Statistical Science
1997, Vol. 12, No. 4, 221–250

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 Environmental Protection Agency 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 U.S. studies and in the global collection of studies.

Key words 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," that is, not due to chance alone, are seen as being established and worth consideration.

However, the use of meta-analysis introduces a situation where studies themselves, both significant and 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 evalua-

Geof Givens is Assistant Professor, David Smith is Graduate Research Assistant and Richard Tweedie is Professor, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523 (e-mail: geof@lamar.colostate.edu, http://www.stat.colostate.edu/~geof).

tions 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 U.S. 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, Glantz and Rennie, 1994, page 133).

In any meta-analysis, a well-documented concern (Hedges, 1992; Dear and Begg, 1992; Sterling, Rosenbaum and Weinkam, 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 10 possible missing negative studies, and a similar number of missing insignificant positive studies. After allowing for this, we see in Section 4 that the 95% posterior credibility interval for relative risk is shifted downward toward the null hypothesis of no effect; more important, perhaps, the actual estimate of excess risk is cut by approximately one-third.

When applied to studies in the United States, which the EPA used in its final meta-analysis, a very similar picture emerges: only some four or five 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 previously 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

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

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 2×2 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% and 20% of lung cancer cases occur in nonsmokers (Schneiderman, Davis and Wagener, 1989; Alavanja, Brownson and Boice, 1992). Until the early 1980s, epidemiological studies had not reported any noticeable increase in the incidence of lung cancer among nonsmokers who were exposed to ETS. This changed starting in 1981 when a case–control study in Greece by Trichopoulos and coworkers (Trichopoulos, Kalandidi, Sparros and MacMahon, 1981; Trichopoulos, Kalandidi and Sparros, 1983) and a cohort study in Japan by Hirayama (1981, 1984) reported an association between lung cancer and exposure to ETS in nonsmoking 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, 1990); after receiving comments, this was issued as the final EPA Report (EPA, 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 U.S. studies, after receiving arguments on the validity of non-U.S. studies in forming an estimate of relative risk to be used in the U.S. context.

Since that time a small number of other studies have appeared in the United States. The ETS meta-analysis data that we shall use consists of 35 studies that assess the risk of lung cancer in nonsmoking 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, Tweedie and Biggerstaff (1995) and Tweedie, Mengersen and Eccleston (1994), and represent a complete set of such studies as far as could be determined at the time of preparation of Tweedie, Mengersen and Eccleston (1994).

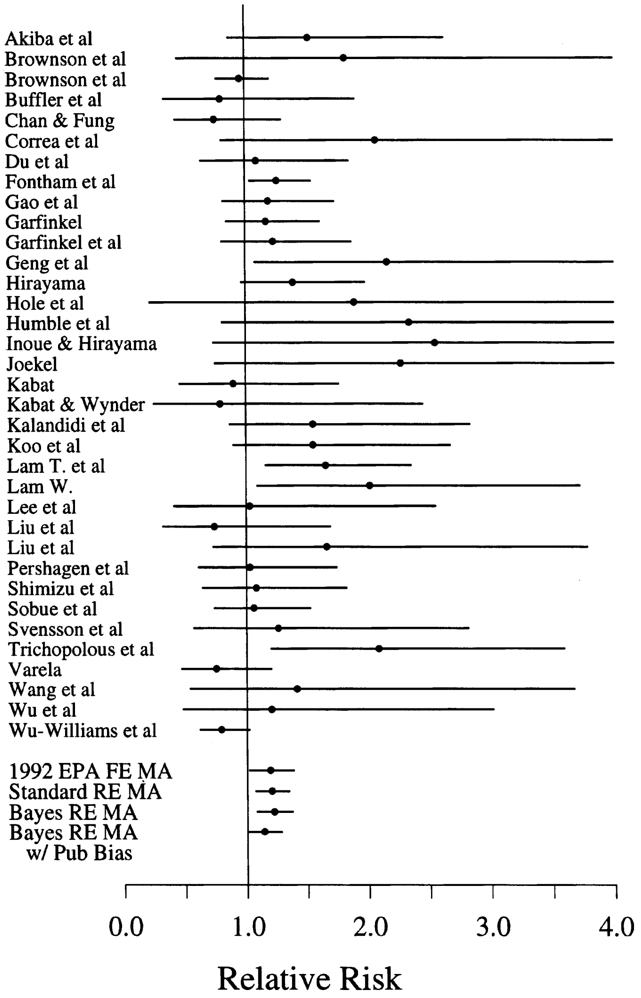

FIG. 1. *Confidence intervals and relative risks for the 35 ETS studies, the EPA fixed effects meta-analysis (based on U.S. studies only), the standard random effects meta-analysis, the standard Bayesian meta-analysis and the Bayesian meta-analysis accounting for potential publication bias.*

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 (OSHA, 1994) and the Australian NH&MRC Draft Report (NH&MRC, 1995)]. As noted in Biggerstaff, Mengersen and Tweedie (1994) and Tweedie, Scott, Biggerstaff and Mengersen (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 to 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 (EPA, 1992), although in comparison with the RE model it has a number of limitations, discussed in some detail in the National Research Council (NRC) Report (NRC, 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, that is, a health benefit.

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

$$Y_j = \Delta + \varepsilon_j \tag{1}$$

where $\varepsilon_j \sim N(0, \sigma_j^2)$, so that Δ is interpreted as the overall relative risk.

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

$$Y_j = \Delta + \beta_j + \varepsilon_j, \tag{2}$$

where $\beta_j \sim N(0, \tau^2)$ is introduced to account for heterogeneity between studies, and $\varepsilon_j \sim N(0, \sigma_j^2)$ represents within-study variability of study j as before. We write $\boldsymbol{\sigma}^2 = \{\sigma_j^2\}$ for these variances.

The RE approach has been argued (NRC, 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

$$\widehat{\Delta} = \frac{\sum Y_j(\sigma_j^2 + \tau^2)^{-1}}{\sum (\sigma_j^2 + \tau^2)^{-1}} \tag{3}$$

with

$$\mathrm{Var}[\widehat{\Delta}] = \frac{1}{\sum (\sigma_j^2 + \tau^2)^{-1}}. \tag{4}$$

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 τ^2 (Biggerstaff and Tweedie, 1996); in both models it is assumed that the σ_j^2 are known, either from estimates based on the raw data in the individual papers or from published estimates in those papers.

TABLE 1
Results from meta-analyses of ETS data

Model	Relative risk	Confidence interval
Fixed effects	1.17	(1.08, 1.26)
Random effects	1.20	(1.07, 1.34)
Bayesian hierarchical	1.22	(1.08, 1.37)

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 $\widehat{\tau}^2 = 0.023$ 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 heterogeneity in the Y_j (expressed through $\tau^2 > 0$) is the use of studies from different countries. The analysis of the ETS data in Mengersen, Tweedie and Biggerstaff (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 heterogeneity has been criticized on these grounds.

Following such comments on the EPA use of FE models, and of amalgamating over different countries, 11 studies relating to the U.S. were used in a FE meta-analysis in the final EPA Report (EPA, 1992). We analyze in more detail in Section 4.3 the 14 studies currently available in the United States. Recent tests for $\tau^2 = 0$ have been developed by Biggerstaff and Tweedie (1996), and applying these to this U.S. data set indicates that in this case the difference between FE and RE models is almost nonexistent: 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, 1992, Table 5-9). Thus the EPA would be reasonably justified, at least in using current U.S. 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, 1992, page 5-31).

2.3 Bayesian Hierarchical Models

In the random effects model, Δ, τ^2 and $\boldsymbol{\sigma}^2$ 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, Δ, τ^2 and $\boldsymbol{\sigma}^2$ 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 Δ, τ^2 and $\boldsymbol{\sigma}^2$ 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 Δ, 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, page 146). Standard Bayesian analyses might use independent conjugate priors, which for this problem are normal for Δ and inverse gamma for $\boldsymbol{\sigma}^2$ and τ^2. The specific priors we adopt are detailed in Section 4.

With these choices, the posterior distribution for Δ 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, Thomas, Best and Gilks, 1996).

Table 1 shows that in the ETS data set 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, e.g., Mosteller and Chalmers, 1992; Felson, 1992; Chalmers, 1991; NRC, 1992; Thompson and Pocock, 1991; Mengersen, Tweedie and Biggerstaff, 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 Ph.D. or M.S. 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, page 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, Rosenbaum and Weinkam (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, Begg and Louis, 1989; Hedges, 1992; Dear and Begg, 1992) and the number of missing studies (Gleser and Olkin, 1996; 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, Tweedie and Biggerstaff, 1995), which is a graphical method to display possible publication bias. It shows the relationship between the estimated value of Δ and the size of the study, measured by, say, the inverse of the standard error, σ_j^{-1}, 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 Δ 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 Δ.

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 Δ 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.

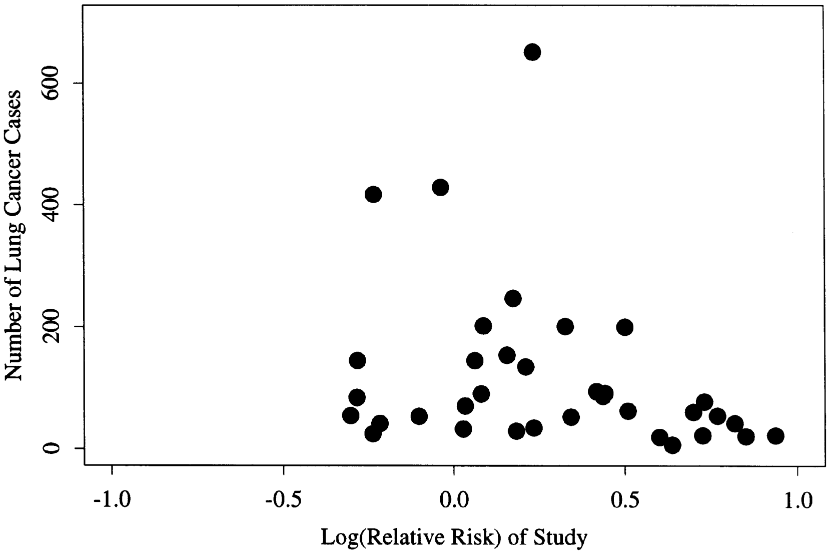

FIG. 2. *Funnel plot of* 35 *ETS studies.*

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, Tweedie and Biggerstaff (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, Tweedie and Biggerstaff, (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, Givens and Tweedie, 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 Sections 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" data set 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, Laird and Rubin, 1977) and the IP algorithm (Tanner and Wong, 1987). The premise of data augmentation is that the "observed data" $\mathbf{Y}$ can be thought of as a partial realization of the random variable $\mathscr{X} = (\mathscr{Y}, \mathscr{Z})$, where a complete realization $\mathbf{X}$ of $\mathscr{X}$ is called the *complete data*, and a realization $\mathbf{Z}$ of $\mathscr{Z}$ is called the *missing* or *latent data*. We assume that the distribution of $\mathscr{X}$ depends on parameters of interest $\boldsymbol{\theta}$ through the family $p(\mathbf{X} \mid \boldsymbol{\theta})$, which gives a marginal distribution $p(\mathbf{Y} \mid \boldsymbol{\theta})$ for the observed data. This framework is most useful when inference about $\boldsymbol{\theta}$ based on $p(\mathbf{Y} \mid \boldsymbol{\theta})$ is difficult, but would be more straightforward using the complete data likelihood $p(\mathbf{X} \mid \boldsymbol{\theta})$.

