Adjustment for Publication and Quality Bias in Bayesian Meta-analysis

D. D. Smith, Geof H. Givens, and R. L. Tweedie
Department of Statistics, Colorado State University
Fort Collins, Colorado 80523-1877, U.S.A.

Summary

Meta-analysis reviews, collects, and synthesizes individual sample surveys to estimate an overall effect size. If the studies for a meta-analysis are chosen through a literature review, an inherent selection bias may arise, since in particular, studies may tend to be published more readily if they are statistically significant, or deemed to be of higher quality. Here, 'quality' depends on sample characteristics, study design elements such as blinding and control, and many other objective and subjective factors.

Within a Bayesian hierarchical model allowing stratification on quality, we develop a data augmentation technique to estimate and adjust for the numbers and outcomes of missing studies after allowing for such effects. We apply this method to a meta-analysis of studies of cervical cancer rates associated with use of oral contraceptives.

1 Introduction

There has been an enormous recent increase in the use of meta-analysis as a statistical technique for combining the results of many individual analyses (Hedges and Olkin, 1985; Olkin, 1992; Cooper and Hedges, 1994). While the combined analysis may have increased inferential power over any individual study, there are several drawbacks to meta-analysis (see, for example, Thompson and Pocock, 1991; the NRC Report, 1992; Mengersen et al., 1995). One well-documented concern is the need to collect all studies, both published and unpublished, relevant to the meta-analysis if the subsequent inferences are to be valid (Iyengar and Greenhouse, 1988; Dear and Begg, 1992; Hedges, 1992).

A meta-analysis based on only a subset of all relevant studies may result in biased conclusions. It is a common belief that studies are not uniformly likely to be published in

Key words: Meta-analysis; Publication bias; Missing studies; Quality; Cervical cancer; Oral contraceptives; File drawer problem; Data augmentation; Bayesian model.

scientific journals (Dickersin, Min and Meinert, 1992). Easterbrook et al. (1991) suggests that statistical significance is a major determining factor of publication. Some researchers (e.g. students with Masters' or Ph.D. theses) may not submit a nonsignificant result for publication, and editors may not publish nonsignificant results even if they are submitted (British Medical Journal, 1983). Therefore, there may be a non-representative proportion of significant studies in the scientific literature. Moreover, studies which are of poor quality may be differentially published; a 'high quality' paper, even if it does not exhibit statistical significance, may well be accepted where a 'low quality' and insignificant result will fare less well (as perhaps it should).

This becomes problematic for a meta-analysis whose data comes solely from the published scientific literature. A non-representative proportion of significant studies in the literature will lead to a non-representative proportion of significant studies in the meta-analysis data set. A standard meta-analysis model will then result in a conclusion biased toward significance. This phenomenon is known as 'publication bias', or the 'file-drawer problem' (Iyengar and Greenhouse, 1988).

There are several methods which may be used to detect and estimate the impact of publication bias on a meta-analysis. Funnel plots and related graphical methods are useful for determining visually the existence of missing studies (Light and Pillemer, 1984; Vandenbroucke, 1988). There are also several quantitative methods which estimate the number of missing studies and explicitly model the probability of publication (Dear and Begg, 1992; Hedges, 1992; Paul, 1995). Many of these methods are limited to factors such as significance level or effect size. They often fail to account for study quality, which may be an important consideration in determining publication.

In this paper we extend a Bayesian method described in Givens, Smith and Tweedie (1997), which covers the situation where publication is due solely to significance levels, to a stratified model which allows for other aspects to be taken into account. Estimation uses the data augmentation principle (Tanner, 1991); specifically, we construct an algorithm which imputes latent sets of missing studies into a meta-analysis, in accordance with the probabilities that they are missing in given significance ranges, or quality ranges, or the like.

We then apply this model to a set of studies collected by Delgado-Rodriguez et al. (1992), who examine studies on the association between the use of oral contraceptives and cervical cancer. We investigated the effect of publication bias due to significance levels alone in LaFleur et al., (1996). Here we are able to consider the effect of the quality assessments made by Delgado-Rodriguez et al. (1992), and show that this makes a considerable difference in the final evaluation of the data set. Specifically, after allowing for these biases the estimated relative risk of cervical cancer from oral contraceptive use is reduced considerably, indicating that a suppression bias against studies of insignificance or poor quality may seriously distort the results of an ordinary meta-analysis of these data.

2 Adjusting for Publication Bias

2.1 Hierarchical Bayes Models for Observed and Latent Data

In what follows we shall consider situations as in Givens et al. (1997) where the measure of association in the meta-analysis is the relative risk (RR). The RR is commonly used to measure the association between a potentially toxic agent and a disease endpoint (Cooper and Hedges, 1994), although our work could equally apply to other measures such as risk differences or mortality ratios. For distributional reasons, we work on a log scale, and so let $\Delta = \log RR$. If $\Delta = 0$ then exposure to the agent is associated with no change in health risk; $\Delta > 0$ implies that exposure is associated with an increased health risk and $\Delta < 0$ implies that exposure is associated with a health benefit.

The augmentation algorithm we shall use is an extension of that in Givens et al. (1997) to a multi-tier situation; frequentist models in the single tier case were also considered by Dear and Begg (1992), Hedges (1992), and Paul (1995).

We assume that studies belong to s different classes, strata, or groups, which we shall call 'tiers', with n_i observed studies belonging to each tier, $i = 1, \ldots, s$. For our purposes, tiers will refer to the quality classifications of the studies, although clearly they could refer to some other study characteristic, such as national origin of study, or to clusters of data identified on the basis of multivariate data. Furthermore, our models are easily extended to additional hierarchical levels when more structured tier assumptions are appropriate.

The individual studies produce estimates of Δ, say Y_{ij}, for $i = 1, \ldots, s$ and $j = 1, \ldots, n_i$. We let p_{ij} equal the one-sided p-value of the $(i,j)^{th}$ study for testing $\Delta > 0$.

The hierarchical model for observed studies is

$$Y_{ij} = \Delta + \alpha_i + \beta_{j(i)} + \epsilon_{ij}, \tag{1}$$

where $\alpha_i \sim N(0, \eta^2)$ represents heterogeneity between tiers, $\beta_{j(i)} \sim N(0, \tau_i^2)$ represents heterogeneity between studies within tier i, and $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$ represents within-study variability of study (i,j). The α_i, $\beta_{j(i)}$, and ϵ_{ij} are assumed to be mutually independent. We write $\boldsymbol{\sigma^2} = \{\sigma_{ij}^2\}$ and $\boldsymbol{\tau^2} = \{\tau_i^2\}$.

Using the hierarchical model in (1), the likelihood of the observed data $\mathbf{Y} = (Y_{11}, \ldots, Y_{sn_s})$ is

$$p(\mathbf{Y} \mid \Delta, \eta^2, \boldsymbol{\tau^2}, \boldsymbol{\sigma^2}) \propto \prod_{i=1}^{s} \prod_{j=1}^{n_i} \exp\left(-\frac{1}{2} \frac{(Y_{ij} - \Delta)^2}{(\eta^2 + \tau_i^2 + \sigma_{ij}^2)} \right) / \sqrt{\eta^2 + \tau_i^2 + \sigma_{ij}^2}. \tag{2}$$

Although the number of tiers, s, must be fixed in advance in our model, we allow for tiers that contain no observed data: that is, for tiers with $n_i = 0$. However, to fully implement such a situation it is necessary to make specific assumptions about the characteristics of the

studies within any tier for which $n_i = 0$ and about the nature of publication bias within such a tier.

To account for publication bias, we assume that in addition to the n observed studies from the s groups, there are an additional m studies from these same groups which were not observed, due to publication bias. The number m and the relative risks which might have been found from these m studies are unknown and must be estimated. Uncertainty about these estimates must be reflected in the final meta-analysis inference, and we do this by treating them as parameters in a Bayesian analysis.

