Biometrics 65, 487–496
June 2009
DOI: 10.1111/j.1541-0420.2008.01068.x

Meta-Analysis of Studies with Missing Data

Ying Yuan[1,*] and Roderick J. A. Little[2]

[1]Department of Biostatistics, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, U.S.A.

[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

[*]*email:* yyuan@mdanderson.org

Summary. Consider a meta-analysis of studies with varying proportions of patient-level missing data, and assume that each primary study has made certain missing data adjustments so that the reported estimates of treatment effect size and variance are valid. These estimates of treatment effects can be combined across studies by standard meta-analytic methods, employing a random-effects model to account for heterogeneity across studies. However, we note that a meta-analysis based on the standard random-effects model will lead to biased estimates when the attrition rates of primary studies depend on the size of the underlying study-level treatment effect. Perhaps ignorable within each study, these types of missing data are in fact not ignorable in a meta-analysis. We propose three methods to correct the bias resulting from such missing data in a meta-analysis: reweighting the DerSimonian–Laird estimate by the completion rate; incorporating the completion rate into a Bayesian random-effects model; and inference based on a Bayesian shared-parameter model that includes the completion rate. We illustrate these methods through a meta-analysis of 16 published randomized trials that examined combined pharmacotherapy and psychological treatment for depression.

Key words: Bias; Nonignorable missing data; Patient-level missing data; Random-effects model.

1. Introduction

Over the last two decades, greater emphasis on evidence-based medicine and the need for reliable summaries of the vast and expanding volume of clinical research (Altman, 2000) have led to an upsurge in the application of meta-analysis to medical research. Because missing data are common in clinical studies, it is important that these meta-analyses make appropriate adjustments for missing data.

A typical meta-analysis includes the results of k independent studies, where the ith study gives an estimate T_i of some underlying treatment effect parameter θ_i. We are interested in combining $\{T_i\}$ to make inferences about the population treatment effect. In the context of meta-analysis, we can classify missing data into three broad categories according to the level of the missing data (Sutton et al., 2000): (i) whole studies missing. A typical example is publication bias where relevant studies are not published because of a lack of significant results, and hence cannot be included in a meta-analysis; (ii) data missing at the study level. For example, investigators may not report estimates of treatment effect size and/or study-level covariates in a publication; and (iii) data missing at the individual patient level, such as nonresponse or failure to record patient outcomes. The first two types of missing data are analogous to unit nonresponse and item nonresponse, respectively, in the survey sampling literature.

Most of the research on missing data in meta-analysis focuses on the situations described in categories (i) and (ii), which are problems commonly regarded as being unique to meta-analysis (Pigott, 1994). Concerning publication bias, Sterling (1959) reported that a large proportion of published studies had rejected the null hypotheses, and later studies include Rosenthal (1978, 1979), Dear and Begg (1992), Hedges (1992), Copas (1999), Duval and Tweedie (2000), Baker and Jackson (2006), Henmi, Copas, and Eguchi (2007), among others. A comprehensive discussion of publication bias is Rothstein, Sutton, and Borenstein (2005). Some research has also been conducted for missing data problems in category (ii). For example, Bushman and Wang (1995, 1996) describe methods to combine studies in which data to calculate an effect size and its standard deviation are not available from all studies.

In contrast, the situation described in category (iii) has received very little attention. This omission is partially due to the presumption that there is no need to adjust for patient-level missing data in a meta-analysis, given that each study has already addressed the issue of missing data and reported consistent estimated treatment effects. Researchers may also assume that they can combine valid study-specific estimates directly, using standard methods of meta-analysis based on random-effects models, to obtain a consistent estimate of the population treatment effect size. We show that these assumptions can be wrong. In particular, we show that if the patient attrition rate depends on the (underlying) true treatment effect size of the primary studies, then a meta-analysis based on a standard random-effects model may lead to biased estimates, even though the estimate of treatment effect size from each study is individually valid. We propose several methods to correct this bias by utilizing the information on attrition rates from the primary studies.

Considerable research has been conducted to investigate the relationship of baseline risk to treatment effect size as a possible explanation of between-study heterogeneity in meta-analysis. Brand and Kragt (1992) developed a simple linear regression to examine the relationship between the effect size and baseline risk rate in the control groups, without considering the measurement errors associated with the observed effect size and baseline risk rate. Several methods have been proposed to take into account the measurements errors in the estimates. McIntosh (1996) proposed a bivariate normal model for the underlying true treatment effect size and the true baseline risk measure, together with an approximate normal measurement error model. Walter (1997) assumed a linear function relationship between true treatment effect and true baseline risk, coupling with an approximate normal measurement error model. Arends et al. (2000) extended the approach of McIntosh (1996) by using a more flexible mixture of two normal distributions to model the baseline risk. This article focuses on the bias caused by the relationship between the attrition rate and treatment effect size. This relationship can also be explored as a part of an explanation of between-study heterogeneity.

Rubin (1976) and Little and Rubin (2002) classify missing data mechanisms into three types: missing completely at random (MCAR), where the probability of missingness does not depend on observed or unobserved data measures; missing at random (MAR), where the probability of missingness depends only on observed data measures; and nonignorable missing data (NI), where the probability of missingness depends on the unobserved data measures. Surprisingly, when attrition rates are associated with study-specific treatment effect sizes, the missing data are nonignorable in the context of a meta-analysis based on a random-effects model, even though the missing data are ignorable (MCAR or MAR) within each primary study. This issue was pointed out previously in the context of multistage sampling by Yuan and Little (2007).

In Section 2, we investigate the bias of a meta-analysis based on a random-effects model, in which the attrition rates depend on treatment effect sizes across studies, and propose several methods to correct the bias. In Section 3, we describe a simulation study comparing the different approaches. In Section 4, we illustrate the methods through a meta-analysis of data from 16 published randomized trials that examined combined pharmacotherapy and psychological treatment for depression. We provide concluding remarks in Section 5.

2. Methods

2.1 *Random-Effects Model Based Meta-Analysis*

Suppose that data to be combined arise from a series of k independent studies with sample sizes $n_1, \ldots, n_k$, in which the ith study consists of r_i completers and $n_i - r_i$ incompleters, with the observed completion rate $\hat{\phi}_i = r_i/n_i$. Let θ_i denote the true underlying effect size for the ith study, and assume that the study reports an estimate of θ_i, say T_i, with associated estimated sampling variance $\hat{\sigma}_i^2$, based on an analysis of the r_i completers, for $i = 1 \ldots k$. We are interested in estimating the population effect size θ.

