J. R. Statist. Soc. A (2005)
168, *Part* 2, *pp.* 267–306

Multiple-bias modelling for analysis of observational data

Sander Greenland

University of California, Los Angeles, USA

[*Read before* The Royal Statistical Society *on Wednesday, September 29th, 2004, the President*, Professor A. P. Grieve, *in the Chair*]

Summary. Conventional analytic results do not reflect any source of uncertainty other than random error, and as a result readers must rely on informal judgments regarding the effect of possible biases. When standard errors are small these judgments often fail to capture sources of uncertainty and their interactions adequately. Multiple-bias models provide alternatives that allow one systematically to integrate major sources of uncertainty, and thus to provide better input to research planning and policy analysis. Typically, the bias parameters in the model are not identified by the analysis data and so the results depend completely on priors for those parameters. A Bayesian analysis is then natural, but several alternatives based on sensitivity analysis have appeared in the risk assessment and epidemiologic literature. Under some circumstances these methods approximate a Bayesian analysis and can be modified to do so even better. These points are illustrated with a pooled analysis of case–control studies of residential magnetic field exposure and childhood leukaemia, which highlights the diminishing value of conventional studies conducted after the early 1990s. It is argued that multiple-bias modelling should become part of the core training of anyone who will be entrusted with the analysis of observational data, and should become standard procedure when random error is not the only important source of uncertainty (as in meta-analysis and pooled analysis).

Keywords: Bayesian statistics; Confidence profile method; Confounding; Epidemiologic methods; Leukaemia; Magnetic fields; Meta-analysis; Meta-statistics; Monte Carlo methods; Observational data; Odds ratio; Relative risk; Risk analysis; Risk assessment; Sensitivity analysis

1. Introduction

1.1. *The problem*

In their discussion of observational data analysis, Mosteller and Tukey (1977), page 328, said standard errors ‘cannot be expected to show us the indeterminacies and uncertainties we face’. More recently, a prize winning paper by Maclure and Schneeweiss (2001) described how random error is but one component in a long sequence of distortive forces leading to epidemiologic observations and is often not the most important. Yet conventional analyses of observational data in the health sciences (as reviewed, for example, in Rothman and Greenland (1998), chapters 12–17) can be characterized by a two-step process that quantifies only random error—

(a) employ frequentist statistical methods based on the following assumptions, which may be grossly violated in the application but are not testable with the data under analysis:

Address for correspondence: Sander Greenland, Departments of Epidemiology and Statistics, University of California at Los Angeles, 22333 Swenson Drive, Topanga, CA 90290, USA.
E-mail: lesdomes@ucla.edu

0964–1998/05/168267

(i) the study exposure is randomized within levels of controlled covariates (sometimes replaced by a practically equivalent assumption of 'no unmeasured confounders' or 'ignorability of treatment assignment');

(ii) selection, participation and missing data are random within levels of controlled covariates;

(iii) there is no measurement error (occasionally, an unrealistically restrictive error model is used to make a correction, which can do more harm than good; see Wacholder *et al.* (1993));

(b) address possible violations of assumptions (i)–(iii) with speculative discussions of how each might have biased the statistical results. If they like the statistical results from the first step, researchers will argue that these biases are inconsequential, rarely offering evidence to that effect (Jurek *et al.*, 2004). However if they dislike their results they may focus on possible biases and may even write whole articles about them (e.g. Hatch *et al.* (2000)).

In practice, the second step is often skipped or fails to address more than one or two assumptions (Jurek *et al.*, 2004). The assumptions in the first step can be replaced by the slightly weaker assumption that any biases from violations of (i)–(iii) cancel, but appeal to such cancellation seems wishful thinking at best.

Paul Meier (personal communication) and others have defended conventional results (derived under step (a)) as 'best case' scenarios that show the absolute minimum degree of uncertainty that we should have after analysing the data. Unfortunately, the above assumptions are far too optimistic, in that they produce misleadingly narrow interval estimates precisely when caution is most needed (e.g. in meta-analyses and similar endeavours with potentially large policy impact, as illustrated below). Worse, the illusory precision of conventional results is rarely addressed by more than intuitive judgments based on flawed heuristics; see Section 4.3.

Another defence is that conventional results merely quantify random error. This defence overlooks the fact that such quantification is hypothetical and hence questionable when no random sampling or randomization has been employed and no natural random mechanism has been documented. Conventional (frequentist) statistics are still often touted as 'objective', even though in observational epidemiology and social science they rarely meet any criterion for objectivity (such as derivation from a mechanism that is known to be operative in the study). This belief has resulted in an unhealthy obsession with random error in both statistical theory and practice. A prime example, which is often lamented but still very much a problem, is the special focus that most researchers give to 'statistical significance'—a phrase whose very meaning in observational studies is unclear, owing to the lack of justification for conventional sampling distributions when random sampling and randomization are absent.

The present paper is about the formalization of the second step to free inferences from dependence on the highly implausible assumptions that are used in the first step and the often misleading intuitions that guide the second step. Although I limit the discussion to observational studies, the bias problems that I discuss often if not usually arise in clinical trials, especially when non-compliance or losses occur, and the methods described below can be brought to bear on those problems.

1.2. An overview of solutions

An assessment of uncertainty due to questionable assumptions (uncertainty analysis) is an essential part of inference. Formal assessments require a model with parameters that measure departures from those assumptions. These parameters govern the bias in methods that rely on the original assumptions; hence I shall call the parameters *bias parameters*, the model for departures a *bias model* and departures from a particular assumption a *bias source*.

Statistical literature on bias models remains fragmented; most of it deals with just one bias source, and the bias model is often used only for a sensitivity analysis (which displays bias as a function of the model parameters), although occasionally it becomes part of a Bayesian analysis. In contrast, the literature on risk assessment and decision analysis has focused on accounting for all major sources of uncertainty (Morgan and Henrion, 1990; Crouch *et al.*, 1997; Vose, 2000; Draper *et al.*, 2000). Most notable in the health sciences are the confidence profile method (Eddy *et al.*, 1992), which incorporates bias models into the likelihood function, analyses based on non-ignorable non-response models with unknown bias parameters (Little and Rubin, 2002), and Monte Carlo sensitivity analysis (MCSA), which samples bias parameters and then inverts the bias model to provide a distribution of 'bias-corrected' estimates (Lash and Silliman, 2000; Powell *et al.*, 2001; Lash and Fink, 2003; Phillips, 2003; Greenland, 2003a, 2004a; Steenland and Greenland, 2004).

1.3. Outline of paper

The next section gives some general theory for bias modelling that encompasses frequentist (sensitivity analysis), Bayesian and MCSA approaches. The theory gives a formal perspective on MCSA and suggests ways to bring it closer to posterior sampling. In particular, it operationalizes the sequential bias factor approach (Maclure and Schneeweiss, 2001) in a form that approximates Gibbs sampling under certain conditions and explains the similarity of Bayesian and MCSA results that are seen in published examples (Greenland, 2003a; Steenland and Greenland, 2004). Section 3 analyses 14 studies of magnetic fields and childhood leukaemia, extending a previous analysis (Greenland, 2003a) by adding new data, providing more detail in illustration and extending the bias model to include classification error. Classification error is a large source of uncertainty due to an absence of data on which to base priors, and due to the extreme sensitivity of results over reasonable ranges for the bias parameters. Section 4 discusses some problems in interpreting and objections to bias modelling exercises; it argues that many of the criticisms apply with even more force to conventional analyses, and that the status of the latter as expected and standard practice in observational research is unwarranted. That section can be read without covering Sections 2 and 3, and I encourage readers who are uninterested in details to skim those two sections and to focus on Section 4.

