Publication Bias in Meta-Analysis: 
A Bayesian Data-Augmentation Approach to Account for Issues Exemplified in the Passive Smoking Debate

Geof H. Givens, D. D. Smith, and R. L. Tweedie*

Abstract

'Publication bias' is a relatively new statistical phenomenon that only arises when one attempts through a meta-analysis to review all studies, significant or insignificant, in order to provide a total perspective on a particular issue. This has recently received some notoriety as an issue in the evaluation of the relative risk of lung cancer associated with passive smoking, following legal challenges to a 1992 EPA analysis which concluded that such exposure is associated with significant excess risk of lung cancer.

We introduce a Bayesian approach which estimates and adjusts for publication bias. Estimation is based on a data augmentation principle within a hierarchical model, and the number and outcomes of unobserved studies are simulated using Gibbs sampling methods. This technique yields a quantitative adjustment for the passive smoking meta-analysis. We estimate that there may be both negative and positive but insignificant studies omitted, and that failing to allow for these would mean that the estimated excess risk may be overstated by around 30%, both in US studies and in the global collection of studies.

Keywords and phrases: Meta-analysis, publication bias, missing data, data augmentation, Markov chain Monte Carlo, MCMC, Gibbs sampling, environmental tobacco smoke, ETS, passive smoking, lung cancer, file drawer problem.

1 Introduction

1.1 The publication bias problem

Publication bias, or the 'file-drawer problem' (Iyengar and Greenhouse, 1988), is in some sense a new statistical phenomenon which runs counter to the way in which the scientific method has developed over the past century.

One of the key historical contributions of statistical thinking has been a move away from a context where possibly random observations were acceptable, to one where only those results which are 'statistically significant', i.e. not due to chance alone, are seen as being established and worth consideration.

*Geof Givens is Assistant Professor, David Smith is Graduate Research Assistant, and Richard Tweedie is Professor and Chair, all at the Department of Statistics, Colorado State University, Fort Collins, CO 80523
However, the use of meta-analysis introduces a situation where studies themselves, both significant or insignificant, form the basic population of interest, so that this paradigm ceases to be valid. Meta-analysis seeks to combine the analyses from all relevant individual studies into a single statistical analysis with an overall estimate and confidence interval for effect size (Cooper and Hedges, 1994; Hedges and Olkin, 1985). Ideally, greater statistical power can be achieved through meta-analysis than through any one individual study, since data from a greater number of subjects are used, and in recent years there has been an enormous increase (see e.g. Olkin, 1992) in the use of meta-analysis in many areas in order to obtain overall evaluations of association when individual studies are equivocal.

Studies for a meta-analysis are usually collected through a review of the literature. Since insignificant studies are, by the very nature of the scientific process, published less frequently (if at all), such a process is inherently subject to bias introduced from being based on only one part of the real population.

This problem has recently received considerable notoriety in the debate on passive smoking, or exposure to environmental tobacco smoke (ETS). The US Environmental Protection Agency (EPA) issued, in December 1992, a report concluding that ETS is a class A human carcinogen. This was based largely on an argument by analogy with data on the relationship between lung cancer and active smoking, but also included a meta-analysis of 31 studies on the association of lung cancer in never smokers with ETS exposure through spousal smoking. After this assessment was published, several tobacco companies filed a lawsuit against the EPA, claiming that "... various sources of bias, including publication bias ... could explain any association claimed by the EPA between ETS and lung cancer" (Bero et al., 1994, p. 133).

In any meta-analysis, a well documented concern (Hedges, 1992; Dear and Begg, 1992; Sterling et al., 1995) is the need to have available all relevant information. It is clearly crucial to attempt to collect at least all published studies, and if possible, one should also search for unpublished studies such as dissertations and technical reports. After doing so, however, it then seems appropriate to assess not only the existence, but also the possible extent, of the potential biasing effect of unpublished or uncollected studies, to attempt to quantify claims such as that against the EPA evaluation.

No such attempt was made by the EPA, and this exemplifies the need for the type of methodology we will consider.

In this paper, we develop a new Bayesian approach and use it to examine the existing ETS data. The method is based on a Bayesian hierarchical model for meta-analysis that combines the estimated effect sizes from heterogeneous individual studies after estimating and adjusting for potential publication bias. We use a data augmentation technique that is related to the frequentist model of Hedges (1992), which assumes that studies are missing with probabilities that are a function of their lack of statistical significance. Our analysis indicates that world-wide, there may be around ten possible missing negative studies, and a similar number of missing insignificant positive studies. After allowing for this, as we see in Section 4 that the 95% posterior credibility interval for relative risk is shifted downward towards the null hypothesis of no effect; more importantly, perhaps, the actual estimate of excess risk is cut by approximately one third.

When applied to studies in the US, which the EPA used in its final meta-analysis, a very
similar picture emerges: only some 4-5 studies are estimated as missing but the effect is now to lower the Bayesian overall relative risk estimate from 1.17 with a 95% posterior credibility interval of (1.02,1.33) to 1.10 with a 95% interval of (0.95,1.29).

The ETS issue is destined to be only one of many important public debates in which meta-analysis is emerging as a useful tool to provide an overview of multiple and perhaps disparate studies. Although one of our goals is to quantify, in this specific context, an issue that has previous been approached in largely qualitative terms, the methodology we develop is clearly applicable to the wider range of situations in which this same question arises.

2 The ETS debate

2.1 Studies of lung cancer and ETS exposure

Epidemiological studies such as those related to exposure to ETS are carried out to try to confirm or quantify the health risk associated with exposure to some possible toxic agent. The investigators collect prospective or retrospective data in order to estimate relative risk, which we denote by \( RR \). Conceptually, relative risk is the ratio

\[
RR = \frac{Pr[\text{getting disease | exposure}]}{Pr[\text{getting disease | no exposure}]}.
\]

Estimates of the relative risk also lead to estimates of the 'excess risk' given by \( RR - 1 \), which is often used also as a measure of the impact of the exposure on the disease incidence.

In general, epidemiological studies are necessarily observational, rather than controlled experiments. In the two most common study designs, cohort and case-control studies (Mausner and Kramer, 1985), subjects are categorized in a \( 2 \times 2 \) cross-classification table. Each subject is classified as either exposed to the possible toxic agent or not exposed. Each subject is also classified based on disease status, with those diagnosed with the disease being 'cases', and those without the disease being 'controls'. The relative risk is then estimated as the ratio of the incidence rate among the exposed population to the incidence rate among the unexposed population. In our ETS modeling, the substance—environmental tobacco smoke—is a potential carcinogen, the disease is lung cancer, and the hypothesis of concern is \( RR > 1 \).

Between 9% to 20% of lung cancer cases occur in non-smokers (Schneiderman et al., 1989; Alavanja et al., 1992). Until the early 1980’s, epidemiological studies had not reported any noticeable increase in the incidence of lung cancer among non-smokers who were exposed to ETS. This changed starting in 1981 when a case-control study in Greece by Trichopoulos et al. (1981; 1983) and a cohort study in Japan by Hirayama (1981; 1984) reported an association between lung cancer and exposure to ETS in non-smoking spouses of smokers.

During the next 15 years, a large number of such epidemiological studies were conducted to address the health effects of ETS. In 1990, the Environmental Protection Agency of the United States published a draft evaluation of the association of ETS exposure with lung cancer (EPA Draft Report, 1990); after receiving comments, this was issued as the final EPA Report (1992), and concluded that exposure to ETS was a class A human carcinogen. Much of the argument in
that paper was based on biological and toxicological studies which considered the similarities and differences between ETS exposure and active smoking. However, a key component of the EPA report was a meta-analysis of epidemiological studies. The EPA initially considered 31 studies, but changed in the 1992 Report to using, for most purposes, a formal combined estimate based only on 11 US studies, after receiving arguments on the validity of non-US studies in forming an estimate of relative risk to be used in the US context.