Two standard approaches for combining the estimated treatment effects $T_1, \ldots, T_k$ are the fixed-effects model and the random-effects model. The fixed-effects model assumes that all the studies estimate a common underlying treatment effect $\theta = \theta_1 = \cdots = \theta_k$; a widely used estimate of θ is then $\sum_{i=1}^k T_i u_i / \sum_{i=1}^k u_i$ with $u_i = 1/\hat{\sigma}_i^2$ (Cochran, 1937; Shadish and Haddock, 1994). However, the assumption of common underlying treatment effects across studies rarely holds in practice, because heterogeneity almost always exists among studies due to various factors, such as different study designs, different treatment protocols, different within-study matching protocols, or differences in characteristics of study participants (Bailey, 1987). Tests of homogeneity are often used to test the heterogeneity among studies. However, these tests often have little power as the number of studies in meta-analysis is often small. When the test of homogeneity is not significant, there can still be a nonnegligible amount of heterogeneity among studies. Not being able to reject the null hypothesis of zero between-study variation (i.e., homogeneity) is not the same as assuming it is zero. Using fixed-effects models under conditions of heterogeneity underestimates the uncertainty about the treatment effect size, leading to invalid confidence intervals and tests.

In this article, we focus on the random-effects meta-analysis model, which has become increasingly popular in medical research (DerSimonian and Laird, 1986; Fleiss, 1993; Mosteller and Colditz, 1996; Sutton et al., 2000). Unlike a fixed-effects model, the random-effects model accounts for heterogeneity among studies by assuming that θ_i is not fixed, but is itself random and has its own distribution, namely,

$$\begin{aligned} T_i &= \theta_i + e_i \\ \theta_i &= \theta + e_i^*, \end{aligned}$$

where e_i and e_i^* are independent random variables with mean 0 and variance σ_i^2 and τ^2. Typically, we assume that T_i is sufficient for θ_i and independent when conditional on random effects θ_i. The random-effects model becomes a fixed-effects model if $\tau^2 = 0$. Let $\hat{\tau}^2$ denote a consistent estimate of τ^2, and assume that the sample sizes of primary studies are sufficiently large that we can ignore uncertainty in the within-study variance estimates $\hat{\sigma}_i^2$. This assumption is standard in meta-analysis. Then a widely used estimate of the population effect size θ is given by DerSimonian and Laird (1986):

$$\hat{\theta}_{DL} = \sum_{i=1}^k w_i T_i \Big/ \sum_{i=1}^k w_i, \tag{1}$$

where $w_i = (\hat{\sigma}_i^2 + \hat{\tau}^2)^{-1}$, and $\hat{\tau}^2$ is given by

$$\hat{\tau}^2 = \begin{cases} 0 & \text{if} \quad Q \le k-1 \\ [Q-(k-1)]/U & \text{if} \quad Q > k-1, \end{cases}$$

where

$$Q = \sum_{i=1}^k w_i \left(T_i - \frac{\sum_{i=1}^k w_i T_i}{\sum_{i=1}^k w_i} \right)^2,$$

$$U = (k-1)\left(\frac{1}{k}\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2 - \frac{1}{k}\left(\sum_{i=1}^{k} w_i\right)^2}{(k-1)\sum_{i=1}^{k} w_i}\right).$$

A simple estimate of the variance of the DerSimonian–Laird (DL) estimate (ignoring uncertainty in the estimated variances) is

$$\mathrm{Var}(\hat{\theta}_{DL}) = 1 \Big/ \sum_{i=1}^{k} w_i.$$

Commonly, the random variables e_i and e_i^* are assumed to follow normal distributions, yielding the basic normal random-effects model

$$T_i \,\big|\, \theta_i, \sigma_i^2 \sim N\big(\theta_i, \sigma_i^2\big)$$
$$\theta_i \,\big|\, \theta, \tau^2 \;\sim N\big(\theta, \tau^2\big).$$

In this case, maximum likelihood methods can be used to estimate θ, σ_i^2, and τ^2 simultaneously (Hedges, 1981; DerSimonian and Laird, 1986). The DL estimate is consistent (when missing data are MAR), but it ignores the sampling variance of $\hat{\sigma}_i^2$ and $\hat{\tau}$, leading to a confidence interval that is too narrow. This problem was discussed by Louis and Zelterman (1994), Hardy and Thompson (1996), and Biggerstaff and Tweedie (1997). In particular, Louis and Zelterman (1994) noted that a fully Bayesian random-effects model would automatically take into account the extra uncertainty induced by estimating τ.

The above meta-analysis only involves summary statistics $\{T_i, \hat{\sigma}_i^2\}$ from k primary studies, and does not require patient-level data. Provided the estimates $\{T_i, \hat{\sigma}_i^2\}$ from the analysis of completers are valid—in particular, the completer analysis is not subject to bias and the standard error $\hat{\sigma}_i$ is based on the completer sample size r_i—one might think that the missing data do not have any effect on the meta-analysis, and $\hat{\theta}_{DL}$ is a consistent estimate of θ. However, this is not the case. The random-effects model-based estimate $\hat{\theta}_{DL}$ may be biased when the underlying completion rate of the ith primary study, say ϕ_i, depends on the underlying effect size θ_i. In practice, some study-level characteristics, such as geographic location of the study, budget of the study, quality of the health care provided, the amount of experience of the treating physicians, and many others, are often associated with both the response rate and the effect size. If these study-level variables are not measured and controlled in the meta-analysis, then the response rate ϕ_i will depend on the effect size θ_i. In these cases, the missing data are nonignorable for the meta-analysis because the random effects θ_i are unobserved. A consequence of the nonignorability of the missing data is that the DL estimate is biased (Little and Rubin, 2002). This interesting fact has not been previously noted in the meta-analysis literature.

The DL estimate is generally biased as the attrition rate depends on the study-specific effect size. The bias does not depend on the statistical methods used to adjust the missing data within each individual study. For example, multiple imputation (MI) is a popular method to deal with missing data. Even we conducted MI using an appropriate model and obtained consistent estimates of θ_i based on the multiply imputed datasets via MI combination rules, the DL estimate is still subject to bias if the attrition rate depends on the study-specific effect size. In general, we assume that estimates of effect sizes from each study are consistent. We do not impose any restrictions on the methods used to obtain these consistent estimates.