2. Some theory for observational statistics

2.1. Model expansion to include bias parameters

To review formal approaches to the bias problem, suppose that the objective is to make inferences on a target parameter $\theta = \theta(\alpha)$ of a population distribution parameterized by α, using an observed data array A. Conventional inference employs a model $L(A;\alpha)$ for the distribution of A given α and some background assumptions that are sufficient to identify θ from A, such as 'randomization of units to treatment', 'random sampling of observed units and of data on those units' and 'no measurement error' (step (a)(i)–(a)(iii) above). Most statistical methodology concerns extensions of basic models, tests and estimators to complex sampling, allocation and measurement structures. The identification of θ is retained by treating these structures as known or as jointly identifiable with θ from A under the assumed model, making assumptions as necessary to ensure identifiability (e.g. assumptions of 'no unmeasured confounders', 'missing at random' and 'ignorable non-response').

On the basis of assumptions about their operation, the effects of bias sources on L may be modelled by using a bias parameter vector η, so that the data distribution is represented by an expanded model $L(A; \alpha, \eta)$. Examples include most models for uncontrolled confounding and response bias (such as non-ignorable treatment assignment models and non-response models with unknown parameters) (e.g. Leamer (1974, 1978), Rubin (1983), Copas and Li (1997), Robins *et al.* (1999), Gelman *et al.* (2003), Little and Rubin (2002), Rosenbaum (2002) and Greenland (2003a)). Typically, α is not even partially identified without information on η, in that every distinct distribution in the family $L(A; \alpha, \eta)$ can be generated from a given α by finding a suitable η, and a prior that is uniform in η leads to $p(\alpha|A) \approx p(\alpha)$. Thus, inferences about α are infinitely sensitive to η, and $L(A; \alpha, \eta)$ is uninformative for α (and hence θ) without prior information on η. In the same manner, η is not identified without information on α, so that a prior that is uniform in α leads to $p(\eta|A) \approx p(\eta)$.

I shall consider only large sample behaviour. 'Unbiased' will be mean $\sqrt{n}$-consistent uniformly asymptotically unbiased normal, as is customary in much epidemiologic statistics. For simplicity I shall assume that α fully specifies the population distribution, but the theory can be extended to semiparametric models by using familiar modifications of likelihood (conditional, partial, etc.). I shall also assume that any necessary regularity conditions hold, e.g. the joint parameter space of (α, η) is the product of the marginal spaces of α and η, and all models and functions are smooth in their arguments and parameters. Given these conditions, conventional estimators of θ extend naturally to the expanded model. For example, suppose that $\hat{\theta}_\eta$ and $\hat{s}_\eta$ are the maximum likelihood estimator (MLE) of θ and its estimated standard error obtained from $L(A; \alpha, \eta)$ when the bias parameter is fixed at a known value η. Under the models that are used here, $\hat{\theta}_\eta$ is unbiased for θ and $\hat{\theta}_\eta \pm 1.96\hat{s}_\eta$ is a large sample 95% confidence interval when the model and value of η that are used in it are correct. Parameterizing L so that $\eta = 0$ corresponds to no bias, $L(A; \alpha, 0)$ then represents the conventional analysis distribution, $\hat{\theta}_0$ is the conventional estimator, $E(\hat{\theta}_0) - \theta$ is its (asymptotic) bias and $\hat{\theta}_0 - \hat{\theta}_\eta$ is its estimated bias given η.

2.2. Sensitivity analysis

Because η is unknown and not identified, the preceding results are of little use by themselves. Sensitivity analyses display how statistics like $\hat{\theta}_\eta$ and derived confidence limits and P-values vary with η. Epidemiologic examples date back at least to Cornfield *et al.* (1959), and since then the methods have been extended to many settings (e.g. Eddy *et al.* (1992), Greenland (1996), Copas and Li (1997), Copas (1999), Robins *et al.* (1999), Little and Rubin (2002) and Rosenbaum (2002)). Yet sensitivity analysis remains uncommon in health and medical research reports. This is not surprising, given the lack of motivation for its use and its relative unfamiliarity: sensitivity analysis is mentioned in few journal instructions or statistics text-books. As with informal discussions, those sensitivity analyses that are published rarely examine more than one bias at a time and so overlook interactions, such as those that arise from covariate effects on classification errors (Greenland and Robins, 1985; Flegal *et al.*, 1991; Lash and Silliman, 2000).

To address this concern we can use a model $L(A; \alpha, \eta)$ that incorporates multiple bias sources; indeed, my thesis is that realistic uncertainty analyses of observational data must do so. Nonetheless, the difficulty of examining a grid beyond three dimensions necessitates some sort of summarization over the sensitivity results. If (as here) the net bias in the conventional estimator $\hat{\theta}_0$ is not constrained by the data, any reasonable summary will be determined entirely by the choices of values for η and so will be arbitrarily sensitive to those choices (Greenland, 1998). Furthermore, apparent data constraints on bias and hence on inference can depend entirely on the structure of the data model and can disappear after only minor elaboration (Poole and Greenland, 1997).

These problems lead to another obstacle for the adoption of sensitivity analysis: its potential for arbitrary or nihilistic output. In the present setting, for any preselected value θ_v for θ, we can find a value for η that yields $\hat{\theta}_\eta = \theta_v$; thus, any output pattern can be produced by manipulating η. Although an arbitrary or purely manipulative analysis (one that displays a pattern that is preselected by the analyst) might be obvious in a simple case, it might not be so obvious with multiple bias sources.

To summarize: sensitivity analysis only describes the dependence of statistics on η. η is often of high dimension. The complexity of the dependence can render sensitivity analyses difficult to present without drastic (and potentially misleading) simplifications. Furthermore, sensitivity analysis may exclude no possible value for θ: results can be infinitely sensitive to η, and hence without some constraint on η the analysis will only display this fact. The constraints chosen can play a pivotal role in the appearance of the results, and informed choices essentially correspond to a prior for η (Greenland, 1998, 2001a).

2.3. Bayesian analysis and Monte Carlo sensitivity analysis

One way to address the limits of sensitivity analysis is to specify explicitly a prior density $p(\alpha, \eta)$ and base inferences for $\theta = \theta(\alpha)$ on the marginal posterior

$$p(\alpha|A) \propto \int L(A; \alpha, \eta)\, p(\alpha, \eta)\, \mathrm{d}\eta$$

(Leamer, 1974, 1978; Eddy *et al.*, 1992; Graham, 2000; Little and Rubin, 2002; Gustafson, 2003; Greenland, 2001a, 2003a). To account for shared prior information (and the resulting prior correlations) between components of η, the bias parameter η may itself be modelled as a function of known covariates and unknown hyperparameters β, resulting in a hierarchical bias model (Greenland, 2003a), as below. None-the-less, many health researchers reject formal Bayesian methods as too difficult if not philosophically objectionable: analytic solutions for $p(\theta|A)$ involving just one bias source can appear formidable (Eddy *et al.*, 1992; Graham, 2000; Gustafson, 2003), and sampler convergence remains crucial yet extremely technical (Gelman *et al.*, 2003; Gustafson, 2003).

An easier alternative specifies only a marginal prior $p(\eta)$ for the bias parameters, samples η from this prior, computes $\hat{\theta}_\eta$ and $\hat{s}_\eta$ from each sample and summarizes the resulting distribution of $\hat{\theta}_\eta$ and of statistics derived from $\hat{\theta}_\eta$ and $\hat{s}_\eta$. The $\hat{\theta}_\eta$ that are generated by this MCSA have various uses. The distribution of $\hat{\theta}_0 - \hat{\theta}_\eta$ estimates the distribution of net bias under $p(\eta)$, and the distribution of $\hat{\theta}_\eta$ can be compared with the sampling distribution of $\hat{\theta}_0$ to measure the relative importance of bias uncertainty and random error. Standard errors shrink as data accumulate and hence bias uncertainty grows in importance and eventually dominates uncertainty due to random error. As will be illustrated, the comparison of bias uncertainty and random error can reveal that the benefits of study replication diminish far below those indicated by conventional power calculations, for the latter ignore bias uncertainty.