Since that time a small number of other studies have appeared in the US. The ETS meta-analysis data that we shall use consists of 35 studies that assess the risk of lung cancer in non-smoking women exposed to spousal smoking. These studies with their relative risks and associated confidence intervals are given in the top part of Figure 1. The studies are enumerated and described by Lee (1992), Mengersen et al., (1995) and Tweedie et al., (1994), and represent a complete set of such studies as far as could be determined at the time of preparation of Tweedie et al., (1994).

FIGURE 1 NEAR HERE

We note that one of the real issues in the ETS area is the relevance of these data to exposure to ETS in the workplace, where many of the regulations on ETS exposure are being proposed (cf. the recent OSHA Draft Regulation (1994) and the Australian NH&MRC Draft Report (1995)). As noted in Biggerstaff et al., (1994) and Tweedie et al., (1996) there is now a reasonable amount of data relevant to workplace exposure, but we will not consider such studies in more detail here, merely noting that the methods we propose could be applied to the workplace data set also.

2.2 Frequentist Models for Meta-analysis

We first need to outline the models for meta-analysis which we use without considering publication bias, and sketch their application in the ETS context.

The most commonly used frequentist models for meta-analysis of relative risk (Cooper and Hedges, 1994) are the so-called 'fixed effects' (FE) and 'random effects' (RE) models. The FE model was used in the EPA Report (1992), although in comparison with the RE model it has a number of limitations, discussed in some detail in the NRC Report (1992) on combining data.

Both models assume that there is a true underlying value of $RR$ across all studies. In order to use normal theory, it is common to work on the log scale and we take $\Delta = \log RR$ as the response variable of interest. If $\Delta = 0$ then exposure is associated with no change in health risk; $\Delta > 0$ implies that exposure is associated with an increased risk, and $\Delta < 0$ implies that exposure is associated with a decreased risk, i.e. a health benefit.

We assume we have $n$ individual studies which produce estimates of $\Delta$, say $Y_j$, for $j = 1, \ldots, n$. The FE model treats these results from the individual studies as data, and models them by

$$Y_j = \Delta + \epsilon_j$$

(1)

where $\epsilon_j \sim N(0, \sigma^2_\epsilon)$, so that $\Delta$ is interpreted as the overall relative risk.

The random effects model has an extra term compared with (1), namely

$$Y_j = \Delta + \beta_j + \epsilon_j$$

(2)

where $\beta_j \sim N(0, \tau^2)$ is introduced to account for heterogeneity between studies, and $\epsilon_j \sim N(0, \sigma^2_\epsilon)$ represents within-study variability of study $j$ as before. We write $\sigma^2 = \{\sigma^2_\epsilon\}$ for these variances.
<table>
<thead>
<tr>
<th>Model</th>
<th>Relative Risk</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
<td>1.19</td>
<td>(1.02, 1.38)</td>
</tr>
<tr>
<td>Random Effects</td>
<td>1.20</td>
<td>(1.07, 1.34)</td>
</tr>
<tr>
<td>Bayesian Hierarchical</td>
<td>1.22</td>
<td>(1.08, 1.37)</td>
</tr>
</tbody>
</table>

Table 1: Results from meta-analyses of ETS data.

The RE approach has been argued (NRC Report, 1992) to be preferable to the FE model which essentially assumes that any heterogeneity between studies is purely random. In the special case where $\tau^2 = 0$, indicating such homogeneity between studies, the RE model (2) reduces to the FE model.

This frequentist meta-analysis then leads through normal theory to the estimate

$$\hat{\Delta} = \frac{\sum Y_j (\sigma_j^2 + \tau^2)^{-1}}{\sum (\sigma_j^2 + \tau^2)^{-1}}$$

with

$$\text{Var}[\hat{\Delta}] = \frac{1}{\sum (\sigma_j^2 + \tau^2)^{-1}}.$$  \hspace{1cm} (4)

in the FE model we take $\tau^2 = 0$ in these equations and in the RE model there are various moment-based and maximum likelihood approaches giving estimates of $\tau^2$ (Biggerstaff and Tweedie, 1996).

The results of meta-analyses using (3) and (4) are given in Table 1, based on the 35 studies in Figure 1. Note that the RE analysis does make a difference to the 95% CI although not in any meaningful way to the estimate of $RR$ itself; the estimate of $\hat{\tau}^2 = 0.023$ in this case is insufficient to make a great deal of difference (Tweedie et al., 1996).

Clearly one source of between study variation that might lead to a requirement for $\tau^2 > 0$ is the use of studies from different countries. The analysis of the ETS data in Mengersen et al. (1995) clearly shows this to be a real concern with ETS data, and the initial use of FE approaches by the EPA without allowing for this has been criticized on these grounds.

Following such comments on their use of FE models, and of amalgamating over different countries, 11 studies relating to the US were used in a FE meta-analysis in the final EPA Report (1992). We analyse in more detail in Section 4.3 the 14 studies currently available in the US. Recent tests for $\tau^2 = 0$ have been developed by Biggerstaff and Tweedie (1996), and applying these to this US data set indicates that in this case the difference between FE and RE models is almost non-existent: both lead to an estimate of $RR = 1.16$ with a 95% CI of (1.04, 1.31) for the RE and (1.03, 1.30) for the FE model. These are quite close to the EPA values of $RR = 1.19$ with a 95% CI of (1.04, 1.35) (EPA Report, 1992, Table 5-9). Thus the EPA would be reasonably justified, at least in using current US data, in maintaining its stance that “... it is implicitly assumed that studies within a country ... are sufficiently homogeneous to warrant combining their statistical results into a single estimate for the country” (EPA Report, 1992, p. 5-31).
2.3 Bayesian Hierarchical Models

In the random effects model, $\Delta$, $\tau^2$, and $\sigma^2$ are presumed to be fixed parameters. We will also consider a Bayesian analysis of this model, using methods described in detail by DuMouchel (1990). In the general hierarchical Bayesian scheme, $\Delta$, $\tau^2$, and $\sigma^2$ are also treated as random variables. The distributions of these quantities are specified *a priori* according to the application. In our approach, Bayesian methods will not primarily be used to describe prior information in any strong sense. Rather, the prior distributions for $\Delta$, $\tau^2$, and $\sigma^2$ can be viewed as more detailed descriptions of the way in which the studies might be heterogeneous. This allows one to account explicitly for greater variability in the underlying collection of studies than is done in the fixed or even the random effects models.

Typically an 'uninformative' prior is chosen for $\Delta$, since even with a small number of studies, "the combined data become relatively informative about the location of the effect-size prior distribution" (Carlin, 1992, p. 146). Standard Bayesian analyses might use independent conjugate prior, which for this problem are normal for $\Delta$ and inverse gamma for $\sigma^2$ and $\tau^2$. The specific priors we adopt are detailed in Section 4.

With these choices, the posterior distribution for $\Delta$ is a normal distribution centered at the weighted average of the mean relative risks from the prior and from the individual studies; the weights in the average are proportional to the inverse of the variance of the prior and the variances of the individual studies. In this formulation, other posterior distributions become quite complicated, leading DuMouchel (1990) to make approximations to normality for computational convenience. In contrast, we use Markov chain Monte Carlo (MCMC) methods to carry out the analysis of our extension of this model, implemented using the Gibbs sampling routines in BUGS (Spiegelhalter et al., 1996).

Table 1 shows that in the ETS dataset in Figure 1 the Bayesian methodology does not make a large difference to the estimates of RR given by the RE models, as indicated in more detail in Tweedie et al. (1996).

2.4 Publication Bias and the Funnel Plot

A large number of discussion papers have appeared which assess the benefits, drawbacks and problems of meta-analysis techniques (see for example Mosteller and Chalmers, 1992; Felson, 1992; Chalmers, 1991; NRC Report, 1992; Thompson and Pocock, 1991; Mengersen et al., 1995). One of the most frequently considered aspects is the need for collection of all studies, especially taking into account the possibility that some studies might not get to the peer reviewed publication stage.