To see the bias of random-effects meta-analysis more directly, let T_i^* and $\hat{\sigma}_i^{*2}$ denote the estimate of θ_i and the associated sampling variance *without* missing data. For ease of exposition, we assume that the sampling variance of T_i^* is proportional to n_i^{-1}, i.e., $\hat{\sigma}_i^{*2} = \hat{\sigma}^2/n_i$ and $\hat{\sigma}_i^2 = \hat{\sigma}^2/r_i$. Without missing data, the DL estimate is given by

$$\hat{\theta}_{DL}^* = \frac{\sum_{i=1}^{k}\left(\hat{\sigma}^2/n_i + \hat{\tau}^2\right)^{-1} T_i^*}{\sum_{i=1}^{k}\left(\hat{\sigma}^2/n_i + \hat{\tau}^2\right)^{-1}},$$

which is a consistent estimate of the population treatment effect size. In the presence of $n_i - r_i$ dropouts, the estimate of the treatment effect size becomes

$$\begin{aligned}\hat{\theta}_{DL} &= \frac{\sum_{i=1}^{k}\left(\hat{\sigma}^2/r_i + \hat{\tau}^2\right)^{-1} T_i}{\sum_{i=1}^{k}\left(\hat{\sigma}^2/r_i + \hat{\tau}^2\right)^{-1}} \\ &= \frac{\sum_{i=1}^{k}\left(\hat{\phi}_i^{-1}\hat{\sigma}^2/n_i + \hat{\tau}^2\right)^{-1} T_i}{\sum_{i=1}^{k}\left(\hat{\phi}_i^{-1}\hat{\sigma}^2/n_i + \hat{\tau}^2\right)^{-1}}.\end{aligned} \tag{2}$$

Based on our assumptions, T_i is a consistent estimate of θ_i; however, the weight assigned to T_i is distorted by $\hat{\phi}_i^{-1}$. As a result, $\hat{\theta}_{DL}$ is biased. For example, if a higher value of the study-specific treatment effect size θ_i is associated with a higher value of the study response rate ϕ_i, $\hat{\theta}_{DL}$ is inflated, as studies with high treatment effect sizes are overweighted in equation (2). The degree of the bias depends on the ratio of the within-study variance σ_i^2 versus the between-study variance τ^2 (i.e., heterogeneity among studies). If this ratio is large, then the influence of $\hat{\phi}_i$ is also large, leading to a relatively large bias.

2.2 *Methods to Correct Bias*

2.2.1 *Reweighted DL (RWDL) estimate.* The foregoing bias analysis motivates a simple approach to correct the bias of the DL estimate by modifying w_i to reflect the correct weight. Letting $\nu_i = \hat{\phi}_i\hat{\sigma}_i^2$, we modify w_i in (1) to

$$w_i = \left(\nu_i + \hat{\tau}^2\right)^{-1}.$$

This RWDL estimate corrects the bias of the DL estimate, but has two drawbacks. First, the resulting confidence interval tends to be narrow because, like the DL estimate, it ignores the uncertainty from the estimation of the between-study variance τ^2. Second, it implicitly assumes that the variance of T_i is proportional to n_i^{-1}. If this assumption does not hold, the

reweighting estimate may be biased, but the bias is expected to be smaller than that of the DL estimate.

2.2.2 *Reweighted Bayesian random-effects (RWRE) model.* One way to address the underestimation of sampling variance of the DL estimate is to take a fully Bayesian approach (Louis and Zelterman, 1994). We thus propose the following RWRE model to overcome the drawback of the RWDL estimate:

$$\begin{aligned} T_i \,|\, \theta_i, \sigma_i^2 &\sim N\left(\theta_i, \hat{\phi}_i \sigma_i^2\right) \quad i = 1, \ldots, k, \\ \theta_i \,|\, \theta, \tau^2 &\sim N(\theta, \tau^2) \\ \theta, \tau &\propto Constant. \end{aligned} \tag{3}$$

In model (3), we assign noninformative priors to θ and τ. Informative priors can also be used to incorporate prior information about these parameters. Incorporating prior information may substantially improve the inference for meta-analyses involving a small number of studies, where data contain very limited information to estimate τ. When lacking prior information, care is needed in specifying the prior for τ^2. Various noninformative priors have been proposed for modeling variance parameters. A uniform prior distribution on $\log(\tau)$ would seems natural—working with the logarithm of a parameter that must be positive—but it results in an improper posterior distribution. A popular alternative is a vague inverse-gamma prior distribution in which the shape and scale parameters are set at a small value δ, say $\delta = 0.001$. The inverse-gamma prior is conditionally conjugate for normal hierarchical models and convenient to use, but it is problematic for random-effects models with a small number of clusters (i.e., studies) because inference may become sensitive to the value of δ (Gelman, 2006). In the model (3), we adopted the noninformative prior $\tau \propto$ Constant, which generally performs well unless the number of studies k is very low, say $k < 5$. In the case that k is very low, the half-Cauchy prior may be a useful alternative (Gelman, 2006).

The estimate of θ based on the RWRE model automatically takes into account the uncertainty associated with estimating τ^2, addressing the first drawback of the RWDL estimate. However, it still makes the assumption that the variance of T_i is proportional to the inverse of sample size, which often holds in practice but not always.

2.2.3 *Bayesian shared-parameter (SP) model.* When the study response rate depends on the study effect size, as we noted in Section 2.1, the missing data are nonignorable. From a modeling perspective, a systematic way to deal with nonignorable missing data is to jointly model the outcome process and the missing data mechanism (Little and Rubin, 2002). To this end, we propose the Bayesian SP model

$$\begin{aligned} T_i \,|\, \alpha_i, \varphi_i, \beta, \sigma_i^2 &\sim N\left(\alpha_i + \beta\varphi_i, \sigma_i^2\right) \quad i = 1, \ldots, k, \\ r_i \,|\, \varphi_i &\sim \text{Binom}(n_i, \Phi(\varphi_i)), \\ \alpha_i \,|\, \alpha, \tau^2 \sim N(\alpha, \tau^2) \quad & \varphi_i \,|\, \varphi, \omega^2 \sim N(\varphi, \omega^2), \\ \alpha, \beta, \tau, \varphi, \omega &\propto \text{Constant}, \end{aligned} \tag{4}$$

where $\text{Binom}(n_i, \Phi(\varphi_i))$ denotes a binomial distribution for n_i trials with a success probability $\Phi(\varphi_i)$ with $\Phi(\cdot)$ the cumulative density function of the standard normal distribution. Our model can be viewed as a variant of the SP model proposed in the context of longitudinal data with nonignorable (or informative) dropout (Wu and Carroll, 1988; De Gruttola and Tu, 1994; Follmann and Wu, 1995, Ten Have et al., 1998; Albert and Follmann, 2000). Specifically, in model (4), the first equation specifies the measurement process, and the second equation models the dropout mechanism. By sharing common random effects φ_i, the study-specific effect size is linked with the completion rate.

The Bayesian SP model explicitly takes into account the dependence between the study effect size and the completion rate, thereby correcting the bias of the DL estimate if the model is correctly specified. It also incorporates the uncertainty induced by estimating τ^2 by taking a fully Bayesian approach. Again, as we discussed in the previous section, special attention is needed when specifying noninformative (or weakly informative) prior to the variance parameters τ^2 and ω^2 in meta-analysis with a small number of studies.