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 $\mathbf{Y} = (Y_1, \dots, Y_n)$ is

$$p(\mathbf{Y} \mid \Delta, \tau^2, \boldsymbol{\sigma}^2) \propto \prod_{j=1}^{n} \exp\left(-\frac{1}{2}\frac{(Y_j - \Delta)^2}{(\tau^2 + \sigma_j^2)}\right) \bigg/ \sqrt{\tau^2 + \sigma_j^2}. \tag{5}$$

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 jth missing study be denoted as Z_j for $j = (n+1), \dots, (n+m)$, and let $\mathbf{Z} = \{Z_j\}$. We will also denote the complete set of estimated log relative risks for all studies, both observed and missing, by $\mathbf{X} = \{X_j\}$ 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,

$$Z_j = \Delta + \beta_j + \varepsilon_j, \tag{6}$$

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

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 $I_1, \dots, I_c$. A p-value from any individual study must fall into one of these intervals. Now let

$$w^k = \Pr\big[\text{a study with } p\text{-value in } I_k \text{ is published}\big], \quad k = 1, \dots, c, \tag{7}$$

and let $\mathbf{w} = \{w^k\}$. For consistency with model extensions by Smith, Givens and Tweedie (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 (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 $\mathbf{m} = \{m^k\}$. We adopt a negative binomial model for the number of missing studies with p-values in I_k:

$$m^k \mid \mathbf{w} \sim \text{negative binomial}(n^k, w^k). \tag{8}$$

Note that (8) depends on knowing the weight vector $\mathbf{w}$. Hedges (1992) and Dear and Begg (1992) present a maximum likelihood method for estimating the w^k from a meta-analysis data set, 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 $\mathbf{Y}$ 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

$$p(\mathbf{Y} \mid \Delta, \tau^2, \boldsymbol{\sigma}^2, \mathbf{w}) \propto \prod_{j=1}^{n} \prod_{k=1}^{c} \mathbf{1}_{\{p_j \in I_k\}} \frac{\exp(-\frac{1}{2}(Y_j - \Delta)^2/(\tau^2 + \sigma_j^2))}{\sqrt{\tau^2 + \sigma_j^2}}. \tag{9}$$

The latent data are the outcomes $\mathbf{Z}$ of the unobserved studies, and the numbers of such studies $\mathbf{m}$. 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.

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

$$p(\mathbf{X} \mid \Delta, \tau^2, \boldsymbol{\sigma}^2, \mathbf{m}) \propto \prod_{j=1}^{n+m} \prod_{k=1}^{c} \mathbf{1}_{\{p_j \in I_k\}} \frac{\exp(-\frac{1}{2}(X_j - \Delta)^2/(\tau^2 + \sigma_j^2))}{\sqrt{\tau^2 + \sigma_j^2}}. \tag{10}$$

We stress that (10) 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

$$\begin{aligned} p(\mathbf{X}, \mathbf{m} \mid \Delta, \tau^2, \boldsymbol{\sigma}^2, \mathbf{w}) &\propto p(\mathbf{X} \mid \Delta, \tau^2, \boldsymbol{\sigma}^2, \mathbf{m}) \\ &\quad \cdot \prod_{k=1}^{c} \binom{n^k + m^k - 1}{m^k} (w^k)^{n^k} (1 - w^k)^{m^k}. \end{aligned} \tag{11}$$

In our Bayesian analysis, we adopt independent prior distributions $p(\Delta)$, $p(\tau^2)$, $p(\boldsymbol{\sigma}^2)$, $p(\mathbf{w})$ and $p(\mathbf{Z})$ for the model and latent data treated as nuisance parameters. Since $\mathbf{m}$ and $\mathbf{w}$ are related through (8), 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 may take σ_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 $\mathbf{w}$ 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 (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, Taylor, Smith and Tweedie, 1996).

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

$$p(\Delta \mid \cdot) \propto \frac{p(\Delta)}{A(\Delta)} \prod_{j=1}^{n+m} \exp\left(-\frac{1}{2} \frac{(X_j - \Delta)^2}{(\tau^2 + \sigma_j^2)}\right), \tag{12}$$

$$\begin{aligned} p(\tau^2 \mid \cdot) &\propto \frac{p(\tau^2)}{A(\tau^2)} \\ &\quad \cdot \prod_{j=1}^{n+m} \left[\exp\left(-\frac{1}{2} \frac{(X_j - \Delta)^2}{(\tau^2 + \sigma_j^2)}\right) \Big/ \sqrt{\tau^2 + \sigma_j^2}\right], \end{aligned} \tag{13}$$

where here and below A is a normalizing function $A(\Delta, \tau^2, \boldsymbol{\sigma}^2, \mathbf{w})$, 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:

$$\begin{aligned} p(\mathbf{Z}, \boldsymbol{\sigma}^2 \mid \cdot) &\propto \frac{p(\mathbf{Z}, \boldsymbol{\sigma}^2)}{A(\boldsymbol{\sigma}^2)} \\ &\quad \cdot \prod_{j=1}^{n+m} \prod_{k=1}^{c} \frac{\exp(-\frac{1}{2}(X_j - \Delta)^2 / (\tau^2 + \sigma_j^2))}{\sqrt{\tau^2 + \sigma_j^2}} \\ &\quad \cdot \mathbf{1}_{\{p_j \in I_k\}}. \end{aligned} \tag{14}$$

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

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

$$p(\mathbf{m} \mid \mathbf{w}) \propto \prod_{k=1}^{c} \binom{n^k + m^k - 1}{m^k} (w^k)^{n^k} (1 - w^k)^{m^k}, \tag{15}$$

since we have no prior on $\mathbf{m}$, as discussed above.

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

$$p(\mathbf{w} \mid \cdot) \propto \frac{p(\mathbf{w})}{A(\mathbf{w})} p_1(\mathbf{w} \mid \cdot), \tag{16}$$

where $p_1(\mathbf{w} \mid \cdot)$ is the conditional probability density function that results when the conditional probability density function of $\mathbf{w} \times \max_k w^k$ 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 Δ 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: $(\mathbf{m}, \mathbf{Z}, \boldsymbol{\sigma}^2)$, $\mathbf{w}$, Δ and τ^2. We update $\mathbf{m}$, $\mathbf{Z}$ and $\boldsymbol{\sigma}^2$ 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 σ_j^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^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 $p(\mathbf{Z} \mid \boldsymbol{\sigma}^2, \cdot)$. Note also that in our examples below we assume that σ_j^2 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, Flannery, Teukolsky and Vetterling, 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, Δ, we calculate the corresponding sample quantiles from a collection of values of Δ obtained via simulation. Iteration length, burn-in and subsampling are discussed in Sections 3.5 and 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, β_j and ϵ_j in (2) are nonidentifiable. 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 σ_j^2 is drawn is degenerate at $\hat{\sigma}_j^2$ for the jth study when j indexes an observed study, but the remaining σ_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 $\hat{\sigma}_j^2 + \tau^2$, where $\tau^2 = 0.03$. 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 et al., 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 $\mathbf{w}$, these meta-analyses all had identical priors given by $\Delta \sim \text{normal}(0, 0.4^2)$, $\tau^2 \sim \text{inverse gamma}(\text{shape} = 32, \text{mean} = 1/32)$ 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 1,000 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 (0.889, 1.156) for $RR = \exp \Delta$, which indicates that the sample we had drawn was not an aberrant one.

3.5(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 envelops the true probability of being published while not reflecting strong beliefs about the amount of publication bias present in the data set.

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.

3.5(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.

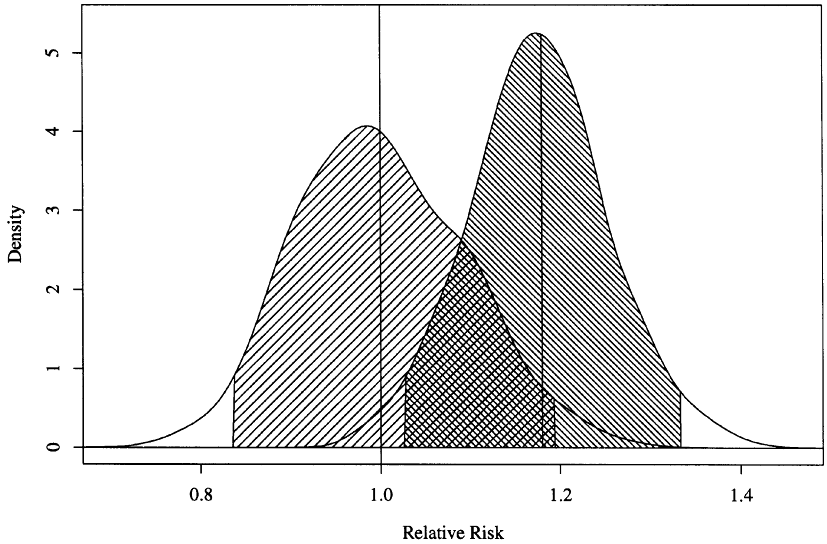

FIG. 3. *Relative risk posteriors for the simulated data set with suppression of negative studies. The posterior on the left was calculated using data augmentation, and the one on the right assumes no publication bias. The truth is $RR = 1$.*

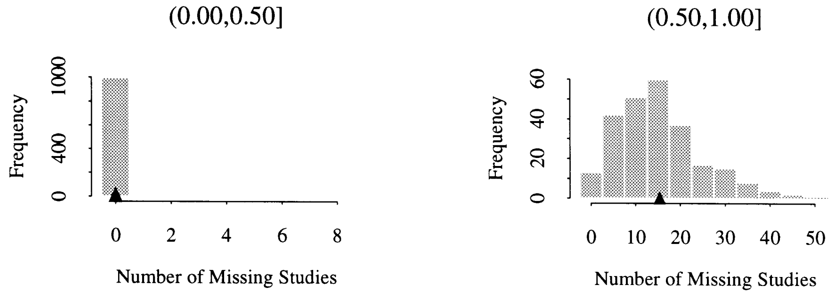

FIG. 4. *Frequency histograms of the numbers of studies augmented in the p-value intervals* [0.00, 0.50] *and* (0.50, 1.00]. *The true number missing in* [0.00, 0.50] *is* 0, *and the true number missing in* (0.50, 1.00] *is* 18. *The black triangle represents the mean of the number of studies augmented.*

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.

3.5(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 were 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 nontrivial probability of a study being unpublished.