The studies that are to be combined in a meta-analysis to obtain an overall estimate of relative risk are usually compiled by review of scientific journals. Even if the search is effective or even exhaustive, this selection process may introduce an important source of bias, since not all studies submitted for publication are accepted, and not all studies conducted are even submitted.

There are many reasons why simple searches might not turn up all studies. One widely believed publication bias hypothesis is that scientific journals prefer to publish articles that show statistically significant results. Another potential source of bias in the same direction could be the possible decision by scientists not to submit for publication manuscripts describing the results of their
studies because the results were not statistically significant.

There are other sources of potential publication bias, even against significant studies. For example, some students leave the academic arena and do not publish their PhD or MS dissertations; or studies are suppressed by those who do not wish to have results appear that are against their own vested interests, political beliefs, or funding source’s interests (see Crossen, 1994, p. 19). With these possible reasons for publication bias, it is clearly hard to ensure that all studies will be found even by diligent search procedures. Sterling et al. (1995) discuss recent indications that publication bias may be pervasive in the scientific literature and can create potentially severe distortions in meta-analyses.

Publication bias is not incorporated in the combined estimates in Table 1. A number of ways of attempting to assess the possibility of missing studies (Berlin et al., 1989; Hedges, 1992; Dear and Begg, 1992), and the number of missing studies (Gleiser and Olkin, 1993; Eberly and Casella, 1996) based on such data have been proposed but perhaps the most common is the funnel plot (Light and Pillemer, 1984; Vandenbroucke, 1988; Thompson, 1993; Mengersen et al., 1995), which is a graphical method to display possible publication bias. It shows the relationship between the estimated value of \( \Delta \) and the size of the study, measured by, say, the inverse of the standard error, \( \sigma_j^{-2} \), or the number of lung cancer cases in the studies. If there is no publication bias then one expects to get a typical inverted funnel shape, since the estimates of \( \Delta \) for small studies at the bottom of the graph are more variable, whereas the estimates from larger studies near the top of the graph are more concentrated, but both should center around the common true value of \( \Delta \).

FIGURE 2 NEAR HERE

Figure 2 shows a funnel plot of the data in Figure 1. For this ETS funnel plot, most of the studies are clustered to the right of zero, suggesting that \( \Delta \) may be positive. However, the funnel shape of Figure 2 is asymmetric: the lower left corner of the graph appears to be missing a number of points. This suggests publication bias may be present, because these missing points would correspond to studies that would have shown nonsignificant risk, or even a negative result, for ETS exposure. The funnel plot suggests there are fewer of these studies published than one would expect. Other graphical indicators used in Mengersen et al. (1995) support this conclusion; this contrasts with Vandenbroucke (1988) who decided, using an early subset of these data, that there was some indication at that time of missing male studies but no such indication of missing female studies.

The sensitivity to possible publication bias of point estimates and associated confidence intervals as given in Table 1 cannot be overlooked. Mengersen et al. (1995), using an ad hoc method based on Figure 2, estimated that the possible impact of allowing for this publication bias would be to reduce the RE estimate of \( RR \) from 1.20 (95% CI (1.07, 1.34)) to 1.12 (95% CI (1.01, 1.24)). This would indicate that as much as 40% of the observed excess risk could be due to publication bias.

None of the frequentist or Bayesian models above account for such a possibility. We now develop the components of a formal statistical model for meta-analysis data which incorporates potential publication bias. Our approach may be generalized to account for other selection biases, such as those based on differing study quality, for covariates influencing selection bias, and for additional hierarchical strata in the model; we pursue this elsewhere (Smith et al., 1997). Clearly, this approach will also be applicable in many areas other than the epidemiological context in which
we illustrate it.

3 Meta-Analysis Allowing for Publication Bias

3.1 The Data Augmentation Approach

If it were somehow possible to discover all missing studies, meta-analysis would be straightforward using any of the models described in Section 2.2–2.3. The approach we develop in this paper to account for potential publication bias relies on the ideas of missing data and data augmentation: using a Bayesian model we augment the observed data by simulating the outcomes for the missing studies, thus creating a ‘complete’ dataset for analysis.

Data augmentation is a technique which has proven useful in a range of Bayesian and likelihood problems, including applications of the EM algorithm (Dempster et al., 1977) and the IP algorithm (Tanner and Wong, 1987). The premise of data augmentation is that the ‘observed data’ $Y$ can be thought of as a partial realization of the random variable $X = (Y, Z)$, where a complete realization, $X_c$, of $X$ is called the ‘complete data’, and a realization, $Z$, of $Z$ is called the ‘missing’ or ‘latent data’. We assume that the distribution of $X$ depends on parameters of interest $\theta$ through the family $p(X \mid \theta)$, which gives a marginal distribution $p(Y \mid \theta)$ for the observed data. This framework is most useful when inference about $\theta$ based on $p(Y \mid \theta)$ is difficult, but would be more straightforward using the complete data likelihood $p(X \mid \theta)$.

In our case, we treat both the number and outcomes of unpublished studies as latent data to augment the observed study outcomes, using the model described in Section 3.2. Completing the data in this manner allows us to obtain posterior distributions for quantities of interest which are then marginalized across the latent random variables.

Problems with genuinely missing data are natural candidates for data augmentation. It is also possible to recast other problems as if they involved latent data. In these cases, the latent data are only an artifact of the analysis methodology. Our situation is somewhat between these two extremes. The latent studies are missing data in the sense that they possibly exist and we have not observed them. However they are also essentially an artifact to construct a meta-analysis which adjusts for publication bias, since a sampling scheme for observing the complete set of studies is inconceivable.

In the next two sections we describe the formal structure of the meta-analysis problem, and how to augment this structure to consider the possible existence of missing studies resulting from publication bias.

3.2 A Model for Selection Bias

We now formalize the approach described above. Using the random effects model in (2), the likelihood of the observed data $Y = (Y_1, \ldots, Y_n)$ is

$$p(Y \mid \Delta, \tau^2, \sigma_j^2) \propto \prod_{j=1}^{n} \exp \left( \frac{-1}{2} \frac{(Y_j - \Delta_j)^2}{\tau^2 + \sigma_j^2} \right) / \sqrt{\tau^2 + \sigma_j^2}. \tag{5}$$
In order to extend this model to account for publication bias, we assume that in addition to the \( n \) observed studies, there are an additional \( m \) studies 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 the \( j^{th} \) missing study be denoted as \( Z_j \) for \( j = (n + 1), \ldots, (n + m) \), and let \( Z = \{Z_j\} \). We will also denote the complete set of estimated log relative risks for all studies, both observed and missing, by \( X = \{X_j\} \) for all \( j \), where \( X_j = Y_j \) when \( j \) indexes an observed study and \( X_j = Z_j \) when \( j \) indexes a missing study.

We assume that the same random effects model in (2) holds for the outcomes of the missing studies, namely

\[
Z_j = \Delta + \beta_j + \epsilon_j
\]

where \( \beta_j \sim N(0, \tau^2) \), and \( \epsilon_j \sim N(0, \sigma^2_j) \) are mutually independent. Note that now \( \sigma^2 \) includes the variances of the latent studies as well as those of the observed studies.

There are various selection mechanisms that one might now consider when trying to model publication bias. Following Hedges (1992) and Dear and Begg (1992), we assume here that the selection criterion for a study is based solely 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 held view that statistically significant studies are more likely to be published than insignificant studies.

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 individual study must fall into one of these intervals. Now let

\[
w^k = \Pr[\text{a study with } p\text{-value in } I_k \text{ is published}], \quad k = 1, \ldots, c
\]

and let \( w = \{w^k\} \). For consistency with model extensions by Smith et al. (1997), we adopt notation where superscripts index \( p \)-value intervals and subscripts index studies.