Let the estimated log relative risks from these missing studies be denoted as Z_{ij} for $i = 1, \ldots, s$ and $j = (n_i + 1), \ldots, (n_i + m_i)$ where m_i is the number of missing studies in tier i and $m = \sum_i m_i$, and let $\mathbf{Z} = \{Z_{ij}\}$. For notational convenience, we will also denote the complete set of estimated log relative risks for all studies, both observed and missing, by $\mathbf{X} = \{X_{ij}\}$ for all i and j, where $X_{ij} = Y_{ij}$ when (i, j) indexes an observed study and $X_{ij} = Z_{ij}$ when (i, j) indexes a missing study.

We assume that the same random effects model in (1) holds for the outcomes of the missing studies, namely

$$Z_{ij} = \Delta + \alpha_i + \beta_{j(i)} + \epsilon_{ij} \tag{3}$$

where $\alpha_i \sim N(0, \eta^2)$, $\beta_{j(i)} \sim N(0, \tau_i^2)$, and $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$ are mutually independent. Note that now $\boldsymbol{\sigma^2}$ includes the variances of the latent studies as well as those of the observed studies.

2.2 A Model for Selection Bias

There are various selection criteria that one might consider when trying to model publication bias. Following Hedges (1992), Dear and Begg (1992) and Givens et al. (1997), we assume that within each tier the selection mechanism for a study is based on the study's p-value for rejecting the null hypothesis that $\Delta \leq 0$ in favor of the alternative hypothesis $\Delta > 0$. This mechanism is compatible with the widely believed possibility that statistically significant studies are more likely to be published than insignificant studies; but here we will also assume that the selection might be based on the tier in which the study falls, so that for example, a non-significant high quality study may be accepted with a higher probability than a non-significant low quality study.

To make this dependence explicit, we consider a partition of the unit interval into c interval segments, say $I_1, \ldots, I_c$. A p-value from any study must fall into exactly one of these intervals, and we assume that publication is governed by the probabilities

$$w_i^k = \Pr[\text{a tier } i \text{ study with } p\text{-value in } I_k \text{ is published}].$$

We let $\mathbf{w} = \{w_i^k\}$ for all i and k; n_i^k be the number of tier i studies observed with p-values in I_k; and m_i^k be the number of missing tier i studies with (unobserved) p-values

in I_k. Henceforth, superscripts refer to p-value interval and subscripts refer to tiers. Then $n_i = \sum_k n_i^k$ and $m_i = \sum_k m_i^k$, where the n_i^k are known and the m_i^k are unknown. Let $\mathbf{m} = \{m_i^k\}$ for all i and k.

We adopt the following model for the number of tier i missing studies with p-values in I_k:

$$m_i^k \mid \mathbf{w} \sim \begin{cases} \text{Negative Binomial}\left(n_i^k, w_i^k\right) & \text{if } n_i^k > 0, \\ \text{Logarithmic}\left(w_i^k\right) & \text{if } n_i^k = 0, \end{cases} \tag{4}$$

where U has a logarithmic (λ) distribution if the probability mass function of U is $p(U = u \mid \lambda) = (1-\lambda)^u/(-u \log \lambda)$ on $u = 1, 2, \ldots$ (see Mood, Graybill and Boes, 1963). Each tier of studies is treated independently.

Note that (4) depends on knowing the weight vector $\mathbf{w}$. In the single tier context, Hedges (1992) and Dear and Begg (1992) present a maximum likelihood method for estimating the w_i^k from a meta-analysis dataset, but we pursue a Bayesian approach in this paper following that of Givens et al. (1997).

We do not include here a systematic investigation of the case when an entire tier is unobserved ($n_i = 0$ for some i).

2.3 The Complete Data Likelihood and Conditional Posterior Distributions

The observed data are the outcomes, $\mathbf{Y}$, of the observed studies, and we condition on the numbers of observed studies in each tier, n_i. Using (2) we write the likelihood for the observed data under this conditioning as

$$p(\mathbf{Y} \mid \Delta, \eta^2, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2) \propto \prod_{i=1}^{s} \prod_{j=1}^{n_i} \prod_{k=1}^{c} \mathbf{1}_{\{p_{ij} \in I_k\}} \frac{\exp\left(-\frac{1}{2}\frac{(Y_{ij}-\Delta)^2}{(\eta^2+\tau_i^2+\sigma_{ij}^2)}\right)}{\sqrt{\eta^2+\tau_i^2+\sigma_{ij}^2}}. \tag{5}$$

The latent data are the outcomes, $\mathbf{Z}$, of the unobserved studies, and the numbers of such studies in each tier, $\mathbf{m}$. Together, $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{m}$ comprise the complete data $(\mathbf{X}, \mathbf{m})$. At times, it is convenient to consider the latent data $(\mathbf{Z}, \mathbf{m})$ as nuisance parameters to be marginalized out of final inference about Δ.

Under models (1) and (3) for the observed and missing studies, the partial conditional likelihood for the study outcomes is

$$p(\mathbf{X} \mid \Delta, \eta^2, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2, \mathbf{m}) \propto \prod_{i=1}^{s} \prod_{j=1}^{n_i+m_i} \prod_{k=1}^{c} \mathbf{1}_{\{p_{ij} \in I_k\}} \frac{\exp\left(-\frac{1}{2}\frac{(X_{ij}-\Delta)^2}{(\eta^2+\tau_i^2+\sigma_{ij}^2)}\right)}{\sqrt{\eta^2+\tau_i^2+\sigma_{ij}^2}}. \tag{6}$$

We stress that (6) is conditional on knowing $\mathbf{m}$. Treating $\mathbf{m}$ as unknown latent data and conditioning instead on the parameter $\mathbf{w}$, the complete data likelihood is

$$p(\mathbf{X}, \mathbf{m} \mid \Delta, \eta^2, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2, \mathbf{w}) \;\propto\; p(\mathbf{X} \mid \Delta, \eta^2, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2, \mathbf{m}) \prod_{i=1}^{s} Q_i(\mathbf{w}, \mathbf{m}), \tag{7}$$

where

$$Q_i(\mathbf{w}, \mathbf{m}) = \prod_{k=1}^{c} \left[\left[\begin{pmatrix} n_i^k + m_i^k - 1 \\ m_i^k \end{pmatrix} \left(w_i^k\right)^{n_i^k} \left(1 - w_i^k\right)^{m_i^k} \mathbf{1}_{\{m_i^k \in \{0,1,2,\ldots\}\}} \right]^{\mathbf{1}_{\{n_i^k > 0\}}} \times \left[\frac{-(1-w_i^k)^{m_i^k}}{(m_i^k \log w_i^k)} \mathbf{1}_{\{m_i^k \in \{1,2,\ldots\}\}} \right]^{\mathbf{1}_{\{n_i^k = 0\}}} \right]$$

(cf. Givens et al. 1997). In our Bayesian analysis, we adopt independent prior distributions $p(\Delta)$, $p(\eta^2)$, $p(\boldsymbol{\tau}^2)$, $p(\boldsymbol{\sigma}^2)$, $p(\mathbf{w})$, and $p(\mathbf{Z})$. Since $\mathbf{m}$ and $\mathbf{w}$ are related through (4), no separate prior for $\mathbf{m}$ is needed since its conditional distribution is known once $\mathbf{w}$ is known. Degenerate priors are allowed, and for example, we take σ_{ij}^2 to be known for individual observed studies.

Note that (7) is an extension of (2) but now includes parameters $\mathbf{w}$ which can be used to model publication bias. Hedges (1992) considered only the observed data and used an observed data likelihood of a form analogous to (7). For identifiability, he assumed that the probability of publication equalled 1.0 in the most significant p-value interval and considered maximum likelihood estimation only up to a multiplicative constant. Following this approach, we also scale the w_i^k, as shown below, although we do not assume that the maximum publication probability corresponds to the most significant p-value interval. However such a monotonicity constraint is straightforward to enforce in our context, and in Section 4, we discuss the effect on our inferences of constraining the w_i^k to be monotonically increasing as the p-value decreases.