With modification, MCSA can also provide approximate posterior inferences. Suppose that, for each η, $\hat{\theta}_\eta$ is approximately efficient (e.g. is the MLE), $p(\alpha|\eta)$ is approximately uniform and $p(\eta|A) \approx p(\eta)$. We then have approximately $p(\alpha, \eta) \propto p(\eta)$ and

$$p(\alpha|A, \eta) \propto L(A; \alpha, \eta)\, p(\alpha, \eta)/p(\eta|A) \propto L(A; \alpha, \eta),$$

and

$$p(\theta|A, \eta) \propto \int_{\theta(\alpha)=\theta} L(A; \alpha, \eta)\, \mathrm{d}\alpha$$

with the latter approximately normal$(\hat{\theta}_\eta, \hat{s}^2_\eta)$ (Gelman *et al.* (2003), chapter 4). Thus, the MCSA procedure can be modified to approximate sampling from

$$p(\theta|A) = \int p(\theta|A,\eta)\, p(\eta|A)\, \mathrm{d}\eta$$

by

(a) drawing η from $p(\eta)$
(b) computing $\hat{\theta}_\eta$ and $\hat{s}^2_\eta$, and
(c) redrawing $\hat{\theta}_\eta$ from a normal$(\hat{\theta}_\eta, \hat{s}^2_\eta)$ distribution or (equivalently) adding a normal$(0, \hat{s}^2_\eta)$ disturbance to $\hat{\theta}_\eta$ (Greenland, 2003a).

If η partitions into η_k that are estimable given α and the remaining components η_{-k}, the algorithm generalizes to arbitrary $p(\alpha,\eta)$ by cycling among $p(\alpha|A,\eta)$ and the $p(\eta_k|A,\alpha,\eta_{-k})$, drawing from an approximate normal distribution at each step, whence it can be seen as a large sample approximation to Gibbs sampling.

To avoid normal approximations, some researchers resample the data as well as η at each trial (Lash and Fink, 2003). Naïve resampling (i.e. bootstrapping from the empirical data distribution) does, however, have its own small sample problems (Efron and Tibshirani, 1993); for example, in tabular data it leaves empty observed cells as 0s and so will never visit some points in the support of the sampling distribution. To remove these 0s, we may resample the data from a smoothed table, then smooth each resample with the same procedure as that used to smooth the original data.

2.4. Some useful specializations for discrete data

Suppose now that A represents a count vector for a multiway cross-classification of the data. Conventional approaches model A with a distribution $L\{A; E(A;\alpha)\}$ that depends on the population parameters only through the expected counts $E(A;\alpha)$. Suppressing α in the notation, one extension takes $E_\eta \equiv E(A;\alpha,\eta)$, with $E_0 = E(A;\alpha,0)$ the counts expected in the absence of bias, so that the expanded model can be written $L(A;E_\eta)$. Note that E_η is an estimable quantity even though η and E_0 are not separately identified. For example, with no constraint on η or E_0, the multinomial MLE of E_η is the observed A. Hence we can model E_η directly, as will be done below for smoothing purposes.

For some models, $E_\eta = G_\eta(E_0)$ where the 'bias function' G_η is a family of mappings indexed by η and G_0 is the identity; $\hat{\theta}_\eta$ may then be taken as the MLE of θ from $L\{A; G_\eta(E_0)\}$. These models can be especially simple. For example, if the only bias source is non-response and η_R is the vector of log-response-rates within cells of the observed cross-classification, $E_\eta = G_\eta(E_0;\eta_R) = \mathrm{diag}\{\exp(\eta_R)\}E_0$; η_R thus becomes a log-linear offset to the conventional model for E_0, and $\hat{\theta}_\eta$ is the MLE from the offset model for E_η. Confounding can also be represented by a log-linear offset η_C which can be added to the response–bias offset, although this offset is a non-linear function of unmeasured covariate effects (see the example below); in that case $E_\eta = G_\eta(E_0;\eta) = B_\eta E_0$ where $B_\eta = \mathrm{diag}\{\exp(\eta_C + \eta_R)\}$ (Greenland, 2003a).

Next, suppose that q_{ij} is the probability that a unit will be classified in cell i of A given that it should be in cell j. With Q the matrix of q_{ij} and $\eta_M = \mathrm{vec}(Q)$, one model for confounding and non-response followed by misclassification would be $E_\eta = G_\eta(E_0;\eta) = B_\eta E_0$ where $B_\eta = Q\ \mathrm{diag}\{\exp(\eta_C + \eta_R)\}$. Alternatively, suppose that p_{ij} is the probability that a unit should be in cell i given that it is classified in cell j; with P the matrix of p_{ij} and $\eta_M = \mathrm{vec}(P)$ we have $E_\eta = G_\eta(E_0;\eta) = B_\eta E_0$ where $B_\eta = P^{-1}\ \mathrm{diag}\{\exp(\eta_C + \eta_R)\}$.

Without validation data identifying P, acceptable assumptions or priors about misclassification more often concern Q than P, as below. An important difference between the Q- and P-models is that A is informative for Q even without information about α; hence use of Q may lead to $p(\eta|A) \neq p(\eta)$. For example, a non-zero observed cell i implies that $q_{ij} > 0$ for some j; further assumptions can lead to stronger constraints on Q. In contrast, A alone does not constrain P. Thus, the above arguments for MCSA as an approximation to posterior sampling do not strictly apply under the Q-model unless the support of the prior $p(Q)$ falls within the identified bounds.

2.5. Sequential correction

If G_η is invertible, a 'bias-corrected' estimator $\hat{\theta}_{0\eta}$ of θ can be obtained by applying a conventional estimator $\hat{\theta}_0$ to the 'corrected data' $F(A;\eta) \equiv G_\eta^{-1}(A)$, where $F(A;0) = A$ (Lash and Fink, 2003). $\hat{\theta}_{0\eta}$ is an unbiased estimator of θ given that η and the model are correct, but the 'standard error' $\hat{s}_{0\eta}$ for $\hat{\theta}_{0\eta}$ that is obtained by applying a conventional variance estimator to $F(A;\eta)$ is not generally valid (see below).

$F(A;\eta)$ is typically derived from separate correction formulae in conventional sensitivity analyses (Rothman and Greenland (1998), chapter 19). There are many formulae $F_{\mathrm{C}}(\cdot;\eta_{\mathrm{C}})$, $F_{\mathrm{R}}(\cdot;\eta_{\mathrm{R}})$ and $F_{\mathrm{M}}(\cdot;\eta_{\mathrm{M}})$ that correct for confounding, response bias and misclassification. For example, with η_{R} the vector of log-response-rates, a correction that adjusts the relative frequencies for non-response is $F_{\mathrm{R}}(A;\eta_{\mathrm{R}}) = \mathrm{diag}\{\exp(c - \eta_{\mathrm{R}})\}A$, where c is a log-normalizing-constant vector to preserve A-margins that are fixed by design. With P and Q as above, correction formulae for misclassification include $F_{\mathrm{M}}(A;\eta_{\mathrm{M}}) = PA$ and $F_{\mathrm{M}}(A;\eta_{\mathrm{M}}) = Q^{-1}A$; these formulae automatically preserve fixed margins if (as is often the case) misclassification can only occur within the strata that are defined by those margins, for then $p_{ij} = q_{ij} = 0$ when i and j are in different strata, and hence P and Q are block diagonal when the indices are ordered by stratum.