Let \( n^k \) be the number of studies observed with \( p \)-values in \( I_k \). Similarly, let \( m^k \) be the number of missing studies with \( (\text{unobserved}) \) \( p \)-values in \( I_k \). Let \( p_j \) equal the \( p \)-value of study \( j \) corresponding to \( H_0 : \Delta \leq 0 \). Then \( n = \sum_k n^k \) and \( m = \sum_k m^k \), where the \( n^k \) are known and the \( m^k \) are unknown; we write \( m = \{m^k\} \). We adopt the a negative binomial model for the number of missing studies with \( p \)-values in \( I_k \):

\[
m^k | w \sim \text{Negative Binomial } \left(n^k, w^k\right).
\]

Note that (8) depends on knowing the weight vector \( w \). Hedges (1992) and Dear and Begg (1992) present a maximum likelihood method for estimating the \( w^k \) from a meta-analysis dataset, but we pursue a Bayesian approach in this paper.
3.3 The Complete Data Likelihood and Conditional Posterior Distributions

The observed data are the outcomes, $Y$, of the observed studies, and we condition on the numbers of observed studies $n$. Using (5) we write the likelihood for the observed data under this conditioning as

$$p(Y | \Delta, \tau^2, \sigma^2, w) \propto \prod_{j=1}^{n} \prod_{k=1}^{c} 1_{\{x_{ij} \in I_{jk}\}} \exp\left(-\frac{1}{2} \frac{(y_{ij} - \Delta)^2}{\tau^2 + \sigma_j^2}\right) \frac{1}{\sqrt{\tau^2 + \sigma_j^2}}.$$  \hspace{1cm} (9)

The latent data are the outcomes, $Z$, of the unobserved studies, and the numbers of such studies $m$. At times, it is convenient to consider the latent data $(Z, m)$ as nuisance parameters to be marginalized out of final inference about $\Delta$.

This model has a partial conditional likelihood for the complete set of outcomes $X$ given by

$$p(X | \Delta, \tau^2, \sigma^2, m) \propto \prod_{j=1}^{n+m} \prod_{k=1}^{c} 1_{\{x_{ij} \in I_{jk}\}} \exp\left(-\frac{1}{2} \frac{(X_{ij} - \Delta)^2}{\tau^2 + \sigma_j^2}\right) \frac{1}{\sqrt{\tau^2 + \sigma_j^2}}.$$  \hspace{1cm} (10)

We stress that (10) is conditional on knowing $m$. Treating $m$ as unknown latent data and conditioning instead on the parameter $w$, the complete data likelihood is

$$p(X, m | \Delta, \tau^2, \sigma^2, w) \propto p(X | \Delta, \tau^2, \sigma^2, m) \times \prod_{k=1}^{c} \left( \binom{n_k + m_k - 1}{m_k} \right) (w^k)^{n_k} (1 - w^k)^{m_k}.$$  \hspace{1cm} (11)

In our Bayesian analysis, we adopt independent prior distributions $p(\Delta), p(\tau^2), p(\sigma^2), p(w)$, and $p(Z)$ for the model and latent data treated as nuisance parameters. Since $m$ and $w$ are related through (8), no separate prior for $m$ is needed since its conditional distribution is known once $w$ is known. Degenerate priors are allowed, and for example, we may take $\sigma_j^2$ to be known for individual observed studies; see Section 3.5.

Note that (11) is an extension of (5) but now includes parameters $w$ 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 (11). For identifiability, Hedges (1992) assumed that $w^1 = 1$, and considered maximum likelihood estimation only up to a multiplicative constant. Following Hedges (1992), we also scale the $w^k$, as shown below, and 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 4.2, we discuss the effect on ETS inferences of constraining the $w^k$ to be monotonically decreasing as the $p$-value increases. Such a constraint is much harder to put in place in the frequentist setting (Dear, 1995), and we note that in other circumstances we have found that it seems to be worth enforcing (LaFleur et al., 1996).

Using prior distributions and the complete data likelihood, univariate conditional posterior distributions can be derived. We use $p(q | \cdot)$ to represent the conditional posterior distribution of
any parameter $q$ given all other parameters. The univariate conditionals for $\Delta$, and $\tau^2$ are then easily found from (11) as

$$
p(\Delta \mid \cdot) \propto \frac{p(\Delta)}{A(\Delta)} \prod_{j=1}^{n+m} \exp \left( -\frac{1}{2} \frac{(X_j - \Delta)^2}{\tau^2 + \sigma_j^2} \right),
$$

$$
p(\tau^2 \mid \cdot) \propto \frac{p(\tau^2)}{A(\tau^2)} \prod_{j=1}^{n+m} \left[ \exp \left( -\frac{1}{2} \frac{(X_j - \Delta)^2}{\tau^2 + \sigma_j^2} \right) \right], \quad \sqrt{\tau^2 + \sigma_j^2},
$$

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

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

$$
p(Z, \sigma^2 \mid \cdot) \propto \frac{p(Z, \sigma^2)}{A(\sigma^2)} \prod_{j=1}^{n+m} \prod_{k=1}^{c} \frac{\exp \left( -\frac{1}{2} \frac{(X_j - \Delta)^2}{\tau^2 + \sigma_j^2} \right)}{\sqrt{\tau^2 + \sigma_j^2}} 1_{\{p_j \in I_k\}}.
$$

We consider $Z$ and $\sigma^2$ in a bivariate form since for any new study the values of $Z_j$ and $\sigma_j^2$ must be chosen to ensure the constraint $1_{\{p_j \in I_k\}}$ is satisfied.

If we consider $m$ as a nuisance parameter, then its conditional posterior distribution is merely

$$
p(m \mid w) \propto \prod_{k=1}^{c} \left( \binom{n^k + m^k - 1}{m^k} w^k (1 - w^k)^{m^k} \right),
$$

since we have no prior on $m$, as discussed above.

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

$$
p(w \mid \cdot) \propto \frac{p(w)}{A(w)} p_1(w \mid \cdot)
$$

where $p_1(w \mid \cdot)$ is the conditional probability density function that results when the conditional probability density function of $w \times \max_k w^k$ is proportional to (15).

### 3.4 Gibbs Sampling Methods

The model above is more complex than the standard hierarchical Bayesian model, and the posterior for $\Delta$ can no longer be derived in a tractable analytical form. Instead, numerical techniques must be used, and we use a Gibbs sampling strategy (Geman and Geman, 1984) to obtain approximate samples from the desired posterior distribution. Gibbs sampling techniques, which have been very successful at solving a wide variety of similar problems in Bayesian estimation (Smith and Roberts, 1993; Besag and Green, 1993), can be used to obtain a sample from a desired distribution by
simulating realizations from a Markov chain whose stationary distribution is equal to the target distribution.

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.

We iterate Gibbs steps in the following sequence: $(\mathbf{m}, \mathbf{Z}, \sigma^2), \mathbf{w}, \Delta$, and $\tau^2$. We update $\mathbf{m}$, $\mathbf{Z}$, and $\sigma^2$ jointly to ensure that the number of missing study outcomes is always equal to the number of missing studies. In practice, given $\mathbf{m}$, for each $k$ it is efficient to draw $m^k$ missing study variances, $\sigma^2_j$, from $p(\sigma^2 | \cdot)$ with no constraint on the outcomes or $p$-values of the missing studies, then simulate the $m^k$ missing study $p$-values, $p_j$, uniformly on $I_k$, and finally calculate the corresponding $Z_j = \sigma_j \Phi^{-1}(p_j)$. This effectively draws the $m^k$ values of $Z_j$ from their conditional density which is proportional to $p(Z | \sigma^2, \cdot)$. Note also that in our examples below we assume that $\sigma^2_j$ are fixed for the observed studies, which corresponds to taking their priors as degenerate at the observed values.

In our case, the univariate conditional posteriors derived in Section 3.3 are not easily sampled, and we use an inverse CDF method (Press et al., 1986) to perform this numerically.

The Gibbs sampling results in a large collection of approximate realizations from the joint posterior. The distribution of sampled points converges to the posterior distribution as iterations increase, because the procedure generates an aperiodic Markov chain which is irreducible since the conditionals in equations (12)–(16) assign positive probability to the entire parameter space that may be supported by the posterior.