Using prior distributions and the complete data likelihood, we derive the univariate conditional posterior distributions. We use $p(q \mid \cdot)$ to represent the conditional posterior distribution of q given all other parameters. The univariate conditionals for Δ, η^2 and $\boldsymbol{\tau}^2$ are then easily found from (7) as

$$p(\Delta \mid \cdot) \;\propto\; \frac{p(\Delta)}{A(\Delta)} \prod_{i=1}^{s} \prod_{j=1}^{n_i+m_i} \exp\left(-\frac{1}{2} \frac{(X_{ij} - \Delta)^2}{\left(\eta^2 + \tau_i^2 + \sigma_{ij}^2\right)} \right), \tag{8}$$

$$p(\eta^2 \mid \cdot) \;\propto\; \frac{p(\eta^2)}{A(\eta^2)} \prod_{i=1}^{s} \prod_{j=1}^{n_i+m_i} \left[\exp\left(-\frac{1}{2} \frac{(X_{ij} - \Delta)^2}{\left(\eta^2 + \tau_i^2 + \sigma_{ij}^2\right)} \right) \Big/ \sqrt{\eta^2 + \tau_i^2 + \sigma_{ij}^2} \right], \tag{9}$$

$$p(\boldsymbol{\tau}^2 \mid \cdot) \;\propto\; \frac{p(\boldsymbol{\tau}^2)}{A(\boldsymbol{\tau}^2)} \prod_{i=1}^{s} \prod_{j=1}^{n_i+m_i} \left[\exp\left(-\frac{1}{2} \frac{(X_{ij} - \Delta)^2}{\left(\eta^2 + \tau_i^2 + \sigma_{ij}^2\right)} \right) \Big/ \sqrt{\eta^2 + \tau_i^2 + \sigma_{ij}^2} \right], \tag{10}$$

where here and below A is a normalizing function $A(\Delta, \eta^2, \boldsymbol{\tau^2}, \boldsymbol{\sigma}^2, \mathbf{w})$, which we write in varying notation to emphasize its dependence on each parameter of interest.

The conditional density for the pair $(\mathbf{Z}, \boldsymbol{\sigma^2})$ is also straightforward:

$$p(\mathbf{Z}, \boldsymbol{\sigma^2} \mid \cdot) \quad \propto \quad \frac{p(\mathbf{Z}, \boldsymbol{\sigma^2})}{A(\boldsymbol{\sigma^2})} \prod_{i=1}^{s} \prod_{j=1}^{n_i+m_i} \prod_{k=1}^{c} \frac{\exp\left(-\frac{1}{2}\frac{(X_{ij}-\Delta)^2}{\left(\eta^2+\tau_i^2+\sigma_{ij}^2\right)}\right)}{\sqrt{\eta^2+\tau_i^2+\sigma_{ij}^2}} \mathbf{1}_{\{p_{ij} \in I_k\}}. \tag{11}$$

We consider $\mathbf{Z}$ and $\boldsymbol{\sigma^2}$ in a bivariate form since for any new study the values of Z_{ij} and σ_{ij}^2 must be chosen to ensure the constraint $\mathbf{1}_{\{p_{ij} \in I_k\}}$ is satisfied.

In practice, for each (i, k) it is efficient to draw m_i^k missing study variances, σ_{ij}^2, from $p(\boldsymbol{\sigma^2} \mid \cdot)$ with no constraint on the outcomes or p-values of the missing studies, then simulate the m_i^k missing study p-values, p_{ij}, uniformly on I_k, and finally calculate the corresponding $Z_{ij} = \sigma_{ij}\Phi^{-1}(p_{ij})$.

Since we have no prior on $\mathbf{m}$ as discussed above, its conditional distribution is merely $p(\mathbf{m} \mid \mathbf{w}) \propto \prod_{i=1}^{s} Q_i(\mathbf{w}, \mathbf{m})$.

Finally, because of the scaling we impose on the weights $\mathbf{w}$, the posterior conditional distribution of $\mathbf{w}$ is given by

$$p(\mathbf{w} \mid \cdot) \propto \frac{p(\mathbf{w})}{A(\mathbf{w})} \prod_{i=1}^{s} p_1(\mathbf{w}_i \mid \cdot), \tag{12}$$

where for any fixed i, $\mathbf{w}_i = (w_i^1, \ldots, w_i^c)$ and $p_1(\mathbf{w}_i \mid \cdot)$ is the conditional probability density function of $\mathbf{w}_i \times \max_k w_i^k$ that results when $\mathbf{w}_i$ has density proportional to $Q_i(\mathbf{w}, \mathbf{m})$.

The posterior for Δ given here has a form which prohibits an analytical solution. Instead, numerical techniques must be used, and we use a Gibbs sampling (Geman and Geman, 1984) strategy to obtain approximate samples from the desired posterior distribution. Gibbs techniques can be used to obtain a sample from a desired target distribution by simulating realizations from a Markov chain whose stationary distribution is equal to the target.

Here, the target distribution is the joint posterior distribution implied by the priors and complete data likelihood for our model. This target is then marginalized to obtain the observed data posterior, from which inference is drawn. By sequentially sampling from the univariate conditional posterior distributions of the parameters, we can simulate approximate realizations from the joint posterior. The distribution of sampled points converges to the posterior distribution as iterations increase because the conditionals in equations (8)–(12) assign positive probability to the entire parameter space that may be supported by the posterior, thus producing an aperiodic irreducible Markov chain (Smith and Roberts, 1993). We use the quantile method to obtain posterior interval estimates. The details of the implementation, sampling, burn-in, simulation length, and subsampling are similar to those in Givens et al. (1997) and we do not repeat them here.

3 Performance on Simulated Data Sets

3.1 Simulated Data Sets

Performance of the single-tier model in Givens et al. (1997) was shown to be satisfactory based on a number of simulation scenarios. Here we investigate the performance of the method on data sets with more than one tier (i.e. where $s > 1$).

We consider studies that fall into $s = 3$ tiers, and evaluate performance when (a) there is no overall association ($\Delta = 0, RR = 1.0$), and (b) when there is a positive association ($\Delta = 0.4, RR \approx 1.5$). We are also interested in the effect that heterogeneity within tiers has on the results. To that end, we simulated data sets where within-tier studies were homogeneous ($\tau_1^2 = \tau_2^2 = \tau_3^2 = 0$) and where there were differences in the heterogeneity of each tier ($\tau_1^2 = 0$, $\tau_2^2 = 0.3$, $\tau_3^2 = 0.6$). In tabled results, we denote these as "Zero" and "Nonzero", respectively. We drew the individual study variances, $\hat{\sigma}_{ij}^2$, from (i) $\hat{\sigma}_{ij}^2 \sim$ Gamma (shape = 3, mean = 0.33), for all i, j, and (ii) $\hat{\sigma}_{1j}^2 \sim$ Gamma (shape = 3, mean = 0.1), $\hat{\sigma}_{2j}^2 \sim$ Gamma (shape = 3, mean = 0.33), and $\hat{\sigma}_{3j}^2 \sim$ Gamma (shape = 3, mean = 0.6). We denote these as "Ident" and "Diff", respectively. In all simulated data sets we set η^2 equal to zero for simplicity.

We simulated 60 studies for each of the three tiers and each of the eight parameter combinations above. We split the studies' p-values into three intervals, which were the same for all tiers:

$$I_1 = [0, 0.10], \quad I_2 = (0.10, 0.50], \quad I_3 = (0.50, 1].$$

It should be noted that we calculated one-sided p-values of the hypothesis test corresponding to $\Delta > 0$ using only X_{ij} and $\hat{\sigma}_{ij}^2$. In practice, one would not have knowledge of the values of the other variance components when calculating p-values.