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,

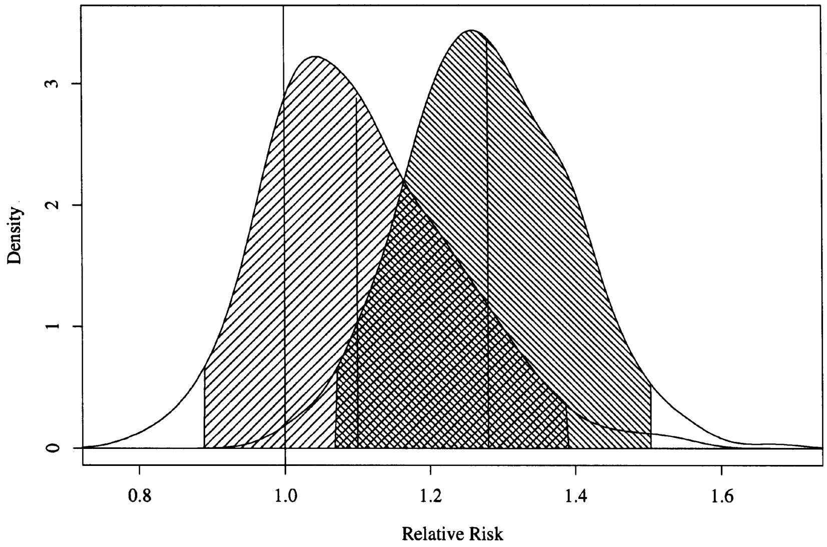

FIG. 5. *Relative risk posteriors for the simulated data set with heavy suppression of insignificant studies. The posterior on the left was calculated using our data augmentation procedure, and the one on the right assumes no publication bias. The truth is* $RR = 1$.

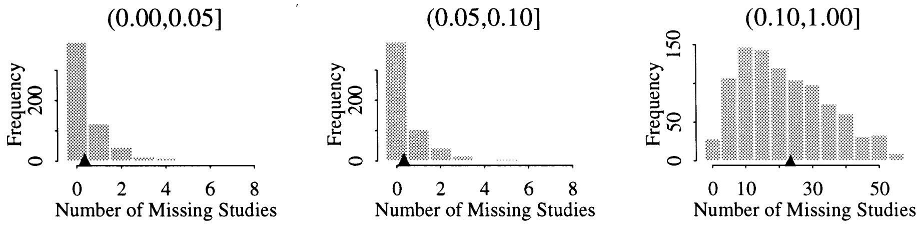

FIG. 6. *Frequency histograms of augmented studies in three p-value intervals*: $[0.00, 0.05]$, $(0.05, 0.10]$ *and* $(0.10, 1.00]$. *The true numbers of missing studies were* 0, 0, *and* 31, *respectively. The black triangle represents the mean of the number of studies augmented.*

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.

3.5(d). *Strong suppression applied with different priors*. The prior used in Section 3.5(c) 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 Section 3.5(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 Section 3.5(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 $w^3 = 0.3$.

These results indicates that, although the method is somewhat sensitive to choice of prior on $\mathbf{w}$, 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 Δ 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 $I_1, \dots, I_c$ to stratify p-values. We use the intervals $[0, 0.01]$, $(0.01, 0.05]$, $(0.05, 0.10]$, $(0.10, 0.50]$ and $(0.50, 1.00]$ 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 m^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 σ_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 analyze 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 into the model.

4. PUBLICATION BIAS IN THE ETS DATA SET

4.1 The Possible Effect of Bias

We now apply the methods above to the ETS data set 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 satisfyingly close to the ad hoc value of 1.12 with 95% posterior probability interval (1.01, 1.24) found in Mengersen, Tweedie and Biggerstaff (1995).

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 Δ, 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 σ_j^2 corresponding to a missing study, based on the 35 published variances. We also assume that each individual study variance $\hat{\sigma}_j^2$ is exactly correct, so $p(\boldsymbol{\sigma}^2)$ is degenerate for observed studies. For τ^2, we use an exponential prior with mean 0.031, based on a meta-analysis of studies on workplace ETS (Biggerstaff, Mengersen and Tweedie, 1994). We take an improper uniform prior for $\mathbf{Z}$ and our initial prior for $\mathbf{w}$ (before being scaled by the largest) is that of three

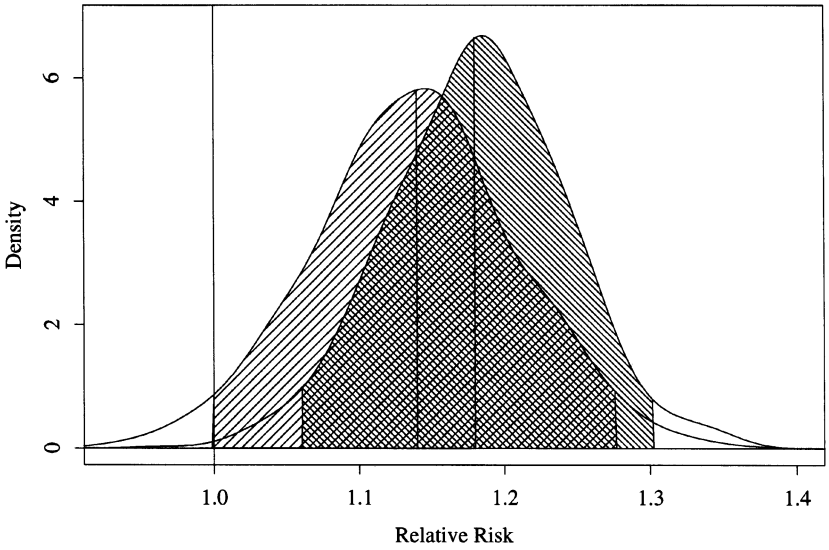

FIG. 7. *Estimated posterior of relative risk and 95% posterior probability region for the ETS example. The posterior on the left was calculated using our data augmentation procedure which accounts for publication bias, and the one on the right assumes no publication bias.*

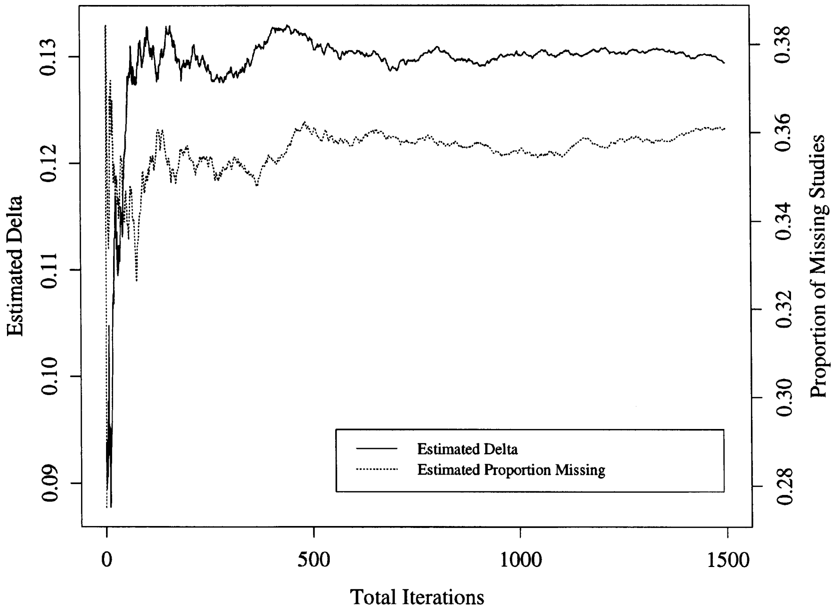

FIG. 8. *The convergence behavior of estimates of* Δ *and the proportion of missing studies for the ETS example.*

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.

We used a burn-in of 500 iterations and stored an additional 1,000 iterations. Figure 8 shows the convergence behavior of the Gibbs approach with respect to Δ 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 Δ. For the 0.0125 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 4,020 realizations for the extreme 0.005 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 1,000 to 5,000.

Figure 9 shows histograms of the total numbers of missing studies simulated in each p-value interval

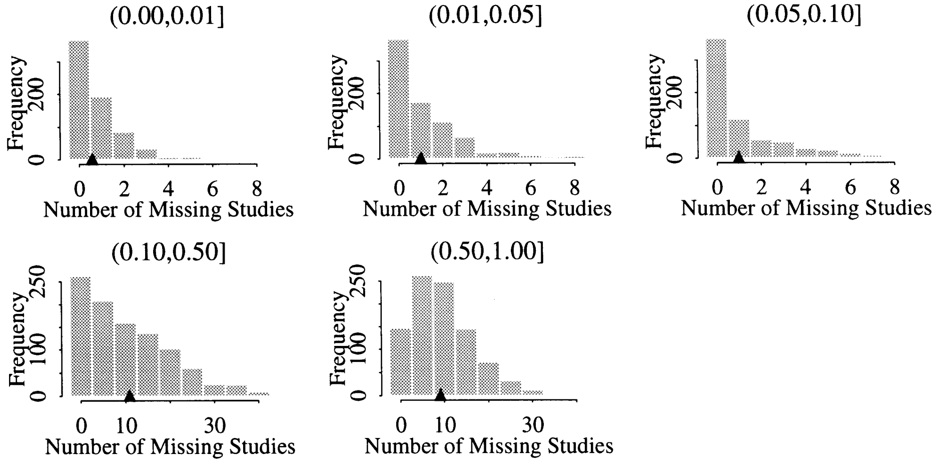

FIG. 9. *Frequency histograms of the numbers of missing studies simulated in each p-value interval for the ETS example. The black triangle represents the mean of the number of studies augmented.*

TABLE 2
Results of sensitivity analysis on five classes

	$\max_k w^k = 1$		$1 = w^1 \geq w^2 \geq w^3 \geq w^4 \geq w^5$	
Prior for Δ	**Posterior mean of *RR***	**95% posterior prob. int.**	**Posterior mean of *RR***	**95% posterior prob. int.**
$N(0,\ \mathrm{SD} = 0.1)$	1.12	(1.01, 1.25)	1.12	(1.01, 1.24)
$N(0,\ \mathrm{SD} = 0.15)$	1.14	(1.00, 1.28)	1.14	(1.03, 1.26)
$N(0.1133,\ \mathrm{SD} = 0.11)$	1.15	(1.02, 1.26)	1.14	(1.03, 1.25)
$N(0,\ \mathrm{SD} = 0.4)$	1.15	(1.03, 1.31)	1.14	(1.02, 1.27)

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 the following:

(a) a variety of priors for Δ, namely, (i) A restrictive or more informative $N(0, 0.1^2)$ prior, (ii) the prior used in the main analysis, $N(0, 0.15^2)$, (iii) an empirical $N(0.1133, 0.11^2)$ prior based on workplace exposure to ETS (Biggerstaff, Mengersen and Tweedie, 1994) and (iv) a broader $N(0, 0.4^2)$ 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 τ^2 and $\boldsymbol{\sigma}^2$, 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 Δ, and slightly more sensitive to the choice of four versus five p-value intervals (and the corresponding change in the prior on $\mathbf{w}$). Sensitivity to the prior for Δ is understandable because this prior provides information not only about Δ 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 Δ is probably not centered around 1.0, even after assessing the possibility of publication bias and attempting to account for it.