In the Bayesian SP model (4), the regression parameter β is of particular interest because it controls how the completion rate affects the measurement process. If $\beta = 0$, the missing data are ignorable, and conventional methods, such as the DL method, yield consistent estimates without any missing data adjustments. Unfortunately, the observed data often contain limited information to estimate β precisely in SP models (Ten Have et al., 1998). In this case, a sensible strategy is performing sensitivity analysis. For the Bayesian SP model, one type of sensitivity analysis is to set β at a series of fixed values (rather than estimating β based on the data), and then evaluate the sensitivity of the inference to the value of β (Rotnitzky, Robins, and Scharfstein, 1998; Rotnitzky et al., 2001). Within the Bayesian paradigm, we propose to study the sensitivity of our model by assigning informative priors on β. In particular, we assign one (or several) informative priors on β based on subject matter, and then evaluate how it influences the results. If the inference shows no essential change, the interpretation of results is straightforward. Otherwise, there is some residual ambiguity of interpretation.

2.3 *Model Assessment*

We use the posterior predictive method (Gelman, Meng, and Stern, 1996) to assess the goodness of fit of proposed models. In this approach, a discrepancy measure $D(\boldsymbol{T}, \boldsymbol{\theta})$, which is a function of data $\boldsymbol{T}$ and parameters $\boldsymbol{\theta}$, is chosen to summarize the model deviation of interest. The observed value of the discrepancy measure is then compared to values of the discrepancy measure evaluated at replicate observations simulated from the posterior-predictive density to assess the adequacy of the model.

In the posterior predictive approach, it is important to choose an appropriate discrepancy measure to reflect inferential interests. In the RWRE model (3) and the Bayesian SP model (4), the parameter of interest is the population treatment effect size, and we are not particularly interested in making inferences about the random effects. Therefore, we chose discrepancy measures based on marginal models obtained by integrating out random effects from these models. In particular, for the RWRE model, we propose the discrepancy measure

$$D_{RE}(\boldsymbol{T}, \boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{(T_i - \theta)^2}{\tau^2 + \hat{\phi}_i \sigma_i^2};$$

and for the Bayesian SP model, we propose the discrepancy measure

$$D_{SP}(\boldsymbol{T}, \boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{(T_i - \alpha - \beta\varphi)^2}{\tau^2 + \beta^2\omega^2 + \sigma_i^2}.$$

It is easy to see that, given $\boldsymbol{\theta}$, both $D_{RE}(\boldsymbol{T}, \boldsymbol{\theta})$ and $D_{SP}(\boldsymbol{T}, \boldsymbol{\theta})$ follow a χ_k^2 distribution.

Let $\boldsymbol{\theta}^j$, $j = 1, \ldots, J$, denote a set of posterior draws of $\boldsymbol{\theta}$ conditional on $\boldsymbol{T}$. To assess the fitness of proposed models, we simply compare the realized discrepancies $D_{RE}(\boldsymbol{T}, \boldsymbol{\theta}^j)$ and $D_{SP}(\boldsymbol{T}, \boldsymbol{\theta}^j)$ to their reference distribution χ_k^2. Gelman et al. (1996) recommend making the scatterplot of realized discrepancies against the reference distribution. Alternatively, we can calculate the posterior predictive p-value as

$$p = \int \Pr\left(X_k^2 \le D_{RE}(\boldsymbol{T}, \boldsymbol{\theta})\right) \Pr(\boldsymbol{\theta} \mid \boldsymbol{T})\, d\boldsymbol{\theta},$$

where X_k^2 represents a chi-squared random variable with k degrees of freedom. Once posterior draws from $\Pr(\boldsymbol{\theta} \mid \boldsymbol{T})$ are obtained, the computation of the posterior predictive p-value is straightforward to approximate using Monte Carlo draws.

2.4 *Estimation*

Fitting the RWRE is straightforward using the Gibbs sampler (Gelfand et al., 1990). The Bayesian SP model can also be conveniently estimated using the Gibbs sampler by introducing latent variables (Albert and Chib, 1993). The details can be found in the Web Appendix.

3. Simulation Study

We simulated a meta-analysis of 16 independent studies, each of which consists of $n = 100$ paired observations (x_{ij}, y_{ij}) for $i = 1, \ldots, 16$ and $j = 1, \ldots, 100$, where x_{ij} and y_{ij} are pre- and posttreatment measurements, respectively. For simplicity, we assume that the pretreatment measures $\{x_{ij}\}$ are always observed in the studies, but that varying percentages of the posttreatment measures $\{y_{ij}\}$ are subject to data missingness. These data were generated using the following model:

$$\begin{aligned} x_{ij} \mid \sigma^2 &\sim N(0, \sigma^2/2), \\ y_{ij} \mid \alpha_i, \varphi_i, \sigma^2 &\sim N\left(\alpha_i + \beta\varphi_i, \sigma^2/2\right), \\ z_{ij} \mid \varphi_i &\sim N(\varphi_i, 1), \\ r_{ij} &\sim \begin{cases} 1 & \text{if} \quad z_{ij} > 0 \\ 0 & \text{if} \quad z_{ij} < 0, \end{cases} \\ \alpha_i \mid \alpha, \tau^2 \sim N(\alpha, \tau^2) \qquad & \varphi_i \mid \varphi, \omega^2 \sim N(\varphi, \omega^2). \end{aligned} \tag{5}$$

The treatment effect size of interest is the change in the outcome measurement after treatment. Within each study, given a constant probability of being missing, the missing data are MCAR and an estimate of the effect size, T_i, is simply the observed mean difference $\bar{u}_i = \sum_{j=1}^{r_i} u_{ij}/r_i$, where $u_{ij} = y_{ij} - x_{ij}$. An estimate of the sampling variance of T_i is $\mathrm{Var}(T_i) = \sum_{j=1}^{r_i}(u_{ij} - \bar{u}_i)^2/[r_i(r_i - 1)]$. We assume that the only data published and thus available for meta-analysis are T_i, $\mathrm{Var}(T_i)$ and the completion rate r_i/n (not the individual patient data $\{x_{ij}, y_{ij}, r_{ij}\}$).

In the population model (5), β controls the missing data mechanism at the level of the meta-analysis. If $\beta = 0$, the missing data are MCAR; if $\beta \neq 0$, the response rates are associated with the underlying study-specific treatment effect sizes, and the missing data are nonignorable. The degree of nonignorability depends on the value of β, i.e., a larger value of β causes more nonignorability. We simulated various values of β, ranging from 0 to 10. As discussed in Section 2, the bias of the random-effects model also depends on the ratio of sampling variance of T_i(i.e., σ^2/n) to the variance $\tau^2 + \beta^2\omega^2$. In the simulations, we set $\omega^2 = 1$, $\tau^2 = 4$ and varied $\sigma^2/\{n(\tau^2 + \beta^2\omega^2)\}$ from 0.1 to 5. We controlled the overall completion rate at 80% by setting the value of φ at 1.2. Values of other parameters were chosen so that the population effect size is 5. Under each setting, we generated 1000 samples and imputed the estimate of the treatment effect size from each method proposed in Section 2, including DL estimate, RWDL estimate, RWRE model, and Bayesian SP model. We also obtained the DL estimate before deleting the missing data for comparison.