In each tier, we applied one of three different schemes of suppression, $(1 - w_i^k)$ to the studies within the tier. These were: no suppression ("N"), light suppression ("L") and heavy suppression ("H"), given respectively for any i by

$$\begin{aligned}
&\text{N: } w_i^1 = w_i^2 = w_i^3 = 1.00 \\
&\text{L: } w_i^1 = 1.00 \quad w_i^2 = 0.75 \quad w_i^3 = 0.50 \\
&\text{H: } w_i^1 = 1.00 \quad w_i^2 = 0.50 \quad w_i^3 = 0.25.
\end{aligned}$$

To create overall suppression schemes, we combined these choices in three ways: "NNN", "LLH" and "LHH", where, for example, "LLH" is

$$\begin{aligned}
&\text{Tier 1: } w_1^1 = 1.00 \quad w_2^2 = 0.75 \quad w_2^3 = 0.50 \\
&\text{Tier 2: } w_2^1 = 1.00 \quad w_2^2 = 0.75 \quad w_2^3 = 0.50 \\
&\text{Tier 3: } w_3^1 = 1.00 \quad w_3^2 = 0.50 \quad w_3^3 = 0.25
\end{aligned}$$

We performed two Bayesian meta-analyses for each simulated data set. The first was a standard Bayesian meta-analysis which does not adjust for publication bias; we performed the second meta-analysis using the augmentation method in Section 2. Both meta-analyses had identical priors. The prior on Δ was flat; the priors on τ_i^2 for all i were exponentials with mean 3.0; the prior on η^2 was an exponential with mean 0.3; the priors on the within-study population variances, σ_{ij}^2, were degenerate at $\hat{\sigma}_{ij}^2$, the observed study variances; the priors on the missing within-study population variances were empirical priors based on the observed study variances. The priors on the missing studies' log relative risks were improper (uniform on the real line).

We consider two methods of comparing a standard meta-analysis with a publication bias adjusted meta-analysis. The first is to compare the posterior credibility intervals for Δ for both methods. The second compares the square root of expected squared-error loss, or risk, given by

$$\mathcal{R}(\Delta^\star) = \sqrt{\int_{-\infty}^{\infty} p(\Delta \mid \cdot)(\Delta - \Delta^\star)^2 \, d\Delta},$$

where $p(\Delta \mid \cdot)$ is the marginal posterior of Δ and $\Delta^\star$ is the true value of the simulated data set's log RR, which is either 0 or 0.4 in our simulations.

We define $\mathcal{R}_0(\Delta^\star)$ as the square root of risk of a standard Bayesian meta-analysis and $\mathcal{R}_{\mathrm{A}}(\Delta^\star)$ as the square root of risk of the augmentation method described in Section 2. For each simulated data set, we calculate the risk ratio $\mathcal{R}_{\mathrm{A}}(\Delta^\star)/\mathcal{R}_0(\Delta^\star)$. Therefore, risk ratios less than 1.0 favor the augmentation method and ratios greater than 1.0 favor the standard Bayesian meta-analysis.

3.2 Results of the Simulations

Table 1 shows the results of the risk calculations. Table 2 shows the posterior mean and a 95% credible interval (CI) for Δ in each parameter/suppression combination when $\Delta = 0$. Table 3 gives these results for $\Delta = 0.4$.

Table 1 indicates that the posterior of Δ resulting from a standard Bayes meta-analysis is concentrated near the true Δ when there are no studies missing. In those cases where publication bias is present, the meta-analysis adjusted for publication bias decreases the ratio of the risks. In those data sets where there is no suppression, risk ratio favors a standard Bayesian meta-analysis over the augmentation algorithm. This is partly due to the augmentation algorithm generally underestimating the true log RR in these cases and partly due to the fact that the credible intervals for the standard Bayesian meta-analysis are more narrow than those of the augmentation algorithm. However, Table 1 shows that $\mathcal{R}_{\mathrm{A}}(\Delta^\star)$ is more consistent across the suppression schemes than $\mathcal{R}_0(\Delta^\star)$. The $\mathcal{R}_0(\Delta^\star)$ posterior risk increases as more studies are suppressed, which is the expected effect of publication bias.

Δ	Suppr.	$\boldsymbol{\tau^2}$	$\boldsymbol{\sigma^2}$	$\mathcal{R}_{\mathrm{A}}(\Delta^\star)$	$\mathcal{R}_0(\Delta^\star)$	$\mathcal{R}_{\mathrm{A}}(\Delta^\star)/\mathcal{R}_0(\Delta^\star)$
0	NNN	Zero	Ident.	0.11	0.05	2.30
0	NNN	Zero	Diff.	0.08	0.04	2.21
0	NNN	Nonzero	Ident.	0.10	0.05	1.79
0	NNN	Nonzero	Diff.	0.11	0.05	1.96
0	LLH	Zero	Ident.	0.10	0.12	0.83
0	LLH	Zero	Diff.	0.09	0.11	0.81
0	LLH	Nonzero	Ident.	0.14	0.21	0.64
0	LLH	Nonzero	Diff.	0.10	0.12	0.86
0	LHH	Zero	Ident.	0.10	0.16	0.60
0	LHH	Zero	Diff.	0.09	0.15	0.62
0	LHH	Nonzero	Ident.	0.15	0.29	0.53
0	LHH	Nonzero	Diff.	0.11	0.18	0.62
0.4	NNN	Zero	Ident.	0.19	0.08	2.29
0.4	NNN	Zero	Diff.	0.25	0.08	2.94
0.4	NNN	Nonzero	Ident.	0.27	0.09	3.02
0.4	NNN	Nonzero	Diff.	0.25	0.08	3.01
0.4	LLH	Zero	Ident.	0.12	0.24	0.49
0.4	LLH	Zero	Diff.	0.14	0.12	1.16
0.4	LLH	Nonzero	Ident.	0.15	0.24	0.64
0.4	LLH	Nonzero	Diff.	0.13	0.17	0.75
0.4	LHH	Zero	Ident.	0.11	0.27	0.41
0.4	LHH	Zero	Diff.	0.12	0.15	0.81
0.4	LHH	Nonzero	Ident.	0.15	0.30	0.49
0.4	LHH	Nonzero	Diff.	0.11	0.22	0.49

Table 1: Risk calculations of a meta-analysis which fails to adjust for publication bias, $\mathcal{R}_0(\Delta^\star)$, and the augmentation method, $\mathcal{R}_{\mathrm{A}}(\Delta^\star)$, on simulated data in three tiers.

	Suppr.	τ^2	σ^2	Post. Mean Δ	Lower 95% Bnd.	Upper 95% Bnd.
Aug	NNN	Zero	Ident.	-0.08	-0.24	0.06
Std	NNN	Zero	Ident.	-0.02	-0.10	0.06
Aug	NNN	Zero	Diff.	-0.05	-0.19	0.09
Std	NNN	Zero	Diff.	-0.01	-0.08	0.06
Aug	NNN	Nonzero	Ident.	-0.03	-0.20	0.14
Std	NNN	Nonzero	Ident.	0.02	-0.07	0.11
Aug	NNN	Nonzero	Diff.	-0.08	-0.24	0.08
Std	NNN	Nonzero	Diff.	-0.04	-0.12	0.04
Aug	LLH	Zero	Ident.	0.00	-0.17	0.19
Std	LLH	Zero	Ident.	0.09	-0.01	0.21
Aug	LLH	Zero	Diff.	0.01	-0.12	0.18
Std	LLH	Zero	Diff.	0.09	0.01	0.18
Aug	LLH	Nonzero	Ident.	0.08	-0.09	0.25
Std	LLH	Nonzero	Ident.	0.18	0.06	0.29
Aug	LLH	Nonzero	Diff.	0.03	-0.13	0.20
Std	LLH	Nonzero	Diff.	0.09	-0.01	0.21
Aug	LHH	Zero	Ident.	0.04	-0.11	0.20
Std	LHH	Zero	Ident.	0.13	-0.01	0.26
Aug	LHH	Zero	Diff.	0.05	-0.07	0.19
Std	LHH	Zero	Diff.	0.12	-0.01	0.23
Aug	LHH	Nonzero	Ident.	0.11	-0.02	0.27
Std	LHH	Nonzero	Ident.	0.24	0.09	0.37
Aug	LHH	Nonzero	Diff.	0.06	-0.08	0.21
Std	LHH	Nonzero	Diff.	0.12	-0.02	0.25

Table 2: Results of a meta-analysis which fails to adjust for publication bias ("Std") and the augmentation method ("Aug") on simulated data in three tiers. The mean and 95% credible limits come from the marginal posterior of Δ. The true Δ is 0.00.