4.3 United States Data

The initial EPA Draft Report (EPA, 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 (EPA, 1992) the studies were grouped into different geographical areas and the EPA focussed largely on the FE meta-analysis of U.S. studies in drawing conclusions about the public health aspects relevant to the United States.

A recent review in California (OEHHA, 1996) also used 14 U.S. studies, updating the EPA Report to include a further 3 studies. The data are given in Table 4; details of these studies are in either the EPA Report (EPA, 1992) or the OEHHA Draft Review (OEHHA, 1996). The values we have used are adjusted for various covariates, and all relate to

TABLE 3
Results of sensitivity analysis on four classes

	$\max_k w^k = 1$		$1 = w^1 \geq w^2 \geq w^3 \geq w^4$	
Prior for Δ	**Posterior mean of *RR***	**95% posterior prob. int.**	**Posterior mean of *RR***	**95% posterior prob. int.**
$N(0,\ \mathrm{SD} = 0.1)$	1.09	(0.98, 1.22)	1.10	(1.00, 1.22)
$N(0,\ \mathrm{SD} = 0.15)$	1.11	(1.00, 1.24)	1.13	(1.01, 1.25)
$N(0.1133,\ \mathrm{SD} = 0.11)$	1.12	(1.01, 1.25)	1.12	(1.02, 1.24)
$N(0,\ \mathrm{SD} = 0.4)$	1.12	(1.00, 1.25)	1.12	(1.00, 1.26)

TABLE 4
Individual studies and meta-analyses of studies of U.S. nonsmoking women (except Janerich et al., 1990; see text) exposed to spousal ETS

Study	*RR*	95% CI
Brownson (1987)	1.68	(0.39, 6.90)
Brownson (1992)†	1.00	(0.80, 1.20)
Buffler (1984)	0.80	(0.34, 1.90)
Butler (1988)	2.02	(0.48, 8.56)
Correa (1983)	2.07	(0.81, 5.25)
Fontham (1994)†	1.29	(1.04, 1.60)
Garfinkle (1981)	1.17	(0.85, 1.61)
Garfinkle (1985)	1.23	(0.81, 1.87)
Humble (1987)	2.20	(0.80, 6.60)
Kabat (1984)	0.79	(0.25, 2.45)
Kabat (1995)†	1.08	(0.60, 1.94)
Janerich et al. (1990)	0.93	(0.55, 1.57)
Stockwell (1992)†	1.60	(0.80, 3.00)
Wu (1985)	1.20	(0.48, 3.01)
EPA FE analysis*	1.19	(1.04, 1.35)
RE Meta-analysis	1.16	(1.04, 1.31)
Bayesian analysis	1.17	(1.02, 1.33)
Publication bias analysis	1.10	(0.95, 1.29)

*EPA did not use studies marked with a dagger (†) (although they used an earlier version of Fontham, 1994) and used a 90% CI.

studies of never-smoking females in the U.S. exposed to spousal ETS, except for Janerich et al. (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 et al. (1990) but rather those in our Figure 1 taken from the original Varela thesis (Varela, 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, Tweedie and Biggerstaff (1995) with regard to these data sets.

Table 4 also gives various relevant meta-analyses, including for comparison purposes the FE 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 τ^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.

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 $p > 0.5$, that is, with $RR < 1$. Overall, there is an estimated mean number of 4.5 studies missing.

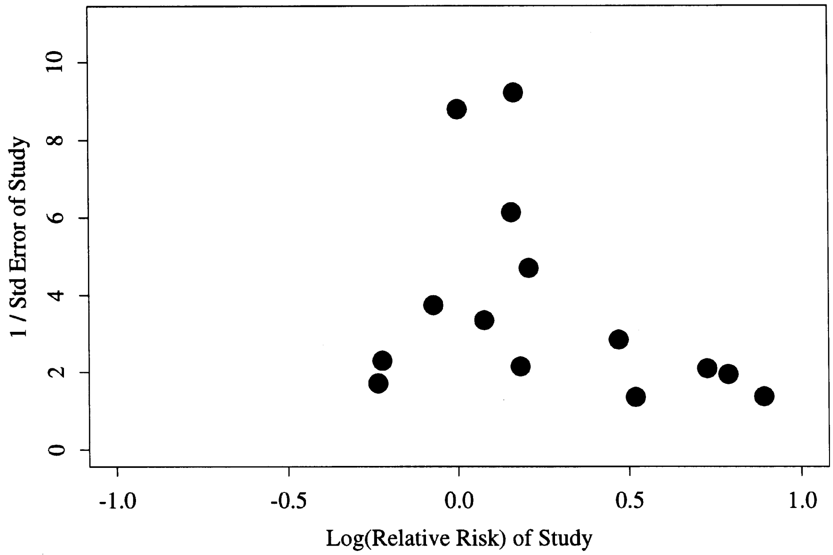

FIG. 10. *Funnel plot of* 14 *U.S. ETS studies.*

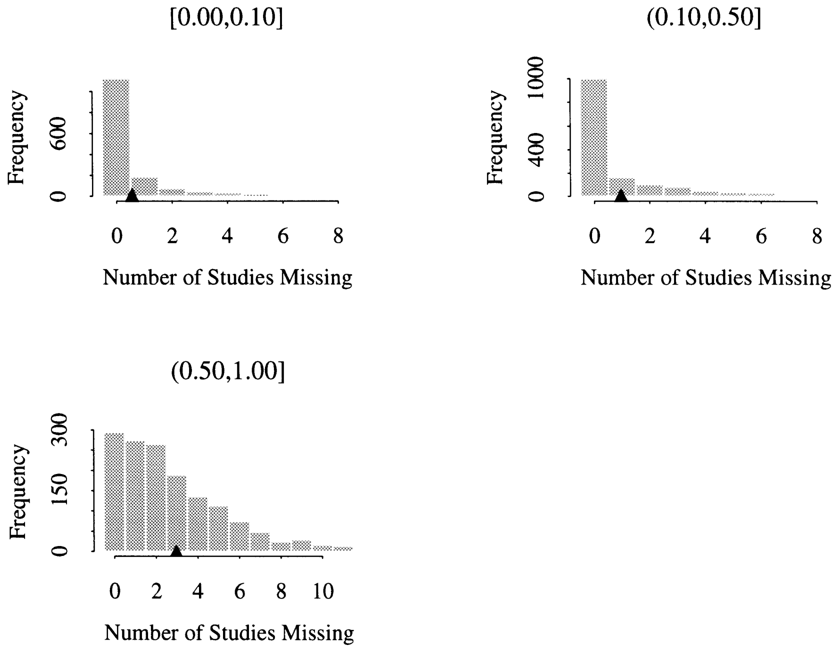

FIG. 11. *Frequency histograms of the numbers of missing studies simulated in each p-value interval for the U.S. ETS studies. The black triangle represents the mean of the number of studies augmented.*

The effect on the posterior distribution in Figure 12 is, however, quite noticable: 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

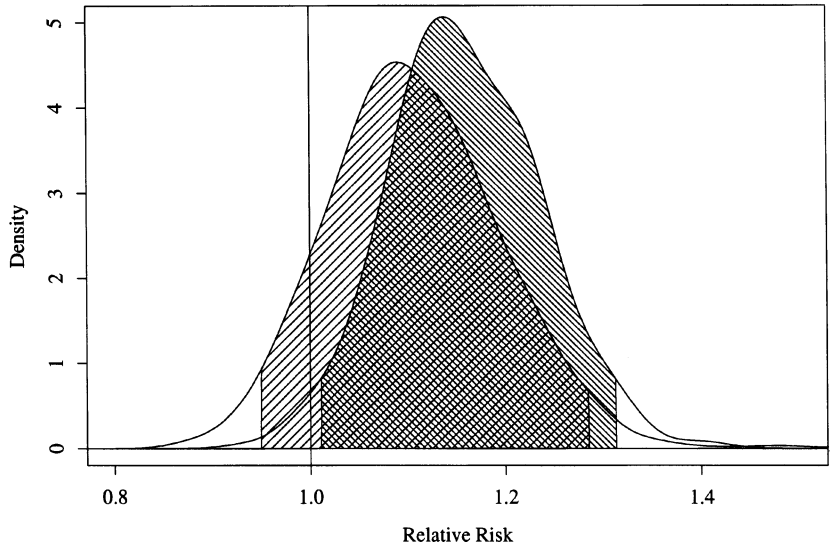

FIG. 12. *Estimated posterior of relative risk and 95% posterior probability region for U.S. ETS studies. The posterior on the left was calculated using our data augmentation procedure which accounts for publication bias, and the one on the right assumes no publication bias.*

just a potential problem which has to be overlooked for lack of any remedies. Second, 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 noticable differences, especially in estimated excess risk.

An alternative approach to the problem, quite different from that we have used, is to search the literature for clues that might lead to missing papers. This was carried out in Bero, Glantz and Rennie (1994), and they found at that time "five unpublished negative studies" not cited in the EPA Report (EPA, 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 U.S. data set, and we have shown that even this degree of omission can have a serious effect on relative risk estimates.

Nonetheless, Bero, Glantz and Rennie (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 (EPA, 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 OEHHA Report (OEHHA, 1996, pages 9–10) but ignored (largely on the basis of the findings of Bero, Glantz and Rennie, 1994), and similarly is mentioned by the Australian NH&MRC Draft Report (NH&MRC, 1995, page 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, Glantz and Rennie (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 (NH&MRC, 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, Mengersen and Eccleston, 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 (EPA, 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, Givens and Tweedie (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. Simes (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 toward improving that situation.

ACKNOWLEDGMENTS

We are indebted to the referees for suggesting that the paper be reorganized 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 United States (or their representatives) and research partly funded by them, but the opinions expressed here are entirely those of the authors.

REFERENCES

ALAVANJA, M. C. R., BROWNSON, R. C. and BOICE, J. D., JR. (1992). Pre-existing lung disease and lung cancer among nonsmoking women. *American Journal of Epidemiology* **6** 623–632.

BERLIN, J. A., BEGG, C. B. and LOUIS, T. A. (1989). An assessment of publication bias using a sample of published clinical trials. *J. Amer. Statist. Assoc.* **84** 381–392.

BERO, L. A., GLANTZ, S. A. and RENNIE, D. (1994). Publication bias and public health policy on environmental tobacco smoke. *Journal of the American Medical Association* **272** 133–136.

BESAG, J. E. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 25–37.

BIGGERSTAFF, B. J., MENGERSEN, K. L. and TWEEDIE, R. L. (1994). Passive smoking in the workplace: a classical and Bayesian meta-analysis. *International Archives of Occupational and Environmental Health* **66** 269–277.

BIGGERSTAFF, B. J. and TWEEDIE, R. L. (1996). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16** 753–768.

