Statistical Science
2004, Vol. 19, No. 3, 450–471
DOI 10.1214/088342304000000350

Adjusted Likelihoods for Synthesizing Empirical Evidence from Studies that Differ in Quality and Design: Effects of Environmental Tobacco Smoke

Robert L. Wolpert and Kerrie L. Mengersen

Abstract. Methods are introduced and illustrated for synthesizing evidence from case-control and cohort studies, and controlled trials, accounting for differences among the studies in their design, length of follow-up and quality. The methods, based on hierarchical but nonexchangeable Bayesian models, are illustrated in a synthesis of disparate information about the health effects of passive exposure to environmental tobacco smoke.

Key words and phrases: Adjustment, environmental tobacco smoke, meta-analysis, nonexchangeable Bayesian hierarchical model, quality of studies.

1. INTRODUCTION

The hazards of combining evidence naïvely from multiple sources are well known. Even among sources intended to be homogeneous, such as multi-center controlled medical trials, there will be some variation in patient attributes and treatment regimens. This variation is strongly exacerbated in epidemiological studies because of differing study designs and methods, and differing population characteristics: designs may vary according to their retrospective or prospective nature, their case-control, cohort or cross-sectional structure, length of follow-up, admission and classification criteria, control of biases, consideration of confounders and data collection procedures. Study populations vary in age structure, diet and other lifestyle factors, genetic makeup, and environmental exposure to competing and contributing risks. Investigators also vary with respect to experience, motivation and skill. All of these issues threaten the validity of methods which simply pool the data. Methods which are sensitive to the probability distributions that govern the individual studies' observed outcomes can make more complete use of study information and so can lead to more powerful statistical tests and to shorter credible intervals for uncertain parameters.

Beginning with Fisher's (1934) early attempt to avoid the problem of heterogeneity in field trials by combining individual p values to assess the overall significance level, a number of authors have introduced novel statistical methods intended to quantify an overall "effect size" and assess its variation. These methods, referred to collectively as *meta-analysis*, have been described and compared by many authors, including Hedges and Olkin (1985) and the National Research Council review (1992).

More recently Bayesian and Bayesian-inspired methods have been brought to bear on the meta-analysis problem, starting with the landmark papers by DuMouchel and Harris (1983), who introduced the idea of constructing hierarchical Bayesian models to synthesize information from five types of environmental studies of the effect on human and animal subjects of exposure to nine related environmental agents, and Dempster, Selwyn and Weeks (1983), who synthesized historical and contemporary clinical evidence. Subsequent authors who offered broad guides to the use of Bayesian hierarchical models for

Robert L. Wolpert is Professor of Statistics and Decision Sciences and Professor of the Environment and Earth Sciences, Duke University, Durham, North Carolina 27708-0251, USA (e-mail: wolpert@stat.duke.edu). Kerrie L. Mengersen is Professor of Statistics, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia.

synthesizing evidence include Berry (1990), Carlin (1992), DuMouchel (1990, 1996), Higgins and Spiegelhalter (2002), Liao (1999), who combined 2×2 tables, Morris and Normand (1992), Normand (1999), Smith, Spiegelhalter and Parmar (1996), Spiegelhalter, Thomas, Best and Lunn (2002, in the BUGS software examples) and Sutton and Abrams (2001). All of these approaches make at least some accommodation for study heterogeneity, commonly through inclusion of a single extra variance term in the hierarchical model structure, but each treats all the studies as exchangeable (at least within groups).

In many meta-analytic applications the simplifying assumption of exchangeability (de Finetti, 1930) is simply untenable—studies differ too much in their designs, subject selection criteria or other details for any analysis that ignores these differences to be convincing. One approach would be to quantify how similar the different studies are to each other (Draper, Hodges, Mallows and Pregibon, 1993) or how similar they are to the target circumstances, and then exclude or discount in ad hoc ways those studies regarded as too dissimilar. Another approach, due to Eddy (1989), is to replace each study's likelihood function with a subjective estimate of "what would have been observed" had that study followed the target circumstances exactly. Eddy coined the term "adjusted likelihood function" for these replacements, a term we employ in our own (related) approach introduced in Section 2 below.

We propose a different approach: the construction of a hierarchical Bayesian model with submodels, for each study or other source of evidence, that reflect and accommodate important study-specific differences. In this coherent approach the investigator begins by specifying in detail the target conditions—for example, the subject population, treatment or exposure details, and case or outcome details. Each individual study offers *direct* evidence through its likelihood function about the parameters that govern that particular study, with its specific design, selection criteria, and so forth. We construct an *adjusted likelihood function* that describes the *indirect* evidence offered by each study about the questions of interest to the investigator under the specified target conditions. Studies conducted under conditions quite similar to the target conditions lend strong evidence; studies less similar lend more uncertain evidence, leading in a natural way for them to be discounted appropriately in an overall synthesis.

1.1 Evidence from Individual Studies

Epidemiological studies offer direct evidence about the conditional probabilities of outcome status (for cohort studies) or exposure (for case-control studies) for the study populations and, through these, indirect evidence about the quantity of interest: usually, some measure of the association between exposure and outcome within the entire population. Evidence is often available about the size and impact of varying age distributions, misclassification rates, and other confounding features and possible biases, but this evidence is rarely used in analyses. The methodology we present offers an opportunity to exploit this evidence.

In a cohort study (CHS) subjects are drawn from a specified population and classified as exposed or unexposed (Breslow and Day, 1987). They are then followed for a specified length of time and classified with respect to their disease status. Such studies may be conducted prospectively or retrospectively. Evidence bears directly on the conditional probabilities of becoming a case, $p_{c|e}$ and $p_{c|\bar{e}}$, in the exposed and unexposed study arms, respectively, through the likelihood function (Berger and Wolpert, 1988). In the simple case where age and other covariates are not considered this is

$$\begin{aligned} &L_{\mathrm{CHS}}(p_{c|e}, p_{c|\bar{e}}) \\ &\quad\propto (p_{c|e})^{n_{ce}}\,(1-p_{c|e})^{n_{\bar{c}e}}(p_{c|\bar{e}})^{n_{c\bar{e}}}\,(1-p_{c|\bar{e}})^{n_{\bar{c}\bar{e}}}, \end{aligned} \tag{1}$$

where n_{ce} and $n_{\bar{c}e}$ are the numbers of cases and noncases among the n_e exposed subjects, and $n_{c\bar{e}}$ and $n_{\bar{c}\bar{e}}$ are the numbers of cases and noncases among the $n_{\bar{e}}$ unexposed subjects, respectively. Occasionally these conditional probabilities may themselves be useful, but more often interest lies in some measure of their difference quantifying the association between outcome and exposure, such as the simple difference $\varepsilon_{\mathrm{SD}} = p_{c|e} - p_{c|\bar{e}}$, log relative risk $\varepsilon_{\mathrm{LRR}} = \log(p_{c|e}/p_{c|\bar{e}})$ or log odds ratio

$$\varepsilon_{\mathrm{LOR}} = \log(p_{c|e}\,p_{\bar{c}|\bar{e}}/p_{\bar{c}|e}\,p_{c|\bar{e}}).$$

Each of these measures will be positive if greater exposure is associated with higher levels of disease and zero if they are unrelated, and each can be estimated in a cohort study by its maximum likelihood estimator (MLE) such as $\hat{\varepsilon}_{\mathrm{LOR}} = \log(n_{ce}\,n_{\bar{c}\bar{e}}/n_{c\bar{e}}\,n_{\bar{c}e})$. These and other measures have been discussed, illustrated and compared by many authors, including Breslow and Day (1980, Chapter 2), Cox (1970, page 20), Freeman (1987, pages 66 and 95) and Wolpert (1986).

A variety of methods (see Berger, Liseo and Wolpert, 1999) have been proposed for extracting information about a quantity of interest like ε amid nuisance parameters; the Bayesian approach is to select a conditional prior distribution $\pi(dp\,|\varepsilon)$ for $p = (p_{c|e}, p_{c|\bar{e}})$ and summarize the evidence about ε alone by $L_{\mathrm{CHS}}(\varepsilon) = \int L_{\mathrm{CHS}}(p)\pi(dp|\varepsilon)$. A change of variables to the log odds ratio $\varepsilon = \varepsilon_{\mathrm{LOR}}$ and independent Jeffreys' reference priors for both case probabilities $p_{c|e}$ and $p_{c|\bar{e}}$ leads to the likelihood

$$\begin{aligned} L_{\mathrm{CHS}}(\varepsilon) &\propto \varepsilon^{-1}(e^{\varepsilon}-1)e^{n_{ce}\varepsilon} \\ &\quad \cdot \int_0^1 \frac{p^{n_c}(1-p)^{n_{\bar{c}}}}{[1-p(1-e^{\varepsilon})]^{n_e+1}}\,dp \\ &\propto \varepsilon^{-1}(e^{\varepsilon}-1)e^{n_{ce}\varepsilon} \\ &\quad \cdot {}_2F_1(n_e+1, n_c+1; n+2; 1-e^{\varepsilon}), \end{aligned} \tag{2}$$

where ${}_2F_1(a, b; c; z)$ is the confluent hypergeometric function (Abramowitz and Stegun, 1964, page 558). With the implied prior $\pi(\varepsilon) = \varepsilon/2\pi^2 \sinh(\varepsilon/2)$, the log odds ratio $\varepsilon_{\mathrm{LOR}}$ has posterior mean and variance

$$\begin{aligned} \mu &= \psi(n_{ce}+1/2) - \psi(n_{\bar{c}e}+1/2) \\ &\quad - \psi(n_{c\bar{e}}+1/2) + \psi(n_{\bar{c}\bar{e}}+1/2) \\ &\approx \log(n_{ce}n_{\bar{c}\bar{e}}/n_{\bar{c}e}n_{c\bar{e}}), \end{aligned} \tag{3}$$

$$\begin{aligned} \sigma^2 &= \psi'(n_{ce}+1/2) + \psi'(n_{\bar{c}e}+1/2) \\ &\quad + \psi'(n_{c\bar{e}}+1/2) + \psi'(n_{\bar{c}\bar{e}}+1/2) \\ &\approx 1/n_{ce} + 1/n_{\bar{c}e} + 1/n_{c\bar{e}} + 1/n_{\bar{c}\bar{e}}, \end{aligned} \tag{4}$$

where $\psi(z+1/2) = \log z + \mathcal{O}(z^{-2})$ and $\psi'(z+1/2) = 1/z + \mathcal{O}(z^{-3})$ are the digamma and trigamma functions (Abramowitz and Stegun, 1964, Section 6.3.4).

For example, Figure 1 shows $L_{\mathrm{CHS}}(\varepsilon)$ for the famous 1985 extracorporeal membrane oxygenation (ECMO) trial (see Ware, 1989), in which 6 of 10 subjects survived in the control group and 9 of 9 survived who were "exposed" to ECMO. The solid line is $L_{\mathrm{CHS}}(\varepsilon)$ from (2), the dashed line is the marginal prior density function $\pi(\varepsilon)$ and the vertical bar at $\varepsilon_{\mathrm{LOR}} = 0$ marks the null effect. The posterior probability of greater risk for ECMO is $\mathrm{P}[\varepsilon_{\mathrm{LOR}} > 0] \approx 0.0107$ with this prior (see Kass and Greenhouse, 1989 and Lavine, Wasserman and Wolpert, 1991, for discussions of prior distributions for this problem).

For less extreme contingency tables, including all of those in the environmental tobacco smoke (ETS) case study considered below, the distribution is very well approximated by a normal with the same mean and variance or with the approximate means and variances given in (3) and (4) above. Similar expressions are available for the other measures of exposure effect in cohort studies.

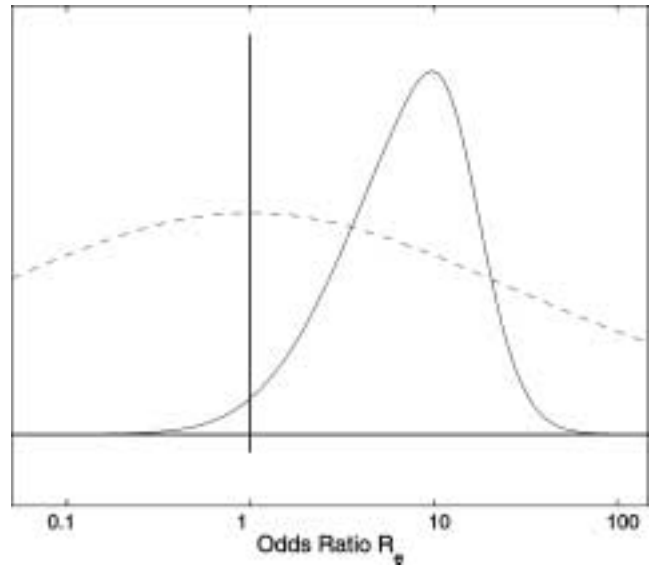

FIG. 1. *Marginal likelihood* (*solid*) *and prior density* (*dashed*) *for odds ratio of survival* $R = \exp(\varepsilon_{\mathrm{LOR}})$ *in the 1985 ECMO trial* (*on log scale*).

Similarly, in a case-control study (CCS) some number n_c of cases are matched with $n_{\bar{c}}$ noncases (controls) on the basis of demographic variables and other covariates (Breslow and Day, 1980), and then within each of these groups subjects are further classified into those who are exposed (n_{ce}, $n_{\bar{c}e}$) and unexposed ($n_{c\bar{e}}$, $n_{\bar{c}\bar{e}}$). Such a study can offer direct evidence only about the conditional probabilities of exposure within case and noncase groups, $p_{e|c}$ and $p_{e|\bar{c}}$, through the likelihood function

$$\begin{aligned} &L_{\mathrm{CCS}}(p_{e|c}, p_{e|\bar{c}}) \\ &\quad \propto (p_{e|c})^{n_{ce}}(1-p_{e|c})^{n_{ce}}(p_{e|\bar{c}})^{n_{\bar{c}e}}(1-p_{e|\bar{c}})^{n_{\bar{c}\bar{e}}}. \end{aligned} \tag{5}$$

Only indirect evidence is given about the remaining probabilities. Independent reference priors for the unobserved exposure probabilities $p_{e|\bar{c}}$ and $p_{e|c}$ lead to an adjusted likelihood function

$$\begin{aligned} L_{\mathrm{CCS}}(\varepsilon) &\propto \varepsilon^{-1}(e^{\varepsilon}-1)e^{n_{ce}\varepsilon} \\ &\quad \cdot {}_2F_1(n_c+1, n_e+1; n+2; 1-e^{\varepsilon}), \end{aligned} \tag{6}$$

again approximately normal with mean $\mu \approx \log(n_{ce}n_{\bar{c}\bar{e}}/n_{\bar{c}e}n_{c\bar{e}})$ and variance $\sigma^2 \approx 1/n_{ce} + 1/n_{c\bar{e}} + 1/n_{\bar{c}e} + 1/n_{\bar{c}\bar{e}}$.