	Suppr.	τ^2	σ^2	Post. Mean Δ	Lower 95% Bnd.	Upper 95% Bnd.
Aug	NNN	Zero	Ident.	0.28	0.14	0.41
Std	NNN	Zero	Ident.	0.42	0.32	0.51
Aug	NNN	Zero	Diff.	0.23	0.10	0.35
Std	NNN	Zero	Diff.	0.37	0.28	0.45
Aug	NNN	Nonzero	Ident.	0.22	0.09	0.38
Std	NNN	Nonzero	Ident.	0.38	0.28	0.48
Aug	NNN	Nonzero	Diff.	0.23	0.10	0.36
Std	NNN	Nonzero	Diff.	0.38	0.28	0.47
Aug	LLH	Zero	Ident.	0.39	0.25	0.55
Std	LLH	Zero	Ident.	0.53	0.42	0.64
Aug	LLH	Zero	Diff.	0.33	0.21	0.45
Std	LLH	Zero	Diff.	0.46	0.36	0.55
Aug	LLH	Nonzero	Ident.	0.36	0.17	0.52
Std	LLH	Nonzero	Ident.	0.53	0.41	0.65
Aug	LLH	Nonzero	Diff.	0.36	0.23	0.50
Std	LLH	Nonzero	Diff.	0.49	0.38	0.59
Aug	LHH	Zero	Ident.	0.40	0.28	0.54
Std	LHH	Zero	Ident.	0.56	0.44	0.67
Aug	LHH	Zero	Diff.	0.34	0.23	0.46
Std	LHH	Zero	Diff.	0.48	0.38	0.58
Aug	LHH	Nonzero	Ident.	0.39	0.18	0.57
Std	LHH	Nonzero	Ident.	0.57	0.43	0.69
Aug	LHH	Nonzero	Diff.	0.38	0.26	0.50
Std	LHH	Nonzero	Diff.	0.52	0.39	0.63

Table 3: Results of a meta-analysis which fails to adjust for publication bias ("Std") and the augmentation method ("Aug") on simulated data in three tiers. The mean and 95% credible limits come from the marginal posterior of Δ. The true Δ is 0.40.

In those data sets where $\Delta = 0$, the suppression of studies causes the standard meta-analysis to elevate the estimate of Δ. As more studies are suppressed, the bias of the standard meta-analysis increases. However, in all cases, the augmentation algorithm adjusted the mean log RR toward 0 in those data sets where studies were suppresed. Four of the standard meta-analyses failed to include 0 in their CI's under this choice of parameters, and yet all of the posteriors of Δ using the standard Bayes meta-analysis exhibited a low degree of variability. The augmentation algorithm's credible intervals are wider than those of the standard meta-analysis, since the latent studies tend to be imputed into the nonsignificant p-value intervals. The augmentation algorithm's credible intervals, however, always included the true Δ.

In those data sets where $\Delta = 0.4$, the overall number of studies suppressed was smaller compared to those where $\Delta = 0$ since there were fewer studies in I_2 and I_3. The credible intervals of standard Bayesian meta-analyses failed to cover the true Δ in four cases. The widths of these credible intervals were approximately the same as those of the standard Bayesian meta-analysis in Table 2. The credible intervals of the meta-analyses adjusted for publication bias failed to cover the truth in three data sets where there was no suppression. The augmentation algorithm performed remarkably well in those data sets with publication bias.

The results in Tables 1 to 3 are derived from a simulation design that is biased against the data augmentation approach for the following reason. To avoid a confounding influence on the results, we retained the same priors on suppression rate in all cases. This prior for $\mathbf{w}$ covered the same wide range of suppression rates in all trials, with the understandable result of degraded performance in the NNN trials. In real applications, we would reflect our *a priori* belief of low suppression (based on diagnostics such as funnel plots) through less broad priors on $\mathbf{w}$. If we had designed simulations to mimic this behavior, thus varying the $\mathbf{w}$ prior across simulations, the performance of the augmentation approach would have improved.

3.3 Coverage of Standard Bayes Credible Intervals

We note that in Table 2 and 3, the credible intervals for the standard Bayesian meta-analyses are more narrow than those of the augmentation algorithm's. Since they fail to include the true value of Δ in several data sets with publication bias, we suspect that these intervals' coverage is less than 95%.

We chose one parameter combination from Table 2 ($\Delta = 0$, "'Nonzero", "Ident.") and examined the coverage of the credible intervals for a standard Bayesian meta-analysis under LLH suppression. We simulated 12 independent data sets under this scheme and performed a standard Bayes meta-analysis with identical priors on each one. If π is the coverage probability of each meta-analysis, 12 independent Bernoulli trials can distinguish between $H_0 : \pi \geq 0.95$ versus $H_1 : \pi = 0.5$ with approximately 95% power.

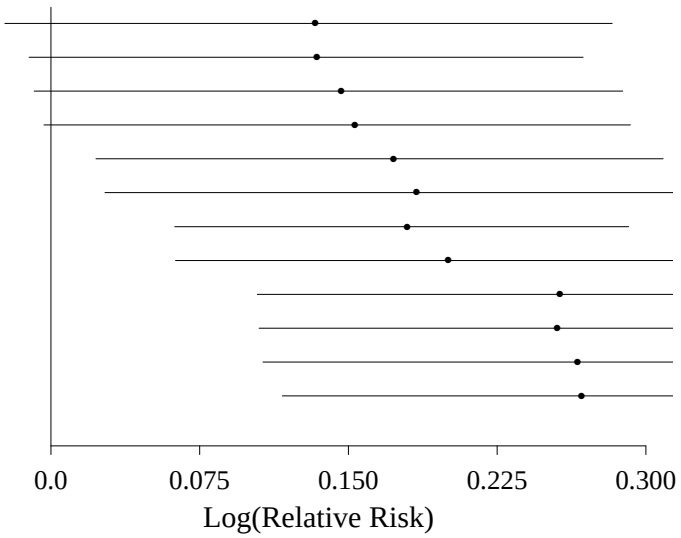

Figure 1: Coverages of unadjusted meta-analyses on twelve independent data sets with parameter combination ($\Delta = 0$, "'Nonzero", "Ident.", LLH). Eight of the twelve credible intervals failed to cover $\Delta = 0$.

The results of these 12 meta-analyses appear in Figure 1. Only four of the twelve covered the true $\Delta = 0$. The p-value for the test $H_0 : \pi \geq 0.95$ is roughly 2×10^{-8}. Therefore we conclude that the coverage of credibility intervals of a standard Bayesian meta-analysis under this parameter combination is likely less than 95%. Our experience in a variety of simulations has been that coverage rates for the standard Bayesian analysis can be quite poor. The coverage rates are generally much better for the augmentation approach, despite the fact that for NNN it has poor coverage for $\Delta = 0.4$.

4 Cervical Cancer and Oral Contraceptive Use

4.1 Data and One-Tier Meta-analyses

In this section, we apply the the meta-analysis model described in Section 2 to a data set with tier structure and potential publication bias.