Use of the observed counts in the formulae corresponds to using $E_\eta = A$, which is a saturated model for E_η. Some formulae (like that based on Q) can yield impossible (e.g. negative) corrected counts for certain values of η, especially if 0s are present in A, which lead to breakdown (division by 0) or wild behaviour of $\hat{\theta}_{0\eta}$. These problems can often be avoided by preliminary smoothing of A to remove non-structural 0s, e.g. by averaging A with a model-fitted count (which generalizes adding a constant to each cell (Bishop *et al.* (1975), chapter 12, and Good (1983), section 9.4), or by replacing A with a count that is expected under a nearly saturated model that preserves data patterns regardless of the statistical significance of the patterns (Greenland, 2004b).

Formulae can be applied in sequence to correct multiple biases, although the order of corrections is important if the formulae do not commute (Greenland, 1996). One can imagine each correction moving a step from the biased data back to the unbiased structure, as if hypothetically 'unwrapping the truth from the data package'. For example, suppose that the data generation process is one in which causal effects (including effects of unmeasured confounders) generate population associations, subjects are sampled in a manner that is subject to non-response and finally the responding subjects are classified subject to error. This chronology suggests that we should correct misclassification first, then non-response, and then uncontrolled confounding. With $\eta = (\eta_{\mathrm{C}}, \eta_{\mathrm{R}}, \eta_{\mathrm{M}})$, the resulting bias-corrected counts $F(A;\eta)$ are $F_{\mathrm{C}}[F_{\mathrm{R}}\{F_{\mathrm{M}}(A;\eta_{\mathrm{M}});\eta_{\mathrm{R}}\};\eta_{\mathrm{C}}]$. If the bias model is $E_\eta = B_\eta E_0$, we have $F(A;\eta) = B_\eta^{-1}A$; for example, under the Q-model above,

$$B_\eta^{-1} = \mathrm{diag}\{\exp(c - \eta_{\mathrm{C}} - \eta_{\mathrm{R}})\}\,Q^{-1}.$$

The sequential correction approach can be simpler both conceptually and computationally than fully Bayesian or likelihood-based sensitivity approaches. We just plug the sampled bias parameters $(\eta_{\mathrm{C}}, \eta_{\mathrm{R}}, \eta_{\mathrm{M}})$ into their respective formulae, apply the resulting corrections in proper sequence and compute $\hat{\theta}_{0\eta}$ from the resulting $F(A;\eta)$, possibly replacing A by a smoothed count. If $\hat{\theta}_0$ has a closed form (e.g. a Mantel–Haenszel estimator) then $\hat{\theta}_{0\eta}$ will also be of closed form, resulting in a very rapid Monte Carlo procedure. In some examples such as that below, confounding and response bias corrections simplify to division of conventional stratum-specific odds ratio estimates by independent bias factors free of the data, leading to an even simpler and more rapid procedure. Finally, as with Bayesian analyses, in some simple cases the entire MCSA distribution has a closed form approximation (Greenland, 2001a).

2.6. Sequential correction and posterior sampling

The earlier arguments for MCSA as approximate posterior sampling hinge on the use of the MLE or an equivalent $\hat{\theta}_\eta$ derived under the expanded model $L(A;E_\eta)$ and so do not extend to the use of $\hat{\theta}_{0\eta}$. Suppose that under the conventional model $\hat{\theta}_0$ is asymptotically equivalent to $\theta(\hat{\alpha})$, where $\hat{\alpha}$ is an inverse variance weighted least squares estimator of α from a regression of A on the classification axes, e.g. as when α comprises log-linear model parameters and $\hat{\theta}_0$ is the conventional MLE of a log-odds ratio. Then, if $\eta=0$, $\hat{\theta}_0$ is asymptotically efficient and first order equivalent to the conventional MLE. Because $\hat{\theta}_{0\eta}$ treats $F(A;\eta)$ as the observed counts, however, when $\eta \neq 0$ the implicit weights are no longer the correct inverse variances and $\hat{\theta}_{0\eta}$ is no longer efficient.

As an example, using maximum likelihood log-linear Poisson regression, the weight matrix for $\ln\{F(A;\eta)\}$ which is implicit in $\hat{\theta}_{0\eta}$ is $W_{0\eta} = \mathrm{diag}\{F(E_\eta;\eta)\}$. The asymptotic inverse covariance matrix for $\ln\{F(A;\eta)\}$ is, however,

$$W_\eta = \{D'_\eta \,\mathrm{diag}(E_\eta) D_\eta\}^{-1}$$

where $D_\eta = \partial[\ln\{F(E_\eta;\eta)\}]/\partial E_\eta$. If $F(A;\eta) = B_\eta^{-1}A$, then $D_\eta^{-1} = B_\eta \,\mathrm{diag}(E_0)$ and hence $W_\eta = B_\eta \,\mathrm{diag}(E_0^2/E_\eta) B'_\eta \neq W_{0\eta} = \mathrm{diag}(B_\eta^{-1}E_\eta) = \mathrm{diag}(E_0)$ unless B_η is the identity. Furthermore, when B_η is diagonal (as when only confounding and response bias are corrected), W_η reduces to $\mathrm{diag}(E_\eta)$, the weight matrix for the uncorrected regression, rather than to $W_{0\eta}$.

Because $W_{0\eta}$ does simplify to W_η when $\eta=0$, the sequential estimator using $\hat{\theta}_{0\eta}$ can be viewed as an approximation to the MLE $\hat{\theta}_\eta$ in a neighbourhood of $\eta=0$. The Monte Carlo distribution of $\hat{\theta}_{0\eta}$ over $p(\eta)$ might thus be reasonably expected to approximate that of the MLE $\hat{\theta}_\eta$ if $p(\eta)$ is centred on zero and is not too dispersed. Alternatively, if $\hat{\theta}_{0\eta}$ has an explicit weighted form, we could just estimate W_η directly and use that to compute $\hat{\theta}_{0\eta}$. Unfortunately, after misclassification correction, W_η is not diagonal, does not readily simplify along with the bias corrections and must be recomputed for each η. To avoid these problems, we could just use the diagonal matrix of uncorrected weights, which under the models that are used here would approximate the correct weights in a neighbourhood of $\eta_{\mathrm{M}}=0$ rather than just $\eta=0$. In examples based on the data below and with similar priors, the latter estimator augmented by a normal$(0,\hat{s}_0^2)$ disturbance very closely approximated posterior distributions (Greenland, 2003a), so only this modified sequential approach will be illustrated.

3. Magnetic fields and childhood leukaemia

3.1. The data

The example data in Table 1 are taken from a pooled analysis of 12 pre-1999 case–control studies of residential magnetic fields and childhood leukaemia (Greenland, Sheppard, Kaune, Poole and Kelsh, 2000), plus two additional studies unpublished at the time that the analysis was done

Table 1. Summary data from 14 case–control studies of magnetic fields and childhood leukaemia

Reference	*Country*	*Number of cases*		*Number of controls*		*Odds ratio (95% limits)*
		>3 mG	*Total*	*>3 mG*	*Total*	
Coghill *et al.* (1996)	England	1	56	0	56	∞
Dockerty *et al.* (1998)	New Zealand	3	87	0	82	∞
Feychting and Ahlbom (1993)	Sweden†	6	38	22	554	4.53 (1.72,12.0)
Kabuto (2003)	Japan	11	312	13	603	1.66 (0.73,3.75)
Linet *et al.* (1997)	USA‡	42	638	28	620	1.49 (0.91,2.44)
London *et al.* (1991)	USA‡	17	162	10	143	1.56 (0.69,3.53)
McBride *et al.* (1999)	Canada‡	14	297	11	329	1.43 (0.64,3.20)
Michaelis *et al.* (1998)	Germany	6	176	6	414	2.40 (0.76,7.55)
Olsen (1993)	Denmark†	3	833	3	1666	2.00 (0.40,9.95)
Savitz *et al.* (1988)	USA‡	3	36	5	198	3.51 (0.80,15.4)
Tomenius (1986)	Sweden	3	153	9	698	1.53 (0.41,5.72)
Tynes and Haldorsen (1997)	Norway†	0	148	31	2004	0
UK Childhood Cancer Study Investigators (1999)	UK§	5	1057	3	1053	1.66 (0.40,6.98)
Verkasalo *et al.* (1993)	Finland†	1	32	5	320	2.03 (0.23,18.0)
Totals§§		115	4025	146	8740	1.69 (1.28,2.23)

†Calculated fields (the others are direct measurement).
‡120 V–60 Hz systems (the others are 220 V–50 Hz).
§Comparison of >4 mG *versus* ⩽2 mG, excluding 16 cases and 20 controls at 2–4 mG.
§§The final column is the MLE of the common odds ratio (lower $P=0.0001$; homogeneity $P=0.24$).