1.2 Exchangeable Combination of Evidence

If study subjects comprise a simple random sample from some population, then both cohort studies and case-control studies give evidence about the same vector $\theta = (\theta_{ce}, \theta_{c\bar{e}}, \theta_{\bar{c}e}, \theta_{\bar{c}\bar{e}})$ of probabilities that an eligible subject drawn randomly from the target population will be an exposed or unexposed case, exposed or unexposed noncase, respectively, through their respective likelihood functions $L_{\mathrm{CHS}}(p_{c|e}, p_{c|\bar{e}})$ (since $p_{c|e} = \theta_{ce}/\theta_e$, $p_{c|\bar{e}} = \theta_{c\bar{e}}/\theta_{\bar{e}}$) and $L_{\mathrm{CCS}}(p_{e|c}, p_{e|\bar{c}})$ (since $p_{e|c} = \theta_{ce}/\theta_c$, $p_{e|\bar{c}} = \theta_{\bar{c}e}/\theta_{\bar{c}}$). Hence, it is convenient to express all studies' evidence in terms of θ before trying to combine them across study types.

Evidence about a common parameter vector θ may be captured from a set of I *independent* studies through the joint likelihood function $L^I(\theta) = \prod_{i\in I} L_i(\theta)$. For case-control studies, for example, this is equivalent to observing $n^+_{ce} \equiv \sum_{i\in I} n^i_{ce}$ exposed and $n^+_{c\bar{e}} \equiv \sum_{i\in I} n^i_{c\bar{e}}$ unexposed cases, $n^+_{\bar{c}e} \equiv \sum_{i\in I} n^i_{\bar{c}e}$ exposed and $n^+_{\bar{c}\bar{e}} \equiv \sum_{i\in I} n^i_{\bar{c}\bar{e}}$ unexposed noncases in a single pooled study. However, in most cases, the underlying assumption of exactly the same conditional probabilities in all I studies is just not reasonable. The studies almost always differ in important respects that affect their probabilities of exposure and outcome, and hence their evidence about θ. The effect of ignoring this variation is systematically to underrepresent uncertainty in the posterior distribution and sometimes to distort the location of the distribution as well.

A first step toward accommodating study variation is to allow the conditional case (resp., exposure) probabilities $p_{c|\bar{e}}, p_{c|e}$ (resp., $p_{e|c}, p_{e|\bar{c}}$) to vary across studies, while still treating the measure of interest ε as constant across all studies. Changing variables to $\varepsilon = \varepsilon_{\mathrm{LOR}}$ and $p^i_{c|\bar{e}}$ for cohort studies and $p^i_{e|\bar{c}}$ for case-control studies leads to

$$L^I_{\mathrm{CHS}}(\varepsilon, p^I_{c|\bar{e}}) \propto e^{\varepsilon n^+_{ce}} \prod_{i\in I} [(p^i_{c|\bar{e}})^{n^i_c}(1-p^i_{c|\bar{e}})^{n^i_{\bar{c}}} \cdot [1 - p^i_{c|\bar{e}}(1-e^{\varepsilon})]^{-n^i_e}], \tag{7}$$

$$L^I_{\mathrm{CCS}}(\varepsilon, p^I_{e|\bar{c}}) \propto e^{\varepsilon n^+_{ce}} \prod_{i\in I} [(p^i_{e|\bar{c}})^{n^i_e}(1-p^i_{e|\bar{c}})^{n^i_{\bar{e}}} \cdot [1 - p^i_{e|\bar{c}}(1-e^{\varepsilon})]^{-n^i_c}] \tag{8}$$

and in each case to $L^I(\varepsilon) = \int L(\varepsilon, p^I)\pi(dp^I|\varepsilon)$ for evidence about ε (for independent Jeffreys' prior distributions, e.g.) to

$$L^I_{\mathrm{FE}}(\varepsilon) \propto \frac{\varepsilon^{|I|} e^{n^+_{ce}\varepsilon}}{(e^{\varepsilon}-1)^{|I|}} \cdot \prod_{i\in I} {}_2F_1(n^i_c+1, n^i_e+1; n^i+2; 1-e^{\varepsilon}), \tag{9}$$

which is approximately normal again, but now with mean $\mu_{\mathrm{FE}} \approx \sum_{i\in I} \mu_i\sigma_i^{-2} / \sum_{i\in I} \sigma_i^{-2}$ and variance $\sigma^2_{\mathrm{FE}} \approx 1/\sum_{i\in I}\sigma_i^{-2}$, the precision-weighted mean of the sample means $\mu_i \approx \log(n^i_{ce}n^i_{\bar{c}\bar{e}}/n^i_{\bar{c}e}n^i_{c\bar{e}})$ using variances $\sigma_i^2 \approx [1/n^i_{ce} + 1/n^i_{\bar{c}e} + 1/n^i_{c\bar{e}} + 1/n^i_{\bar{c}\bar{e}}]$. This is the familiar fixed-effects model. The best choice of $\pi(dp|\varepsilon)$ depends on the application. In the example presented in Section 3, both reference and informed prior distributions are illustrated.

While some invariance may be offered by using a measure of exposure effect that is thought to be relatively insensitive to study variation (e.g., using $\varepsilon_{\mathrm{LOR}}$ if the variations are expected to scale incidence *odds* by equal factors; see Wolpert, 1986), possible variation across studies in the effect measures ε^i should be modeled explicitly. If we take $\varepsilon^i \sim \mathrm{No}(\varepsilon, \sigma^2)$, for example, independent normal variates with common unknown mean ε and known variance σ^2, and ascribe a diffuse conjugate prior distribution $\varepsilon \sim \mathrm{No}(\mu, \tau^2)$, the earlier normal approximations to $L_{\mathrm{CHS}}(\varepsilon_{\mathrm{LOR}})$ and $L_{\mathrm{CCS}}(\varepsilon_{\mathrm{LOR}})$ lead to the familiar random-effects model whose approximately normal posterior distribution has mean $\mu_{\mathrm{RE}} \approx \sum_{i\in I}\mu_i(\sigma^2+\sigma_i^2)^{-1}/(\tau^{-2}+\sum_{i\in I}(\sigma^2+\sigma_i^2)^{-1})$ and variance $\sigma^2_{\mathrm{RE}} \approx 1/(\tau^{-2} + \sum_{i\in I}(\sigma^2+\sigma_i^2)^{-1})$, with the same study-specific μ_i and σ_i^2 as above. Uncertainty about hyperparameters τ^2 and σ^2 is reflected through prior distributions in the case study presented in Section 3. Note that the fixed-effects model may be recovered in the limit as $\sigma^2 \to 0$, $\tau^2 \to \infty$.

It is a small generalization of the random-effects model to describe the variation in study-specific parameters p_i explicitly. Modeling them as exchangeable is equivalent (by de Finetti's theorem) to treating them as conditionally independent identically distributed random vectors, given an overall hyperparameter θ. If the exposure effect ε can be written as a function of θ (or, more generally, if ε is conditionally independent of the $\{p_i\}$, given θ), then the joint prior distribution for ε, θ, and all the $\{p_i\}$ can be factored as $\pi(d\varepsilon)\pi(d\theta|\varepsilon)\prod_{i\in I}\pi(dp_i|\theta)$, giving the marginal likelihood for the exposure effect the representation

$$L^I_{\mathrm{EHM}}(\varepsilon) = \int \left[\prod_{i\in I}\int L_i(p_i)\pi(dp_i|\theta)\right]\pi(d\theta|\varepsilon). \tag{10}$$

This reduces to the random-effects model for $\theta \equiv (\mu, \tau, \sigma)$ and normal logistic conditional distributions $\pi(dp_i|\theta)$; to the fixed-effects model for $\theta \equiv \varepsilon$ and beta conditional distributions $\pi(dp^i_{c|\bar{e}}|\theta)$, $\pi(dp^i_{e|\bar{c}}|\theta)$; and to simple pooling for $\theta \equiv (p_{ce}, p_{c\bar{e}}, p_{\bar{c}e}, p_{\bar{c}\bar{e}})$ with unit point masses $\pi(dp_i|\theta)$ at $p_i = \theta$.

1.3 Partially Exchangeable Combination of Evidence

Sometimes studies can be regarded as exchangeable within known groups, but the groups may differ systematically among themselves. Health care at urban and rural facilities may differ, for example, or educational performance at public and private schools or universities. The possible effects of ETS exposure in the example in Section 3 can be expected to differ across different countries for any number of reasons (competing risks, differing levels of exposure, geographically varying genetic propensities, etc.). In the simplest form the groups may themselves be treated as exchangeable, leading to a multistage *partially exchangeable* hierarchical model of the form

$$\begin{aligned} \varepsilon &\sim \pi(d\varepsilon), & \theta &\sim \pi(d\theta|\varepsilon), \\ \theta_g &\sim \pi(d\theta_g|\theta), & g &\in G, \\ \theta^i &\sim \pi(d\theta^i|\theta_g), & i &\in g, \end{aligned}$$

where $g \in G$ indexes the groups and $i \in g$ indexes the studies within a group, and where each given distribution is (implicitly) conditionally independent of all parameters higher in the hierarchy. Now

$$\begin{aligned} &L^I_{\mathrm{PEM}}(\varepsilon) \\ &\quad = \int \left\{ \prod_{g\in G} \int \left[\prod_{i\in g} \int L_i(\theta^i)\pi(d\theta^i|\theta_g) \right] \pi(d\theta_g|\theta) \right\} \\ &\qquad \cdot \pi(d\theta|\varepsilon). \end{aligned}$$

For example, normal prior distributions at each stage could lead to the model $\varepsilon \sim \mathrm{No}(0, \tau^2)$, $\varepsilon^g \sim \mathrm{No}(\varepsilon, \sigma_g^2)$, $g \in G$, $\varepsilon^i \sim \mathrm{No}(\varepsilon^g, \sigma_i^2)$, $i \in g$, for the exposure effect ε (with suitable conditional distributions for θ^i, given ε^i), where σ_g^2 and σ_i^2 are the group-level and individual-level prior variances and τ^2 is the large variance of the diffuse prior for the overall effect ε, allowing for greater homogeneity among studies within groups.

2. SYNTHESIZING HETEROGENEOUS EVIDENCE

Studies vary in the degree to which confounding effects and possible biases have been recognized and accommodated, and in the levels of similarity between their subject populations and the study population. The five synthesis methods presented in Section 1 (simple pooling, fixed- and random-effect models, and fully and partially exchangeable hierarchical models) allow for increasing degrees of heterogeneity among studies, but they do not make use of collateral evidence about the size and impact of varying age and exposure distributions, misclassification rates and other differences that might be expected to affect the evidence.

Several methods have been proposed for discounting evidence thought to be less reliable or to apply less directly to the questions of interest in meta-analysis, some of which were first developed to address the similar problems that arise in synthesizing the opinions of several experts (Makridakis and Winkler, 1983; Genest and Zidek, 1986; Wolpert, 1989). These include *threshold exclusion*, in which studies thought to be less applicable are simply excluded from the analysis; *weighted likelihood functions*; *block mixtures*, in which groups of studies of comparable quality or applicability are successively included in the analysis; *mixtures*, in which individual studies are included in the analysis with probabilities based on their relative applicability; and *hierarchical models*, in which treatment effects are estimated separately for homogeneous groups of studies, to reveal systematic variation of apparent effects across groups.

This paper investigates another alternative, intended to extract, reconcile and synthesize information about a quantity of interest despite study variations and flaws: the systematic *adjustment* of the different studies within the Bayesian paradigm to accommodate (and, where possible, correct for) their differences from one another and from the intended object of study.

2.1 Adjustment

A common problem in synthesizing evidence is that of making inference about the parameter θ^0 that would govern an ideal (or *paradigm*) study for a particular purpose—one for the population of interest to the investigator, without misclassification or other weaknesses, on the basis of nonideal studies whose conditions vary in important ways from the ideal. If the ith study offers direct evidence about a parameter θ^i through a likelihood $L_i(\theta^i)$, and if each θ^i (including θ^0) is related to a hyperparameter θ through a known functional relationship $\theta^i = \phi_i(\theta)$, then we can "adjust" the evidence from the ith study to bear directly on θ (and hence on θ^0) through the relationship

$$L_i^{\mathrm{Adj}}(\theta) = L_i(\phi_i(\theta)), \tag{11}$$

where the arbitrary function $\phi_i(\theta)$ represents the value of θ^i when the hyperparameter value is θ.

If the analyst studying a series of clinical trials believes that the success probability in the ith trial should be only a fraction (say, half) of the success probability θ^0 under paradigm conditions, perhaps because half of the subjects in that trial were noncompliant, then s/he might take $\theta^0 = \phi_0(\theta) = \theta$ and set $\theta^i = \phi_i(\theta) = \theta/2$ for that trial. Of course, this choice implies that the joint prior probability distribution for the pair (θ, θ^i) is concentrated on the one-dimensional set $\{(\theta, \theta^i) : 0 \le \theta^i \le 1/2,\ \theta = 2\theta^i\}$ and in particular that $\theta^i \le 1/2$. Similarly, one analyzing a series of cohort studies might use $\phi_i(\theta) = 0.9\theta + 0.1\theta_*$ if 10% of the subjects were thought to be misclassified from a group with event probability θ_* rather the paradigm probability $\theta^0 = \theta$.

A useful generalization is the *parametric adjustment model* in which the adjustment function $\theta^i = \phi(\theta, \alpha_i)$ depends explicitly on a parameter α_i, leading to

$$L_i^{\mathrm{Adj}}(\theta) = L_i(\phi(\theta, \alpha_i)). \tag{12}$$

For example, we may allow an arithmetic shift in the binomial success probability parameter θ by setting $\phi(\theta, \alpha_i) = \theta + \alpha_i$, or a logistic shift by specifying $\phi/(1-\phi) = e^{\alpha_i}\theta/(1-\theta)$, that is, $\phi(\theta, \alpha_i) = \theta e^{\alpha_i}/[\theta e^{\alpha_i} + 1 - \theta]$.