The studies that we will consider were collected by Delgado-Rodriguez et al. (1992) using Medline and Indice Medico Español searches, and relate to the association between the use of oral contraceptives and cervical cancer. There has been an abundance of research on this

	Group I	Group II
Dysplasia	1.31 (1.24, 1.38)	1.52 (1.27, 1.82)
	(20 studies)	(6 studies)
Carcinoma in situ	1.29 (1.18, 1.41)	1.52 (1.31, 1.76)
	(28 studies)	(10 studies)
Invasive cancer	1.13 (0.99, 1.27)	1.21 (1.06, 1.37)
	(16 studies)	(10 studies)

Table 4: RR and 95% CI's of the Delgado-Rodriguez et al. (1992) fixed effects meta-analyses of cervical neoplasia among ever users of oral contraceptives. Two Group I studies failed to report measures of variability.

topic around the world, but the studies' conclusions range from a very strong associations to negative associations, making meta-analysis appropriate.

Delgado-Rodriguez et al. (1992) evaluate 62 published relative risks from 51 published papers. The data set was thought to be complete up to 1990. They classify the results into two groups: Group I is the set of all 62 results, and Group II is a set of 26 'methodologically acceptable' results or higher quality studies, although the rationale for being acceptable is not spelled out in any detail. They also split outcomes into three categories corresponding to indicators of cervical cancer (dysplasia, carcinoma *in situ* and invasive cancer). Table 4 summarizes their findings: these are based on fixed effects meta-analyses on each of the three outcomes, despite finding evidence of heterogeneity among the three populations. Note that the exclusion of 'low quality' studies in Group II substantially increases all estimates of RR.

When the results are stratified by type of cervical cancer, inference based on such small samples becomes questionable. An advantage of meta-analysis is lost when we subdivide: inferential power decreases. On a biological basis, conversely, it may be inappropriate to combine studies done on specific cancer types if we attempt to estimate the association of 'cervical cancer in general' with oral contraceptive use. Nonetheless, we will assume that meta-analysis can validly address the question "Is there an association between cervical cancer and oral contraceptive use?", even if we do not specify cancer types, or other covariates such as brand of oral contraceptives, explicitly. We must however interpret any biological relevance of these meta-analyses with caution, even though the random effects and Bayesian methods are designed to allow for variation due to heterogeneous populations.

In LaFleur et al. (1996) we present fixed effects, random effects and Bayesian random effects meta-analyses on this data set, grouping all three outcomes so that inferential power is increased. These results appear in Table 5. Again note that the inclusion of all Group I studies decreases the estimate of RR, and that the use of random effects or Bayesian models also lowers the estimate in Group I, which one might take as a serious indication of the impact of heterogeneity in this group.

Type of Analysis	Group I	Group II
Overall Fixed Effects	1.30 (1.24, 1.35)	1.37 (1.26, 1.49)
Overall Random Effects	1.15 (1.10, 1.30)	1.38 (1.17, 1.63)
Overall Bayesian Model	1.13 (0.95, 1.34)	1.46 (1.08, 1.94)

Table 5: RR and 95% CI's of LaFleur et al. (1996) meta-analyses of cervical neoplasia among ever users of oral contraceptives.

We first apply the publication bias method using only one-tier to the Group II or 'high quality' studies, and find that there appears to be publication bias present, as is suggested by the funnel plot in Figure 2 where Group II is depicted by solid circles. The posterior RR mean and 95% CI is 1.29 $(1.07, 1.53)$, compared with a standard Bayesian meta-analysis on the same studies and same priors results in a RR and 95% CI of 1.46 (1.08, 1.94). Hence, the publication bias adjustment reduces the estimated excess risk while narrowing the posterior CI.

This analysis uses the following p-value intervals for a one-tier analysis:

$$I_1 = [0, 0.01], \quad I_2 = (0.01, 0.05], \quad I_3 = (0.05, 1];$$

we take the priors on Δ and τ_1^2 as non-informative and the priors on $\{\sigma_{1j}^2\}$ as degenerate at their corresponding study estimates, $\{\hat{\sigma}_{1j}^2\}$, following Givens et al. (1997). The priors on w_1^2 and w_1^3 are Uniform on $[0.5, 1.0]$; we also enforce the monotonicity constraint $1 = w_1^1 \geq w_1^2 \geq w_1^3 \geq 0$.

The number of observed studies in each of the p-value intervals are $n_1^1 = 4, n_1^2 = 7$ and $n_1^3 = 15$, and the posterior means of m_1^1, m_1^2 and m_1^3 are $0, 1.76$ and 8.81 respectively.

4.2 Multi-tier Meta-analysis

There is claimed to be a clear distinction between the quality of studies included in Group II and the studies that are excluded from Group II in Delgado-Rodriguez et al. (1992). We now analyze these data using the multi-tier approach described in Section 2. We are motivated by the assumption that there is sufficient information in the low-quality studies that we should not lightly omit them from a meta-analysis, but that it may be reasonable to assume that the low-quality studies have a different chance of being published compared to the high-quality Group II studies. Grouping the studies into quality tiers allows us to make publication bias adjustments within each tier, with the hierarchical structure of the model reflecting these adjustments in the overall RR.

We group the 62 studies into two tiers: Tier 1 (low-quality studies) will refer to those studies which are in Group I but are excluded from Group II; Tier 2 (high-quality) studies

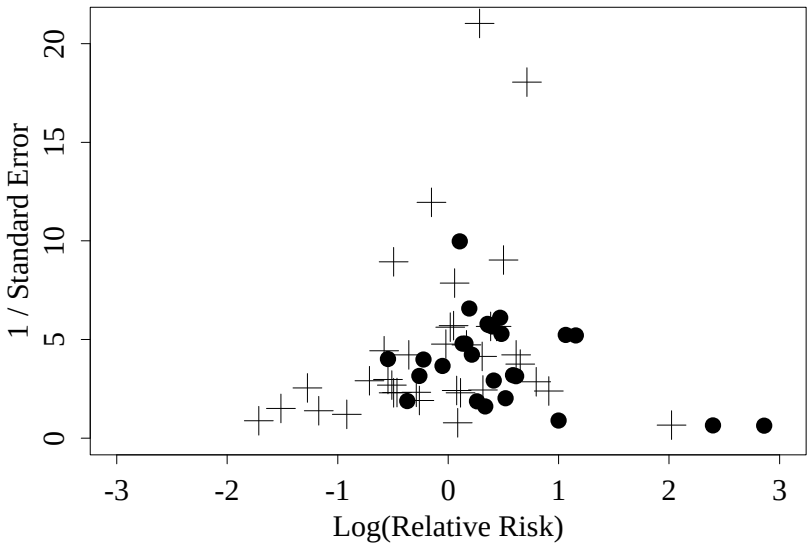

Figure 2: Funnel plot of all 62 studies in Delgado-Rodriguez et al. (1992) that measure the association between oral contraceptive use and cervical cancer. The 26 Group II ('high-quality' Tier 2) studies are denoted by filled circles.

will refer to those studies in Group II. The high-quality studies appear in Figure 2 as dark circles; the low-quality studies appear as crosses. Note that the Tier 1 studies appear to cluster around a lower within-tier log RR than the Tier 2 studies.

We first performed a standard Bayesian meta-analysis on all the Group I studies, treated as a two-tier but complete data set. The estimated RR and 95% CI of this meta-analysis is 1.16 (0.97, 1.34). The posterior means of τ_1^2 and τ_2^2 are 0.11 and 0.09 respectively, and the posterior mean of $\eta^2 = 0.14$. Thus, although this meta-analysis failed to adjust for publication bias, the results suggested a weak, nonsignificant association between cervical cancer and oral contraceptive use. Note that the two-tier analysis is rather similar to the non-tier analysis in Table 5.

In our two-tier analysis, the prior for Δ was a non-informative Normal; the priors for η^2 and $\boldsymbol{\tau^2}$ were non-informative exponentials.

We use the p-value intervals $I_1 = [0, 0.10]$, $I_2 = (0.10, 0.50]$, $I_3 = (0.50, 1]$. and we place the same priors on the publication weights in both tiers: the prior distribution of weights in I_1 and I_2 are Uniform (0.5, 1.0), and the prior distribution of weights in I_3 are Uniform (0.2, 1.0). We again assume that the priors on the observed study variances are degenerate. The priors that we place on the unobserved study variances are broader than the empirical distribution of the observed study variances.