(Kabuto, 2003; UK Childhood Cancer Study Investigators, 1999). Because the UK childhood cancer study did not supply individual data, its estimate compares the published categories of greater than 4 mG *versus* less than or equal to 2 mG. It is included here on the basis of several considerations. First, it appears to be sufficiently consistent with the remainder to pool. Second, a reanalysis of the pre-1999 studies using the cut point at 4 mG changed the pooled estimate by only 5% (Greenland, Sheppard, Kaune, Poole and Kelsh, 2000), suggesting that the use of 4 rather than 3 mG is of little importance (apart from increasing instability). Third, as will be discussed further, the classifications are at best a surrogate for a true unknown measure, and there are other measurement differences among the studies of potentially much greater importance. Fourth, as with most of the previous studies, covariate adjustment had almost no effect on the estimates.

Two other recent studies were excluded. Green *et al.* (1999) presented only analyses based on quartile categories, resulting in upper cut points of only 1.3–1.5 mG. This study was excluded because the use of such low cut points strongly influenced estimates from earlier studies (Greenland, Sheppard, Kaune, Poole and Kelsh, 2000); it did, however, report positive associations on contrasting the top and bottom quartiles. Schüz *et al.* (2001) had only three highly exposed cases; this study was excluded because of evidence of severe upward bias (twofold or threefold, with odds ratios from 5 to 11) in the reported estimates due to sparse data (Greenland, Schwartzbaum and Finkle, 2000), and because of insufficient reporting of raw data to allow further evaluation.

3.2. A conventional analysis

Leukaemia is a very rare disease and the usual justifications for interpreting the observed odds ratios as rate ratio estimates apply (Rothman and Greenland (1998), chapter 7). The odds ratios are remarkably consistent across studies (homogeneity $P = 0.24$), and the pooled MLE suggests a 70% higher leukaemia rate among children with estimated average exposure above 3 mG. Much like the ML results in Table 1, a Mantel–Haenszel analysis produces an estimated odds ratio for the field–leukaemia association of 1.68, with 95% confidence limits of (1.27, 2.22) and a lower deviance P-value of 0.0001. Adding study-specific random effects, the usual moment-based overdispersion estimates are 0 owing to the homogeneity, leaving the summary odds ratio and limits virtually unchanged. Under a model with a common rate ratio Ω across the underlying study populations, no bias and a uniform prior for $\theta = \ln(\Omega)$, the lower P-value can be interpreted as $p(\theta < 0|A)$, the posterior probability that $\theta < 0$.

The association is not explained or modified by any known study characteristic or feature of the available data. The results are unchanged by using finer categories (e.g. contrasting greater than 3 mG *versus* less than or equal to 1 mG) or continuous field measurements, and there is no evidence of publication bias (Greenland, Sheppard, Kaune, Poole and Kelsh, 2000). None-the-less, taking the statistics in Table 1 as unbiased for the field effect is equivalent to assuming that each study reported an experiment in which children were randomized to known residential field levels, were never switched from their initial assignment and were followed until either leukaemia, selection as a control or random censoring occurred.

Put another way, the statistics in Table 1 ignore every source of uncertainty other than random error, including

(a) possible uncontrolled shared causes (confounders) of field exposure and leukaemia,
(b) possible uncontrolled associations of exposure and disease with selection and participation (sampling and response biases) and
(c) magnetic field measurement errors.

Regarding (a), several confounders have been suggested (especially social factors) but there are fewer data on most of these factors than on magnetic fields, and their estimated associations with leukaemia are mostly less than that observed for magnetic fields (to account for the association a factor must by itself have a much stronger association with leukaemia) (Langholz, 2001; Brain *et al.*, 2003). Regarding (b), data suggest that there has been control selection bias in several studies that used direct field measurement (Hatch *et al.*, 2000; Electric Power Research Institute, 2003). Regarding (c), no one doubts that measurement errors must be large. Unfortunately there is no reference measure ('gold standard') for calibration or validation of the measures, in part because no-one knows what an aetiologically relevant magnetic field exposure would be (if one exists). There is only a 'surrogacy' hypothesis that the *known* covariate, contact current, is the 'true' (aetiologically relevant) exposure that is responsible for the observed associations (Kavet and Zaffanella, 2002; Brain *et al.*, 2003), and that magnetic fields are simply an indirect measure of this covariate. Studies are under way to address this hypothesis.

3.3. Initial simplifications

Because of the great uncertainty about the bias sources, the inferential situation is very complex and several defensible simplifications will be made. One simplification restricts attention to the dichotomization of field measurements at 3 mG, which greatly eases specification. It was suggested by the repeated observation of almost no field–leukaemia association below 3 mG and was justified by the small changes in conventional statistics that were obtained from a continuous or more finely categorized exposure model (Greenland, Sheppard, Kaune, Poole and Kelsh, 2000; Kabuto, 2003; UK Childhood Cancer Study Investigators, 1999).

Of the three covariates that were uniformly defined and measured on most subjects (study source, age and sex), only study source (modelled as an indicator vector S) is used here. On prior grounds, age and sex among preschool children should be weakly related or unrelated to the exposure and the disease; for example, as noted at least 300 years ago, the sex ratio of births is highly invariant across all factors (Stigler, 1986); also, the rate of leukaemia is only 20% higher among males than females (Brain *et al.*, 2003). Thus, since nearly all the subjects are preschool children, it appears that age and sex can be ignored, and the data conform closely to this expectation; for example, Table 4 of Greenland, Sheppard, Kaune, Poole and Kelsh (2000) shows the small changes in conventional statistics on age–sex adjustment, and adjustment also has little effect in the later studies (Kabuto, 2003; UK Childhood Cancer Study Investigators, 1999). With one exception (London *et al.*, 1991), race is nearly homogeneous within studies and so is automatically controlled by including S in the models. If further covariate modelling were desired, however, one would expand the vector S to include other covariates.

Misclassification, non-response and confounding are the bias sources that are believed important by most investigators in this area and will be the only sources that we model. Another source which is often important is publication bias (Copas, 1999), but in the present context such bias is thought to be highly unlikely because of the great public interest in null results (Greenland, Sheppard, Kaune, Poole and Kelsh, 2000). The parameters of the three modelled sources will be given independent priors, so that the full prior covariance matrix is block diagonal with three blocks. This block independence greatly simplifies specification of the prior; misclassification effects on prior information about non-response and confounding will not be considered.