Although (12) offers no real increase in generality over the nonparametric adjustments made in (11), it is often easier (in our experience) to elicit expert opinion about parameter values than about transformation functions. Several examples illustrate the parametric adjustment approach below.

If the parameter α_i in (12) is regarded as uncertain and, therefore (in the Bayesian context), random with a prior probability distribution $\pi_i^{\alpha}(d\alpha_i|\theta)$, then we can form a conditional distribution for θ^i given θ by averaging (12) over the possible values of α_i,

$$\pi_i(d\theta^i|\theta) = \int \delta\big(\theta^i - \phi_i(\theta, \alpha_i)\big)\pi_i^{\alpha}(d\alpha_i|\theta),$$

and an adjusted likelihood function

$$\begin{aligned} L_i^{\mathrm{Adj}}(\theta) &= \int L_i(\theta^i)\pi_i(d\theta^i|\theta) \\ &= \int L_i(\phi(\theta, \alpha_i))\pi_i^{\alpha}(d\alpha_i|\theta). \end{aligned} \tag{13}$$

It is important *not* to use a noninformative prior for α_i, since this ordinarily results in a constant likelihood function $L_i^{\mathrm{Adj}}(\theta)$ that lends no evidence whatsoever about θ. This is the mathematical reflection of the fact that an instrument whose bias or scale is entirely unknown can give no evidence about a measured quantity.

2.2 Specific Examples of Adjustment

In this section we examine two specific examples of adjustments, each of the parametric or uncertain type described above, to illustrate the concepts.

In both case-control and cohort studies, each subject is classified three times: once when assessing eligibility, once when assessing exposure and once when assessing outcome (case or noncase status). Thus three fundamentally different sorts of classification error may affect the analysis: eligibility violation, exposure misclassification and case misclassification.

Anticipating the passive smoking application of Section 3, in which current or former smokers are ineligible, we denote by $p^i_{jk\ell}$ the (true) fraction of the ith study population with case status $j \in \mathcal{C} \equiv \{c, \bar{c}\}$ (case and noncase), exposure status $k \in \mathcal{E} \equiv \{e, \bar{e}\}$ (exposed and unexposed) and eligibility status $\ell \in \mathcal{S} \equiv \{s, \bar{s}\}$ (ineligible, i.e., ever-smoker, and eligible, i.e., never-smoker).

Interest centers on the vector $\theta^i = (\theta^i_{ce}, \theta^i_{c\bar{e}}, \theta^i_{\bar{c}e}, \theta^i_{\bar{c}\bar{e}})$ of true classification probabilities $\{\theta^i_j \equiv p^i_{j|\bar{s}}\}_{j \in \mathcal{C}\mathcal{E}}$ of eligible (never-smoking) subjects, but through study designs and classification errors the studies give direct evidence only on the apparent classification probabilities $\{q^i_{ce}, q^i_{c\bar{e}}, q^i_{\bar{c}e}, q^i_{\bar{c}\bar{e}}\}$ of the ostensibly eligible subjects admitted to the study, including both those who are truly eligible and those who are not.

2.2.1 *Eligibility violation.* Under both CCS and CHS designs, eligibility criteria may be subjective or may be based on information which is potentially inaccurate. Eligibility violations can distort evidence, particularly if violation rates differ across study arms. If the probability $\alpha^i_{\bar{s}|js}$ that an ineligible individual in the ith population of true classification $j \in \mathcal{C}\mathcal{E}$ will appear to be eligible is greater than 0 or if the probability $\alpha^i_{\bar{s}|j\bar{s}}$ that an eligible individual will be recognized as eligible is less than 1, then the classification probabilities $\{q^i_j\}$ for ostensibly eligible subjects will differ from those $\{\theta^i_j\}$ for truly eligible ones:

$$\begin{aligned} q^i_j &= \alpha^i_{\bar{s}|js}p^i_{js} + \alpha^i_{\bar{s}|j\bar{s}}p^i_{j\bar{s}} \\ &= \alpha^i_{\bar{s}|js}p^i_{js} + \alpha^i_{\bar{s}|j\bar{s}}\theta^i_j p^i_{\bar{s}}. \end{aligned} \tag{14}$$

This is a parametric adjustment $\theta^i = \phi(\theta, \alpha_i)$ of the form of (12) for the parameter vector $\alpha_i \equiv (\alpha^i_{\bar{s}|j\ell},$

$p^i_{js})_{j\in\mathcal{CE},\ell\in\mathcal{S}}$. The experimenter may specify particular values for the eligibility misclassification probabilities and ineligible-subject classification probabilities or, as in (13), may reflect uncertainty about them through informative or even reference conditional prior probability distributions.

2.2.2 *Misclassification of exposure and disease.* In studies of both CHS and CCS design a true case might be misclassified as a noncase or vice versa. Similar errors may arise in ascertaining or recording a subject's exposure status. Thus a subject in any of the four case–exposure classes $\mathcal{CE} = \{ce, c\bar{e}, \bar{c}e, \bar{c}\bar{e}\}$ in a CHS or CCS might be misclassified in any of the other classes. Denote by $\alpha^i_{j|k}$ the conditional probability that a subject in study i of true case–exposure class k will be classified to class j, for each $j, k \in \mathcal{CE}$. If the true case–exposure class probabilities for randomly drawn subjects from the population under study in study i are $p^i \equiv (p^i_{ce}, p^i_{c\bar{e}}, p^i_{\bar{c}e}, p^i_{\bar{c}\bar{e}})$, the study will give only indirect evidence about p^i through likelihood function $L^i(q^i)$, which gives direct evidence about the apparent classification probabilities $q^i \equiv (q^i_{ce}, q^i_{c\bar{e}}, q^i_{\bar{c}e}, q^i_{\bar{c}\bar{e}})$ given by

$$q^i_j = \sum_k \alpha^i_{j|k} p^i_k. \tag{15}$$

The 8×8 classification probability matrix $\alpha^i_{j|k}$, $j, k \in \mathcal{CES}$, would be the identity matrix for a study without classification errors, but in general must be expected to have some nonzero off-diagonal elements. If multiple classification errors are regarded as negligibly likely, then $\alpha^i_{j|k}$ will require that up to eight misclassification probabilities be specified for each study; we illustrate this for the ETS example in Sections 3.2.3 and 3.2.4.

3. IS EXPOSURE TO ETS ASSOCIATED WITH LUNG CANCER IN FEMALE NONSMOKERS?

Whether or not exposure to other people's tobacco smoke, or *passive smoking*, is harmful to health is an issue which has been widely debated over the past decade and which has broad implications in current tort and public policy decisions. An important open question remains about the impact of *study quality* on individual study results and on the overall body of evidence. Because lung cancer is a rare disease among never-smokers, and the possible effects under study are small, possible biases and misclassification have not been ruled out entirely as explanations for any observed increase in relative risk of deleterious health effects for those thought to be exposed to environmental tobacco smoke (ETS). Three main quality issues are commonly identified as influential in the analysis of the association between ETS exposure and lung cancer in adults who have never smoked:

1. *Misclassification of ever-smokers as never-smokers*. Lung cancer rates are known to be much higher among ever-smokers than among never-smokers (whether or not exposed to ETS) and ever-smokers are more likely to have smoking spouses than never-smokers (the so-called marriage concordance). Thus there is concern that the inclusion of active or former smokers in a study of never-smokers may lead to a systematic overstatement of the effect of ETS exposure. The debate over the effect of this issue has been lively. While some investigators such as Wu (1999) and Boffetta et al. (1998) argued that the bias is unlikely to explain the observed excess risk, others such as Lee and Forey (1999) disagreed.
2. *Misclassification of disease*. The degree to which disease classification is verified histologically differs markedly among studies, and errors in diagnosis of lung cancer deaths based on death certificates or clinical diagnoses are widely recognized (Lee, 1992, pages 128–129; Table 23.3, page 87). McFarlane, Feinstein and Wells (1986) reported that such misclassification is differential between active smokers and nonsmokers, but Lee (1992, page 129) countered that these differentials have not been established for those exposed or unexposed to ETS.
3. *Misclassification of exposure*. The measure of exposure to ETS is most often indirect and inadequate, leading to speculation about under- or overreporting and possible misclassification as "exposed" or "unexposed." Most studies in our data set adopt "married to a smoker" as a surrogate measure of exposure and do not use any objective measure of exposure (such as cotinine analysis, which itself is prone to criticism since it measures only recent exposure). LeVois and Switzer (1998) suggested that variation of this misclassification rate with case status may lead to spuriously high observed relative risk.

The Environmental Protection Agency (EPA Review, 1992) and Lee (1992) provide comprehensive discussions and literature reviews of these problems for 31 studies available at that time, of which relevant data

are available for 29. As a result, we focus on this data set and exclude more recent studies for which such data are unavailable. In Section 5 we show that this does not limit the relevance of our conclusions to the present debate.

The 29 studies are summarized in Table 1, based on the EPA Review (1992, Tables 5-1 and 5-2) and Lee (1992, Table 3.13), where primary references appear. Studies are categorized into quality tiers (1 = best; 4 = worst) by the EPA Review to reflect the study-specific level of care taken to control for various quality issues. The tier for each study was based on a sum of penalty points (ranging from −0.5, a bonus, to as high as +2.5) awarded in each of 19 categories, including eligibility (never-smoker status confirmation), explicit ETS-exposure criteria, lung cancer indication, interview type, proxy respondents, follow-up, design issues and analysis issues (control of age, control of other confounders, statistical methods). Studies in the lowest (fourth) tier were regarded by the EPA Review as unsuitable for inclusion in a meta-analysis.

Study heterogeneity was acknowledged in the EPA Review by undertaking meta-analyses *within* but not *across* country groups (Greece, Hong Kong, Japan, United States, Western Europe, China). They employed a fixed-effects model, with the explicit assump-

TABLE 1
Twenty-nine studies of the association between lung cancer and exposure to spousal smoking among never-smoking females

			Cases		Noncases			
			Exp	Unexp	Exp	Unexp	Qual	Ctry
Studies		**Years**[a]	n_{ce}	$n_{c\bar{e}}$	$n_{\bar{c}e}$	$n_{\bar{c}\bar{e}}$	**tier**[b]	**grp**[b]
Case control								
1	Akiba et al. (1986)	71–80	73	21	188	82	2	JP
2	Brownson et al. (1987)	79–82	4	15	6	41	3	US
3	Buffler et al. (1984)	76–80	33	8	164	32	3	US
4	Chan and Fung (1982)	76–77	34	50	66	73	4	HK
5	Correa et al. (1983)	79–82	14	8	61	72	2	US
6	Fontham et al. (1991)	85–88	294	126	492	288	1	US
7	Gao et al. (1987)	84–86	189	57	276	99	3	CN
8	Garfinkel et al. (1985)	71–81	91	43	254	148	2	US
9	Geng et al. (1988)	83–83	34	20	41	52	4	CN
10	Humble et al. (1987)	80–84	15	5	91	71	2	US
11	Inoue and Hirayama (1988)	73–83	18	4	30	17	4	JP
12	Kabat and Wynder (1984)	61–80	13	11	15	10	2	US
13	Kalandidi et al. (1990)	87–89	65	26	74	46	1	GR
14	Koo et al. (1987)	81–83	51	35	66	70	1	HK
15	Lam, T. et al. (1987)	83–86	115	84	152	183	2	HK
16	Lam, W. (1985)	81–84	37	23	64	80	3	HK
17	Lee et al. (1986)	79–82	22	10	45	21	2	EU
18	Liu et al. (1991)	85–86	45	9	176	26	4	CN
19	Pershagen et al. (1987)	61–80	37	33	153	141	1	EU
20	Shimizu et al. (1988)	82–85	52	38	91	72	2	JP
21	Sobue et al. (1990)	86–88	80	64	395	336	2	JP
22	Svensson et al. (1989)	83–85	24	10	114	60	2	EU
23	Trichopoulos et al. (1983)	78–80	53	24	116	109	3	GR
24	Wu et al. (1985)	81–82	19	10	38	24	2	US
25	Wu-Williams et al. (1990)	85–87	205	212	331	271	4	CN
Cohort								
26	Butler (1988)	76–82	3	5	3128	6071	2	HK
27	Garfinkel (1981)	59–72	88	65	94792	81794	3	US
28	Hirayama (1984)	65–81	163	37	69482	21858	2	JP
29	Hole et al. (1989)	72–85	5	1	1290	488	1	EU

SOURCE: EPA Review (1992), Tables 5-1 and 5-2; Lee (1992), Table 3.13.
[a]Years refers to each study's case accrual period.
[b]Described in the text.

tion that studies within country groups are relatively homogeneous with respect to exposure, incidence, risk and other confounders (EPA Review, 1992, pages 5–31). The impact of study quality is investigated through cumulative meta-analyses of studies in the various quality tier ranges (Tier 1, Tiers 1–2, Tiers 1–3), again within country groups. For some country groups, the overall relative risk was inflated with the addition of poorer quality studies, while for other groups the reverse occurred. For example, an estimate of 1.92 (90% CI 1.13–3.23) was reported for Tier 1–2 studies in Greece, compared to 2.01 (1.42–2.84) based on all studies in this country group, while respective figures for the USA country group were 1.23 (1.04–1.42) and 1.19 (1.04–1.35).

In further analyses the EPA Review pays specific attention to the quality issue of misclassification of active smokers as never-smokers. Unlike previous adjustments (Wald, Nanchahal, Thompson and Cuckle, 1986; National Research Council Committee on Passive Smoking, 1986), in which only the overall relative risk was adjusted following analysis, the EPA Review (1992) computed "corrected" relative risk estimates for each study and the consequent bias expressed as a ratio of the corrected and uncorrected estimates (EPA Review, 1992, Table 5-8). The underlying methodology, due to Wells and Stewart (Wells, 1990), is based on misclassification rates found in a small number of cotinine studies and studies of discordant answers. Different (and larger) biases were found by Lee (see discussion pages B-3–B-4, EPA Review, 1992), using essentially the same methodology, but different baseline estimates.

These observations confirm the need to account for inhomogeneity among studies and variations in study quality in a meta-analysis of these studies. In this paper we investigate alternative methods for doing this, based not on crude overall or individual study adjustment, but on integrating information about quality issues into the likelihood itself.