The RR and 95% CI of the augmented meta-analysis that accounts for publication bias was 1.06 (0.78, 1.42). Figure 3 compares the posterior distribution of this meta-analysis with the posterior of a Bayesian meta-analysis that does not adjust for publication bias. We also performed a meta-analysis with the monotonicity constraint enforced within both tiers, and this resulted in a RR and 95% CI of 1.01 (0.80, 1.27). These results appear in Table 6. The posterior means of the τ_1^2, τ_2^2, and η^2 variance components are similar to the standard Bayesian meta-analysis; the meta-analysis with augmentation but with no monotonicity constraint gives the posterior means of τ_1^2, τ_2^2, and η^2 as 0.14, 0.11, and 0.06, respectively.

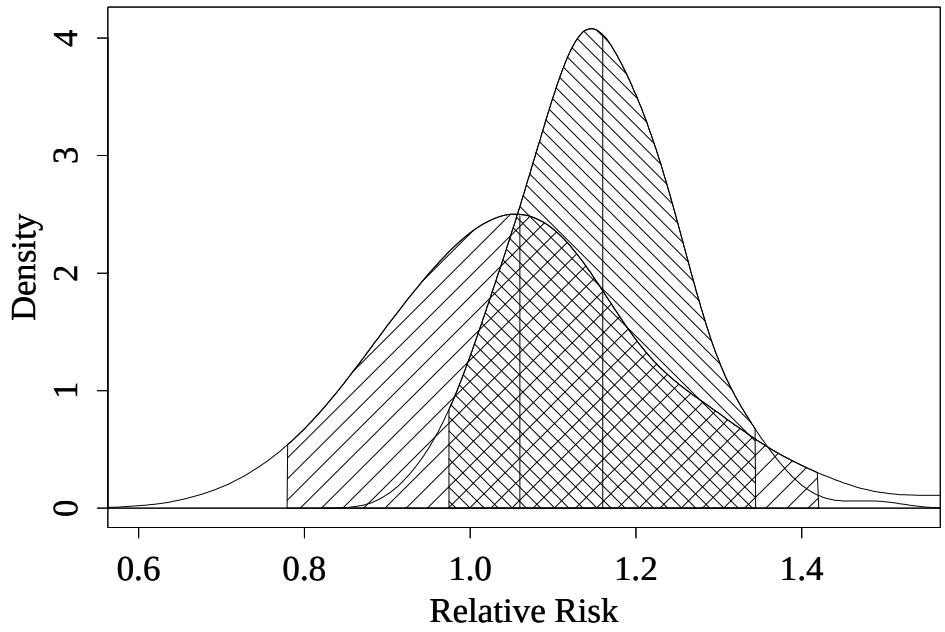

Figure 3: Relative risk posteriors for a two-tier analysis of the cervical cancer data. The leftmost posterior was calculated using data augmentation, and the right one assumes no publication bias. The 95% credible intervals are shaded.

Figures 4 and 5 show relative frequency histograms of the number of missing studies in each interval, with and without monotonicity restrictions. Note that in the former case, the algorithm placed many fewer missing studies into the insignificant range, and only about half as many altogether. This does accord with the intuition one might have from the funnel plot in Figure 2.

It appears rather conclusive, then, that publication bias seems to have the potential to make a considerable impact on the inference in this data set.

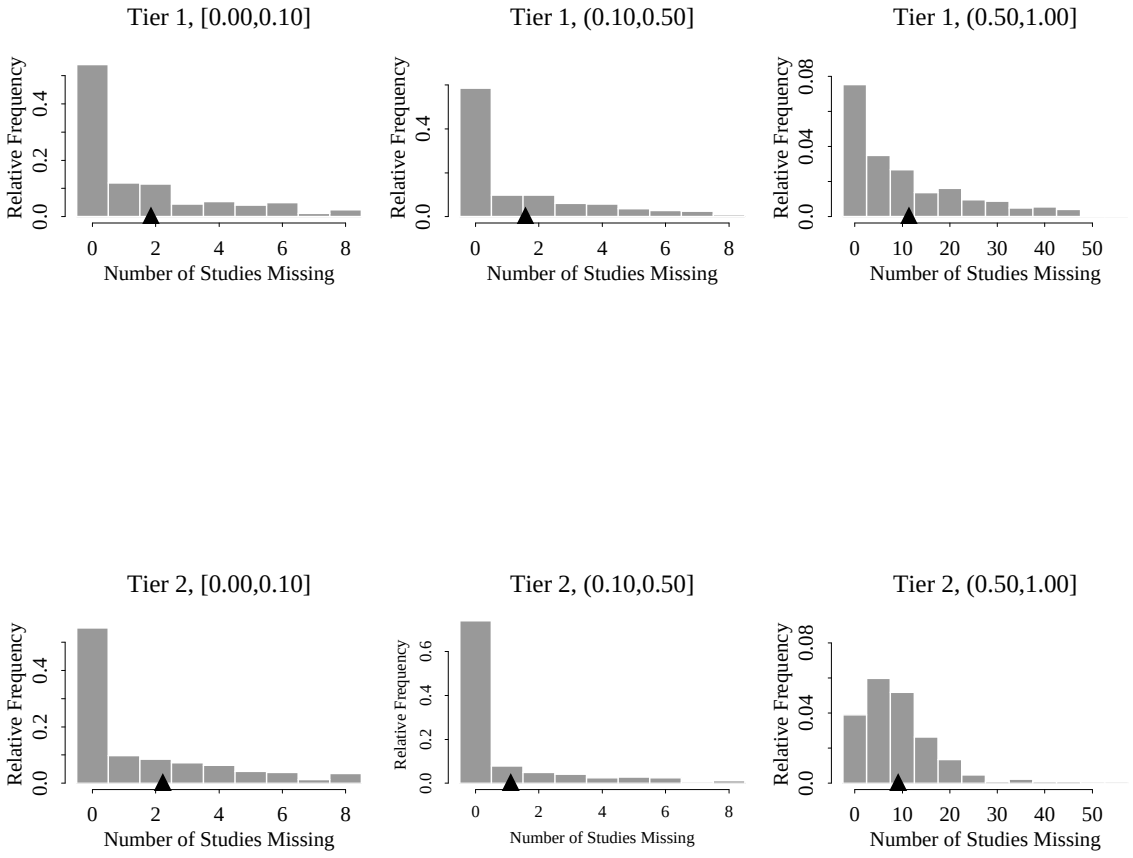

Figure 4: Histograms of the number of missing studies, m_i^k, in each of three p-value divisions. The posterior mean is marked by a dark triangle.

