tolerance Interval and Stress-Strength Reliability," *Technometrics*, 50, 69–78. [559]

Malzahn, U., Bohning, D., and Holling, H. (2000), "Nonparametric Estimation of Heterogeneity Variance for the Standardized Difference Used in Meta-analysis," *Biometrica*, 87, 619–632. [557]

Mantel, N., and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *Journal of National Cancer Institute*, 22, 19–48. [555]

Paule, R., and Mandel, J. (1982), "Consensus Values and Weighting Factors," *Journal of Research of National Bureau Standard*, 87, 377–385. [558]

Rihal, C. S., Raco, D. L., Gersh, B. J., and Yusuf, S. (2003), "Indications for Coronary Artery Bypass Surgery and Percutaneous Coronary Intervention in Chronic Stable Angina: Review of the Evidence and Methodological Considerations," *Circulation*, 108, 2439–2445. [563]

Rosamond, W., Flegal, K., Friday, G., Furie, K., Go, A., et al. (2007), "Heart Disease and Stroke Statistics-2007 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee," *Circulation*, 115, 69–171. [563]

Schomig, A., Mehilli, J., Waha, A. D., Seyfarth, M., Pache, J., and Kastrati, A. (2008), "A Meta-Analysis of 17 Randomized Trials of a Percutaneous Coronary Intervention-Based Strategy in Patients With Stable Coronary Artery Disease," *Journal of the American College of Cardiology*, 52, 894–904. [563,564]

Self, S., and Liang, K. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610. [559]

Shapiro, A. (1985), "Asymptotic Distribution of Test Statistics in the Analysis of Moment Structures Under Inequality Constraints," *Biometrika*, 72, 133–144. [559]

Shuster, J. J. (2010), "Empirical vs Natural Weighting in Random Effects Meta-Analysis," *Statistics in Medicine*, 29, 1259–1265. [555,556,557]

Sidik, K., and Jonkman, J. (2005), "Simple Heterogeneity Variance Estimation for Meta-Analysis," *Journal of Royal Statistical Society,* Series C, 54, 367–384. [557]

——— (2007), "A Comparison of Heterogeneity Variance Estimators in Combining Results of Studies," *Statistics in Medicine*, 26, 1964–1981. [557]

Sweeting, M., Sutton, A., and Lambert, P. (2004), "What to Add to Nothing? The Use and Avoidance of Continuity Corrections in Meta-Analysis of Sparse Data," *Statistics in Medicine*, 23, 1351–1375. [555,556,560,564]

Timmis, A. D., Feder, G., and Hemingway, H. (2007), "Prognosis of Stable Angina Pectoris: Why We Need Larger Population Studies With Higher Endpoint Resolution," *Heart*, 93, 786–791. [563]

Viechtbauer, W. (2007), "Confidence Intervals for the Amount of Heterogeneity in Meta-Analysis," *Statistics in Medicine*, 26, 37–52. [557]

Whitehead, A., and Whitehead, J. (1991), "A General Parametric Approach to the Meta-Analysis of Randomized Clinical Trials," *Statistics in Medicine*, 10, 1665–1677. [563]