Therefore, for example, to obtain the overall median and 95% interval for relative risk, $\Delta$, we calculate the corresponding sample quantiles from a collection of values of $\Delta$ obtained via simulation. Iteration length, burn in, and subsampling are discussed in Sections 3.5 and 3.4. Estimation from this sample reflects the combined results from all studies and accounts for estimated publication bias.

### 3.5 Simulation Studies

It is important to evaluate the reasonableness of this method before using it to address a real analysis such as that of lung cancer and ETS. Readers who want to jump straight to the ETS results may prefer to skip this section.

We assessed the method on a range of simulation studies. We first generated 50 studies with mean $\Delta = 0$ and suppressed some of them according to the various criteria described below. The studies not suppressed were assumed to be observed. Without further data, $\beta_j$ and $\epsilon_j$ in (2) are non-identifiable. However, each 'observed' study has not only a published outcome $Y_j$ but also a published variability estimate, say $\hat{\sigma}_j^2$. We assume here that each individual study variance, $\hat{\sigma}_j^2$, is exactly correct. Hence, the prior distribution from which each $\sigma_j^2$ is drawn is degenerate at $\hat{\sigma}_j^2$ for the $j^{th}$ study when $j$ indexes an observed study, but the remaining $\sigma_j^2$ are random.

We generated the original variances, $\hat{\sigma}_j^2$, for the 50 studies from a gamma distribution with a
shape parameter of 3 and a mean of 1/3. Each of the 50 relative risks, $X_j$, was drawn from a normal distribution with mean 0 and variance $\tilde{\sigma}_j^2 + \tau^2$, where $\tau^2=.03$. This gives data which are not dissimilar in structure to the ETS data.

We then applied suppression criteria to simulate publication bias in three different circumstances, as detailed below. In each case, we either partitioned the unit interval into $k = 3$ $p$-value regions given by $I_1 = [0, 0.05), I_2 = [0.05, 0.10), I_3 = [0.10, 1.00)$ or $k = 2$ $p$-value regions with $I_1 = [0, 0.50)$ and $I_2 = [0.50, 1.00]$. In every case we performed two Bayesian meta-analyses on the observed studies after suppression: one standard hierarchical analysis as in (2) using BUGS (Spiegelhalter, 1996) which does not take into account the possibility of missing data, and one using our data augmentation techniques as discussed above. Except for the prior for $w$, these meta-analyses all had identical priors given by $\Delta \sim \text{Normal}(0, .4^2)$, $\tau^2 \sim \text{Inverse Gamma}(\text{shape} = 32, \text{mean} = 1/32)$, and $\sigma_j^2 \sim \text{Inverse Gamma}(\text{shape} = 3.5, \text{mean} = 0.33)$ for any $j$ that indexed a missing study. For all Gibbs sampler runs, we used a burn-in of 500 and ran 1000 additional iterations. Convergence over this period seemed acceptable; formal and graphical assessments of convergence were similar to those discussed in Section 4 for the ETS analysis.

A Bayesian meta-analysis without augmentation on the complete set of 50 simulated data points resulted in a posterior mean and 95% interval of 1.007 (.889, 1.156) for $RR = \exp \Delta$, which indicates that the sample we had drawn was not an aberrant one.

(a) Suppression applied to negative studies

We initially tested the performance of the algorithm when a considerable portion of the data is missing. For this, we used only two $p$-value regions, $I_1 = [0, 0.50]$ and $I_2 = [0.50, 1.00]$ We suppressed no studies in $I_1$, but we suppressed 70% of all studies in $I_2$: that is, we chose $w^1 = 1$ and $w^2 = 0.3$. Out of the 50 simulated studies, 25 each had $p$-values in $I_1$ and $I_2$. After the suppression criteria were applied to $I_2$, our observed data set consisted of 32 studies; specifically, $n_1 = 25$ had $p$-values in $I_1$ and $n_2 = 7$ had $p$-values in $I_2$. The 18 suppressed studies were discarded.

We took the prior on $w^1$ as uniform on $[0.5, 1.0]$ and the prior on $w^2$ as uniform on $[0.2, 1.0]$. Each of these priors envelopes the true probability of being published while not reflecting strong beliefs about the amount of publication bias present in the data set.

Figures 3 - 4 NEAR HERE

Figure 3 shows the posteriors of the two meta-analyses for $RR = \exp \Delta$. These density estimates were obtained from the Gibbs samples using normal kernel density estimation with the maximal smoothing span of Terrell (1990), as were all other density estimates below.

The standard meta-analysis produced a posterior mean and 95% interval for $RR$ of 1.18 and (1.03, 1.33). This interval does not include the null value of 1.00, thus leading to an erroneous inference that $\Delta > 0$. In contrast, the mean of the posterior and 95% posterior probability interval for our Bayesian meta-analysis with data augmentation to account for publication bias was 1.00 and (0.84, 1.19). This interval contains the 95% posterior probability interval from the meta-analysis performed on all 50 studies.

Figure 4 shows histograms of the numbers of missing studies in both $p$-value intervals at each iteration of Gibbs sampling. Although there might seem to be some probability of the algorithm finding studies missing in $I_1$, the weighting of the $w^k$ so the maximum is 1.0 has led to no missing
studies being found in $I_1$. In the $I_2$ interval, the correct number of studies missing was 18. The mean of the posterior distribution of $m^2$ was 15.3, so the algorithm slightly underestimated, on the average, the number of missing studies. The posterior means of the weights are $w^1 = 1.00, w^2 = 0.35$; clearly the prior mean for $w^2$ of 0.6 has been adjusted downward substantially by the data to approach the true value of 0.3.

(b) No suppression

The other extreme we tested was where in fact no studies were suppressed. For the augmented data meta-analysis, we used three intervals and assumed the prior weights were uniform on $[0.5, 1.0]$ for all of $I_1, I_2$ and $I_3$. The algorithm (as one must expect given any priors not degenerate at 1) gave a positive number of predicted missing studies: the mean estimated number of missing studies was 15.7 compared with a prior expected value of 16.6. Clearly there is a quite strong lingering effect of the prior distribution of $w$ in this case. However, the posterior mean and 95% interval for $RR$ were 0.99 (0.83, 1.14), and so the estimated relative risk was largely unaffected by the latent values, which were not distributed in such a way as to affect the meta-analysis unduly.

(c) Heavy suppression applied for insignificant studies

We next consider a situation where a fairly heavy suppression regime was in place for insignificant studies. The weights for acceptance chosen were: for $I_1 = [0, 0.05]$, $w^1 = 1$; for $I_2 = [0.05, 0.10]$, $w^2 = 0.85$; and for $I_3 = [0.10, 1.00]$, $w^3 = 0.3$. Out of the 50 simulated studies, 2, 3, and 45 had $p$-values in $I_1, I_2$ and $I_3$, respectively. After the suppression criteria was applied, we ‘observed’ 19 studies: specifically, $n^1 = 2, n^2 = 3$, and $n^3 = 14$ observed studies had $p$-values in $I_1, I_2$ and $I_3$. Thus, on this sample the suppression rate was almost exactly realized.

For the augmented data meta-analysis, we assumed the prior weights were respectively uniform on $[0.5, 1.0], [0.5, 1.0]$ and $[0.2, 0.7]$. Note that the prior assumes that on $I_3$ there must be a non-trivial probability of a study being unpublished.

Figures 5 – 6 NEAR HERE

Figure 5 shows the posteriors of these meta-analyses for $RR = \exp \Delta$. The posterior mean and 95% posterior probability interval for $RR$ using our Bayesian meta-analysis with data augmentation to account for publication bias were 1.105 and (0.889, 1.391), compared with the standard meta-analysis mean and 95% interval of 1.28 and (1.07, 1.50). Thus our data augmentation procedure shifted the posterior of $RR$ to the left so that the true relative risk, 1.00, is now within the 95% posterior probability interval for $RR$. In contrast, note that the standard meta-analysis produces an interval which does not cover the truth.