Table 1 shows the relative bias with respect to the population effect size, estimated standard error and coverage rate of the 95% confidence interval or credible interval (nominally, we expect 95%) based on four methods under three different values of β and three different ratios of within-study variance versus between-study variance. Across all scenarios, as expected, the DL estimate based on the complete data is consistent, but of course this estimate is not available. In addition, it underestimates the standard error and leads to confidence intervals that are too narrow. In our simulation settings, the coverage rate is around 93%. In comparison, the DL estimate based on observed data is unbiased only when the missing data are MCAR ($\beta = 0$) at the meta-analysis level. When the response rate depends on the study effect size (i.e., $\beta = 5$ or 10), the DL estimate is biased. This bias depends on both the value of β, i.e., how strong the study response rate is associated with the study-specific treatment effect size (Figure 1), and the ratio of within-study variance and between-study variance (Figure 2). In general, the bias increases as (a) the association between the study-specific treatment effect size and the study response rate becomes stronger (the value of β increases); or as (b) the ratio of within-study variance and between-study variance becomes larger. The proposed RWDL estimate corrects the bias of the DL estimate, but like the DL estimate underestimates the standard error, as shown by the below-nominal coverage rates in our simulations. The RWRE model addresses this problem, and yields consistent estimates and reasonable coverage rates. The RWRE model works well in our simulations because the sampling variance of T_i is proportional to the inverse of the sample size. The Bayesian SP model also effectively corrects the bias of the DL estimate and yields coverage rates close to the nominal value.

4. Application

Pampallona et al. (2004) reported a meta-analysis of 16 published randomized clinical trials to determine whether the combined therapy of antidepressant drugs plus psychological treatment was more efficacious than the drug alone in treating depressive disorders. The efficacy improvement of the

Table 1

Relative bias with respect to the population effect size (%), mean square error (MSE), and coverage rate of the 95% confidence (or credible) interval for four methods under three different values of β and ratio of the within-study variance versus the between-study heterogeneity (i.e., $\frac{\sigma^2}{n(\tau^2+\beta^2\omega^2)}$). RWDL denotes the reweighted DL estimate; RWRE denotes the reweighted Bayesian random-effects model based estimate; and SP denotes the Bayesian shared-parameter based estimate. The bias of the DL estimate is shown in bold.

$\frac{\sigma^2}{n(\tau^2+\beta^2\omega^2)}$	β		Before deletion	DL	RWDL	RWRE	SP
0.2	0	Relative bias	−0.11	−0.13	−0.06	−0.17	−0.22
		MSE	0.30	0.32	0.33	0.33	0.32
		Coverage	93.1	93.3	93.0	95.3	96.1
	5	Relative bias	0.15	**8.69**	0.57	−0.05	0.80
		MSE	2.25	2.33	2.45	2.44	2.36
		Coverage	91.4	91.3	91.9	95.4	95.6
	10	Relative bias	3.22	**19.43**	3.56	2.00	4.11
		MSE	8.18	8.59	8.81	8.62	8.63
		Coverage	94.3	91.9	93.2	95.5	96.0
0.6	0	Relative bias	0.46	0.49	0.55	0.35	0.30
		MSE	0.42	0.48	0.51	0.51	0.49
		Coverage	92.4	93.0	92.6	94.5	95.3
	5	Relative bias	−0.63	**12.60**	−0.90	−1.84	−1.08
		MSE	3.12	3.57	3.74	3.74	3.58
		Coverage	92.9	90.4	92.5	94.0	96.2
	10	Relative bias	−1.96	**25.60**	−0.44	−2.58	0.21
		MSE	11.11	13.29	13.75	13.42	12.94
		Coverage	92.8	90.7	90.7	93.8	95.5
2	0	Relative bias	−0.79	−0.41	−0.47	−0.91	−0.99
		MSE	0.78	0.94	1.10	1.10	1.01
		Coverage	94.3	93.3	92.9	94.5	96.4
	5	Relative bias	0.67	**19.00**	2.14	0.36	1.81
		MSE	5.57	7.19	7.57	7.45	6.75
		Coverage	93.9	92.0	94.0	94.7	95.5
	10	Relative bias	2.07	**37.89**	2.57	−1.41	6.75
		MSE	21.13	26.54	27.58	26.09	24.32
		Coverage	92.5	91.5	92.4	94.1	96.1

combined therapy over the pharmacotherapy only was measured by the odds ratio (OR) of the treatment response (patient response to treatment or not). Most studies used the 17-item Hamilton Depression Rating Scale (HDRS) to define response to the treatment (e.g., HDRS < 7), while other studies defined response according to Beck Depression Inventory, Raskin Depression Scale, or Quality of Life Depression Scale. Because all these instruments are well validated, it may be reasonable to pool the ORs of response across studies via meta-analysis (Pampallona et al., 2004). Table 2 displays characteristics of these 16 studies, including sample size, number of dropout, logarithm of ORs, and variance of logarithm of ORs.

The estimated treatment effect size demonstrates substantial heterogeneity across 16 studies. For example, study 9 has the smallest OR of 0.75, and study 15 has the largest OR of 5.02. The χ^2 test of heterogeneity (Cochran, 1937) yielded a marginal significant p-value of 0.06, suggesting some evidence of heterogeneity. Although the test of heterogeneity often has little power, it still provides a useful informal measure of heterogeneity. We use a random-effects model to account for between-study heterogeneity and combine the ORs across studies.

The 16 studies had attrition rates ranging from 2% to 50%. To explore the potential for bias due to dependence between the study effect size and the completion rate, we plotted the completion rate against the estimate of the logarithm of the OR for the 16 primary studies (Figure 3). A clear pattern is evident, namely that studies with lower ORs tend to have higher completion rates, which casts some doubt on the validity of conventional random-effects model based meta-analysis, and motivates us to apply our methods to combine the logarithm of the ORs across the studies.

Table 3 shows estimates of the (population) OR and the corresponding 95% credible (or confidence) intervals from various methods. The DL method yields an estimated OR of 1.87, while the three proposed methods, RWDL estimate, RWRE model, and Bayesian SP model, lead to similar estimates of about 1.94. Because a lower completion rate relates to a larger effect size, and the larger effect size is downweighted in the standard random-effects model, we expect that the DL estimate underestimates the effect size. Confidence intervals for the DL estimate and the RWDL estimate are narrower than credible intervals based on the RWRE model and Bayesian SP model, reflecting the fact that they ignore the uncertainty in estimating the between-study variance.