3.1 Exchangeable Meta-Analyses

Consider simple pooling of the data into two large studies with aggregate counts $(n^+_{ce}, n^+_{c\bar{e}}, n^+_{\bar{c}e}, n^+_{\bar{c}\bar{e}})$ of (1617, 946, 3499, 2424) for the case-control studies and (259, 108, 176 143, 147 192) for the cohort studies. With $\mathrm{Be}(1/2, 1/2)$ reference priors on the pairs of conditional probabilities $(p_{e|c},\ p_{e|\bar{c}})$ and $(p_{c|e},\ p_{c|\bar{e}})$, respectively, the posterior distributions for the exposure log odds ratio $\varepsilon_{\mathrm{LOR}}$, with likelihood given exactly in (2) and (6), are indistinguishable from the normal approximations with means and standard deviations [(3) and (4)] of 0.1690 ± 0.0487 for the case-control studies and 0.6951 ± 0.1146 for the cohort studies.

Fixed- and random-effects analyses, using Markov chain Monte Carlo (MCMC), give overall estimates for $\varepsilon_{\mathrm{LOR}}$ of 0.2072 ± 0.0471 and 0.2849 ± 0.1958, respectively. The mean exposure probabilities and mean case probabilities are similar to those from the pooled estimates, but the standard deviations are substantially larger, reflecting variability among the studies. Posterior distributions of individual study log odds ratios $\varepsilon^i_{\mathrm{LOR}}$ and classification probabilities ($p^i_{e|c}$ and $p^i_{e|\bar{c}}$ for CCS; $p^i_{c|e}$ and $p^i_{c|\bar{e}}$ for CHS) are also available from this methodology.

A partially exchangeable random-effects model can be constructed to accommodate the heterogeneity of country groups simply by employing a hierarchical prior distribution for the study-specific log odds ratios ε_i, allowing for country-group-specific effects that arise from such sources as varying exposure standards, intensities of spousal smoking and ambient air standards. Results from such a model are presented in Section 4.1.

3.2 Nonexchangeable Models: Quality Adjustment

The assumption of exchangeability that is implicit in all the methods of Section 3.1 seems untenable in light of the acknowledged heterogeneity of the 29 ETS studies. We now turn to implementing the new methods for synthesizing heterogeneous evidence presented in Section 2. While the studies differ in many ways, we address the three specific examples of nonexchangeability described in Sections 2.2.1 and 2.2.2 to illustrate how these methods for adjusting likelihoods may be used to improve inference.

Each subject is classified three times: first for eligibility (never-smoking female), then for disease (lung cancer case) and then for exposure (married to a smoker) in an order that depends on the study design (CCS or CHS). Altogether there are eight possible classifications, and through errors any of these could be (mis)classified as any other. In our approach we model explicitly the latent true *population* proportions $p^i_{jk\ell}$ with case status $j \in \mathcal{C} \equiv \{c, \bar{c}\}$ (case and noncase), exposure status $k \in \mathcal{E} \equiv \{e, \bar{e}\}$ (exposed and unexposed), and eligibility status $\ell \in \mathcal{S} \equiv \{s, \bar{s}\}$ (ineligible, i.e., ever-smoker, and eligible, i.e., never-smoker). As in Section 2.2 we denote by $\alpha^i_{j|k}$ the conditional probability that a subject in study i of true case/exposure/

eligibility class k will be misclassified as class j for $j,k \in \mathcal{CES}$, and find that study i gives direct evidence only about (some aspects of) the apparent classification probabilities

$$q^i_j = \sum_{k\in\mathcal{CES}} \alpha^i_{j|k} p^i_k, \quad j \in \mathcal{CE},$$

for apparently eligible subjects, through their individual likelihood functions

$$\begin{aligned} L^i_{\mathrm{CCS}}(q^i) &\propto (q^i_{e|c\bar{s}})^{n^i_{ce}}(1-q^i_{e|c\bar{s}})^{n^i_{c\bar{e}}} \\ &\quad\cdot (q^i_{e|\bar{c}\bar{s}})^{n^i_{\bar{c}e}}(1-q^i_{e|\bar{c}\bar{s}})^{n^i_{\bar{c}\bar{e}}}, \\ L^i_{\mathrm{CHS}}(q^i) &\propto (q^i_{c|e\bar{s}})^{n^i_{ce}}(1-q^i_{c|e\bar{s}})^{n^i_{\bar{c}e}} \\ &\quad\cdot (q^i_{c|\bar{e}\bar{s}})^{n^i_{c\bar{e}}}(1-q^i_{c|\bar{e}\bar{s}})^{n^i_{\bar{c}\bar{e}}}. \end{aligned} \tag{16}$$

We now turn to estimating the classification probabilities $\alpha^i_{j|k}$ needed to make inference about the true population-based classification probabilities p^i_k and, through them, the true relative risk of exposure $p_{c|e\bar{s}}/p_{c|\bar{e}\bar{s}}$, the nearly identical exposure odds ratio $R_e = p_{ce|\bar{s}}p_{\bar{c}\bar{e}|\bar{s}}/p_{c\bar{e}|\bar{s}}p_{\bar{c}e|\bar{s}}$ or its logarithm $\varepsilon_{\mathrm{LOR}} = \log R_e$.

Denote by θ^i the vector $\theta^i = \{\theta^i_{ce}, \theta^i_{c\bar{e}}, \theta^i_{\bar{c}e}, \theta^i_{\bar{c}\bar{e}}\}$ of true classification probabilities for eligible members of each study population (these are just the conditional classification probabilities $\theta^i_{ce} = p^i_{ce|\bar{s}}$, etc.). For the purpose of the present analysis, based on the availability of information about classification reliability in the 29 studies we consider, we make the provisional simplifying assumption that double and triple misclassifications are sufficiently rare to be negligible, and we consider only the possibility of at most one misclassification per subject. Under this simplification only four true classifications contribute to each apparent one, and (14) and (15) lead to a simple expression for the probability of being classified (rightly or wrongly) in the ith study as an eligible exposed case,

$$\begin{aligned} q^i_{ce\bar{s}} &\approx \alpha^i_{\bar{s}|ces}p^i_{ces} + \alpha^i_{ce\bar{s}|ce\bar{s}}p^i_{ce\bar{s}} \\ &\quad + \alpha^i_{e|c\bar{e}\bar{s}}p^i_{c\bar{e}\bar{s}} + \alpha^i_{c|\bar{c}e\bar{s}}p^i_{\bar{c}e\bar{s}} \\ &= \alpha^i_{\bar{s}|ces}p^i_{ces} \\ &\quad + (\alpha^i_{ce\bar{s}|ce\bar{s}}\theta^i_{ce} + \alpha^i_{e|c\bar{e}\bar{s}}\theta^i_{c\bar{e}} + \alpha^i_{c|\bar{c}e\bar{s}}\theta^i_{\bar{c}e})p^i_{\bar{s}}, \end{aligned} \tag{17}$$

as the sum of the probabilities of being in fact an ineligible exposed case or of being an eligible subject who is an exposed case (correctly classified), an unexposed case (with misclassified exposure) and an exposed noncase (with misclassified case status), respectively. Similar expressions are available for $q^i_{c\bar{e}\bar{s}}$, $q^i_{\bar{c}e\bar{s}}$ and $q^i_{\bar{c}\bar{e}\bar{s}}$, and from them, the derived conditional probabilities $q^i_{e|c\bar{s}}$ and so forth that appear in the likelihood functions (16). We now turn to the problem of identifying the parameters needed to make inference about θ^i and R^i_e from (16) and (17): the smoking prevalences p^i_s, smokers' classification probabilities $p^i_{j|s}$ and classification probabilities $\alpha^i_{j|k}$, for $j,k \in \mathcal{CES}$.

3.2.1 *Ever-smoking prevalence and classification.* Estimates of the study-specific population ever-smoking prevalences p^i_s are given in EPA Review (1992, Table B-11). Ever-smokers' exposure probability $p^i_{e|s} = K\theta^i_e/(K\theta^i_e + \theta^i_{\bar{e}})$ is available from the never-smokers' exposure probability $p^i_{e|\bar{s}} = \theta^i_e = \theta^i_{ce} + \theta^i_{\bar{c}e}$ and the marriage concordance $K \equiv (p_{es}\ p_{\bar{e}\bar{s}})/(p_{e\bar{s}}\ p_{\bar{e}s})$, reflecting the propensity of spouses to have similar smoking habits. We follow Lee (1992, pages 158–160 and Table 3.40), who found $K \approx 3$, although estimates as high as 5.52 have been reported in the literature (Ogden et al., 1997).

Because the relative risk of lung cancer associated with active smoking in the ith study, R^i_s, is in general much higher than any that might be associated with passive exposure to environmental tobacco smoke (see EPA Review, 1992, Table B-11) we take ever-smokers' case probabilities $p^i_{c|es}$ and $p^i_{c|\bar{e}s}$ not to depend on exposure status and to be given by $p^i_{c|s} = R^i_s p^i_{c|\bar{s}}$; see Section 4.3 for further comments. Since the overall cancer rate may be written $p^i_c = p^i_{c|s}p^i_s + p^i_{c|\bar{s}}p^i_{\bar{s}} = p^i_{c|\bar{s}}(p^i_s R^i_s + p^i_{\bar{s}})$, ever-smokers' and never-smokers' cancer rates may be computed as

$$\begin{aligned} p^i_{c|s} &= \frac{p^i_c R^i_s}{p^i_s R^i_s + 1 - p^i_s}, \\ p^i_{c|\bar{s}} &= \frac{p^i_c}{p^i_s R^i_s + 1 - p^i_s} \end{aligned} \tag{18}$$

from the reported smoking prevalence p^i_s, smokers' relative risk R^i_s (EPA Review, 1992, Table B-11) and overall cancer rate p^i_c (EPA Review, 1992, Table C-2). Together, these values of p^i_s, $p^i_{e|s}$, $p^i_{c|es}$ and $p^i_{c|\bar{e}s}$ determine all four ever-smokers' classification probabilities p^i_{js}, $j \in \mathcal{CE}$.

Now we turn to the classification probabilities $\alpha^i_{j|k}$, which depend on how well each study addressed the difficulties of correct classification, that is, on study quality.

3.2.2 *First study quality adjustment: Eligibility violations.* Each of the 29 studies we considered requires its eligible subjects to be never-smoking females; nevertheless, for the reasons discussed at the beginning of Section 3, one must expect a small number of current or former smokers to be erroneously included in some of the studies. In this section we implement the method presented in Section 2.2.1 of adjusting the analysis to eliminate the bias and to reflect added uncertainty.

First we need to find the study-specific and case–exposure class-specific probabilities $p_{\cdot\bar{s}'|s}$ that ever-smoking individuals in the population from which the ith study sample is drawn misrepresent their smoking status as "never-smoker." From Lee (1992, Table 3.38) and in accord with other estimates (Tables 3.36 and 3.37 of Lee, 1992; Table B-3 and the discussion on pages B-8–B-13 of EPA Review, 1992), we estimate that about 5% of ever-smokers deny ever smoking. The EPA Review and Lee considered separately the effects of misclassified regular smokers and occasional smokers; in the present analysis we do not distinguish these subclasses of ever-smokers. Lee (1992, pages 156–157) suggested that the denial rate among cancer subjects may be a bit lower, while the EPA Review (1992, page B-10) and (paradoxically) Lee (1992, page 151) reported evidence of "markedly higher" denial rates among lung cancer patients than among the general population. In light of this discrepancy, we take them not to depend on case status (see, however, the discussion in Section 4.3).

There appears to be little evidence about whether smoking misclassification rates vary with exposure status. One might imagine that exposed subjects (i.e., those married to a smoker) might be more likely to discount any former or occasional smoking, and represent themselves as never-smokers. Conversely, one might imagine that unexposed subjects find more social pressure to deny former or occasional smoking. We explore sensitivity to this aspect in Section 4.3, but following the speculation of Lee (1992, page 157), we make the provisional assumption that smoking misclassification rates do not depend on exposure status.

The EPA assigns "penalty points" A_i ranging from -0.5 (a bonus) to $+1.0$ for each study's control of this source of bias (EPA Review, 1992, Table A-2). We take error rates to be approximately 5% for typical studies, and to double with each successive penalty point, leading to

$$\alpha^i_{\bar{s}|ces} = \alpha^i_{\bar{s}|c\bar{e}\bar{s}} = \alpha^i_{\bar{s}|\bar{c}es} = \alpha^i_{\bar{s}|\bar{c}\bar{e}\bar{s}} = 0.05 \times 2^{A_i}. \tag{19}$$

3.2.3 *Second study quality adjustment: Misclassification of exposure.* Next we need to find the study- and case status-specific probabilities $p_{\cdot e'|\bar{e}}$ and $p_{\cdot\bar{e}'|e}$ that eligible (never-smoking female) individuals have their exposure status misclassified. A few studies report estimates of the *other* conditional exposure misclassification probabilities—$p_{\bar{e}|\cdot e'}$, for example, the fraction of women with negligible exposure among those classified as exposed. We use Bayes' theorem to invert the conditioning below. Friedman, Petitti and Bawol (1983) estimated that 47% of currently nonsmoking wives of smokers have negligible (less than 1 per day) exposure at home, and that 40–50% of women with nonsmoking spouses have significant ETS exposure (in the workplace, e.g.). In contrast, Lee (1992, page 130) said that exposure misclassification "does not seem likely to be a major issue," and Jarvis et al. (2001) argued that "married to a smoker" is a good surrogate for "exposed." We interpret Lee's and Jarvis et al.'s remarks as a suggestion that approximately 0–5% of women have misclassified exposure status, and we use nominal values of $p_{\bar{e}|\cdot e'} = 0.25$ and $p_{e|\cdot\bar{e}'} = 0.10$ as a compromise. Although Lee (1992, Table 3.41, page 161) offered an estimate of 22.5% for the overall apparent exposure rate $p_{\cdot e'}$, the empirical rates for the individual studies we considered ranged from 15 to 87% in the EPA Review (1992, Table 5.2, pages 5-6–5-7). Consistent with these values, we take a nominal value of 0.36 for $p_{\cdot e'}$ (approximately the sample median), leading to $p_e = p_{\cdot e'}p_{e|\cdot e'} + p_{\cdot\bar{e}'}p_{e|\cdot\bar{e}'} = 0.3340$, $p_{\bar{e}} = 1 - p_e = 0.6660$, $p_{\cdot\bar{e}'|e} = p_{\cdot\bar{e}'}\ p_{e|\cdot\bar{e}'}/p_e = 0.1916$ and $p_{\cdot e'|\bar{e}} = p_{\cdot e'}\ p_{\bar{e}|\cdot e'}/p_{\bar{e}} = 0.1351$, irrespective of case status. Thus about 14% of never-smoking women with negligible exposure to others' tobacco smoke are married to smokers and so are treated in the studies as "exposed," while almost 20% of women exposed to tobacco smoke do not have smoking spouses, but rather are exposed from other sources. The EPA assigns penalty points B_i ranging from -0.5 (a bonus) to $+2.5$ for each study's control of this source of bias (EPA Review, 1992, Table A-2). We take these rates to apply to those studies with the least control over exposure misclassification, and again take error rates to double with each successive penalty point, leading to study-specific misclassification probabilities of

$$\alpha^i_{\bar{e}|ce\bar{s}} = \alpha^i_{\bar{e}|\bar{c}e\bar{s}} = 0.1916 \times 2^{B_i - 2.5},$$
$$\alpha^i_{e|c\bar{e}\bar{s}} = \alpha^i_{e|\bar{c}\bar{e}\bar{s}} = 0.1351 \times 2^{B_i - 2.5}.$$