The posterior mean numbers of imputed studies were: 0.3 in $I_1$, 0.3 in $I_2$, and 23.4 in $I_3$, corresponding to posterior mean weights of 0.89, 0.93, and 0.42, and to distributions as in Figure 6. Thus in this case the data have not moved the weights in the last interval as close to the true weight of 0.3 as one might hope although all the posterior weights move in the right direction.

(d) Strong suppression applied with different priors

The prior used in (c) above for $w^3$ may seem to lead to a mean number of missing studies rather less than those we actually simulated. To assess sensitivity to such prior choice, we also considered the model in (c) with two different priors applied, one less and one more suitable for the actual
situation.

Our first variation uses prior weights taken respectively as uniform on [0.5, 1.0], [0.5, 1.0] and [0.2, 0.4]: that is, as in (c) except that the prior mean suppression rate in \( I_3 \) in this case is exactly equal to the true suppression rate. In this case the method performed extremely well. The mean number of studies estimated as missing in \( I_3 \) was 30, compared to the actual 31, and the posterior mean and 95\% interval of \( RR \) was 1.07 (0.88, 1.32).

Our second variation uses prior weights which are respectively uniform on [0.5, 1.0], [0.5, 1.0] and [0.3, 0.7], so the true suppression rate on \( I_3 \) was on the boundary of the corresponding prior. Now the mean of the posterior and 95\% posterior probability interval for \( RR \) were 1.14 and (0.92, 1.42). Thus our data augmentation procedure still shifted the posterior of \( RR \) to the left, relative to the standard analysis, so that the true relative risk is within the 95\% posterior probability interval for \( RR \). However, histograms analogous to Figure 6 show that the algorithm typically underestimated the correct number of missing studies in \( I_3 \) in this case, and tended to a posterior mean probability of publication in \( I_3 \) that was greater than the true value of \( w^0 = 0.3 \).

These results indicates that, although the method is somewhat insensitive to choice of prior on \( w \), the impact on the final estimate of \( RR \) is less serious than the impact on the number of imputed studies might indicate.

### 3.6 Methodological Comments

These simulation trials indicate that the method gives an outcome that is usually conservative: not conservative in the number of missing studies, perhaps, but conservative in the adjusted estimate of \( \Delta \) in the final meta-analysis. This helps to obviate the concern that the number of studies assessed as missing is driven to some extent by the prior distribution on the probability of publication in each interval.

The method we have used is based on a fixed set of intervals \( I_1, \ldots, I_c \) to stratify \( p \)-values. We use the intervals \([0, .01], [.01, .05], [.05, .10], [.10, .50], \) and \([.50, 1.00]\) in our ETS example, and similar cutoff points in the simulations. This is based on the idea (Hedges, 1992) that these are the common ranges in which editors and researchers might decide to change the probabilities of publication. Other researchers (Dear and Begg, 1992; Paul, 1995) have considered methods for estimating the endpoints and number of such intervals, rather than fixing them in advance. This may permit a more flexible, data-based determination of how the probability of publication depends on \( p \)-value. However, the intervals and the expected number of missing studies in each interval can be very variable with this approach. The fixed interval approach seems to provide an adequate, stable estimate.

Direct parametric modeling of publication probability has also been proposed (Iyengar and Greenhouse, 1988; Patil and Taillie, 1989). Larose and Dey (1995) survey and compare several alternative parametric models. This parametric approach seems to yield some of the same benefits as our approach, including tighter posterior confidence intervals and a direct model for publication bias and heterogeneity. However, our method should be more robust than a parametric method to changes in the form of the exclusion criteria.

It would be of substantial interest to develop a more advanced model for the \( n^k \) which depends on covariates. In this article we have assumed that the probability of publication depends only
on the p-value, which in turn depends only the study’s estimated relative risk and $\sigma_j^2$. We can extend the dependence of the publication probability to other covariates, such as study quality, study design, sample size, the mode of exposure, the population studied, or other factors. One way to do this is to define the analogues of the classes $I_k$ based on other properties of the studies. The model is then extended to additional hierarchical levels. This extended model is described in Smith et al. (1997) and is used there to analyse a collection of studies relating the relative risk of cervical cancer to use of oral contraceptives, a situation that we have also studied using the simpler models above in LaFleur et al. (1996). Direct modeling of the relationship between covariates and publication probability can also be worked in to the model.

4 Publication Bias in the ETS Dataset

4.1 The Possible Effect of Bias

We now apply the methods above to the ETS dataset in Figure 1. This leads to the posterior density in Figure 7. The posterior mean relative risk is 1.14 and the 95% posterior probability interval (1.00, 1.28), compared with the Bayesian posterior values of 1.22 (1.08, 1.37) ignoring publication bias in Table 1. These results show that the meta-analysis after adjusting for publication bias continues to suggest that exposure to ETS through spousal smoking is associated with an increased risk of lung cancer, but it also appears credible that there is distinct publication bias in this data set. This result is satisfactorily close to the ad hoc value of 1.12 with 95% posterior probability interval (1.01, 1.24) found in Mengersen et al. (1995).

Figure 7 NEAR HERE

The details of the approach are as follows. We take $c = 5$, and we use the intervals $I_1 = [0, 0.01]$, $I_2 = (0.01, 0.05]$, $I_3 = (0.05, 0.1]$, $I_4 = (0.1, 0.5]$, and $I_5 = (0.5, 1]$, by analogy with Hedges (1992). Of the 35 studies under consideration, 2 had p-values in $I_1$, 4 in $I_2$, 7 in $I_3$, 14 in $I_4$, and 8 in $I_5$. Thus, 75% of the observed studies have RR greater than 1, and nearly 40% have significance levels of 0.1 or less. This suggests either that exposure to ETS through spousal smoking elevates lung cancer risk, or that publication bias favors positive and significant studies, or both.

For $\Delta$, we adopt a $N(0, 0.15^2)$ prior distribution. This allows us to cover a reasonable range of relative risk. We use an empirical exponential prior with mean 0.17 independently for each $\sigma_j^2$ corresponding to a missing study, based on the 35 published variances. We also assume that each individual study variance, $\sigma_j^2$, is exactly correct, so $p(\sigma^2)$ is degenerate for observed studies. For $\tau^2$, we use an exponential prior with mean 0.031, based on a meta-analysis of studies on workplace ETS (Biggerstaff et al., 1994). We take an improper uniform prior for $Z$ and our initial prior for $w$ (before being scaled by the largest) is that of 3 uniform random variates on (0.5, 1] for $I_1$, $I_2$, $I_3$, and uniform on (0.3, 1] for $I_4$ and on (0.3, 0.7] for $I_5$: that is, we assume a positive probability of suppression in the least significant class.

Figures 8 – 9 NEAR HERE

We used a burn-in of 500 iterations and stored an additional 1000 iterations. Figure 8 shows the convergence behavior of the Gibbs approach with respect to $\Delta$ and the proportion of missing studies, and indicates that the estimates of each stabilize reasonably quickly and the burn-in period
seems adequate. We based our simulation effort on the methods of Raftery and Lewis (1992a,b; 1995), focussing on the central 95% posterior probability interval for $\Delta$. For the .0125 precision tolerance level recommended by Raftery and Lewis for ordinary situations, we calculate that 650 realizations are needed after a burn-in of 300. This number changed to 4020 realizations for the extreme .005 tolerance limits Raftery and Lewis recommend when the posterior may have severely heavy tails. These diagnostics also suggested that there was no detectable autocorrelation between iterations and that it is not necessary to thin the chain to maintain a roughly independent sample. Results seemed to be insensitive (at least to the second decimal place) to increasing the number of iterations from 1000 to 5000.

Figure 9 shows histograms of the total numbers of missing studies simulated in each $p$-value interval at each iteration of the Gibbs sampling. The modal number of missing studies in each $p$-value interval is zero except for the interval $I_5$. The posterior mean number of missing studies is around 22, with about 10 in each of the two higher groups.

### 4.2 Sensitivity Analyses

We carried out several sensitivity analyses on the analyses above. These include using

(a) a variety of priors for $\Delta$, namely

- A restrictive or more informative $N(0,.1^2)$ prior,
- The prior used in the main analysis, $N(0,.15^2),$
- An empirical $N(.1133,.11^2)$ prior based on workplace exposure to ETS (Biggerstaff et al., 1994), and
- A broader $N(0,.4^2)$ prior that is quite uninformative.