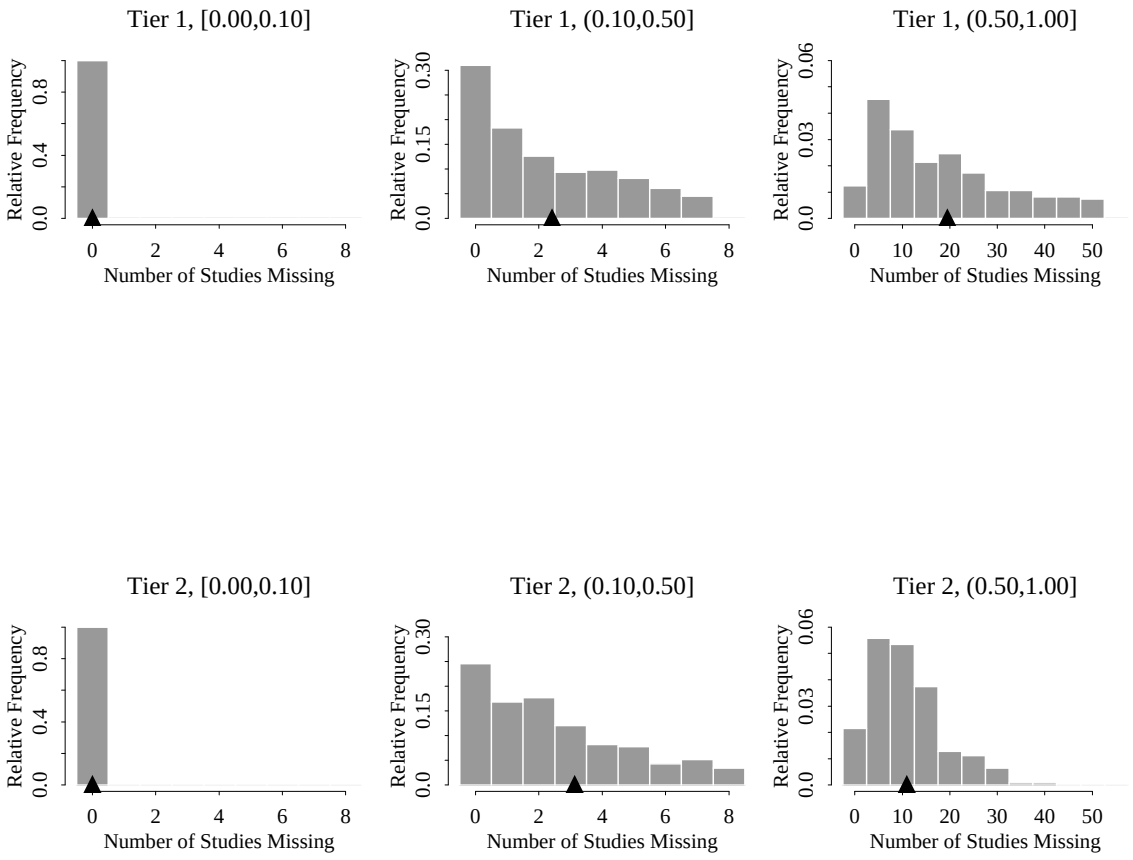

Figure 5: Histograms of the number of missing studies, m_i^k, in each of three p-value divisions. This meta-analysis was performed with a monotonicity constraint on the publication probabilities within both tiers. The posterior mean is marked by a filled triangle.

	Post. RR Mean	95% Cred. Int.
Bayes, no augmentation	1.16	(0.98, 1.37)
Pub. bias adj.	1.06	(0.78, 1.42)
Pub. bias adj., monotone wts.	1.01	(0.80, 1.27)

Table 6: Results of two-tier meta-analyses of the Delgado-Rodriguez et al. cervical cancer data set.

5 Discussion

There have been many schemes proposed to judge the quality of studies for meta-analysis (see, for example, Moher et al., 1995). Longnecker et al. (1988) perform a quality-adjusted meta-analysis by considering quality as a covariate. Other studies group results into quality tiers and then use a weighting scheme to combine them: for example, in the EPA report on environmental tobacco smoke (EPA Review, 1992), whole tiers were either included or excluded from some meta-analyses.

We have found that the multi-tier version of the augmentation method in Givens et al. (1997) provides an alternative quantitative approach which has utility in analyzing meta-analysis data sets with inherent tiers. Our simulation results suggest that the technique performs well in a variety of data sets which suffer from publication bias, although in cases where there is no publication bias, the standard Bayesian analysis with no adjustments yields more satisfactory results.

Care must be taken when either meta-analytic method is used. A standard Bayesian analysis is of course preferable when there are no studies missing in the data set since the augmentation algorithm presented tends to bias relative risk estimates downward. However, the coverage of the standard Bayesian analysis can be far less than the nominal $1 - \alpha$ in the presence of publication bias. The augmentation method performs well in such cases and it adjusts the relative risk credible intervals to recover fully from the bias caused by missing studies. Further sensitivity work can be found in Smith (1997).

Overall, it is clear that the issues of publication bias are real and serious ones. While no method will totally overcome the problem of identifying intrinsically unavailable information, it is apparent that in examples such as that of the cervical cancer-oral contraceptive association, we can make a credible effort to assess and account for the impact of such studies.

Acknowledgements: The authors wish to thank Bonnie LaFleur and Sue Taylor of the University of Colorado Health Sciences Center for bringing the Delgado-Rodriguez et al. (1992) collection to our attention, and for discussions on the epidemiological questions in that application of our methodology.

References

British Medical Journal Editorial Staff (1983). The editor regrets... (editorial). *British Medical Journal* 280:508.

Cooper, H. and Hedges, L. V., eds. (1994). *The Handbook of Research Synthesis.* Russell Sage Foundation, New York, N.Y.

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

Delgado-Rodriguez, M., Sillero-Arenas, M., Martin-Moreno, J., and Galvez-Vargas, R. (1992). Oral contraceptives and cancer of the cervix uteri. *Acta Obstetricia et Gynecologica Scandinavica*, 71:368–376.

Dickersin, K., Min, Y., and Meinert, C. (1992). Factors influencing publication of research results. *Journal of the American Medical Association*, 267:374–378.

Easterbrook, P., Berlin, J., Gopalan, R., and Matthews, D. (1991). Publication bias in clinical research. *Lancet*, 337:867–872.

EPA Review (1992). *Health Effects of Passive Smoking: Assessment of Lung Cancer in Adults and Respiratory Disorders in Children.* National Academy Press, United States EPA, Washington.

Givens, G., Smith, D., and Tweedie, R. (1997). Publication Bias in Meta-analysis: A Bayesian Data-Augmentation Approach to Account for Issues Exemplified in the Passive Smoking Debate. *Statistical Science*, to appear.

Hedges, L. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7:227–236.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-analysis.* Academic Press, New York, N.Y.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3:109–135.

LaFleur, B., Taylor, S., Smith, D., and Tweedie, R. (1996). Bayesian assessment of publication bias in meta-analyses of cervical cancer and oral contraceptives. In *Proceedings of the Joint Statistical Meetings*, Chicago.

Light, R. and Pillemer, D. (1984). *Summing Up: the Science of Reviewing Research.* Harvard Univ. Press.

Longnecker, M. P., Berlin, J. A., Orza, M. J., and Chalmers, T. C. (1988). A meta-analysis of alcohol consumption in relation to risk of breast cancer. *Journal of the American Medical Association*, 260:652–656.

Mengersen, K., Tweedie, R., and Biggerstaff, B. (1995). The impact of method choice in meta-analysis. *Australian Journal of Statistics*, 37:19–44.

Moher, D., Jadad, A., Nichol, G., Penman, M., Tugwell, P., and Walsh, S. (1995). Assessing the quality of randomized controlled trials: an annotated bibliography of scales and

checklist. *Controlled Clinical Trials*, 16:62–73.

Mood, A., Graybill, F., and Boes, D. (1963). *Introduction to the Theory of Statistics.* McGraw-Hill, New York, third edition.

NRC Committee on Applied and Theoretical Statistics (1992). *Combining Information: Statistical Issues and Opportunities for Research.* National Academy Press, Washington.

Olkin, I. (1992). Meta-analysis: methods for combining independent studies. *Statistical Science*, 7:226.

Paul, N. L. (1995). Non-parametric classes of weight functions to model publication bias. Technical Report 622, Department of Statistics, Carnegie-Mellon Univ., Pittsburgh, PA.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1986). *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge.

Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society Series B*, 55:3–23.

Smith, D.D. (1997). *Adjusting for Publication Bias and Quality Effects in Bayesian Random Effects Meta-analysis.* PhD thesis, Colorado State University, Fort Collins, Colorado.

Tanner, M.A. (1991). Tools for Statistical Inference: Observed Data and Data Augmentation Methods. *Lecture Notes in Statistics 67.* J. Berger, S. Feinberg, J. Gani, K. Krickeberg, I. Olkin, B. Singer (eds.). Springer-Verlag, New York.

Thompson, S. and Pocock, S. (1991). Can meta-analyses be trusted? *Lancet*, 338:1127–1130.

Vandenbroucke, J. (1988). Passive smoking and lung cancer: a publication bias? *British Medical Journal*, 296:391–392.