3.2.4 *Third study quality adjustment: Misclassification of cases.* The final classification error we considered is that of misclassification of lung cancer. We need to find the study- and exposure-specific probabilities $p_{\cdot c'|\bar{c}}$ and $p_{\cdot \bar{c}'|c}$ that eligible individuals have their case status misclassified. The EPA Review (1992) attempted to classify studies with respect to their ability to control this source of bias through the use of histological verification. Although lung cancer misdiagnosis rates are known to differ for smokers and nonsmokers, there is no evidence or suggestion that they differ with respect to ETS exposure. Lee (1992, page 129) cited studies in which 30–40% of lung cancers seen at autopsy are missed clinically. We take a nominal false-negative misclassification rate of 0.35, which we reduce for each study by the fraction C_i of cases histologically verified (Table 5-4 of EPA Review, 1992; Table 3.3 of Lee, 1992). The fraction of false positives is less than the overall lung cancer rate, negligibly small in the present context. Thus,

$$\alpha^i_{c|\bar{c}e\bar{s}} = \alpha^i_{c|\bar{c}\bar{e}\bar{s}} \approx 0,$$

$$\alpha^i_{\bar{c}|ce\bar{s}} = \alpha^i_{\bar{c}|c\bar{e}\bar{s}} = 0.350 \times (1 - C_i).$$

3.2.5 *A hierarchical prior distribution.* To complete our Bayesian model formulation it remains only to specify the joint prior distribution of the $\{\theta^i\}$ and any features of interest, such as a measure ε of exposure effect. Since θ^i can be recovered from $\theta^i_c = \theta^i_{ce} + \theta^i_{c\bar{e}}$, $\theta^i_e = \theta^i_{ce} + \theta^i_{\bar{c}e}$ and $\varepsilon^i_{\mathrm{LOR}}$ by solving for θ^i_{ce} in the quadratic relationship

$$\exp(\varepsilon^i_{\mathrm{LOR}}) = \frac{\theta^i_{ce}\theta^i_{\bar{c}\bar{e}}}{\theta^i_{c\bar{e}}\theta^i_{\bar{c}e}} = \frac{\theta^i_{ce}(1 - \theta^i_c - \theta^i_e + \theta^i_{ce})}{(\theta^i_c - \theta^i_{ce})(\theta^i_e - \theta^i_{ce})},$$

we construct the joint distribution for θ^i from that of θ^i_c, θ^i_e and ε^i.

We employ a similar prior normal hierarchical distribution for the logistics $\log(\theta^i_c/\theta^i_{\bar{c}})$ of the cancer rates $\theta^i_c = \theta^i_{ce} + \theta^i_{c\bar{e}}$ for eligible subjects in the studies, centered at conditionally independent country-group levels which are centered in turn at an overall level with prior mean the population-wide lung cancer rate, approximately $25/10^5$, with low (0.5) precision at each level to express very little prior opinion about never-smokers' cancer rates $p^i_{c|\bar{s}}$.

From our earlier estimate of $p_{\cdot e\cdot} \approx 0.36$ and observation that reported (apparent) exposure rates vary from 15 to 87%, we assign independent normal prior distributions with mean $\mu_e \approx \log(0.36/0.64) = -0.57$ and variance $\sigma^2_e = 0.84^2$ to the logistic of each θ^i_e, chosen to ensure that $\mathrm{P}[0.10 < \theta^i_e < 0.75] \approx 90\%$.

We assign conditionally independent normal distributions to the study-specific log odds ratio $\varepsilon^i_{\mathrm{LOR}} = \log(\theta^i_{ce}\theta^i_{\bar{c}\bar{e}}/\theta^i_{c\bar{e}}\theta^i_{\bar{c}e})$, centered at a country-group-specific level ε^g. The country-group means ε^g are drawn from a normal distribution centered at an overall ε, whose distribution in turn is normal centered at zero. The relative risk for active smokers, which reportedly ranges from $R^i_s = 1.66$–16.3 (EPA Review, 1992, Table B-11), is believed to exceed whatever relative risk [approximately $\exp(\varepsilon^i_{\mathrm{LOR}})$] may be associated with ETS exposure. Consistent with this we chose overall country-group-level and individual-level variances $\sigma^2_{\varepsilon O} = \sigma^2_{\varepsilon G} = \sigma^2_{\varepsilon I} = 0.33$ to ensure a marginal probability $\mathrm{P}[\exp(\varepsilon^i_{\mathrm{LOR}}) > 10] \approx 5\%$ with correlation of about $2/3$ within each country group and about $1/3$ for studies in different country groups.

4. QUALITY ADJUSTMENT RESULTS

Posterior distributions for the quantities of interest in our Bayesian hierarchical model are not available in closed form, but are easily approximated using MCMC (Besag, Green, Higdon and Mengersen, 1995; Gelfand and Smith, 1990; Gilks, Richardson and Spiegelhalter, 1996; Tierney, 1994). Full conditional distributions are available for parameters from the overall and group levels of the hierarchy, allowing us to use Gibbs sampling at those levels, while a Metropolis–Hastings approach was used at the individual study level with proposal distributions drawn from a symmetric Gaussian random walk with step sizes chosen to attain an acceptance rate of about 30% for proposed steps. The model is implemented in MatLab (MathWorks, 2002); source code and data sets are available from the authors upon request. Inference is based on 2500 equally spaced samples of nearly independent observations from runs of 10 million steps after a burn-in period of 1 million steps, which appears to be more than adequate to ensure MCMC convergence.

4.1 Unadjusted Results

It is apparent from Figure 2 that the degree of association between lung cancer and ETS varies markedly across country groups.

The model described in Section 3.2 with the misclassification probabilities $\alpha^i_{j|k}$ all set to zero for $j \neq k$, $j, k \in \mathcal{CE}$, reduces to the partially exchangeable model introduced in Section 1.3, a hierarchical model in which study-specific effects ε^i are taken to be similar within country groups. The posterior distributions of the study-specific quantities ε^i are drawn away

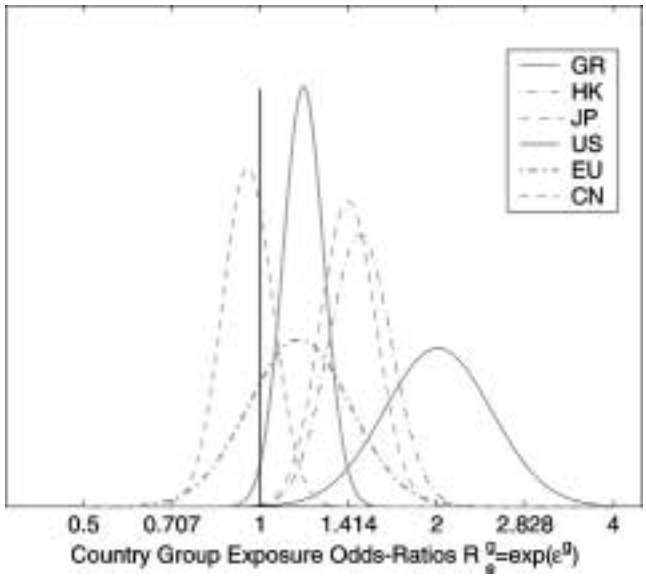

FIG. 2. *Likelihood functions for country-group effects from the EPA Review* (1992, *Table* 5-7) *for Greece, Hong Kong, Japan, United States, Western Europe and China.*

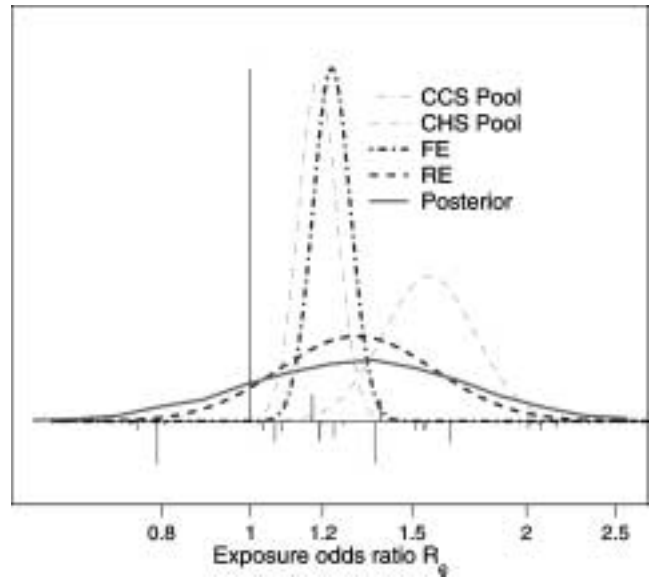

FIG. 4. *Posterior density of overall exposure odds ratio in a hierarchical model without adjustment, with posteriors from naive pooling, from a fixed-effects model and from a random-effects model.*

from their respective maximum likelihood estimates (MLEs) $\hat{\varepsilon}^i = \log(n_{ce}n_{\bar{c}\bar{e}}/n_{c\bar{e}}n_{\bar{c}e})$ (or posterior means $\bar{\varepsilon}^i$ under independent reference prior distributions) toward group level quantities ε^g, for $i \in g$, whose distributions in turn are drawn together toward an overall level ε. This *shrinkage* effect, more pronounced for smaller studies than for larger ones, is an expression of the regression effect. Figure 3 illustrates this phenomenon, with MLEs indicated at the left (A), posterior expectations $\mathrm{E}[\varepsilon^i]$ in the left center (B), group posterior expectations $\mathrm{E}[\varepsilon^g]$ in the right center (C) and the overall posterior mean $\mathrm{E}[\varepsilon]$ at the right (D). Note the wide variability of the posterior distribution for ε, shown as a probability density function.

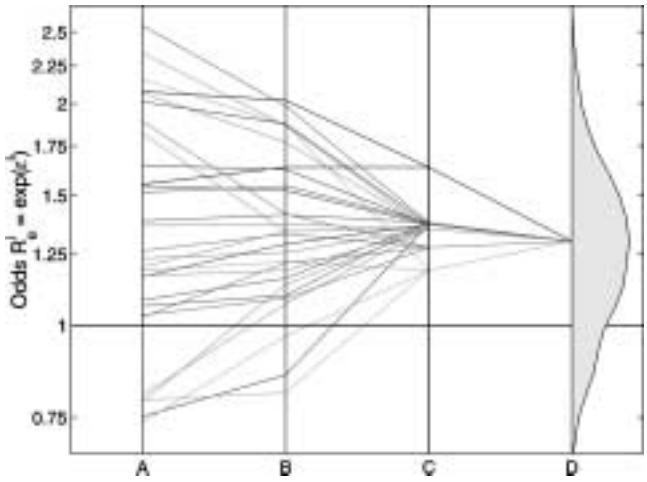

FIG. 3. *Individual exposure odds ratios* $R_e^i = \exp(\varepsilon^i)$: (A) *MLE,* (B) *posterior mean,* (C) *posterior group mean and* (D) *overall effect* (*with posterior pdf*) *in a hierarchical model without quality adjustment* (*all on log scale*).

Figure 4 shows the posterior distribution for the overall exposure odds ratio $R_e \equiv \exp(\varepsilon)$ with no adjustment (thick solid line), overlaid with the posterior distributions from naïve pooling of CCS and CHS studies (dash–dot and dashed thin lines), and from the fixed-effects and random-effects models of Section 3.1 (dash–dot and dashed thick lines). Individual study MLEs are shown on the horizontal axis as downward tickmarks (for CCS) and upward tickmarks (for CHS), with lengths proportional to precision (i.e., larger studies exerting more influence on the overall posterior distributions are indicated by larger symbols). Note the larger variability of the hierarchical random-effects model compared with the simple pooling and fixed-effects models, more accurately reflecting all the sources of variability and uncertainty, and its general similarity with the (still more widely dispersed) posterior density for the unadjusted model.

4.2 Adjusting Study Evidence for Quality Variations

Studies differ in their degrees of effort and levels of success in addressing each of the three types of misclassification discussed in Section 3. We investigate the impact of adjustment for misclassification

in two stages: first for eligibility and then for exposure and case status. The marriage concordance (Lee, 1992, pages 158–160) led many investigators (e.g., Lee, 1992, pages 143–145; EPA Review, 1992, Section 5.2.2, pages 5-22–5-25 and Table 5-8) to expect that disproportionately many exposed subjects in case-controlled studies would be ever-smokers misclassified as never-smokers, creating a bias that elevated the apparent association between lung cancer and ETS. We thus anticipated that properly adjusting for eligibility misclassification would reduce this bias and show a smaller degree of association.

Figure 5 shows the posterior means for the study-specific exposure odds ratios $\mathrm{E}[R_e^i \equiv \exp(\varepsilon^i)]$ within the unadjusted model (as in Figure 4) on the left (A), adjusted for eligibility in the middle (B) and adjusted for all misclassifications on the right (C). Evidently, our adjustment for eligibility misclassification led to a very slight reduction in apparent association between ETS and lung cancer (the mean dropped from 1.301 to 1.291), while subsequent adjustments for case and exposure misclassification led to slight reductions for some studies and slight increases for others, and generally to slight increase in mean with a much larger variability (95% CI widened and rose from [0.800, 2.066] to [0.845, 2.224]), contrary to our anticipation of a consistent downward trend, but consistent with an earlier analysis (EPA Review, 1992, Table B-11) and the recent observation of Boffetta (2002) that several sources of bias may lead to both overestimation and underestimation of true association.