(b) an alternative model, which enforces a monotonicity constraint on the publication probabilities, namely, $1 = w^1 \geq w^2 \geq w^3 \geq w^4 \geq w^5$. This constraint was implemented by rejection sampling, and reflects a popular belief about the nature of publication bias.

(c) The same variants but with only four classes, amalgamating $I_4$ and $I_5$ and using a single publication probability on the resultant class, with prior uniform on $(0.3, 0.7)$.

Tables 2 and 3 list the results of these analyses. We also carried out some investigations of sensitivity to the priors for $\tau^2$ and $\sigma^2$, and the meta-analysis results seem to be insensitive to the choice of priors for these parameters.

The posterior for $RR$ appears to be only mildly sensitive to the choice of prior for $\Delta$, and slightly more sensitive to the choice of 4 versus 5 $p$-value intervals (and the corresponding change in the prior on $w$). Sensitivity to the prior for $\Delta$ is understandable because this prior provides information not only about $\Delta$ itself, but also implicitly about the likely amount of publication bias present.

Overall, however, the posterior mean remains within the range 1.09–1.15 for our choices of priors, indicating both a real effect of missing studies, and also that the posterior of $\Delta$ is probably not centered around 1.0, even after assessing the possibility of publication bias and attempting to account for it.
<table>
<thead>
<tr>
<th>Prior for $\Delta$</th>
<th>max$_k w^k = 1$</th>
<th>$1 = w^1 \geq w^2 \geq w^3 \geq w^4 \geq w^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posterior Mean of $RR$</td>
<td>$95%$ Posterior Prob. Int.</td>
</tr>
<tr>
<td>$N(0, SD = 0.1)$</td>
<td>1.12 (1.01, 1.25)</td>
<td>1.12 (1.01, 1.24)</td>
</tr>
<tr>
<td>$N(0, SD = 0.15)$</td>
<td>1.14 (1.00, 1.28)</td>
<td>1.14 (1.03, 1.26)</td>
</tr>
<tr>
<td>$N(1.133, SD = 0.11)$</td>
<td>1.15 (1.02, 1.26)</td>
<td>1.14 (1.03, 1.25)</td>
</tr>
<tr>
<td>$N(0, SD = 0.4)$</td>
<td>1.15 (1.03, 1.31)</td>
<td>1.14 (1.02, 1.27)</td>
</tr>
</tbody>
</table>

Table 2: Results of sensitivity analysis on 5 classes

<table>
<thead>
<tr>
<th>Prior for $\Delta$</th>
<th>max$_k w^k = 1$</th>
<th>$1 = w^1 \geq w^2 \geq w^3 \geq w^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posterior Mean of $RR$</td>
<td>$95%$ Posterior Prob. Int.</td>
</tr>
<tr>
<td>$N(0, SD = 0.1)$</td>
<td>1.09 (0.98, 1.22)</td>
<td>1.10 (1.00, 1.22)</td>
</tr>
<tr>
<td>$N(0, SD = 0.15)$</td>
<td>1.11 (1.00, 1.24)</td>
<td>1.13 (1.01, 1.25)</td>
</tr>
<tr>
<td>$N(1.133, SD = 0.11)$</td>
<td>1.12 (1.01, 1.25)</td>
<td>1.12 (1.02, 1.24)</td>
</tr>
<tr>
<td>$N(0, SD = 0.4)$</td>
<td>1.12 (1.00, 1.25)</td>
<td>1.12 (1.00, 1.26)</td>
</tr>
</tbody>
</table>

Table 3: Results of sensitivity analysis on 4 classes

4.3 United States Data

The initial EPA Draft Report (1990) was criticized for not using a random effects model, especially since the overall data set seems to have some identified subgroups such as country groups within it. In the final EPA Report (1992) the studies were grouped into different geographical areas and the EPA focussed largely on the FE meta-analysis of US studies in drawing conclusions about the public health aspects relevant to the US.

A recent review in California (OEHHA Draft Review, 1996) also used 14 US studies, updating the EPA Report to include a further three studies. The data are given in Table 4; details of these studies are in either the EPA Report (1992) or the OEHHA Draft Review (1996). The values we have used are adjusted for various covariates, and all relate to studies of never-smoking females in the US exposed to spousal ETS, except for Janerich (1990) in which there are both males and females. Note that one might choose to exclude this study on those grounds, but we will not address such issues. We also do not wish here to go into related questions such as choice of adjusted or unadjusted data, or the choice of studies used. Note, however that this can require real decisions; for example, in Section 4.1 we have not used the results in Janerich (1990) but rather those in our Figure 1 taken from the original Varela thesis (1987), on which the Janerich paper is based. The statistical issues raised in all such questions are beyond the scope of this paper, although some of them are addressed by Tweedie et al. (1996) and Mengersen et al. (1995) with regard to these data sets.

Table 4 also gives various relevant meta-analyses, including for comparison purposes the FE
<table>
<thead>
<tr>
<th>Study</th>
<th>RR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brownson (1987)</td>
<td>1.68</td>
<td>(0.39,6.90)</td>
</tr>
<tr>
<td>Brownson (1992)*</td>
<td>1.00</td>
<td>(0.80,1.20)</td>
</tr>
<tr>
<td>Buffler (1984)</td>
<td>0.80</td>
<td>(0.34,1.90)</td>
</tr>
<tr>
<td>Butler (1988)</td>
<td>2.02</td>
<td>(0.48,8.56)</td>
</tr>
<tr>
<td>Correa (1983)</td>
<td>2.07</td>
<td>(0.81,5.25)</td>
</tr>
<tr>
<td>Fontham (1994)*</td>
<td>1.29</td>
<td>(1.04,1.60)</td>
</tr>
<tr>
<td>Garfinkle (1981)</td>
<td>1.17</td>
<td>(0.85,1.61)</td>
</tr>
<tr>
<td>Garfinkle (1985)</td>
<td>1.23</td>
<td>(0.81,1.87)</td>
</tr>
<tr>
<td>Humble (1987)</td>
<td>2.20</td>
<td>(0.80,6.60)</td>
</tr>
<tr>
<td>Kabat (1984)</td>
<td>0.79</td>
<td>(0.25,2.45)</td>
</tr>
<tr>
<td>Kabat (1995)*</td>
<td>1.08</td>
<td>(0.60,1.94)</td>
</tr>
<tr>
<td>Janerich (1990)</td>
<td>0.93</td>
<td>(0.55,1.57)</td>
</tr>
<tr>
<td>Stockwell (1992)*</td>
<td>1.60</td>
<td>(0.80,3.00)</td>
</tr>
<tr>
<td>Wu (1985)</td>
<td>1.20</td>
<td>(0.48,3.01)</td>
</tr>
<tr>
<td>EPA FE Analysis**</td>
<td>1.19</td>
<td>(1.04,1.35)</td>
</tr>
<tr>
<td>RE Meta-analysis</td>
<td>1.16</td>
<td>(1.04,1.31)</td>
</tr>
<tr>
<td>Bayesian Analysis</td>
<td>1.17</td>
<td>(1.02,1.33)</td>
</tr>
<tr>
<td>Pub. Bias Analysis</td>
<td>1.10</td>
<td>(0.95,1.29)</td>
</tr>
</tbody>
</table>

Table 4: Individual studies and meta-analyses of studies of US non-smoking women (except Janerich (1990); see text) exposed to spousal ETS. **EPA did not use studies marked * (although they used an earlier version of Fontham (1994)), and used a 90% CI.

analysis carried out by the EPA on 11 studies, and using only a 90% CI (which has been much criticized although clearly it is not hard to convert). The RE analysis uses the adjustments for variability in the estimate of r^2 in Biggerstaff and Tweedie (1996); the Bayesian analysis was carried out using BUGS as in previous sections. All of these analyses give very similar pictures.