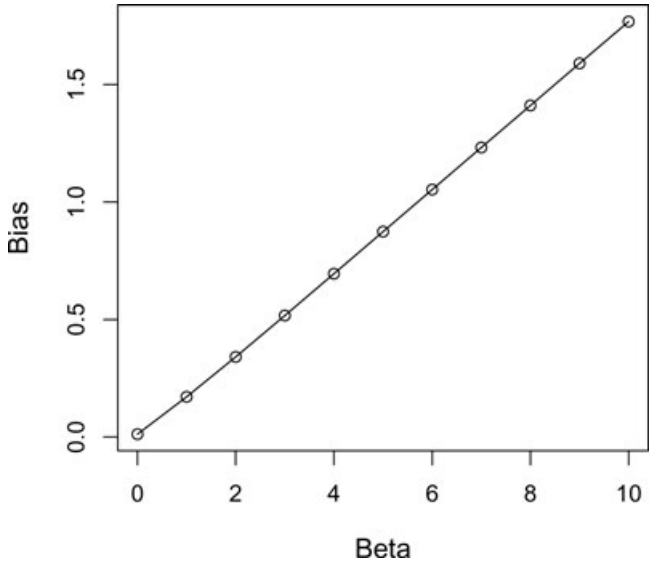

Figure 1. Bias of the DL estimate under various values of β while the within-study variance (σ^2/n) equals the between-study variance ($\tau^2 + \beta^2\omega^2$).

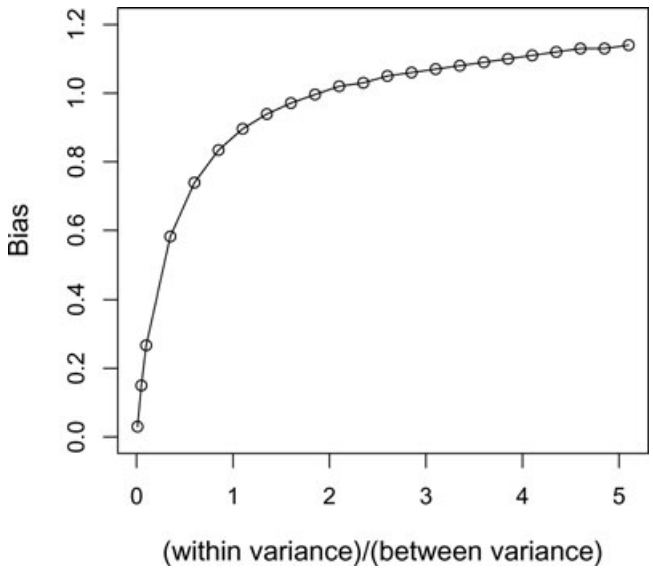

Figure 2. Bias of the DL estimate under various ratios of within-study variance versus between-study variance (i.e., $\sigma^2/n(\tau^2 + \beta^2\omega^2)$) with $\beta = 5$.

For this particular application, the estimates of population average effect based on the proposed methods are similar to that based on the DL method. The attrition often has a larger impact on the (posterior) study-specific estimates. An estimate of θ_i under the standard random-effects model is $\hat{\theta}_i = (\hat{\tau}^2 T_i + \sigma_i^2 \hat{\theta}_{DL})/(\hat{\tau}^2 + \sigma_i^2)$. Following the similar argument at the end of Section 2.1, it is easy to see that $\hat{\theta}_i$ is more susceptible to the influence of attrition because not only the weight σ_i^2 is affected by attrition but also $\hat{\theta}_{DL}$ is biased.

Table 2
Characteristics of 16 *studies included in the meta-analaysis. Log(OR) denotes the logarithm of ORs, and Var(log(OR)) denotes the variance of logarithm of ORs.*

Study	Sample size	Dropout	Log(OR)	Var(log(OR))
1	35	15	0.08	0.65
2	46	14	1.11	0.47
3	47	2	0.57	0.37
4	48	10	−0.07	0.55
5	48	14	1.2	0.37
6	46	14	0.69	0.36
7	48	24	1.5	0.42
8	58	16	0.84	0.29
9	71	12	−0.29	0.24
10	82	34	0.77	0.24
11	96	15	0.08	0.26
12	453	107	0.96	0.05
13	85	14	0.39	0.22
14	472	64	0.09	0.03
15	167	75	1.61	0.21
16	40	5	0.77	0.53

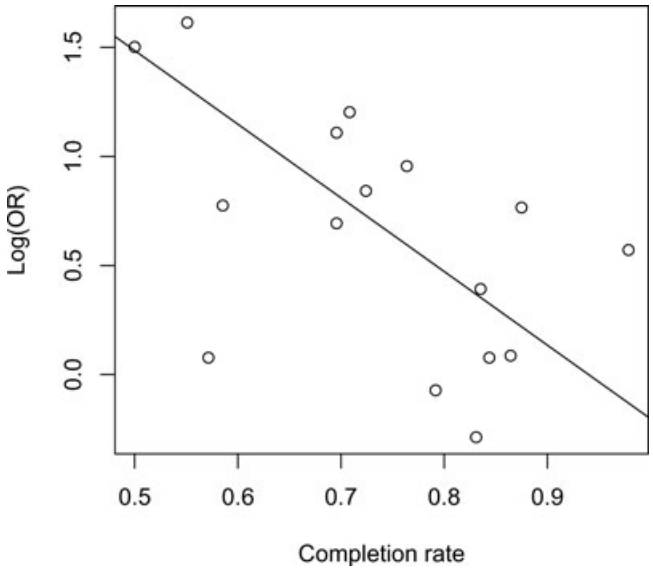

Figure 3. Logarithm of the OR against the estimated completion rate for 16 published randomized clinical trials. The reference line is obtained by weighted least squares.

Although the average effect is often the main interest, sometimes meta-analysis is also interested in finding a particularly efficacious study. The DL, RWDL, RWRE, and SP all identified study 15 as the most efficacious study with estimates of the OR of 2.7, 3.5, 3.7, and 3.4, respectively, showing more noticeable difference between the DL estimate and three proposed estimates.