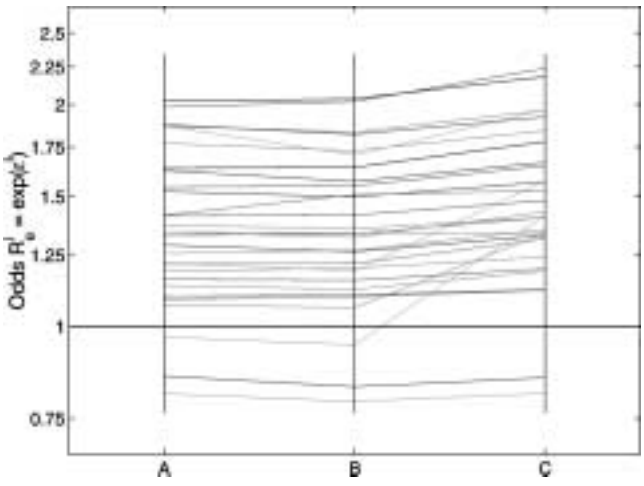

FIG. 5. *Posterior means of exposure odds ratios R_e^i in a hierarchical model* (A) *without adjustment,* (B) *adjusted for eligibility misclassification and* (C) *adjusted for all misclassifications.*

4.2.1 *A tale of two studies.* We illustrate the effects of adjustment by examining how it affects the evidence from two studies, a Tier 4 CCS study (4, CHAN) from country group HK and a Tier 2 CHS study (28, HIRA) from country group JP (quality tiers and country groups assigned by the EPA Review, 1992, are reproduced in Table 1). Case-control studies offer evidence about the population exposure rate θ_e^i, but due to their design, give no evidence about the population case rate θ_c^i. Conversely, cohort studies offer evidence about θ_c^i, but not θ_e^i. Thus the posterior distributions for these quantities differ markedly for the two designs. Data proportions (n_c/n_+, n_e/n_+) are indicated in Figure 6 by solid and dashed vertical lines, respectively. Posterior distributions for the quantities not illuminated by the data (θ_c^i for CCS; θ_e^i for CHS) remain close to their prior distributions (also shown in Figure 6 as unshaded curves).

Study evidence bears directly on the two apparent conditional probabilities that govern the arms of the study—$q_{e|c}^i$, $q_{e|\bar{c}}^i$ for CCS and $q_{c|e}^i$, $q_{c|\bar{e}}^i$ for CHS. Figure 7 shows the likelihood functions for $q_{c|e}$ and $q_{c|\bar{e}}$ for the arms of CHS study 28 (HIRA) as a dotted line, along with their posterior distributions in our hierarchical model (dashed lines) and those of the study-specific true classification probabilities for eligible subjects, $\theta_{c|e}^i$ and $\theta_{c|\bar{e}}^i$ (solid lines). Notice that the likelihood (dotted curve), representing a face-value acceptance of the immediate evidence without considering possible eligibility and classification errors, is far

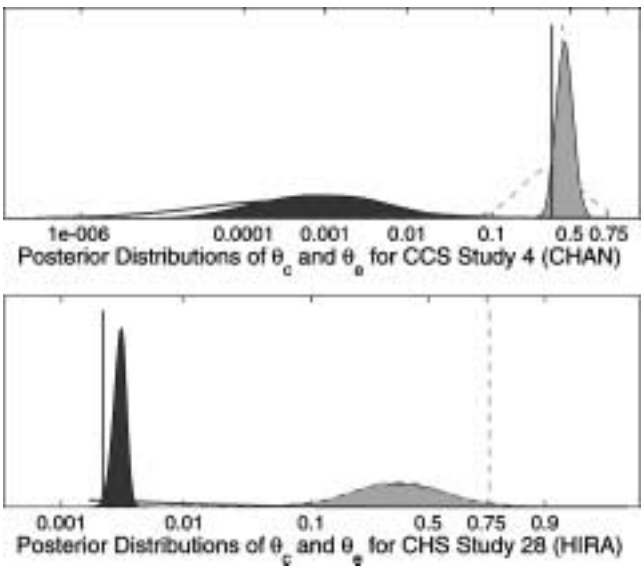

FIG. 6. *Prior and posterior distributions for θ_c^i and θ_e^i for CCS and CHS studies on* (*separate*) *logistic scales.*

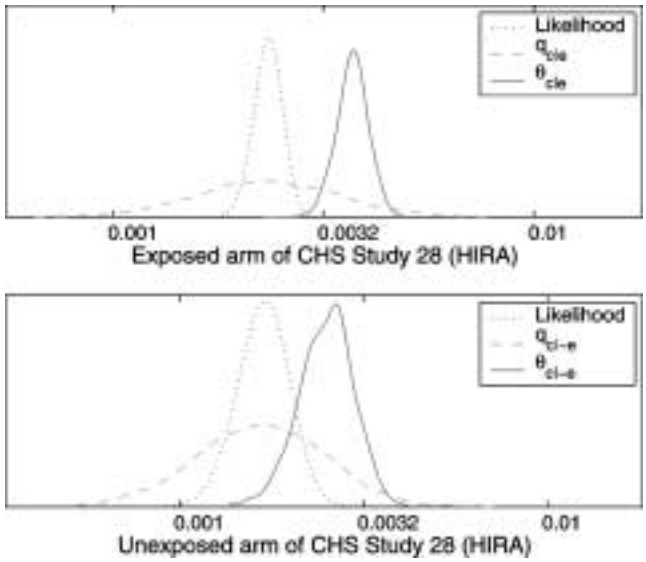

FIG. 7. *Likelihood functions and posterior probabilities for cancer case probabilities within exposed and unexposed arms of a CHS.*

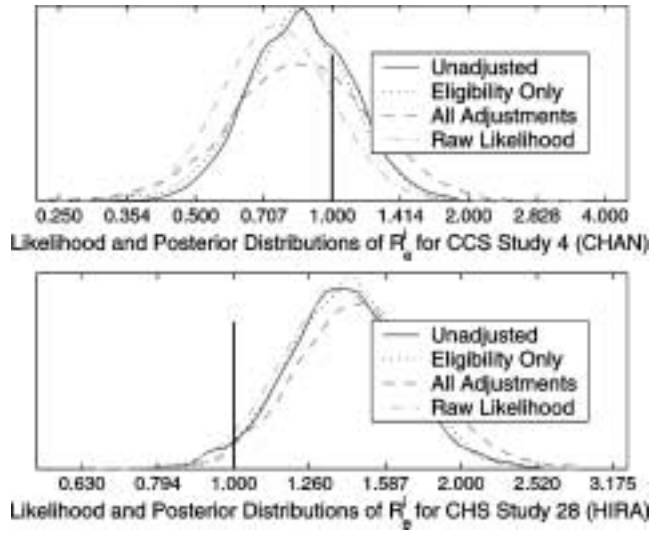

FIG. 8. *Likelihoods and posterior pdf's for ε^i for CCS and CHS studies on (separate) log scales, with different degrees of adjustment.*

narrower than the other curves which reflect the possibility that (through misclassification) the study's sample misrepresents its population and that the apparent probabilities q^i lie closer to the immediate evidence of the likelihoods than do the study-specific estimates of the population-based classification probabilities θ^i. Figures for CCSs and CHSs are similar; the plot for the q^i posterior distributions are broader than their likelihoods, but centered at the same location, while the plots for the θ^i's are very similar to those for the q^i's for studies with low misclassification rates (where adjustment for possible classification errors has little effect) and are often shifted for studies with higher misclassification rates (where adjustment effects are more pronounced).

Posterior distributions for the study-specific log odds ratios $\varepsilon^i_{\mathrm{LOR}}$ are drawn toward the country-group mean for both types of studies. Thus in Figure 8 the "unadjusted" posterior distribution (solid line) is shifted a bit to the right from the likelihood function (dashed line) in the direction of the group mean. Unexpectedly, adjustment for eligibility violations (dotted curves) makes little net change for either of these studies; adjustment for misclassification of case and exposure status (dashed curves) makes a slightly larger impact, broadening the CCS study slightly and shifting the CHS study a bit to the right.

4.3 Sensitivity

The process of building Bayesian models is really dynamic. There is no hope of modeling explicitly *everything* that is uncertain and that affects inference to any degree, however small. The decisions about exactly which features to model as uncertain, requiring the specification of a conditional prior distribution and increasing the dimension of the subsequent posterior integration, are based on judgement or evidence about how sensitive our posterior inference is to their inclusion.

In Section 3.2 we described our approach to finding suitable estimates for the many quantities needed to adjust the 29 studies we consider and make them more nearly comparable—the smoking prevalences p^i_s, smokers' classification probabilities $p^i_{j|s}$, the misclassification rates $\alpha^i_{j|k}$ for $j \neq k$, $j,k \in \mathcal{CES}$, and the means and precisions needed in our three-level hierarchical logistic normal model for $(\varepsilon^i, \theta^i_c, \theta^i_e)$. The ETS case study we present here is unusual in that so much hard work has been done by others (particularly by Lee and by the EPA Review panel) to assess and quantify population and study features in copious detail. Some of these quantities are known with less certainty than others. In this section we explore sensitivity to some of the choices we made— smoking prevalances and some misclassification rates.

In Section 3.2.3 we based estimates of the typical exposure misclassification rates $p_{\cdot\bar{e}\cdot|e} \approx 0.1916$ and $p_{\cdot e\cdot|\bar{e}} \approx 0.1351$, and true exposure rate of $p_e \approx 0.3340$, on nominal literature values of $p_{\cdot e\cdot} \approx 0.36$ for the overall apparent exposure rate and $p_{\bar{e}|\cdot e\cdot} \approx 0.25$ and $p_{e|\cdot\bar{e}\cdot} \approx 0.10$ for the true exposure misclassification

probabilities for those classified as exposed and unexposed, respectively; each of these nominal literature values was a compromise from widely varying literature estimates. If instead we were to take literally the apparent exposure estimate $p_{\cdot e'} \approx 0.225$ from Lee (1992, Table 3.41, page 161; well below the reported exposure rates among our studies), and the misclassification rate estimates of $p_{\bar{e}|\cdot e'} \approx 0.47$ and $p_{e|\cdot \bar{e}'} \approx 0.40$–$0.45$ from Friedman, Petitti and Bawol (1983) (ignoring the contrary evidence of Lee, 1992, page 130, and Jarvis et al., 2001) we would find the far higher error probabilities $p_{\cdot \bar{e}'|e} \approx 0.7474$ and $p_{\cdot e'|\bar{e}} \approx 0.1970$ and higher true exposure rate of $p_e \approx 0.4678$.

This would distort our inference about quantities of interest such as the overall odds ratio $R_e = \exp(\varepsilon_{\mathrm{LOR}})$; point estimates would rise from the value $R_e \approx 1.374$ we find in the present analysis to as high as 1.87 with these (untenable, as we will see) values. CHS study 28 (HIRA) reported $n_{ce} = 163$ cases among $n_e = 69{,}645$ apparently exposed subjects, and $n_{c\bar{e}} = 37$ cases among $n_c = 21{,}895$ apparently unexposed ones. It is impossible to reconcile these with error rates as high as $p_{\cdot \bar{e}'|e} \approx 0.7474$, $p_{\cdot e'|\bar{e}} \approx 0.1970$; the numbers m_j of subjects with true case–exposure classification $j \in \mathcal{CE}$ are related to those of apparent classification $k \in \mathcal{CE}$ by a linear relationship $n_j = \sum \alpha_{j|k} m_j$, whose solution $\hat{m}$ for these error probabilities and case-counts would give $\hat{m}_{ce} \approx 318.2$ cases among $\hat{m}_e \approx 57{,}309$ truly exposed subjects, about double the reported rate, and $\hat{m}_{c\bar{e}} \approx -10$ cases among the $\hat{m}_e \approx 10{,}657$ truly unexposed ones, an obvious impossibility.

In Section 3.2.2 we recounted the contradictory evidence and published opinion about whether and how eligibility misclassification rates might vary with case and exposure status. The results presented in Figure 5 are based on the assumption that eligibility misclassification rates do not vary with status [see (19)]. In a sensitivity analysis we explored the consequences of changing this assumption, taking the eligibility misclassification rates $\alpha_{\bar{s}|js}$ to vary with case–exposure status $j \in \mathcal{CE}$. A surprising sensitivity was revealed, particularly to the possibility of rates that vary with exposure status.

Under the hypothetical assumption of no true association between lung cancer and ETS, $R_e = 1$, a wide variety of eligibility misclassifications will all lead to an apparent positive association $\hat{R}_e > 1$ (Tweedie, Mengersen and Eccleston, 1994). It is perhaps surprising that, for true values $R_e > 1$, differential eligibility misclassification can distort the evidence in either direction. We can illustrate the point with an artificial example.

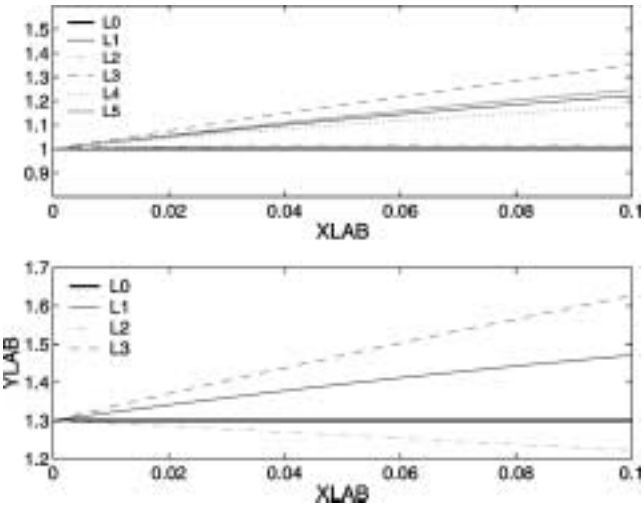

FIG. 9. *Sensitivity of odds ratio inference to exposure-specific eligibility misclassification rates.*

The top portion of Figure 9 illustrates the effect of various possibilities about eligibility misclassification rates for a true odds ratio of $R_e = 1$ (i.e., if ETS and lung cancer were unrelated); each of the curves lies above the horizontal line at $R_e \equiv 1$, indicating that in all cases the effect of misclassification is to inflate the apparent association. However, if the true odds ratio were $R_e \approx 1.3$, a value consistent with many of the estimates reported in the literature, the bottom portion of Figure 9 illustrates that if unexposed subjects were to have double the eligibility misclassification rate of exposed ones (perhaps from social pressure to deny smoking), then the apparent association would fall from $\hat{R}_e \approx 1.30$ to $\hat{R}_e \approx 1.22$ as the exposed-subject eligibility misclassification rate rises from $\alpha_{\bar{s}|ce} = 0$ to 0.10 (dashed line), distorting the apparent relative risk *downward*. Conversely, if exposed subjects were to have half the eligibility misclassification rate of unexposed ones (perhaps because their occasional smoking seems insignificant), then the apparent association would rise from $\hat{R}_e \approx 1.30$ up to the inflated figure of $\hat{R}_e \approx 1.63$ (dotted line in Figure 9). Equal deception rates also show an inflation in apparent association (dash–dot line).