CARLIN, J. B. (1992). Meta-analysis for 2 × 2 tables: a Bayesian approach. *Statistics in Medicine* **11** 141–158.

CHALMERS, T. C. (1991). Problems induced by meta-analysis. *Statistics in Medicine* **10** 971–980.

COOPER, H. and HEDGES, L. V., eds. (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York.

CROSSEN, C. (1994). *Tainted Truth: The Manipulation of Fact in America*. Simon and Schuster, New York.

DEAR, K. B. G. (1995). Personal communication.

DEAR, K. B. G. and BEGG, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statist. Sci.* **7** 237–245.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DOLL, R. (1986). The aetiology of the Spanish toxic shock syndrome: interpretation of the epidemiological evidence. Report to the WHO Regional Office for Europe.

DUMOUCHEL, W. (1990). Bayesian meta-analysis. In *Statistical Methods for Pharmacology* (D. Berry, ed.) 509–529. Dekker, New York.

EBERLY, L. E. and CASELLA, G. (1996). Estimating the number of unseen studies. Technical Report BUM 1308-MA, Biometrics Unit, Cornell Univ.

EPA (1990). Health effects of passive smoking: assessment of lung cancer in adults and respiratory disorders in children. Draft report, United States EPA, Washington, D.C.

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

FELSON, D. T. (1992). Bias in meta-analytic research. *Journal Clinical Epidemiology* **45** 885–892.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

GLESER, L. J. and OLKIN, I. (1996). Models for estimating the number of unpublished studies. *Statistics in Medicine* **15** 2493–2507.

HEDGES, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statist. Sci.* **7** 246–255.

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-analysis*. Academic Press, New York.

HIRAYAMA, T. (1981). Nonsmoking wives of heavy smokers have a higher risk of lung cancer: a study from Japan. *British Medical Journal* **282** 183–185.

HIRAYAMA, T. (1984). Lung cancer in Japan: effects of nutrition and exposure to ETS. In *Lung Cancer: Causes and Preventions* 175–195. Verlag Chemie, Weinheim.

IYENGAR, S. and GREENHOUSE, J. B. (1988). Selection models and the file drawer problem (with discussion). *Statist. Sci.* **3** 109–135.

JANERICH, D. T., THOMPSON, W. D., VARELA, L. R., GREENWALD, P., CHOROST, S., TUCCI, C., ZAMAN, M. B., MELAMED, M. R., KIELY, M. and MCKNEALLY, M. F. (1990). Lung cancer and exposure to tobacco smoke in the household. *New England Journal of Medicine* **323** 632–636.

KAWACHI, I. and COLDITZ, G. C. (1996). Invited commentary: confounding, measurement error, and publication bias in studies of passive smoking. *American Journal of Epidemiology* **144** 909–915.

LAFLEUR, B., TAYLOR, S. J., SMITH, D. D. and TWEEDIE, R. L. (1996). Bayesian assessment of publication bias in meta-analyses of cervical cancer and oral contraceptives. In *Proceedings of the 1996 Epidemiology Section of the Joint Statistical Meetings*, Amer. Statist. Assoc., Alexandria, VA 32–37.

LAROSE, D. T. and DEY, D. K. (1995). Modeling publication bias using weighted distributions in a Bayesian framework. Technical Report 95-02, Dept. Statistics, Univ. Connecticut.

LEE, P. N. (1992). *Environmental Tobacco Smoke and Mortality*. Karger, Basel.

LIGHT, R. J. and PILLEMER, D. B. (1984). *Summing Up: The Science of Reviewing Research*. Harvard Univ. Press.

MAUSNER, J. S. and KRAMER, S. (1985). *Mausner & Bahn Epidemiology—An Introductory Text,* 2nd ed. Saunders, Philadelphia.

MENGERSEN, K. L., TWEEDIE, R. L. and BIGGERSTAFF, B. J. (1995). The impact of method choice in meta-analysis. *Austral. J. Statist.* **7** 19–44.

MOSTELLER, F. and CHALMERS, T. C. (1992). Some progress and problems in meta-analysis of clinical trials. *Statist. Sci.* **7** 227–236.

NH&MRC (1995). *The Health Effects of Passive Smoking: Draft Report*. NH&MRC Working Party, Canberra, Australia.

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

OEHHA (1996). Carcinogenic effects of exposure to environmental tobacco smoke. Excerpt: ETS and lung cancer. Review draft report, Reproductive and Cancer Hazard Assessment Section, Office of Environmental Health Hazard Assessment, CA.

OSHA (1994). Proposed rule on indoor air quality. *Federal Register* **59**(65) 15968–16039. (Draft regulation.)

OLKIN, I. (1992). Meta-analysis: methods for combining independent studies. *Statist. Sci.* **7** 226.

PATIL, G. P. and TAILLIE, C. (1989). Probing encountered data, meta-analysis and weighted distribution methods. In *Statistical Data Analysis and Inference* (Y. Dodge, ed.). North-Holland, Amsterdam.

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

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1986). *Numerical Recipes: The Art of Scientific Computing*. Cambridge Univ. Press.

RAFTERY, A. E. and LEWIS, S. M. (1992a). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 763–773. Oxford Univ. Press.

RAFTERY, A. E. and LEWIS, S. M. (1992b). One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statist. Sci.* **7** 493–497.

RAFTERY, A. E. and LEWIS, S. M. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In *Practical Markov Chain Monte Carlo* (W. R. Gilks, D. J. Spiegelhalter and S. Richardson, eds.). Chapman and Hall, London.

SCHNEIDERMAN, M. A., DAVIS, D. L. and WAGENER, D. K. (1989). Lung cancer that is not attributable to smoking: letter to the editor. *Journal of the American Medical Association* **261** 2635–2636.

SIMES, R. J. (1996). Strategies for minimising bias in systematic reviews of randomised trials. Presented at Sydney International Statistical Congress, Sydney, Australia.

SMITH, D. D., GIVENS, G. H. and TWEEDIE, R. L. (1997). Adjustment for publication and quality bias in Bayesian meta-analysis. Unpublished manuscript.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 3–23.

SPIEGELHALTER, D., THOMAS, A., BEST, N. and GILKS, W. (1996). *BUGS: Bayesian Inference Using Gibbs Sampling, v. 0.50*. MRC Biostatistics Unit, Institute of Public Health, Cambridge.

STERLING, T. D., ROSENBAUM, W. L. and WEINKAM, J. J. (1995). Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *Amer. Statist.* **49** 108–112.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

TAYLOR, S. J. and TWEEDIE, R. L. (1997). Assessing sensitivity to multiple factors in calculating attributable risks. *Environmetrics* **8** 351–372.

TERRELL, G. R. (1990). The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc.* **85** 470–477.

THOMPSON, S. G. (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research* **2** 173–192.

THOMPSON, S. G. and POCOCK, S. J. (1991). Can meta-analyses be trusted? *The Lancet* **338** 1127–1130.

TRICHOPOULOS, D., KALANDIDI, A. and SPARROS, L. (1983). Lung cancer and exposure to ETS. Conclusion of Greek Study. *The Lancet* **ii** 677–678.

TRICHOPOULOS, D., KALANDIDI, A., SPARROS, L. and MACMAHON, B. (1981). Lung cancer and exposure to ETS. *International Journal of Cancer* **27** 1–4.

TWEEDIE, R. L., MENGERSEN, K. L. and ECCLESTON, J. A. (1994). Garbage in, garbage out: can statisticians quantify the effects of poor data? *Chance* **7**(2) 20–27.

TWEEDIE, R. L., SCOTT, D. J., BIGGERSTAFF, B. J. and MENGERSEN, K. L. (1996). Bayesian meta-analysis, with application to studies of environmental tobacco smoke and lung cancer. *Lung Cancer Suppl. 1* **14** S171–S194.

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

VARELA, L. R. (1987). Assessment of the association between passive smoking and lung cancer. Ph.D. dissertation, Yale Univ.

WYNDER, E. L. (1987). Workshop on guidelines to the epidemiology of weak associations. *Preventive Medicine* **16** 139–141.

Comment

Colin B. Begg

The credibility of the statistical analysis of any data set should be influenced, to a considerable extent, by the quality of the data. The recent interest in publication bias is a recognition of a specific data quality problem, and one that is particularly apparent, and theoretically correctable, in the context of meta-analysis. However, this is not the only problem with the data on which the meta-analysis of Givens, Smith and Tweedie is based. My comments will encompass the issue of data quality in some detail, in addition to some purely technical issues regarding the analysis.

DATA QUALITY ISSUES

Givens, Smith and Tweedie advance the thesis that publication bias has influenced substantially the interpretation of the available data on lung cancer risk induced by passive smoking, concluding that the "estimated excess risk may be overstated by around 30%." The notion that selective publication could occur in this context is certainly highly plausible. All but 3 of the 35 studies analyzed are case–control studies of lung cancer. When conducting case-control studies of cancer, epidemiologists will usually collect detailed information on a broad range of risk factors, including diet, alcohol consumption and a variety of other lifestyle and personal factors, in addition to detailed information on smoking history. In fact, smoking is such a pervasive risk factor it is included in the majority of cancer epidemiologic studies, and information on passive smoking will undoubtedly have been collected on an indeterminate number of lung cancer studies not included in this meta-analysis. On completion of a case–control study, the investigators will then typically publish the results in a series of articles, each one dealing with a different risk factor of set of factors. The image of investigators trawling around their factor-rich datasets looking for interesting correlations to publish is a familiar one to statisticians working in this field. An important consideration in sorting out the wheat from the chaff in this setting is the concept of "scientific intent." Usually the study will have been constructed with a stated primary focus, comprising a specific hypothesis or set of hypotheses, and the other factors will be collected either to permit adjustments for potential confounding, or simply for hypothesis-generation purposes. The results pertaining to the primary hypotheses thus have greater credibility than any unexpected or casual observations from the remaining data. In short, results stemming from a hypothesis-driven strategy have more credibility than those stemming from a data-driven strategy.

It is of interest to examine the passive smoking studies from this perspective. In the short time I had available to prepare this commentary I was unable to access the articles from all of the studies. However, I looked at the majority of them and it is clear that most of the articles report subsets of data from larger case–control studies of lung cancer, in which the subsets comprise nonsmoking cases and controls. Indeed, the few cohort studies also appear to be of this variety. It could be argued that the very fact that information on passive smoking was collected is evidence of "scientific intent" with regard to the passive smoking hypothesis, although I am personally skeptical of this in view of the encompassing nature of data collection with regard to risk factors, as outlined above.