Parameter estimates for the Bayesian SP model and RWRE model are displayed in Table 4. For the Bayesian SP model, the regression parameter β has a point estimate of -1.08 with standard error 0.41, suggesting statistically significant evidence that the study effect size depends on the response

Table 3
Estimates of the OR of combined treatment versus pharmacotherapy only and corresponding 95% *credible* (*or confidence*) *interval* (*CI*) *under different methods. RWDL denotes the reweighted DL estimate; RWRE denotes the reweighted Bayesian random-effects model; and SP denotes the Bayesian shared-parameter model.*

	Methods			
Estimates	DL	RWDL	RWRE	SP
OR	1.87	1.94	1.93	1.95
95% CI	(1.37, 2.53)	(1.43, 2.64)	(1.39, 2.70)	(1.37, 2.85)

rate (i.e., $\beta \neq 0$). However, because the magnitude of β is not large, the estimate of the population effect size based on the Bayesian SP model is not very different from the DL estimate. The quantity $(\tau^2 + \beta^2\omega^2)^{1/2}$ measures the between-study standard deviation. The posterior mean and standard error of $(\tau^2 + \beta^2\omega^2)^{1/2}$ are 0.53 and 0.18, respectively, suggesting significant heterogeneity among studies. Compared with the within-study variance listed in Table 2, this heterogeneity is substantial. This conclusion is confirmed by the RWRE model, which yields the estimate of τ of 0.47 with a standard error 0.16.

We assessed the goodness of fit of RWRE model and Bayesian SP model by posterior predictive checking. Figure 4 is the scatterplot of realized discrepancies versus predictive discrepancies, showing substantial agreement. The p-value for the realized discrepancy is 0.61 and 0.81, respectively, under the two models based on 10,000 Monte Carlo draws. This analysis does not yield evidence suggesting lack of fit of the two models.

We also conducted sensitivity analysis for the SP model. We assigned six informative normal priors to β with different means (i.e., -5, -3, -1, 1, 3, and 5), but the same standard deviation of 0.4. The results are displayed in Figure 5, where distribution curves on the bottom of the panel depict the priors of β and top curve denotes value of estimated population effect size. The estimate of the effect size is insensitive to the choice of prior distribution of β.

Some studies in this meta-analysis experienced high attrition, for example, the attrition rate of the studies 7 and 15 were more than 45%. Under such high attrition rates, the missing data mechanism in these studies is probably nonignorable. Our method does not assume any particular missing data mechanism *within* the studies. We only assume that the estimates reported by the studies are consistent. As long as the original studies appropriately took into account the missing data mechanism and reported consistent estimates, our method applies. Because the estimates reported by these studies were obtained by assuming MAR, they are potentially subject to bias if the missing data mechanism actually was nonignorable. Addressing this problem requires obtaining patient-level data, and conducting more detailed patient-level analysis.

5. Conclusion

In meta-analysis of studies with missing data, we have pointed out that the standard random-effects model based DL estimate is biased when attrition rates of studies are associated with study-specific treatment effect sizes, because these types of missing data are nonignorable in the context of meta-analysis. The degree of bias is positively associated with the strength of the association between the study attrition rates and the study-specific treatment effect sizes, and the relative size of the within-study and between-study variance. We have proposed three methods to correct the bias of the DL estimate. The first one, reweighting the DL estimate, is simple, but ignores the sampling variance of estimating the between-study variance, and requires the knowledge of how the within-study variance is dependent on the sample size. The second method is based on a RWRE model, which not only corrects the bias of the DL estimate, but also takes into account the extra uncertainty of estimating the between-study variance. The third method uses a SP model to jointly model the outcome and missing data mechanism.

Given these findings, a practical and important question is how to assess if the completion rate is associated with the study-specific treatment effect size. Plotting the observed study completion rates against the estimates of the treatment effect size is useful to understand the relationship between study-specific treatment effect sizes and study completion rates. If the plot does not suggest a clear relationship, then the use of the DL estimate is justified. If there is association between them, then the proposed methods may be used to correct the potential bias. A precise determination of the relationship between the study completion rate and the study treatment effect size is not necessary. We could conduct formal statistical tests to examine the correlation between the study-specific treatment effect size and the study-specific completion rate. However, when the number of studies in the meta-analysis is small, such formal tests may have low power. More importantly, if the completion rates seem to be related to the treatment effect sizes, we can always apply several analytic methods and compare the results as a form of sensitivity analysis.

Table 4
Estimates of parameters for the Bayesian SP model and RWRE model

	SP						Reweighted random effects	
Model	α	φ	β	τ	ω	$(\tau^2 + \beta^2\omega^2)^{1/2}$	θ	τ
Estimate	1.42	0.70	−1.08	0.20	0.43	0.53	0.66	0.47
Standard error	0.33	0.12	0.41	0.15	0.10	0.18	0.17	0.16

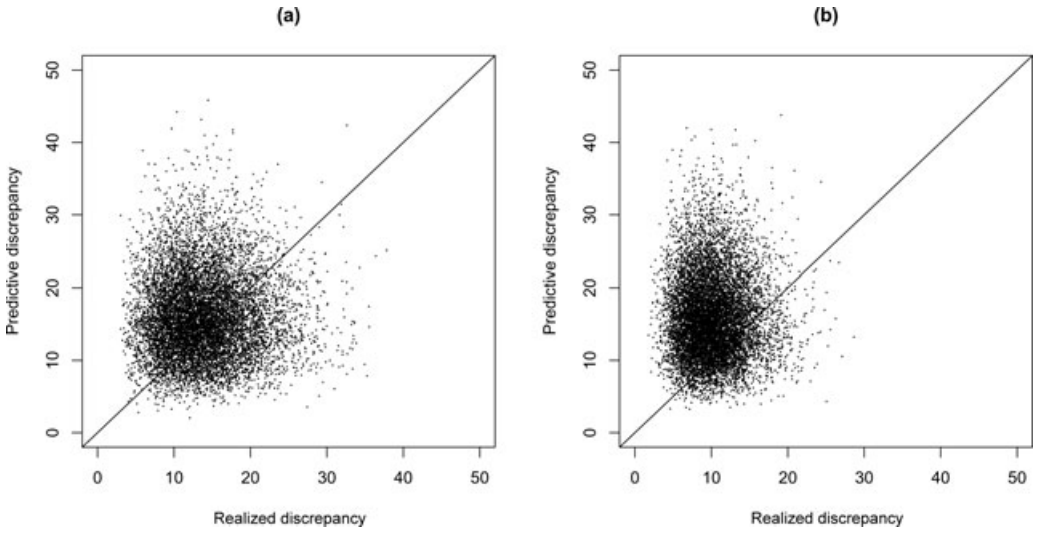

Figure 4. Scatterplot of predictive versus realized discrepancy measures for (a) RWRE model, and (b) Bayesian SP model. The p-value is estimated by the proportion of points above the 45° line.