As in almost all the sensitivity analysis literature, confounding will be modelled via a latent variable U such that the US conditional field–leukaemia association is unconfounded, i.e. conditioning on U removes any confounding that is left after conditioning on S and induces no other confounding. As discussed in Appendix A, the existence of such a sufficient U is guaranteed under certain causal models, and in special cases this U may have as few as three levels. Two practical simplifications are made here: U is further reduced to a binary variable, and the US conditional field–leukaemia odds ratios are assumed homogeneous across U given S. These simplifications greatly ease specification of the prior, do not constrain the amount of confounding in the conventional estimate and appear to have little effect on the results (Greenland, 2003a).

With the above simplifications the classification axes (study variables) are the study, exposure and disease, coded by S, the row vector of all 14 study indicators, X, the indicator of field measurement in the top category, and Y, the leukaemia indicator. The observed data vector A comprises the $14(2^2)=56$ SXY counts of cases and controls in each field category. Define the study level indicators $D_1 \equiv 1$ if a study used direct (compared with calculated) field measurements and $V_1 \equiv 1$ if a study was of 120 V–60 Hz (compared with 220 V–50 Hz) systems. D_1 and V_1 also code study location: $V_1 = 1$ codes North American studies, and $D_1 = 0$ for all Nordic countries except for the study of Tomenius (1986) (Table 1). Finally, let $D \equiv (D_1, 1-D_1)$ and $V \equiv (V_1, 1-V_1)$. D and V are functions of S, so S contains all the covariate data that are used here.

3.4. *Preliminary estimation of uncorrected parameters*

As mentioned earlier, certain sequential estimators break down with zero cell counts. Preliminary smoothing eliminates such counts and has theoretical advantages as well (Bishop *et al.*, 1975; Good, 1983). To minimize alteration of data patterns, the observed count vector A is replaced by a 'semi-Bayes' estimate of E_η (Greenland, 2004b), which averages A (the counts that are expected under a saturated model, regressing X on YS) with those that are expected under a highly saturated model that eliminates 0s, a logistic regression of X on Y, S, YD_1 and YV_1. The E_η-estimates that are used here are penalized likelihood fitted values from a mixed effects logistic regression with fixed effects Y, S, YD_1 and YV_1 and normal$(0, \sigma^2)$ logit residuals (random effects), where $\sigma = \ln(20)/2(1.96) = 0.764$. This estimated E_η is an iterative refinement of averaging the empirical logits (weighted by their inverse empirical variances, with zero weights for undefined logits) with the fixed effects predicted logits (weighted by $1/\sigma^2$); see Greenland (2001b) for a more general example and description of the fitting method. The σ^2-value implies that each exposure odds falls within a 20-fold range of the fixed effects prediction with 95% probability, which is a very weak restriction compared with the fixed-effects-only model. The resulting estimated E_η are very modestly shrunk from A towards the fixed effects predictions: the largest absolute change in a count is 1.01, the mean absolute change is 0.31, the Mantel–Haenszel statistics are unchanged to the third decimal place and the patterns among study-specific odds ratios (e.g. orderings) are unchanged. This data-structured smoothing should be contrasted with adding a constant to each cell, which is equivalent to averaging observed counts with those fitted from an intercept-only model (Bishop *et al.* (1975), chapter 12).

The uncorrected study-specific odds ratios are now the S-specific smoothed sample odds ratios

$$\omega_{XY}(s) \equiv E_{11s} E_{00s} / E_{10s} E_{01s}$$

where E_{xys} is the smoothed count (the estimated E_η-component) at $X = x, Y = y$ and $S = s$. $\hat{\theta}_0$ will be the logarithm of the Mantel–Haenszel weighted average of the $\omega_{XY}(s)$ over the studies:

$$\omega_{\text{MH0}} = \sum_s w_s \, \omega_{XY}(s)/w_+,$$

where $w_s = E_{10s} E_{01s} / \Sigma_{xy} E_{xys}$ and $w_+ = \Sigma_s w_s$; $\hat{s}_0^2$ will be the standard error estimate of $\ln(\omega_{\text{MH0}})$ of Robins *et al.* (1999) (Rothman and Greenland (1998), page 272). In light of the above discussion regarding efficient weighting, the uncorrected weights w_s will be used throughout; this fixed weighting over MCSA trials also avoids adding study reweighting effects to bias correction effects in the distribution of corrected estimates.

An alternative that is often used in meta-analysis, the weighted least squares (Woolf) estimator, averages $\ln\{\omega_{XY}(s)\}$ by using approximate inverse variance weights $(\Sigma_{xy} E_{xys}^{-1})^{-1}$ and so (given $\eta = 0$) is first order efficient for the common odds ratio ω. In contrast, ω_{MH0} is efficient only when $\omega = 1$, although it is only slightly inefficient in realistic examples with $\omega \neq 1$ and exhibits much better behaviour than other estimators when the data are sparse (Breslow, 1981).

3.5. *Classification error*

X will be treated as a misclassified version of a single ‘true’ (but latent) exposure indicator T. Misclassification correction converts the S-specific smoothed sample XY proportions into fitted values for the sample TY odds ratios $\omega_{TY}(s)$,

$$\omega_{TY}(s) \equiv \frac{p(T=1|Y=1,s)/p(T=0|Y=1,s)}{p(T=1|Y=0,s)/p(T=0|Y=0,s)}, \tag{1}$$

where p is used to denote sample probability (expected sample proportion). This conversion requires information on the TX-relationship. Typical prior information concerns the values of the error rates $p(X=x|T=1-x,y,s)$. Let $\varepsilon_0 \equiv p(X=0|T=1,y,s)$ and $\varepsilon_1 \equiv p(X=1|T=0,y,s)$, leaving the dependence on Y and S implicit; then ε_0 and ε_1 are the false negative rates and false positive rates and $1-\varepsilon_0 = p(X=1|T=1,y,s)$ and $1-\varepsilon_1 = p(X=0|T=0,y,s)$ are the sensitivity and specificity of X as a measure of T. Within Q, the ε_x and $1-\varepsilon_x$ are the q_{ij} within blocks defined by S and Y; $q_{ij} = 0$ outside those blocks.

It will be assumed that the error rates satisfy the weak condition $\varepsilon_x < p(X=x|T=x,y,s)$. The quantity $p(X=x|y,s)$ is then an identifiable upper bound on ε_x, and the low exposure prevalences that are seen in Table 1 and in surveys imply that the ε_1 cannot be very large. This does *not* imply that X is probably correct; for example, we may still have (and often do have) $p(T=1|X=1,y,s) < p(T=0|X=1,y,s)$ if the true exposure prevalence $p(T=1|y,s)$ is small. All the studies sought to use identical measurement protocols on cases and controls, and in all the studies very high values for ε_x (especially ε_1) are implausible. Hence it will be further assumed that within studies the same bound applies to cases and controls, and that this upper bound has a user-specified maximum m_x across studies:

$$\varepsilon_x < m(x,s) \equiv \min\{p(X=x|Y=1,s), p(X=x|Y=0,s), m_x\}. \tag{2}$$

This condition is much weaker than the common assumption that the ε_x do not vary with Y (error 'non-differential' with respect to Y), although the ε_x will be given a very high within-study between-Y correlation. The smoothed data estimates of the $p(X=x|y,s)$ will be used to estimate the $m(x,s)$.

The Q correction formula (applied to sample expected counts) is equivalent to the standard conversion formula

$$p(T=x|y,s) = \frac{p(X=x|y,s)-\varepsilon_x}{1-\varepsilon_0-\varepsilon_1} \tag{3}$$

(Rothman and Greenland (1998), chapter 19), which is positive under the above constraints. The sample TY odds ratio at $S=s$ then simplifies to

$$\omega_{TY}(s) = \frac{\{p(X=1|Y=1,s)-\varepsilon_1\}/\{p(X=0|Y=1,s)-\varepsilon_0\}}{\{p(X=1|Y=0,s)-\varepsilon_1\}/\{p(X=0|Y=0,s)-\varepsilon_0\}}. \tag{4}$$

Corrections are computed by replacing the $p(X=x|y,s)$ by the smoothed sample proportions $E_{1ys}/\Sigma_x E_{xys}$, specifying a model for the ε_x, and then sampling the model coefficients from a joint prior distribution. At each sampling, the ε_x are computed from the model; the corrected sample odds ratios are then computed from the resulting ε_x.