Figures 10 – 12 NEAR HERE

Publication bias methods can be applied to this subset of studies, and the outcome of this is shown in Figures 10 – 12. The funnel plot in Figure 10 again shows a clear and classical indication of perhaps three small negative studies which may have been suppressed.

The Bayesian analysis gives strong support to this heuristic. Figure 11 shows posterior distributions of the imputed missing study numbers. There is a weak indication that there might be about one study missing in the positive range, and there is strong indication of 1-5 studies missing in the group with p > 0.5, i.e. with RR < 1. Overall, there is an estimated mean number of 4.5 studies missing.

The effect on the posterior distribution in Figure 12 is however quite noticeable: Table 4 shows that if we allow for these missing studies, the estimate of risk is lowered from around 1.16-1.18 to 1.10, and the credibility interval also now includes the null value.
5 Conclusions

We have tried in this paper to achieve two goals. Primarily, we have wished to show that in meta-analysis, publication bias is a problem which can be addressed using appropriate tools, rather than just a potential problem which has to be overlooked for lack of any remedies. Secondly, we have used the ETS example, currently one of the most visible and contested uses of meta-analysis in the public health arena, to highlight both the use of the techniques and the difference they can make in real terms.

Our approach has been to examine the data for internal consistency and to impute missing studies based on the model used in the meta-analysis itself. We have seen through simulations that this seems to work effectively, and the ETS examples show that such imputation can lead to noticeable differences, especially in estimated excess risk.

An alternative approach to the problem, quite different to that we have used, is to search the literature for clues that might lead to missing papers. This was carried out in Bero et al. (1994), and they found at that time “five unpublished negative studies” not cited in the EPA Report (1992). They imply that the problem is therefore a minor one, although they did not conduct a further meta-analysis using these extra studies. Interestingly, this is very similar to the 4.5 studies our methods show as missing in the US data set, and we have shown that even this degree of omission can have a serious effect on relative risk estimates.

Nonetheless, Bero et al. (1994) stands out as a more serious attempt than usual to attack this problem. It is more common to find that lip-service is paid to the existence of publication bias but that little attempt is made to account for it. The EPA Report (1992) itself makes no attempt to investigate this issue, and as noted in the introduction, is being forced to defend that position. Other reviews of the studies in this area have also swept aside this question: publication bias is mentioned by the Californian OEHHHA Report (1996, pp. 9–10) but ignored (largely on the basis of the Bero et al. (1994) findings), and similarly is mentioned by the Australian NH&MRC Draft Report (1995, p. 89), but again is ignored. Kawachi and Colditz (1996) also review the issue, noting some further studies either completed or located since the EPA Report, but do not carry out any quantitative analyses which shed light on the effect of publication bias. They cite Vandenbroucke (1988) whose funnel-plot analysis is both dated and difficult to sustain, and Bero et al. (1994) again, in asserting that this is not a problem.

The approach we propose is clearly more systematic than merely looking at funnel plots. Even if one would obviously not wish to dismiss any association just because the publication bias meta-analysis indicates lack of formal statistical significance, one would certainly treat it with much more caution. On the other hand, when the publication bias meta-analysis still yields a significant estimate of increased relative risk, the conclusion is even more convincing in light of the cautious assumption of potential missing studies on which the analysis is based. This added strength, in the context of a formal model and analysis, is an important contribution of the approach developed here.

The changes in estimated relative risk when accounting for publication bias might seem to be small perturbations on small numbers. However, the use of meta-analysis as a tool is clearly much more relevant in precisely those areas where the excess risk is small and not well-established. In
such cases, the estimated level of excess risk is of considerable importance. It plays a big part in terms of trying to establish if there is really an association not due to chance, since it relates to strength of association in using, say, the Bradford Hill criteria (see the NH&MRC Draft Report (1995)). Moreover, if the excess risk is small, then there is much more concern about other possible factors that might have led to it than if it is large: the values of the $RR$ (or even of the lower bound on the confidence interval on the $RR$) need to be at least 2 before many authorities will consider them established (Doll, 1986; Wynder, 1987). We have not gone into these issues here, but the possibility that an observed association might be caused by such factors as diet (Lee, 1992) or misclassification bias (Lee, 1992; Tweedie et al., 1994) certainly should deserve more attention if the excess risk is reduced as it seems to be when allowing for publication bias.

The estimate of excess risk is also central in evaluating the problem that the association might cause in the population, and it is used in the EPA Report (1992, Chapter 6) in this way. There are many parameters that need to be taken into account in estimating the attributable number of lung cancer cases that might flow from spousal exposure to ETS, but as shown in Taylor and Tweedie (1997), the value of the relative risk is one of the most sensitive. If the real value is 1.10 rather than 1.19 then this almost halves the estimated attributable number of cases, and this of itself might have a serious impact on how the exposure is viewed.

Most meta-analyses cover relatively small numbers of studies. The 30 or so available in the ETS example, or the similar number on cervical cancer and oral contraceptive use considered in Smith et al. (1997), represent the type of public health study where one might have some confidence that the imputed studies give a credible representation of the truth. What remains to be developed is a method of handling small collections. Sime (1996) estimates that up to half of all studies granted funding do not get to publication. Until we know what these studies showed, or would have shown if completed, we still run grave risks of making decisions based on very limited, and very biased, data. The methods developed here are just one small step towards improving that situation.

Acknowledgements

We are indebted to the referees for suggesting that the paper be reorganised around the public health implications exemplified in the ETS example, to Sue Taylor for comments on the application of the method to cervical cancer and oral contraceptive data, and to Brad Biggerstaff for some of the random effects modeling. Part of the work is based on material and advice provided by the third author to the Tobacco Industry Association of Australia and tobacco companies in the US (or their representatives) and research partly funded by them, but the opinions expressed here are entirely those of the authors.
References


Figure 1: Confidence intervals and relative risks for the 35 ETS studies, the EPA fixed effects meta-analysis (based on US studies only), the standard random effects meta-analysis, the standard Bayesian meta-analysis, and the Bayesian meta-analysis accounting for potential publication bias.
Figure 2: Funnel plot of 35 ETS studies.
Figure 3: Relative risk posteriors for the simulated data set with suppression of negative studies. The posterior on the left was calculated using data augmentation, and the one on the right assumes no publication bias. The truth is $RR = 1$.

Figure 4: Frequency histograms of the numbers of studies augmented in the $p$-value intervals $[0.00, 0.50]$ and $(0.50, 1.00]$. The true number missing in $[0.00, 0.50]$ is 0, and the true number missing in $(0.50, 1.00]$ is 18. The black triangle represents the mean of the number of studies augmented.
Figure 5: Relative risk posteriors for the simulated data set with heavy suppression of insignificant studies. The posterior on the left was calculated using our data augmentation procedure, and the one on the right assumes no publication bias. The truth is RR = 1.

Figure 6: Frequency histograms of augmented studies in three p-value intervals: [0.00, 0.05], (0.05, 0.10] and (0.10, 1.00]. The true numbers of missing studies were 0, 0, and 31 respectively. The black triangle represents the mean of the number of studies augmented.
Figure 7: Estimated posterior of relative risk and 95% posterior probability region for the ETS example. The posterior on the left was calculated using our data augmentation procedure which accounts for publication bias, and the one on the right assumes no publication bias.
Figure 8: The convergence behavior of estimates of $\Delta$ and the proportion of missing studies for the ETS example.
Figure 9: Frequency histograms of the numbers of missing studies simulated in each $p$-value interval for the ETS example. The black triangle represents the mean of the number of studies augmented.
Figure 10: Funnel plot of 14 US ETS studies.
Figure 11: Frequency histograms of the numbers of missing studies simulated in each \( p \)-value interval for the US ETS studies. The black triangle represents the mean of the number of studies augmented.
Figure 12: Estimated posterior of relative risk and 95% posterior probability region for US ETS studies. The posterior on the left was calculated using our data augmentation procedure which accounts for publication bias, and the one on the right assumes no publication bias.