However, there are some studies that are unequivocally hypothesis-driven, and a close examination of the methods used in these studies is instructive about some of the additional problems of studying this issue, unrelated to publication bias. In the study by Garfinkel, Auerbach and Joubert (1985) the investigators identified cases of lung cancer retrospectively in women diagnosed between 1971 and 1981 at four hospitals, and used corresponding cases diagnosed with colorectal cancer as controls, on the presumption that smoking is unrelated to colorectal cancer, a topic that has become more controversial since the publication of the article. The nonsmokers and ex-smokers were identified on the basis of the information in the medical charts, obtained as routine clinical information. The individuals so identified comprise the cases and controls, and only nonsmokers or cases with missing data were followed up. Detailed data on passive smoking were obtained by interviewing

Colin B. Begg, Ph.D., is Chairman, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021-6007.

the proband, or the spouse if the proband was deceased, or the next of kin if the proband was unmarried or no longer living with the spouse (*note*: only 57% of cases were married and living with their husbands at the time of the cancer diagnosis). Passive smoking exposure was based on the habits of the husband or cohabitating relative, whichever was appropriate, and was based on questions relating to issues such as the duration and intensity of smoking, the number of hours per day the proband was exposed and recollected childhood exposures. Women whose "husbands" smoked only occasionally were designated as not exposed. Janerich et al. (1990) pursued a more ambitious strategy using population-based cases and controls, but with similar data collection and case identification methods. Clearly, in these settings, there are numerous opportunities for exposure misclassification, both in absolute and in relative terms. Moreover, these two studies are probably among the best in the meta-analysis with respect to bias control and care in data collection, in addition to their hypothesis-driven intent. In fact in the pioneering study in this genre, the study conducted by Trichopoulous et al. (1981) in Athens, not only was the small series of 51 cases hospital-based, but the controls came from entirely different hospitals, limitations which were recognized and acknowledged by the authors. The bottom line is that the reported results from the component studies in the meta-analysis are individually much "softer" than is reflected in the reported statistical confidence intervals.

There exists a sentiment among some commentators that meta-analysis is simply an inappropriate tool for use in aggregating nonrandomized studies, and that it is especially inappropriate in evaluating "small" effects. Certainly, one has the sense that there is a substantial fudge-factor associated with the studies in this analysis, and that it would not matter how many additional observational studies might be performed. As long as the results continue to produce a summary relative risk in the region of 1.2 or less one can never obtain truly convincing evidence of the causal link between passive smoking and lung cancer based only on these kinds of studies. In fact this is why the arguments supporting EPA classification of passive smoking as a carcinogen rely on analogies and extrapolation of evidence from studies of active smoking, and are fundamentally subjective. My personal view about this particular meta-analysis, and by extension other meta-analyses of observational studies of relatively small effects, is that it is important not to imply a quantitative precision to the statistical analysis that is unsupported by the quality of the data. The authors may feel they are indeed accomplishing this via their adjustments for publication bias. My feeling is that any purely statistical (i.e., quantitative) analysis cannot fully capture the strength of the evidence, or lack thereof, in the data. Moreover, the use of a complex and sophisticated analysis imparts a subliminal message that the analysis is indeed encompassing and definitive. In other words, I am much more in favor of relatively simple analyses and data presentations that allow the data to speak for themselves to the extent possible, and which attempt to provide insights into the quality and strength of the evidence.

TECHNICAL ISSUES

Meta-analysts have long recognized the problem of publication bias, and the funnel plot has been the preferred informal mechanism for identifying its presence (for a review, see Begg and Berlin, 1988). This plot of effect size versus sample size, or more properly the variance of the effect size, should be symmetric in the absence of bias. If there is a systematic preference for publishing data-dependent results favoring (or opposing) the hypothesis of interest, this will have the effect of skewing the graph. A relatively simple significance test can thus be constructed based on the rank correlation between the effect sizes and their variances, suitably standardized to ensure that the studies are i.i.d. (Begg and Mazumdar, 1994). I have performed this test using the data reported in Table 2 of Tweedie et al. (1996), including the study by Butler although it appears to be omitted from the analysis by Givens, Smith and Tweedie. This results in an adjusted rank correlation of 0.18 and a corresponding two-sided p-value of 0.13. The sample size for this test is the number of component studies in the meta-analysis, namely, 36, so its power is limited. Nonetheless the results show a nonsignificant trend supporting the concept of bias, that is, the studies with the smallest p-values tend to be the ones with the smaller sample sizes.

Bias of this nature has a differential impact on the type of analysis performed. Givens, Smith and Tweedie, like many other commentators, favor a random effects approach to the analysis. In fact the traditional random effects method (Dersimonian and Laird, 1986) is much more susceptible to publication bias, in the absence of adjustment for bias, than the fixed effects approach. In both of these methods, the summary effect size is a weighted average of the individual effect sizes; only the weights differ. In the fixed effects approach the weights are inverse variances of the individual estimates, while in the random effects estimator the weights are smoothed out in relation to the extent of between-study heterogeneity, and thus the small (biased) studies are given relatively more weight and the large studies are given relatively less weight. Thus the random effects estimator will have greater bias in the presence of selective publication. Furthermore, since publication bias tends to exaggerate the apparent heterogeneity, this further accentuates the bias in the random effects estimator. In my own analysis of the data from Tweedie et al. (1996). I obtain a fixed effects estimate of 1.17 $[1.08, 1.26]$ and a random effects estimate of 1.21 $[1.09, 1.35]$, and so the additional bias in the random effects estimator caused by this phenomenon would appear to be about 0.04. Although the relative impact of bias on these methods, if it exists, is fairly small in this example, it can be profound if the range of variances is large and the effect of selective publication is strong, as was the case in the important meta-analysis of the risks of cancer due to the chlorination of the water supply (Morris et al., 1992). In general, uncritical use of the random effects method is hazardous, in my opinion.

The use of the funnel graph and the analogous rank correlation test is not the only contrast available for detecting publication bias. In fact this approach is largely dependent for its power on the existence of a broad range of variances among the individual component studies, and this has been examined quantitatively by simulation (Begg and Mazumdar, 1994). A completely different structure for tackling the problem, and one which does not rely in any fundamental way on the variances differing from study to study, is to use "selection modeling," and this is the general framework employed by Givens, Smith and Tweedie in the spirit of earlier work by Iyengar and Greenhouse (1998), Hedges (1992) and Dear and Begg (1992). All of these authors have elected to assume that the selection probability is a function of the p-value. Conceptually, what happens in these models is that the pattern of the distribution of p-values is examined to see if it is consistent with what would be expected in the absence of bias. If there are gaps in the anticipated pattern, then their presence is attributed to missing unpublished studies, the impact of which is imputed to make the bias adjustments. It is easiest to conceptualize this in the context of the null hypothesis of no effect size, that is, $\Delta = 0$. In this setting the p-values should correspond to a uniform distribution on $[0, 1]$. Selective publication of statistically significant studies will lead to a concentration of p-values at the lower end of the sample space. However, this pattern could also be due to the fact that $\Delta \neq 0$. Thus, the effects of a true signal ($\Delta \neq 0$) and of publication bias are hard to disentangle, and the leverage for doing so is entirely bound up in the modeling assumptions, notably the assumption of normal distributions for the observed effect sizes, the assumption of known variances and the nature of the random effects distribution. My own experiences with this kind of approach lead me to believe that it is not a sound basis for making inferences about the true effect size, and that these models are useful only as part of a set of semiformal tools for identifying bias, rather than for correcting it (Dear and Begg, 1992). Indeed the "simulation" studies presented by Givens, Smith and Tweedie, which appear to be simply two applications of the method using data generated from a known model, do not inspire confidence that the model will be reliable in making accurate bias corrections in general.

A final concern I have is with the selection of prior distributions. As so often occurs in the application of fully Bayesian methods the priors appear to be picked out of the thin air without any substantive justification. In the primary analysis, the use of a $N(0, 0.15^2)$ prior is essentially akin to adding a new study to the meta-analysis with effect size zero. That is, this imaginary study has a relative risk of 1 and a confidence interval ranging from 0.75 to 1.34. A glance at Figure 1 of Givens, Smith and Tweedie shows that such a study would be among the larger of the existing studies. Moreover, since it is centered on the null hypothesis, its inclusion clearly tilts the analysis in favor of the null. The sensitivity analyses of this issue are unconvincing to me, since even though the posterior means only range from 1.12 to 1.15 (Table 2), this is quite a large difference in the context of the analysis, especially regarding the conclusion in the Abstract about the 30% overstatement of risk. The only prior that would make any sense to me in this context is the noninformative prior, and the implied advocacy of a highly informative prior centered on the null would seem to me to be very poor advice for any future users of this methodology. The other priors are similarly unappealing to me. The restricted uniform priors on the weights seem contrived, and also tilted in favor of publication bias with no clear rationale. The need to generate study variances via a prior distribution also seems contrived, and tangential to the fundamental goals of the analysis.

In summary, I suspect that the overall conclusions of Givens, Smith and Tweedie may not be too far from the truth, but I am concerned about how the authors got there. There is some suggestive evidence of publication bias, but its impact is probably not especially strong in this meta-analysis, and as a result the apparent overall trend in the data is a small positive effect of passive smoking on lung cancer risk. However, the limitations of data quality, and the apparent weakness of the effect of passive smoking mean that the analysis is far from conclusive, and it is unlikely that additional observational studies could affect this overall conclusion. My feeling is that these are the appropriate conclusions from a relatively simple analysis of these data, comprising a plot of the data as in Figure 1, a funnel-graph as in Figure 2, some rudimentary tests for publication bias and a careful evaluation of the quality of the component studies. Givens, Smith and Tweedie have brought fashionable modern statistical techniques to bear on the issue, with the attendant jargon of Gibbs sampling, burn-in periods, suppression criteria, elaborate prior distributions and all the rest. Does this stuff really add insight to the analysis? I'm afraid my vote is no.

Comment

William DuMouchel and Jeffrey Harris

The paper by Givens, Smith and Tweedie (GST) is a fresh attempt to tackle the "file drawer problem," which at first blush seems insoluble without actually going out and finding some missing studies. For example, the attempt by Iyengar and Greenhouse (1988) seemed to fall short of a solution. The current authors use more sophisticated modeling tools, primarily in their use of a hierarchical random effects model and Gibbs sampling, and perhaps they also have a more fortunate example data set. However, all attempts to assess publication bias beyond simple graphs like the funnel plot seem to involve a *tour de force* of modeling, and as such they are bound to run up against resistance from those who are not statistical modeling wonks. After all, the present analysis is pretty hard to follow, even though the paper is well written, and readers who think they do understand the presentation of the modeling process are likely to be the type who enjoy nit-picking on the details. The following discussion is offered in this latter wonkish spirit.

The random effects model, equation (2) in GST, represents each published study effect as

$$Y_j = \Delta + \beta_j + \varepsilon_j,$$

where the standard deviation of ε_j is σ_j and these standard deviations, usually given as the nominal standard errors presented by the authors of the original studies, play a key role in the detection of publication bias. Some might object that the variance of a study effect involves more than a simple sample size calculation and that, for example, a study that carefully measured exposures and documented lung cancer cases should have a smaller within-study error than a study that did not carefully gauge exposure and relied upon undocumented cancer ascertainment. This raises the question of how and whether measures of study quality can be incorporated into a meta-analysis. If such measures are not available for specific studies, but you suspect that there is a lot of variation in study quality, then the random effect term β_j in the above model provides a handy way to represent such variation. If you desire to incorporate specific information about the quality of particular studies, there are two modeling strategies available. First, you can subjectively inflate the values of σ_j for poor-quality studies. Second, you can incorporate regression terms into the model involving study-level covariates. Both strategies were used in the meta-analysis of biological effects of diesel and related emissions reported in DuMouchel and Harris (1983).