Evidently risk estimates are quite sensitive to the very uncertain feature of possible variation of smoking denial rates with exposure status. Uncertainty about these rates poses an obstacle not only to synthesizing evidence from several studies, but even to interpreting the evidence of a single study. Our observed sensitivity

of the model to "apparently innocuous assumptions" is supported by the findings of Higgins and Spiegelhalter (2002) in a comparison of the inferences that arise from a meta-analysis and those from a megatrial.

It is of course possible to reflect uncertainty about the deception rates within our modeling framework by adding a new level of model hierarchy and introducing a joint prior distribution for the deception rates $\{\alpha^i_{\bar{s}|js}\}$ and misclassification rates $\{\alpha^i_{j|k}\}$ as in (13), but we do not pursue that further here.

5. DISCUSSION

The hierarchical Bayesian approach described in this paper provides a new flexibility in meta-analysis by facilitating the formal adjustment of evidence for variations in important factors such as quality and design at the study-specific level. The nonexchangeable models advocated here accommodate such study-specific heterogeneity through careful specification of joint prior distributions for study-specific parameters that adjust directly the evidence from each study's likelihood. This coherent, likelihood-based approach is arguably superior to alternative approaches such as naively combining reported values without any adjustment, discounting or excluding some or all studies on the basis of perceived problems of quality or study design, or invoking (often unsupportable) exchangeability assumptions in more traditional random effects models or making often arbitrary adjustments at a broad scale. Note that the need for adjustment and the size of its effect do not diminish for large studies—a study with a high misclassification rate will give distorted or biased evidence, no matter how large its sample size.

Other methods for combining the evidence from disparate studies have been suggested in the literature. For example, Thompson and Sharp (1999) proposed weighted regression in place of a hierarchical model for obtaining estimates of effect size for different covariate values. A multivariate meta-regression analogue of this approach could be developed for the ETS case study (as suggested by a referee). Alternatively, the problem could be cast in the form of a measurement error model (Cheng and Van Ness, 1999), in which each study's results are treated as measurements (with error) of model parameters. A related approach is adopted by Aitchison (1977, 1979), who used interclinic calibration data to combine studies from multiple clinics that differ in their methods of measurement of diagnostic features.

The unadjusted estimates in the present meta-analyses of the relative risk of lung cancer associated with ETS exposure are consistent with the results of other published meta-analyses (Wald et al., 1986; National Research Council Committee on Passive Smoking, 1986; Tweedie and Mengersen, 1992; EPA Review, 1992; Lee, 1992; OSHA, 1994; Hackshaw, Law and Wald, 1997; Boffetta et al., 1998; Zhong, Goldberg, Parent and Hanley, 2000; Boffetta 2002). For example, Hackshaw, Law and Wald (1997) estimated $R_e \approx 1.24$ (1.13–1.36), based on 37 studies; Zhong et al. (2000) reported $R_e \approx 1.20$ (1.12–1.29) based on 35 case-control and five cohort studies; Boffetta et al. (1998) reported $R_e \approx 1.16$ (0.93–1.44) in a multicenter European study; and Boffetta (2002) found an overall relative risk of $R_e \approx 1.25$ (1.15–1.37) based on 51 relevant studies that comprised 7369 observed cases of lung cancer. A very recent reanalysis by Enstrom and Kabat (2003) of the large American Cancer Society's first cancer prevention study reported a point estimate less than unity and thus an overall null effect. Moreover, the adjusted point estimates are in broad agreement with those that adjust using less formal methods, although our 95% credible interval (0.845–2.224) is somewhat wider because our hierarchical Bayesian approach is more faithful in representing multiple sources of uncertainty. It is clear that the data set considered here, despite its age, provides a vehicle for making relevant contributions to the current ongoing debate about the association between ETS and lung cancer.

Of course sources of bias other than study quality may influence the results of this or any other meta-analysis. For example, a topical issue is the potential influence of publication bias, that is, the differential tendency to publish small studies that show a positive (whether or not significant) effect, but not to publish small studies that show a negligible or negative effect. An attempt to "control" the quality of included studies by restricting meta-analysis to published studies may exacerbate the effects of publication bias, leading to spuriously high estimates of overall relative risk. Controversy surrounds the potential influence of this source of bias for the issue examined in this paper. Lee (1992, page 166) stated that "overall, it appears that some publication bias has occurred, and that it can explain a part, but by no means all, of the observed association." Givens, Smith and Tweedie (1997) proposed a data-augmentation method to simulate the results of unobserved studies within a hierarchical model and concluded that the overall association between ETS

and lung cancer in nonsmoking women may be overstated by around 30%, in both U.S. and global studies. A trim-and-fill funnel plot approach taken by Duval and Tweedie (2000a, b) led to a similar conclusion of overstatement. Copas and Shi (2000) similarly argued that the overall excess risk of 24% reported by Hackshaw, Law and Wald (1997) could be reduced to 15% after taking into account this form of bias. The ensuing debate over these papers (see the corresponding discussions) has refuted these claims strongly, arguing that they overestimate the number of missing studies and their impact.

Other quality issues not considered in the present analysis include confounding with dietary factors (suggested by Boffetta et al., 1998, but later disputed by Brennan et al., 2000), genetic susceptibility (Bennett et al., 1999) and variable length of exposure (Nyberg et al., 1998), leading to bias through right censoring.

Our approach offers a number of advantages over earlier ones. First, it requires very strict and explicit identification of the factors for which adjustment is to be made. This necessitates a strong audit trail of the rationale behind the proposed adjustment, the source and degree of detail of the evidence at a study-specific level and an acknowledgment of the uncertainty of the proposed adjustment. These in turn guide the expression of the adjustment in the model and the level of the hierarchy to which it will be applied.

A second advantage afforded by the proposed approach is an introspection about the impact of the claimed evidence, at both study-specific and global levels. This was illustrated in the ETS analysis through the sensitivity assessments of Section 4.3. The model naturally afforded a formal assessment of the surprising sensitivity to differential case and exposure misclassification, and the inapplicability at the study-specific level of some published estimates of exposure misclassification. Importantly, a consequence of these requirements for detailed evidence from trusted sources and an understanding of the implications of the choice of evidence is a healthy respect for meta-analysis in general and a careful regard for interpretations made on the basis of it.

Care must be taken in interpreting published estimates of error rates. Usually investigators report estimates of the fraction of the *sample* that is misclassified (e.g., report $\mathrm{P}[\bar{s}|\cdot s\cdot]$), whereas likelihood-based adjustment methods like ours require the fraction of the *population* that is liable to misclassification (e.g., $\mathrm{P}[\cdot\bar{s}\cdot|s]$). These differ whenever $\mathrm{P}[s] \neq \mathrm{P}[\cdot s\cdot]$, as in our example (smokers deny smoking far more often than never-smokers claim to smoke). In Section 3.2.3 we illustrate how to calculate the needed quantities from those usually available.

A third advantage is that we are able to make probabilistic statements impossible under the earlier approach (Lee, 1992; EPA Review, 1992). We can find posterior distributions of any quantities of interest (not just model parameters, but arbitrary functions of them) at any hierarchy level of the model or we can find various point estimates (means, medians, etc.) and statements of associated uncertainty (credible intervals, posterior standard deviations, etc.). In the ETS analysis, for example, the prior probability of any positive association between lung cancer and ETS is $\mathrm{P}[\varepsilon > 0] = 1/2$. Upon shrinking study evidence toward country groups (but not yet adjusting for quality), this rises to $\mathrm{P}[\varepsilon > 0] = 0.851$; upon adjusting for eligibility violations, it rises imperceptibly to $\mathrm{P}[\varepsilon > 0] = 0.856$; and with all quality adjustments, it rises to $\mathrm{P}[\varepsilon > 0] = 0.907$. Cancer odds are unlikely to increase by as much as a factor of 2. However; there is over a 90% chance that the relative risk of cancer associated with never-smoking women married to smokers is less than twice that of never-smoking women married to nonsmokers. Thus on the basis of this analysis the case for an association between ETS and lung cancer is strong, but evidence of a *sizeable* effect is not compelling.

A fourth advantage is that we can use collateral information at whatever is the most appropriate level—individual study, country group or overall—and that our MCMC-based implementation makes it easy to make detailed probabilistic statements about features of inferential importance at any of these levels. Multiple comparisons that are largely inaccessible under other paradigms are achieved here in a straightforward manner. For example, in the ETS model without adjustment for misclassification, $\mathrm{P}[\varepsilon_{\mathrm{US}} = \max_{g\in G}\{\varepsilon_g\}] = 0.11$, whereas $\mathrm{P}[\varepsilon_{\mathrm{GR}} = \max_{g\in G}\{\varepsilon_g\}] = 0.40$ and $\mathrm{P}[\varepsilon_{\mathrm{GR}} > \varepsilon_{\mathrm{US}}] = 0.69$.

The fifth advantage is the approach's simplicity and flexibility. The number and nature of levels of hierarchy can be tailored to the needs of a particular problem, and adjustment (like those we applied for classification errors) may be imposed at any level(s) of that hierarchy. The top level θ of the hierarchy can be any hyperparameter conditional on which the study-specific parameters $\{\theta^i\}$ are independent; in some cases this may be the parameter θ^0 that governs an "ideal" trial for a particular investigator's purposes, as in Section 2.1, but in other cases it may be more abstract.

Our estimates of misclassification rates (leading in our linear parametric adjustment approach to the classification matrix α) are an easily interpreted and public part of our analysis. Investigators who disagree with our choices (e.g., those who disapprove of our simplifying assumption of no multiple misclassifications) are free to substitute their own and explore the same evidence in the light of their different assumptions.

Finally, the type of adjustment itself is also flexible, in that one is able to parameterize whatever variations exist among studies under study and express uncertainty about those variations (perhaps in a nonexchangeable fashion) in the form of joint prior distributions. The computational feasibility of the MCMC algorithm that implements the approach, which in the ETS example translated into the ability to complete million-step runs in under half an hour on a laptop computer, allows the investigator to explore such modifications interactively, with appropriate assessment of convergence and sensitivity. This interactive exploratory model-building with complete representation of uncertainty is a powerful tool for interpreting and synthesizing evidence from multiple sources.

ACKNOWLEDGMENTS

The authors thank Jim Berger, Bill DuMouchel, Ingram Olkin and especially Richard Tweedie for helpful conversations, and thank the Editor and three anonymous referees for helpful suggestions. This work was supported by U.S. Environmental Protection Agency Grants CR-822047-01-0 and R828686-01-0, U.S. National Science Foundation Grants DMS-96-26829 and DMS-01-12069, Australian Research Council Grant C10024120, and the Queensland University of Technology CISSaIM (School of Mathematical Sciences). We also thank the management and staff of P. J. O'Reilly's Pub in Canberra for their support and encouragement. We would *not* like to thank them for their karaoke night, which we are still struggling to forget.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, Washington.

AITCHISON, J. (1977). A calibration problem in statistical diagnosis: The system transfer problem. *Biometrika* **64** 461–472.

AITCHISON, J. (1979). A calibration problem in statistical diagnosis: The clinical amalgamation problem. *Biometrika* **66** 357–366.

AKIBA, S., KATO, H. and BLOT, W. J. (1986). Passive smoking and lung cancer among Japanese women. *Cancer Research* **46** 4804–4807.

BENNETT, W. P., ALAVANJA, M. C. R., BLOMEKE, B., VÄHÄKANGAS, K. H., CASTRÉN, K., WELSH, J. A., BOWMAN, E. D., KHAN, M. A., FLIEDER, D. B. and HARRIS, C. C. (1999). Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women. *J. National Cancer Institute* **91** 2009–2014.

BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28.

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle*: *A Review, Generalizations, and Statistical Implications*, 2nd ed. IMS, Hayward, CA.

BERRY, D. A. (1990). A Bayesian approach to multicenter trials and meta-analysis. *ASA Proc. Biopharmaceutical Section* 1–10. Amer. Statist. Assoc., Alexandria, VA.

BESAG, J., GREEN, P., HIGDON, D. and MENGERSEN, K. L. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10** 3–66.

BOFFETTA, P. (2002). Involuntary smoking and lung cancer. *Scandinavian J. Work, Environment and Health* **28** (Suppl. 2) 30–40.

BOFFETTA, P. et al. (1998). Multicenter case-control study of exposure to environmental tobacco smoke and lung cancer in Europe. *J. National Cancer Institute* **90** 1440–1450.

BRENNAN, P., BUTLER, J., AGUDO, A., BENHAMOU, S., DARBY, S., FORTES, C., JÖCKEL, K.-H., KREUZER, M., NYBERG, F., POHLABELN, H., SARACCI, R., WICHMAN, H.-E. and BOFFETTA, P. (2000). Joint effect of diet and environmental tobacco smoke on risk of lung cancer among nonsmokers. *J. National Cancer Institute* **92** 426.

BRESLOW, N. E. and DAY, N. E. (1980). *Statistical Methods in Cancer Research* **1**. *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

BRESLOW, N. E. and DAY, N. E. (1987). *Statistical Methods in Cancer Research* **2**. *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

BROWNSON, R. C., REIF, J. S., KEEFE, T. J., FERGUSON, S. W. and PRITZL, J. A. (1987). Risk factors for adenocarcinoma of the lung. *American J. Epidemiology* **125** 25–34.

BUFFLER, P. A., PICKLE, L. W., MASON, T. J. and CONTANT, C. (1984). The causes of lung cancer in Texas. In *Lung Cancer*: *Causes and Prevention* (M. Mizell and P. Correa, eds.) 83–99. Verlag Chemie International, New York.

BUTLER, T. L. (1988). The relationship of passive smoking to various health outcomes among Seventh-Day Adventists in California. Ph.D. dissertation, Univ. California, Los Angeles.

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

CHAN, W. C. and FUNG, S. C. (1982). Lung cancer in nonsmokers in Hong Kong. In *Cancer Campaign 6, Cancer Epidemiology* (E. Grundmann, ed.) 199–202. Gustav Fischer, Stuttgart.