There is no quantitative prior information on the error rates. A few studies measured subsets of subjects with different techniques, but the differences in results are highly unstable and there is no evidence on which technique provides a more accurate measure of a true 'high exposure' indicator T. None-the-less, everyone expects considerable heterogeneity in the error rates. Direct and calculated measurements (distinguished by D) are vastly different procedures. North American and European power systems (distinguished by V) differ in ways that could affect error rates. Because measurement protocols varied greatly across studies, other between-study differences in error rates should also be expected.

The hierarchical misclassification (M) model will thus regress the error rates ε_x on S as well as D and V, with Y included to allow for possible differentiality of errors:

$$\begin{aligned}\eta_{\mathrm{M}}(x|y,s) &\equiv \ln[\varepsilon_x/\{m(x,s)-\varepsilon_x\}]\\ &= \sigma_{\mathrm{M}}\{\beta_{\mathrm{M}x} + s\beta_{\mathrm{M}Sx} + d\beta_{\mathrm{M}Dx} + v\beta_{\mathrm{M}Vx} + (y, 1-y)\beta_{\mathrm{M}Yx}\}\end{aligned} \tag{5}$$

for $x=0,1$, where $\eta_{\mathrm{M}}(x|y,s)$ is the logit of the error rate ε_x rescaled to the $\{0, m(x,s)\}$ range of the rate. The model intercepts $\beta_{\mathrm{M}x}$ have variances $\tau^2_{\mathrm{M}x}$. The remaining β-coefficients ($\beta_{\mathrm{M}Sx}$, $\beta_{\mathrm{M}Dx}$, $\beta_{\mathrm{M}Vx}$, $\beta_{\mathrm{M}Yx}$) represent the dependence of the ε_x on the second-stage (group level) covariates S, D, V and Y, and are taken as independent bivariate column vectors whose components are independent with variances (τ^2) such that $\tau^2_{\mathrm{M}x} + \tau^2_{\mathrm{M}Sx} + \tau^2_{\mathrm{M}Dx} + \tau^2_{\mathrm{M}Yx} = 1$. The coefficient scale factor σ_{M} is a known constant which is introduced solely to make the coefficient variances sum to 1, a feature that eases numerical translation of prior correlations between the η_{M} into the variance components τ^2.

Let β_{M} denote the vector of all the unknowns (the β) in the error model, and let $\omega_{TY}(s;\beta_{\mathrm{M}})$ be the study-specific corrected odds ratio that is obtained by substituting this model into $\omega_{TY}(s)$. To reflect the lack of information for ordering the ε_x, I gave all the β_{M}-components zero means and for convenience gave them normal distributions. To reflect that most of the expected heterogeneity of the ε_x is attributed to the type of measurement and study protocol, I assigned τ^2-values to produce a simple but plausible prior correlation structure among the $\eta_{\mathrm{M}}(x|y,s)$: with s_1 and s_0 coding distinct studies (distinct values of S) I wanted correlations of $\eta_{\mathrm{M}}(x|1,s_1)$ with $\eta_{\mathrm{M}}(x|0,s_0)$ ranging from small (0.30) among studies that share neither D nor V to large (0.70) among studies that share both D and V. Within studies, I wanted a nearly perfect case–control correlation (0.95) of $\eta_{\mathrm{M}}(x|1,s)$ with $\eta_{\mathrm{M}}(x|0,s)$; non-differentiality would correspond to perfect correlation, which is equivalent to dropping Y from the model. Because there are arguments for positive and negative correlations of the error rates when $T=1$ compared with when $T=0$, the $\eta_{\mathrm{M}}(1|y,s)$ were left uncorrelated with the $\eta_{\mathrm{M}}(0|y,s)$; to induce a correlation we could introduce components that are shared between the coefficients for these two logits. By back-calculation, these choices require $\tau^2_{\mathrm{M}x}=\tau^2_{\mathrm{M}Vx}=\tau^2_{\mathrm{M}Yx}=0.09$ and $\tau^2_{\mathrm{M}Dx}=0.31$, leaving $\tau^2_{\mathrm{M}Sx}=0.42$.

Unlike with non-response and confounding, there are no data on which to ground the scale factors and maximum upper bounds m_x for the ε_x. Hence the scale factor σ_{M} was set to 2, which makes each ε_x nearly uniform on its support. The m_x are the most arbitrary and so will be the focus of a small meta-sensitivity analysis, followed by an analysis in which they are treated as unknown bias parameters.

3.6. Non-response

Let $R(t,y,s)$ be the response rate among population members with $T=t, Y=y$ and $S=s$. The sample probabilities p are related to the population probabilities P by

$$p(t_0|y,s)=P(t_0|y,s)\;R(t_0,y,s)\Big/\sum_t P(t|y,s)\;R(t,y,s), \tag{6}$$

where the sum is over all possible values of T. The response bias factors are then

$$B_{\mathrm{R}}(s)\equiv\frac{R(T=1,Y=1,s)/R(T=0,Y=1,s)}{R(T=1,Y=0,s)/R(T=0,Y=0,s)}, \tag{7}$$

and hence the population TY odds ratios

$$\Omega_{TY}(s)\equiv\frac{P(T=1|Y=1,s)/P(T=0|Y=1,s)}{P(T=1|Y=0,s)/P(T=0|Y=0,s)}=\frac{P(Y=1|T=1,s)/P(Y=0|T=1,s)}{P(Y=1|T=0,s)/P(Y=0|T=0,s)} \tag{8}$$

can be obtained from the sample TY odds ratios by $\Omega_{TY}(s)=\omega_{TY}(s)/B_{\mathrm{R}}(s)$.

Variables that may have important relationships to the response include continent (coded by V) and idiosyncrasies of the study design and location (coded by S). D has an expected relationship to the response supported by data on X: direct measures ($D_1=1$) require entry to private property, leading to a low response among controls ($Y=0$) and among the exposed (Hatch *et al.*, 2000); in contrast, there is high prior probability that studies with calculated fields ($D_1=0$) have little or no response bias. Hence the model that is used here is

$$\eta_{\mathrm{R}}(s)\equiv\ln\{B_{\mathrm{R}}(s)\}=\sigma_{\mathrm{R}D}(\beta_{\mathrm{R}}+s\beta_{\mathrm{R}S}+d\beta_{\mathrm{R}D}+v\beta_{\mathrm{R}V}) \tag{9}$$

where the β-coefficients represent the dependence of B_{R} on the second-stage (group level) covariates S, D and V. The variance of the intercept β_{R} is denoted τ^2_{R}, and the factor coefficients $\beta_{\mathrm{R}S}$, $\beta_{\mathrm{R}D}$ and $\beta_{\mathrm{R}V}$ are taken as independent bivariate column vectors whose components are independent with variances (τ^2) such that $\tau_{\mathrm{R}}^2+\tau_{\mathrm{R}S}^2+\tau_{\mathrm{R}D}^2+\tau_{\mathrm{R}V}^2=1$. The scale factor $\sigma_{\mathrm{R}D}$ is treated as known but will depend on D.