A key assumption made by GST is that the publication selection criterion is based solely on each study's one-sided p-value for rejecting the null hypothesis $\Delta \leq 0$. Why should this be based on the one-sided p-value? Are the authors assuming that studies showing a significant protective effect of ETS would be discriminated against?

William DuMouchel is with AT&T Labs—Research, 600 Mountain Avenue, Room 2C 271, Murray Hill, New Jersey 07974 (e-mail: dumouchel@research.att.com). Jeffrey Harris is Associate Professor, Department of Economics, Massachusetts Institute of Technology, E52-252, Cambridge, Massachusetts 02139 (e-mail: jeffrey@mit.edu).

More generally, we suspect dependence of the selection criterion on more than the p-value. For example, the sample size, cost or power of a study seem natural additional selection criteria. These could be summarized in the value of σ_j. If studies with high values of σ_j are harder to publish, then of course high p-values would also be underrepresented. If the authors change their model so that (7) reads

$$\begin{aligned}&\Pr\big[\text{a study with standard error } \sigma_j\\ &\quad \text{and a } p\text{-value in } I_k \text{ is published}\big] = w^k/\sigma_j,\end{aligned}$$

their substantive conclusions may be very different.

Section 3.3 of GST, on the definition of the likelihood function, is the technical heart of the paper, and perhaps the hardest section to follow. For example, the authors condition on the numbers of observed studies, $n = \{n^k\}$ in each p-value interval, whereas normally one imagines that the n^k are a function of the Y_k. Maybe it's all right, but there is the appearance of circularity in (11), in that X depends on m, m depends on n and n depends on X. We have a similar difficulty understanding the role of the normalizing functions $A(\cdot)$ in (12), (13), (14) and (16). How exactly are they defined?

In the discussion of the Gibbs sampling steps in Section 3.4, it is stated that the missing p-values are drawn uniformly on the intervals I_k. But is not a uniform distribution for the p_j only appropriate if $\Delta = 0$? Could this be producing a bias in the Gibbs sampling in favor of the null hypothesis?

Considering the simulation experiments, the authors assume that the prior distributions for the selection weights in Section 3.5(b), with no suppression, are uniform on $[0.5, 1]$ for all of the p-value intervals. Yet in the analysis of the ETS data, stronger priors were used. It would be nice to have a comparison assuming identical priors.

Finally, the authors refer to the report of Bero, Glantz and Rennie (1994), who found five unpublished negative studies not cited in the EPA Report (EPA, 1992). What were the values of the σ_j for these new studies? We guess that they are larger than those for most of the first-reported studies.

To summarize our discussion, in spite of what may seem like critical comments we do assume that publication bias is a real phenomenon and that the paper under discussion is a nice contribution to the methodology of detecting and correcting for such bias. Our most serious concern is with the form of the assumed publication bias criterion, and we would like to see whether adding a factor for dependence on the σ_j, as we suggested above, would modify the results of the ETS analysis.

Comment

Annette Dobson and Keith Dear

The culture of meta-analysis has traditionally favored very simple methods, such as weighted averages and the one-step random effects method of Der Simonian and Laird. The same is true of early approaches to publication bias, such as the file drawer of null studies conceived by Rosenthal. Now that meta-analysis is taking a high-profile role in public policy-making and regulatory affairs, it is entirely appropriate that more sophisticated techniques, such as those proposed by Givens, Smith and Tweedie, be developed. In these comments we will concentrate on the methodology, and not on the specific results about the relationship between ETS and lung cancer.

The choice of prior is always an issue in Bayesian analysis and seems to us to be critical here. Three simulations are provided in Section 3.5. In the first simulation [Section 3.5(a)] mild suppression is applied and the prior on the $\mathbf{w^k}$ reflects this by preferring lower probability of publication if the p-value is greater than 0.5 than if it is less than 0.5. This is described as "not reflecting strong beliefs about the amount of publication bias present"; however, it does embody the belief that there is some. In Section 3.5(b) no suppression is applied, and the prior reflects this by having equal priors on all three regions of the p-value scale. Finally, in Section 3.5(c) strong suppression is applied, and publication bias this time is forced into the model by the use of a $U(0.2, 0.7)$ prior for $0.1 < p < 1$. In

Annette Dobson is Professor, and Keith Dear is Senior Lecturer, Department of Statistics, University of Newcastle, NSW 2308, Australia 61-249-215544 (e-mail: stajd@cc.newcastle.edu.au and dear@mail.newcastle.edu.au).

Section 3.5(d) different priors are tried for the data in Section 3.5(c), but always imposing publication bias. In all these simulations, the true mean effect was $\Delta = 0$.

One is inevitably left with several unanswered questions. What would be the effect of applying unprejudiced priors to simulations where suppression was present—would it still be detected? How about prejudiced priors where suppression was absent—would it be spuriously imposed? And, most pertinent, what would be the effect of this last case when $\Delta > 0$? It could be that estimates of Δ close to zero would be returned, as happens with the real ETS data set. The analysis of the real, as opposed to simulated, data uses a prior which prefers low publication probability for $0.1 < p < 0.5$ and insists on it for $0.5 < p < 1$. It would have been valuable to include an assessment of how critically this assumption affects the outcome. The sensitivity analyses of Section 4.2 do not address this particular point.

The simulation trials show that when $\Delta = 0$, but appears positive due to suppression, then the model usually moves the estimate back toward zero but not all the way. It does not necessarily follow that whatever the true degree of suppression, "the method gives an outcome that is usually conservative." It appears to us that some more finely targeted simulations might well have been used to strengthen the argument.

We were glad to see the discussion of possible extensions of the model to include covariates affecting publication probability. Dependence on study size in particular is to be expected—indeed, how much reliance can be placed on a model that omits a covariate known to be important? It may be that while the broad features of such a model are meaningful, such as whether there appears to be publication bias or not, details such as adjusted estimates of relative risk are not to be taken seriously. We look forward to the appearance of Smith, Givens and Tweedie (1997), where a fuller treatment including covariates is promised.

A more aggressive approach also seems possible in the context of data augmentation. If publication probability depends on study size as well as on p-value, then the augmented data set should reflect these patterns and should include studies of different sizes with different frequencies. The expectation is that the missing studies will tend to be small, and this is not sufficiently captured by recreating them based only on the distribution of p-values. Augmenting the data set with a large null study will have a greater moderating effect on the estimate than will adding a small null study, even if both have the same p-value. Consider, for example, a meta-analysis of three studies having $\mathbf{y} = \{1.2, 1.3, 1.4\}$ and $\boldsymbol{\sigma}^2 = \{0.5, 0.6, 0.7\}$. The weighted mean is 1.29. If we now add a fourth study with $\mathbf{y} = 1$, $\boldsymbol{\sigma}^2 = 1$ and therefore a one-sided p-value of 0.16, the mean is reduced only a little to 1.24. This represents adding a small null study, since the within-study variance is relatively large: but if instead we add a large study with $\mathbf{y} = 0.3162$, $\boldsymbol{\sigma}^2 = 0.1$ and therefore again $p = 0.16$, then the revised estimate is drastically reduced to 0.68. Hung, O'Neill, Bauer and Köhne (1997) considered how the distribution of the p-value under the alternative hypothesis depends on sample size.

The paper aims to show, first, that the problem of publication bias is not one "which has to be overlooked for lack of remedies," and, second, that proper adjustment for publication bias can make a difference to the conclusions drawn from a meta-analysis. The proposed new method for accounting for publication bias is not the first such method proposed, and previous methods have not been much used, perhaps because the previous authors (Hedges 1992; Dear and Begg 1992) were more cautious about recommending reliance on estimates emerging from their models, or because of the perceived complexity of the techniques. The value of any new statistical methodology depends, in part, on the extent to which it is adopted. However, the Bayesian approach of Givens, Smith and Tweedie provides valuable insights, not only for the results shown here, but because others may be encouraged to use a similar approach, with modifications, to explore more broadly the practical effects of publication bias.

Rejoinder

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

1. THE GENERAL AND THE PARTICULAR

We are grateful to all the discussants for highlighting a number of the central issues which enter into any approach to the evaluation of meta-analysises, and for their comments on the particular instances relevant to our Bayesian approach to publication bias.

There is some universality of themes in all the discussions, and in particular there are insightful ideas on the questions of how to model the mechanism of publication bias; comments relevant to general Bayesian models on the choice of priors; and some thought-provoking concerns on the practical value of complex models. We address these below.

It is of interest that there is little comment on the specific results of our analysis of the ETS data, other than some supportive analyses by Colin Begg using different methods. In preparing this paper, we were strongly encouraged by editors and referees to focus on the ETS debate because of the almost universal relevance of the subject, and what were perceived as substantive statistical questions that should receive wider review in a journal such as *Statistical Science*. Thus ETS became, not the normal sort of application that we had initially proposed, but rather a more thorough exemplification of the issues facing statisticians commenting on issues in the public arena. In general, however, the discussants have returned the focus rather more to the generic and the mathematical issues, rather than giving strong views on this particular application.

This is refreshing in the light of some of the acrimonious debates in public health journals on such issues. Perhaps optimistically, we hope it signals that statisticians are able to consider the strengths and weaknesses of arguments in a logical rather than a political framework, a role that our discipline is perhaps uniquely designed to play.

2. THE NATURE OF PUBLICATION BIAS

All three discussion papers provide valuable comments of different forms on the nature of publication bias and how it might be modeled.

In our paper we explicitly used the p-value as the sole determining factor in deciding on whether a paper might or might not be published. The argument for this is simple, believable and undoubtedly too simplistic. Begg provides an equally believable account of another source of publication bias, namely, the choice of "interesting" exposures from studies where many sets of data are available but only some are selected by the authors for publication. Some of the authors may have made this choice based on statistical significance, but others may use other criteria. If we add in the work of Simes (1996) and others on research grants that never led to publication, we see that the sources of missing studies are many, and that the parameters describing nonpublication are not easy to select.

Bill DuMouchel and Jeffrey Harris suggest that two parameters, the p-value and the study variance σ_j^2, should be used to decide on the probability of exclusion; Annette Dobson and Keith Dear suggest, essentially equivalently, that p-value and size of study should provide this probability. Certainly it seems probable that large and well-funded studies may be less likely to slip away. But Begg's image, that some of these large studies are still unmined on a variety of topics, cannot be overlooked as an extra complication. As a further alternative, we note that exclusion criteria could be based explicitly on p-values and the size of the relative risk in a study: studies with larger, "interesting," relative risks are published even if insignificant, perhaps. Formally, this latter criterion might be essentially the same as using the size or variance; but the common-sense interpretation might be easier with this parameterization.