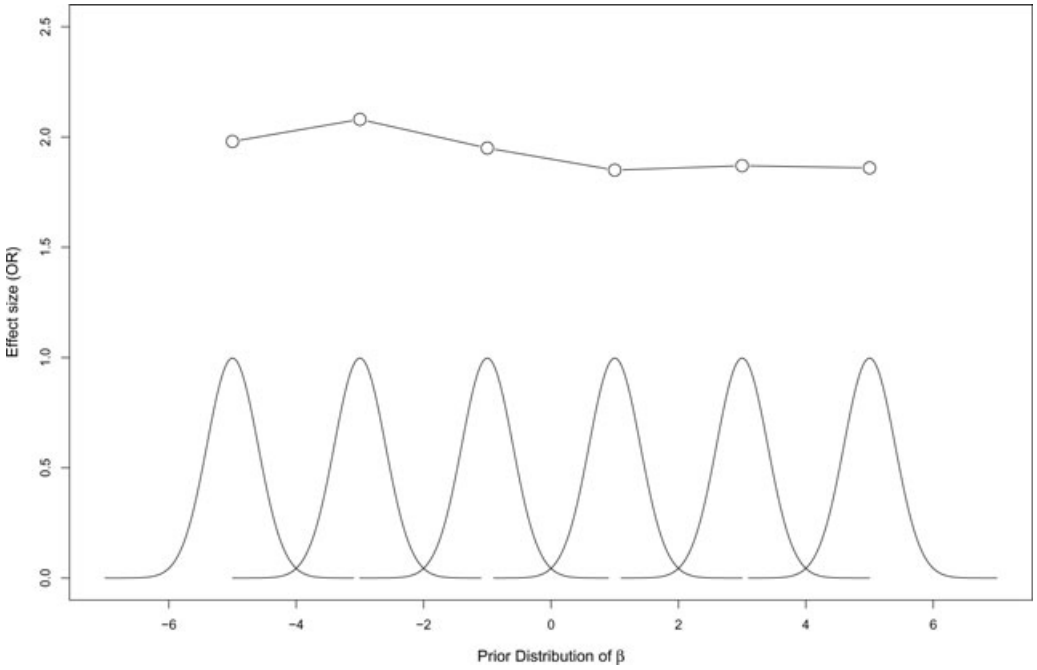

Figure 5. Sensitivity analysis of the SP model for 16 published randomized clinical trials. The curves on the bottom of the panel are prior distributions of β.

6. Supplementary Materials

The Web Appendix referenced in Section 2.4 is available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

We thank the editor and associate editor for their constructive comments that significantly improved this manuscript.

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88,** 669–679.

Albert, P. S. and Follmann, D. (2000). Modeling repeated count data subject to informative dropout. *Biometrics* **56,** 667–677.

Altman, D. G. (2000). Statistics in medical journals: Some recent trends. *Statistics in Medicine* **19,** 3275–3289.

Arends, L. R., Hoes, A. W., Lubsen, J., Grobbee, D. E., and Stijnen, T. (2000). Baseline risk as predictor of treatment benefit: Three clinical meta-re-analyses. *Statistics in Medicine* **19,** 3497–3518.

Bailey, K. R. (1987). Inter-study differences—how should they influence the interpretation and analysis of results. *Statistics in Medicine* **6,** 351–360.

Baker, R. and Jackson, D. (2006). Using journal impact factors to correct for the publication bias of medical studies. *Biometrics* **62,** 785–792.

Biggerstaff, B. J. and Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16,** 753–768.

Brand, R. and Kragt, H. (1992). Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* **11,** 2077–2082.

Bushman, B. J. and Wang, M. C. (1995). A procedure for combining sample correlation coefficients and vote counts to obtain an estimate and confidence interval for the population correlation coefficient. *Psychological Bulletin* **117,** 530–546.

Bushman, B. J. and Wang, M. C. (1996). A procedure for combining sample standardized mean differences and vote counts to estimate the population standardized mean difference in fixed effect models. *Psychological Method* **1,** 66–80.

Cochran, W. G. (1937). Problem arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society Supplement* **4**(1), 102–118.

Copas, J. (1999). What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Series A* **162,** 95–109.

Dear, K. B. G. and Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* **7,** 237–245.

De Gruttola, V. and Tu, X. M. (1994). Modelling progression of CD4 lymphocyte count and its relationship to survival time. *Biometrics* **50,** 1003–1014.

DerSimonian, R and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7,** 177–188.

Duval, S. and Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* **95,** 89–98.

Follmann, D. and Wu, M. C. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51,** 151–168.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research* **2,** 121–145.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs' sampling. *Journal of the American Statistical Association* **85,** 972–985.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1,** 515–533.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6,** 733–807.

Hardy, R. J. and Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* **15,** 619–629.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Education Statistics* **6,** 107–128.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science* **7,** 237–245.

Henmi, M., Copas, J., and Eguchi, S. (2007). Confidence intervals and p-values for meta-analysis with publication bias. *Biometrics* **63,** 475–482.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley & Sons.

Louis, T. A. and Zelterman, D. (1994). Bayesian approaches to research synthesis. In *The Handbook of Research Synthesis*, H. Cooper and L. V. Hedges (eds), 411–422. New York: Russell Sage Foundation.

McIntosh, M. W. (1996). The population risk as an explanatory variable in research syntheses of clinical trials. *Statistics in Medicine* **15,** 1713–1728.

Mosteller, F. and Colditz, G. A. (1996). Understanding research synthesis (meta-analysis). *Annual Review of Public Health* **61,** 714–719.

Pampallona, S., Bollini, P., Tibaldi, G., Kupelnick, B., and Munizza, C. (2004). Combined pharmacotherapy and psychological treatment for depression. *Archives of General Psychiatry* **61,** 714–719.

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In *The Handbook of Research Synthesis*, H. Cooper and L. V. Hedges (eds), 163–176. New York: Russell Sage Foundation.

Rosenthal, R. (1978). Combining the results to independent studies. *Professional Psychology* **17,** 136–137.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin* **86,** 638–641.

Rothstein, H., Sutton, A. J., and Borenstein, M. (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. New York: John Wiley & Sons.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association* **93,** 1321–1339.

Rotnitzky, A., Scharfstein, D., Su, T. L., and Robins, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57,** 103–113.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63,** 581–592.

Shadish, W. R. and Haddock, C. K. (1994). Combining estimates of effect size. In *The Handbook of Research Synthesis*, H. Cooper and L. V. Hedges (eds), 261–281. New York: Russell Sage Foundation.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn test of significance—or vice versa. *Journal of American Statistical Association* **54,** 30–34.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., and Song, F. (2000). *Methods for Meta-Analysis in Medical Research*. New York: John Wiley & Sons.

Ten Have, T. R., Pulkstenis, E., Kunselman, A., and Landis, J. R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative dropout. *Biometrics* **54,** 367–383.

Walter, S. D. (1997). Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine* **16,** 2883–2900.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44,** 175–188.

Yuan, Y. and Little, R. J. A. (2007). Model-based estimates of the finite population mean for two-stage cluster samples with unit nonresponse. *Applied Statistics* **56,** 79–97.

Received July 2007. *Revised March* 2008.
Accepted March 2008.