CHENG, C.-L. and VAN NESS, J. W. (1999). *Statistical Regression with Measurement Error*. Oxford Univ. Press.

COPAS, J. B. and SHI, J. Q. (2000). Reanalysis of epidemiological evidence on lung cancer and passive smoking. *British Medical J.* **320** 417–418.

CORREA, P., FONTHAM, E., PICKLE, L., LIN, Y. and HAENSZEL, W. (1983). Passive smoking and lung cancer. *Lancet* **322** 595–597.

COX, D. R. (1970). *The Analysis of Binary Data*. Methuen, London.

DE FINETTI, B. (1930). Funzione caratteristica di un fenomeno aleatorio. *Memorie della Reale Accademia dei Lincei* **4** 86–133. [Reprinted in de Finetti, B. (1981). *Scritti* (*1926–1930*). Cedam, Padova.]

DEMPSTER, A. P., SELWYN, M. R. and WEEKS, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.* **78** 221–227.

DRAPER, D., HODGES, J. S., MALLOWS, C. L. and PREGIBON, D. (1993). Exchangeability and data analysis (with discussion). *J. Roy. Statist. Soc. Ser. A* **156** 9–37.

DUMOUCHEL, W. H. (1990). Bayesian metaanalysis. In *Statistical Methodology in the Pharmaceutical Sciences* (D. A. Berry, ed.) 509–529. Dekker, New York.

DUMOUCHEL, W. H. (1996). Predictive cross-validation of Bayesian meta-analyses. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 107–127. Oxford Univ. Press, London.

DUMOUCHEL, W. H. and HARRIS, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.* **78** 293–315.

DUVAL, S. J. and TWEEDIE, R. L. (2000a). A non-parametric "trim and fill" method of accounting for publication bias in meta-analysis. *J. Amer. Statist. Assoc.* **95** 89–98.

DUVAL, S. J. and TWEEDIE, R. L. (2000b). Trim and fill: A simple funnel–plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56** 455–463.

EDDY, D. M. (1989). The confidence profile method: A Bayesian method for assessing health technologies. *Oper. Res.* **37** 210–228.

ENSTROM, J. E. and KABAT, G. C. (2003). Environmental tobacco smoke and tobacco related mortality in a prospective study of Californians, 1960–98. *British Medical J.* **326** 1057–1060.

EPA REVIEW (1992). Respiratory health effects of passive smoking: Lung cancer and other disorders. Document EPA/600/6-90/006F, U.S. Environmental Protection Agency. Available at www.epa.gov/nceawww1/ets/etsindex.htm.

FISHER, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* **6** 13–25.

FONTHAM, E. T. H., CORREA, P., WU-WILLIAMS, A., REYNOLDS, P., GREENBERG, R. S., BUFFLER, P. A., CHEN, V. W., BOYD, P., ALTERMAN, T., AUSTIN, D. F., LIFF, J. and GREENBERG, S. D. (1991). Lung cancer in nonsmoking women: A multicenter case-control study. *Cancer Epidemiology Biomarkers and Prevention* **1** 35–43.

FREEMAN, D. H. (1987). *Applied Categorical Data Analysis*. Dekker, New York.

FRIEDMAN, G. D., PETITTI, D. B. and BAWOL, R. D. (1983). Prevalence and correlates of passive smoking. *American J. Public Health* **73** 401–405.

GAO, Y., BLOT, W. J., ZHENG, W., ERSHOW, A. G., HSU, C. W., LEVIN, L. I., ZHANG, R. and FRAUMENI, J. F., JR. (1987). Lung cancer among Chinese women. *International J. Cancer* **40** 604–609.

GARFINKEL, L. (1981). Time trends in lung cancer mortality among nonsmokers and a note on passive smoking. *J. National Cancer Institute* **66** 1061–1066.

GARFINKEL, L., AUERBACH, O. and JOUBERT, L. (1985). Involuntary smoking and lung cancer: A case-control study. *J. National Cancer Institute* **75** 463–469.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GENEST, C. and ZIDEK, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography (with discussion). *Statist. Sci.* **1** 114–148.

GENG, G.-Y., LIANG, Z. H., ZHANG, A. Y. and WU, G. L. (1988). On the relationship between smoking and female lung cancer. In *Smoking and Health 1987* (M. Aoki, S. Hisamichi and S. Tominaga, eds.) 483–486. North-Holland, Amsterdam.

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York.

GIVENS, G. H., SMITH, D. D. and TWEEDIE, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statist. Sci.* **12** 221–250.

HACKSHAW, A. K., LAW, M. R. and WALD, N. J. (1997). The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical J.* **315** 980–988.

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

HIGGINS, J. P. T. and SPIEGELHALTER, D. J. (2002). Being sceptical about meta-analyses: A Bayesian perspective on magnesium trials in myocardial infarction. *International J. Epidemiology* **31** 96–104.

HIRAYAMA, T. (1984). Cancer mortality in nonsmoking women with smoking husbands based on a large-scale cohort study in Japan. *Preventive Medicine* **13** 680–690.

HOLE, D. J., GILLIS, C. R., CHOPRA, C. and HAWTHORNE, V. M. (1989). Passive smoking and cardiorespiratory health in a general population in the west of Scotland. *British Medical J.* **299** 423–427.

HUMBLE, C. G., SAMAT, J. M. and PATHAK, D. R. (1987). Marriage to a smoker and lung cancer risk. *American J. Public Health* **77** 598–602.

INOUE, R. and HIRAYAMA, T. (1988). Passive smoking and lung cancer in women. In *Smoking and Health 1987* (M. Aoki, S. Hisamichi and S. Tominaga, eds.) 283–285. North-Holland, Amsterdam.

JARVIS, M. J., FEYERABEND, C., BRYANT, A., HEDGES, B. and PRIMATESTA, P. (2001). Passive smoking in the home: Plasma cotinine concentrations in non-smokers with smoking partners. *Tobacco Control* **10** 368–374.

KABAT, G. C. and WYNDER, E. L. (1984). Lung cancer in nonsmokers. *Cancer* **53** 1214–1221.

KALANDIDI, A., KATSOUYANNI, K., VOROPOULOU, N., BASTAS, G., SARACCI, R. and TRICHOPOULOS, D. (1990). Passive smoking and diet in the etiology of lung cancer among nonsmokers. *Cancer Causes and Control* **1** 15–21.

KASS, R. E. and GREENHOUSE, J. B. (1989). A Bayesian perspective. Comment on "Investigating therapies of potentially great benefit: ECMO," by J. H. Ware. *Statist. Sci.* **4** 310–317.

KOO, L. C., HO, J. H., SAW, D. and HO, C. Y. (1987). Measurements of passive smoking and estimates of lung cancer risk among nonsmoking Chinese females. *International J. Cancer* **39** 162–169.

LAM, T. H., KUNG, I. T., WONG, C. M., LAM, W. K., KLEEVENS, J. W., SAW, D., HSU, C., SENEVIRATNE, S., LAM, S. Y., LO, K. K. and CHAN, W. C. (1987). Smoking, passive smoking, and histological types in lung cancer in Hong Kong Chinese women. *British J. Cancer* **56** 673–678.

LAM, W. K. (1985). A clinical and epidemiological study of carcinoma of lungs in Hong Kong. Ph.D. dissertation. Univ. Hong Kong.

LAVINE, M. L., WASSERMAN, L. and WOLPERT, R. L. (1991). Bayesian inference with specified prior marginals. *J. Amer. Statist. Assoc.* **86** 964–971.

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

LEE, P. N. (1995). Marriage to a smoker may not be a valid marker of exposure in studies relating environmental tobacco smoke to risk of lung cancer in Japanese non-smoking women. *International Archives of Occupational and Environmental Health* **67** 287–294.

LEE, P. N., CHAMBERLAIN, J. and ALDERSON, M. R. (1986). Relationship of passive smoking to risk of lung cancer and other smoking-associated diseases. *British J. Cancer* **54** 97–105.

LEE, P. N. and FOREY, B. A. (1999). Re: Multicenter case-control study of exposure to environmental tobacco smoke and lung cancer in Europe. *J. National Cancer Institute* **91** 560.

LEVOIS, M. E. and SWITZER, P. (1998). Differential exposure misclassification in case-control studies of environmental tobacco smoke and lung cancer. *J. Clinical Epidemiology* **51** 37–54.

LIAO, J. G. (1999). A hierarchical Bayesian model for combining multiple 2×2 tables using conditional likelihoods. *Biometrics* **55** 268–272.

LIU, Z., HE, X. and CHAPMAN, R. S. (1991). Smoking and other risk factors for lung cancer in Xuanwei, China. *International J. Epidemiology* **20** 26–31.

MAKRIDAKIS, S. and WINKLER, R. L. (1983). Averages of forecasts: Some empirical results. *Management Sci.* **29** 987–996.

MATHWORKS, INC. (2002). *MatLab*, Version 6.5, Release 13.

MCFARLANE, M. J., FEINSTEIN, A. R. and WELLS, C. K. (1986). Necropsy evidence of detection bias in the diagnosis of lung cancer. *Archives of Internal Medicine* **146** 1695–1698.

MORRIS, C. N. and NORMAND, S.-L. T. (1992). Hierarchical models for combining information and for meta-analyses. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) (with discussion) 321–344. Oxford Univ. Press, London.

NATIONAL RESEARCH COUNCIL COMMITTEE ON APPLIED AND THEORETICAL STATISTICS (1992). *Combining Information*: *Statistical Issues and Opportunities for Research*. National Academy Press, Washington.

NATIONAL RESEARCH COUNCIL COMMITTEE ON PASSIVE SMOKING (1986). *Environmental Tobacco Smoke*: *Measuring Exposures and Assessing Health Effects*. National Academy Press, Washington.

NORMAND, S.-L. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18** 321–359.

NYBERG, F., ARGENIUS, V., SVARTENGREN, K., SVENSSON, C. and PERSHAGEN, G. (1998). Environmental tobacco smoke and lung cancer in nonsmokers: Does time since exposure play a role? *Epidemiology* **9** 301–308.

OGDEN, M. W., MORGAN, W. T., HEAVNER, D. L., DAVIS, R. A. and STEICHEN, T. J. (1997). National incidence of smoking and misclassification among the U.S. married female population. *J. Clinical Epidemiology* **50** 253–263.

OSHA (1994). Proposed rule on indoor air quality. *Federal Register* **59** 15,968–16,039. Withdrawn as "no longer necessary" on December 14, 2001.

PERSHAGEN, G., HRUBEC, Z. and SVENSSON, C. (1987). Passive smoking and lung cancer in Swedish women. *American J. Epidemiology* **125** 17–24.

SHIMIZU, H., MORISHITA, M., MIZUNO, K., MASUDA, T., OGURA, Y., SANTO, M., NISHIMURA, M., KUNISHIMA, K., KARASAWA, K., NISHIWAKI, K., YAMAMOTO, M., HISAMICHI, S. and TOMINAGA, S. (1988). A case-control study of lung cancer in nonsmoking women. *Tohoku J. Experimental Medicine* **154** 389–397.

SMITH, T. C., SPIEGELHALTER, D. J. and PARMAR, M. K. B. (1996). Bayesian meta-analysis of randomized trials using graphical models and BUGS. In *Bayesian Biostatistics* (D. A. Berry and D. K. Stangl, eds.) 411–427. Dekker, New York.

SOBUE, T., SUZUKI, R., NAKAYAMA, N., INUBUSHI, C., MATSUDA, M., DOI, O., MORI, T., FURUSE, K., FUKUOKA, M., YASUMITSU, T., KUWABARA, O., ICHIGAYA, M., KURATA, M., NAKAHARA, K., ENDO, S. and HATTORI, S. (1990). Passive smoking among nonsmoking women and the relationship between indoor air pollution and lung cancer incidence–results of a multicenter case controlled study. *Gan No Rinsho* **36** 329–333.

SPIEGELHALTER, D. J., THOMAS, A., BEST, N. and LUNN, D. (2002). WinBUGS User Manual, Version 1.4. Available at www.mrc-bsu.cam.ac.uk/bugs.

SUTTON, A. J. and ABRAMS, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* **10** 277–303.

SVENSSON, C., PERSHAGEN, G. and KLOMINEK, J. (1989). Smoking and passive smoking in relation to lung cancer in women. *Acta Oncologica* **28** 623–629.

THOMPSON, S. G. and SHARP, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine* **18** 2693–2708.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.

TRICHOPOULOS, D., KALANDIDI, A. and SPARROS, L. (1983). Lung cancer and passive smoking: Conclusion of Greek study. *Lancet* **322** 667–668.

TWEEDIE, R. L. and MENGERSEN, K. L. (1992). Lung cancer and passive smoking: Reconciling the biochemical and epidemiological approaches. *British J. Cancer* **66** 700–705.

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.

WALD, N. J., NANCHAHAL, K., THOMPSON, S. G. and CUCKLE, H. S. (1986). Does breathing other people's tobacco smoke cause lung cancer? *British Medical J.* **293** 1217–1222.

WARE, J. H. (1989). Investigating therapies of potentially great benefit: ECMO (with discussion). *Statist. Sci.* **4** 298–340.

WELLS, C. K. (1990). Smoker misclassification does not account for observed passive smoking risk for lung cancer. Submission to Indoor Air Quality and Total Human Exposure Committee, Science Advisory Board, U.S. Environmental Protection Agency.

WOLPERT, R. L. (1986). Choosing a measure of treatment effect. In *Statistical Issues in Combining Environmental Studies* (K. Larntz, ed.) 19–25. EPA Office of Policy, Planning, and Evaluation, Washington.

WOLPERT, R. L. (1989). Eliciting and combining expert judgments about uncertainty. *International J. Technology Assessment in Health Care* **5** 537–557.

WU, A. H. (1999). Exposure misclassification bias in studies of environmental tobacco smoke and lung cancer. *Environmental Health Perspectives* **107** (Suppl. 6) 873–877.

WU, A. H., HENDERSON, B. E., PIKE, M. D. and YU, M. C. (1985). Smoking and other risk factors for lung cancer in women. *J. National Cancer Institute* **74** 747–751.

WU-WILLIAMS, A. H. and SAMET, J. M. (1990). Environmental tobacco smoke: Exposure-response relationships in epidemiologic studies. *Risk Analysis* **10** 39–48.

ZHONG, L. J., GOLDBERG, M. S., PARENT, M. E. and HANLEY, J. A. (2000). Exposure to environmental tobacco smoke and the risk of lung cancer: A meta-analysis. *Lung Cancer* **27**(1) 3–18.