We endorse the desirability of further work to take up these bases for developing exclusion criteria, and indeed in Smith, Givens and Tweedie (1997) we do develop a more complex methodology that could be used to investigate such multiple criteria, as well as those based on exclusion of poor studies (however defined). These further methods might allow some systematic handling of the quality issues raised by Begg or by DuMouchel and Harris, as well as the relationships with other covariates commented on by DuMouchel and Harris or Dobson and Dear.

3. MODELING ISSUES, PRIORS AND BIASES

Even within the framework of a criterion using p-values, there are nontrivial questions. DuMouchel and Harris point out that we use the one-sided p-

value in our criterion. This is deliberate: our experience is that funnel plots are, as in the ETS case, typically missing one rather than two corners, indicating that there is directional discrimination. The model could easily be adjusted to incorporate two-sided p-values if this seemed appropriate.

They also note correctly that in our Gibbs sampling, we introduce a bias by drawing p-values uniformly, implicitly using the distribution appropriate to $\Delta = 0$. Since this is only uniform in each interval I_k, we believe the bias to be small. Within the Gibbs step, the *number* of studies m_k imputed in that interval is based on the current best estimator of Δ (and its current posterior approximation); thus if Δ is truly far from 0, we would expect to place few studies in an inappropriate I_k, even though the consequent augmenting studies have a variance (back-calculated from p) that is biased toward small studies. The exact distribution of p under alternative hypotheses is, as noted by Dobson and Dear, studied by Hung et al. (1997), and with sufficient effort such a distribution could be incorporated in the Gibbs sampler we use.

Conversely, we think there is little cause for the concern of Begg that we might be including augmented studies biased toward the null by use of the prior distribution $N(0, 0.15^2)$ on Δ. This does *not* draw a new study with mean zero and CI (0.75, 1.34), which would indeed be a largish study in the ETS context. It only draws the new mean Δ from that range; the actual mean is modified according to the data, and then the variance of the new study is drawn differently, to fit the p-value criteria. Hence our prior on Δ contains very little information. This is clear in Table 2, where we find that when we increased the variance on the prior by using a $N(0, 0.4^2)$ prior, there was no real change in the results. If we used a prior of $N(0, 0.1^2)$, then the estimate of RR did change, down to 1.12 rather than 1.14, showing that this was perhaps not a sufficiently diffuse prior.

One can use graphical methods to verify these effects, as developed in Smith and Tweedie (1997). Figure 1 shows the characteristic shape of a funnel based on a density estimate of the 50 "observed" studies in simulation example (a) [Section 3.5(a)]. Figure 2 shows a density estimate for the location of the imputed studies in simulation example (a). Note that the imputation typically misses the extreme suppressed studies, although in general it does coincide with the locations of the missing studies; and the augmenting studies are indeed smaller than Begg fears.

All discussants consider questions raised by the use of prior distributions. DuMouchel and Harris, Begg, and Dobson and Dear all found the use of uniform priors on the publication bias weights to present difficulties for different reasons. The option adopted in the similar but simpler model of Eberly and Cassella (1996) was to use beta distributions: this clearly overcomes none of the queries of the discussants. Perhaps all we can do in such a situation is to point out that with our model this aspect is quite subjective and that different users could use other priors more to their taste.

Nonetheless, we recognize that the choice of priors on the w_k is perhaps the most sensitive part of the method, since they always do imply that some degree of publication bias is present; although, if the priors are identical on all I_k, then at least there

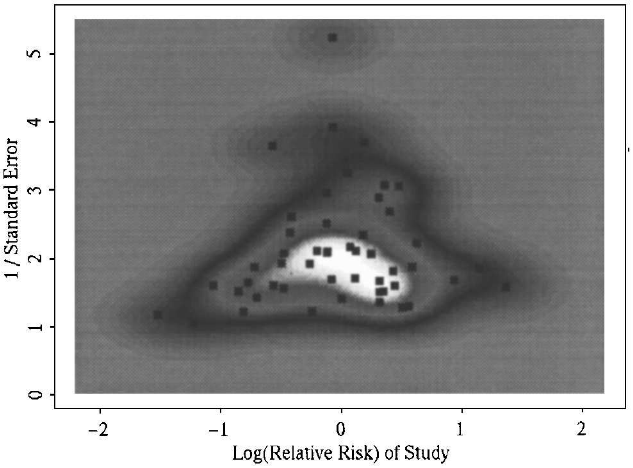

FIG. 1. *Funnel plot and smoothed density of the total data set of* 50 *studies from simulation example* (a) [*Section* 3.5(a)] *before suppression.*

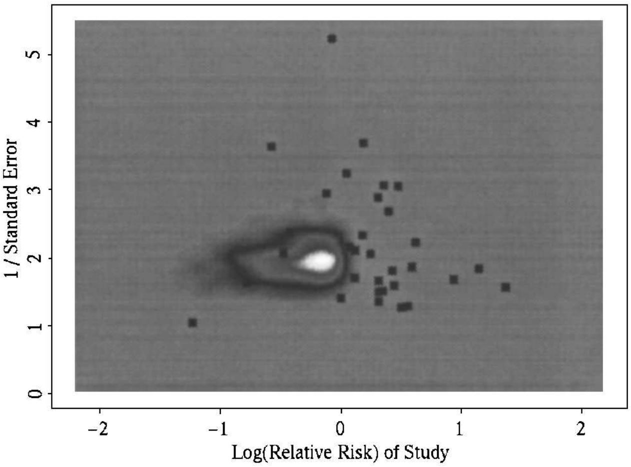

FIG. 2. *Funnel plot of* 32 *"observed" studies and smoothed density of augmented studies from simulation example* (a) [*Section* 3.5(a)]. *Note that the missing studies appear to be imputed near the lower-left corner of the funnel plot, covering essentially the area that has been truncated due to publication bias.*

should be no bias in the location of the augmenting studies.

However, as pointed out by Dumouchel and Harris and by Dobson and Dear, we have used differential priors favoring more bias in low-significance studies in our analysis of the ETS example. Accordingly we have reanalyzed the overall ETS data set with priors uniform on $[0.5, 1]$ for each I_k, as suggested by both these discussants. Table 2 showed that with our original priors the posterior mean on RR was 1.14 with CI $(1.00, 1.28)$. Using the priors identical on each I_k, this changes to 1.19 $(1.06, 1.34)$; using identical priors but with monotonicity constraints we get 1.15 $(1.04, 1.28)$, very comparable to the constrained version of 1.14 $(1.03, 1.26)$ in Table 2.

It thus appears reasonable to conclude that the values we give are not driven too much by the priors if the monotonicity, or ordering of rejection by p-value, is correct: clearly our priors had a tendency for this same preference built in through a different mechanism. We believe this is in general an appropriate constraint, and note that in another application, in LaFleur et al. (1996), we found a similar need to impose monotonicity to reach agreement with the subjective impressions given by the funnel plots.

Finally, we note that Dobson and Dear suggest a number of other questions that might be addressed by simulation. Some of these are indeed considered in Smith (1997); we agree that for the reliable use of this method more such analysis would be useful, but feel in general it is wise for any user to consider the specific shape of the problem (numbers and sizes of studies, etc.) and carry out such simulations in a relevant context.

4. COMPLEX OR SIMPLE SOLUTIONS?

Our remarks about publication bias needing some attention, rather than being "overlooked for lack of remedies," was not in any way intended to ignore the contributions in papers such as Dear and Begg (1992) or Hedges (1992); on the contrary, it was intended as a comment on the way in which some reviews (such as those of ETS) have typically swept the problem aside rather than using such remedies.

Dobson and Dear comment that perhaps the lack of use of those papers was because the authors were more cautious in recommending their adjusted results; but we think they are more accurate with their other possibility, that methods are not used because they are complex. Clearly Begg feels that this is the situation, and we hope his pessimism on this will not prove entirely accurate.

It is indeed regrettable if methodological complexity has discouraged researchers from careful adjustment of meta-analyses for publication bias, and our methods are unlikely to remove such reluctance. We do think that our Bayesian methods, relying on various computer-intensive Monte Carlo methods, are not only fashionable, but are easier to implement than the methods in Dear and Begg (1992) or Hedges (1992); but they are not trivial.

What can be done about more simple methods? We disagree with the philosophy in Begg's comments, that one should use simple methods to identify publication bias even if no adjustment can be carried out. It is pleasing that his rank-correlation method largely confirms, even with its low power, the likely presence of bias in the ETS data; but clearly one should not throw out the ETS data just because such a bias (of unknown size and importance) exists, any more than one should feel comfortable accepting the original values once one confirms the likely presence of bias.

The rather subjective method given in Mengersen, Tweedie and Biggerstaff (1995), of trimming funnel plots of unmatched studies rather than trying to augment them, does at least "disentangle the true signal and the publication bias" to some extent. It does seem vastly preferable to the surprisingly frequently used Rosenthal (1979) "fail-safe" method, which enjoys considerable popularity because it is simple even if it answers quite the wrong problem: it is hard to conceive of a situation where one really wants to know how many exactly null studies are needed to reverse one's conclusions, rather than the degree of bias in the set of studies one actually has!

Perhaps the best balance lies in ensuring that some complex methods are available when needed.

We think DuMouchel and Harris are correct in that, to address this at first sight insoluble problem, one must have some *tours de force* of modeling, whether they be Bayesian or frequentist; and that these will rarely become standard techniques, even among statistical wonks. But in cases such as those of ETS, where serious public health or legal issues are being debated, it must be valuable to be able to revisit *ad hoc* judgements, made by (for example) looking at simple funnel plots, and by using more difficult but in principle more rigorous methods, to provide some level of confirmation or contradiction of the initial conclusions.

ADDITIONAL REFERENCES

BEGG, C. B. and BERLIN, J. A. (1988). Publication bias: a problem in interpreting medical data (with discussion). *J. Roy. Statist. Soc. Ser. A* **151** 419–463.

Begg, C. B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50** 1088–1101.

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

Garfinkel, L., Auerbach, O. and Joubert, L. (1985). Involuntary smoking and lung cancer: a case–control study. *Journal of the National Cancer Institute* **75** 463–469.

Hung, H. M. J., O'Neill, R. T., Bauer, P. and Köhne, K. (1997). The behavior of the *P*-value when the alternative hypothesis is true. *Biometrics* **53** 11–22.

Morris, R. D., Audet, A. M., Angelillo, I. F., Chalmers, T. C. and Mosteller, F. (1992). Chlorination, chlorination by-products and cancer: a meta-analysis. *American Journal of Public Health* **82** 955–963.

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

Smith, D. D. (1997). Accounting for publication bias and quality differences in Bayesian random effects meta-analytic models. Ph.D. dissertation, Colorado State Univ.

Smith, D. D. and Tweedie, R. L. (1997). A density estimation diagnostic for publication bias adjusted meta-analysis. In preparation.