Now let β_{R} denote the vector of all unknowns (β) in this specification, and $B_{\mathrm{R}}(s;\beta_{\mathrm{R}})$ the response bias model that is obtained by substituting the specification into the bias factor $B_{\mathrm{R}}(s)$. To reflect lack of information on response bias sources apart from measurement type, I gave all β_{R}-components mean 0, except that component 1 of $\beta_{\mathrm{R}D}$ was given mean $\ln(1.2)/\sigma_{\mathrm{R}D}$ on the basis of the elevated non-response among exposed controls that was observed by Hatch *et al.* (2000) and others (Electric Power Research Institute, 2003). For convenience I gave the components normal distributions. To reflect the high prior correlation for non-response across studies, I assigned τ^2-values to produce $\eta_{\mathrm{R}}(s)$ correlations ranging from moderate (0.60) between studies with different D and V to very high (0.90) between studies with the same D and V; these are produced by $\tau_{\mathrm{R}}^2=\tau_{\mathrm{R}D}^2=0.36$ and $\tau_{\mathrm{R}V}^2=0.09$, leaving $\tau_{\mathrm{R}S}^2=0.19$. To reflect the greater uncertainty about the amount of response bias in studies with direct measurement, I specified prior 50th, 5th and 95th percentiles for $B_{\mathrm{R}}(s)$ of 1.20, 0.90 and 1.60 (widely dispersed around 1.2) when $D_1=1$, and prior 50th, 5th and 95th percentiles for $B_{\mathrm{R}}(s)$ of 1.00, 0.91 and 1.10 (concentrated around 1) when $D_1=0$. These percentiles result from assigning $\sigma_{\mathrm{R}D}=\ln(1.33)/1.645$ when $D_1=1$ and $\sigma_{\mathrm{R}D}=\ln(1.10)/1.645$ when $D_1=0$.

3.7. Confounding

Let

$$\Omega_U(s)\equiv P(U=1|T=Y=0,s)/P(U=0|T=Y=0,s)$$

be the population odds of the latent confounder U among unexposed non-cases (which compose over 95% of populations in this example), let $\Omega_{TU}(y,s)$ be the (population) TU odds ratio given YS, let $\Omega_{TY}(u,s)$ be the TY odds ratio given US and let $\Omega_{UY}(t,s)$ be the UY odds ratio given TS. As mentioned earlier, I assumed that there is no three-way TUY-interaction given S, so that these S-specific odds ratios are constant over T, U and Y respectively. The change in the S-specific TY odds ratio from ignoring U is then $B_{\mathrm{C}}(s)\equiv\Omega_{TY}(s)/\Omega_{TY}(u,s)$.

Given disease rarity, $B_{\mathrm{C}}(s)$ is also the degree of TY-confounding by U (bias from ignoring U) when $S=s$. The correction formula is thus $\Omega_{TY}(u,s)=\Omega_{TY}(s)/B_{\mathrm{C}}(s)$, where

$$B_{\mathrm{C}}(s)=\frac{\{\Omega_{TU}(y,s)\,\Omega_{UY}(t,s)\,\Omega_U(s)+1\}\{\Omega_U(s)+1\}}{\{\Omega_{TU}(y,s)\,\Omega_U(s)+1\}\{\Omega_{UY}(t,s)\,\Omega_U(s)+1\}} \tag{10}$$

(Yanagawa, 1984). By analogy with response bias we could model $\ln\{B_{\mathrm{C}}(s)\}$ directly (Robins *et al.*, 1999). None-the-less, typical prior information refers instead to the odds ratios in equation (10) and considers those parameters *a priori* independent, which makes it easier to model the Ω directly. The models that are used here are

$$\eta_{TU}(s)\equiv\ln\{\Omega_{TU}(s)\}=\sigma_T(\beta_T+s\beta_{TS}+d\beta_{TD}+v\beta_{TV}), \tag{11}$$

$$\eta_U(s)\equiv\ln\{\Omega_U(s)\}=\sigma_U(\beta_U+s\beta_{US}+d\beta_{UD}+v\beta_{UV}), \tag{12}$$

$$\eta_{UY}(s)\equiv\ln\{\Omega_{UY}(s)\}=\sigma_Y(\beta_Y+s\beta_{YS}+d\beta_{YD}+v\beta_{YV}). \tag{13}$$

As with the earlier models, for convenience the linear predictors are rescaled by specified factors σ_T,σ_U and σ_Y so that the variances (τ^2) of their random (β-) components sum to 1.

Let β_{C} be the vector of all the β in these three formulae, and let $B_{\mathrm{C}}(s;\beta_{\mathrm{C}})$ be the confounding model that is obtained by substituting the specification into the bias factor $B_{\mathrm{C}}(s)$. The prior that is used here is intended to address vague suggestions that some sort of biologically and physically independent leukaemia risk factor may be associated with fields. To reflect the lack of information on specific confounding sources, I gave all β_{C}-components mean 0 and for convenience gave them normal distributions. Effects of unmeasured factors on leukaemia (parameterized by the η_{UY}) would be heavily determined by human cancer biology, which is expected to vary little with location, although the distribution of those factors could easily vary. In contrast, the associations of those factors with fields (η_{TU}) and even more the background prevalences of those factors (whose logits are the η_U) would be heavily affected by local conditions such as wiring practices. Hence, I assigned τ^2-values to produce higher correlations between the $\eta_{UY}(s)$ than between the $\eta_{TU}(s)$, and higher correlations between the $\eta_{TU}(s)$ than between the $\eta_U(s)$. For $\eta_{UY}(s)$ the correlations ranged from 0.85 between studies that share neither D nor V to 0.95 between studies that share both D and V, produced by $\tau_Y^2 = 0.72$ and $\tau_{YD}^2 = \tau_{YV}^2 = 0.09$; for $\eta_{TU}(s)$ the correlations ranged from 0.60 between studies that share neither D nor V to 0.90 between studies that share both D and V, produced by $\tau_T^2 = \tau_{TD}^2 = 0.36$ and $\tau_{TV}^2 = 0.09$; and for $\eta_U(s)$ the correlations ranged from 0.50 between studies that share neither D nor V to 0.70 between studies that share both D and V, produced by $\tau_U^2 = 0.25$ and $\tau_{UD}^2 = \tau_{UV}^2 = 0.12$.

The scale factors were based on results of Langholz (2001), who studied 13 factors associated with household wiring in a survey by Bracken *et al.* (1998). Those data exhibited factor prevalences from very low to very high, so σ_U was set to 2 to produce a nearly uniform distribution for $\Pr(U = 1 | x, y, s)$. The same data also exhibited odds ratios as high as 5.3, which suggests that $\sigma_T = \ln(6)/1.645$ is reasonable (because this choice makes 6.0 the 95th percentile of the $\omega_{TU}(s)$ distribution). There are no analogous data on which to base σ_Y, although general observations on the size of composite effects in cancer epidemiology suggest that the symmetrical choice $\sigma_Y = \sigma_T = \ln(6)/1.645$ is reasonable.

3.8. Results from single corrections

The results in Table 2 are based on 250000 trials for each case and so have Monte Carlo 95% limits within the level of precision that is displayed; hence I shall refer to the observed proportions as probabilities. Before combining corrections, it is instructive to see the effect of each one alone. As a reference point, the first row of Table 2 provides percentiles of the estimated sampling distribution of the uncorrected estimate

$$\omega_{\mathrm{MH0}} = \sum_s w_s\, \omega_{XY}(s)/w_+,$$

which is the distribution of estimates corrected for random error only by drawing a normal$(0, \hat{s}_0^2)$ error and subtracting it from $\ln(\omega_{\mathrm{MH0}})$ (Greenland, 2003a). From the 'proportion < 1' column, there is only a 0.01% chance that the random error in the original summary estimate exceeds ω_{MH0} (i.e. that random error alone could have moved the summary from less than or equal to 1 to ω_{MH0}).

In an analogous fashion, the second row provides percentiles of estimates $\Sigma_s w_s\, \omega_{XY}(s)/B_{\mathrm{R}}(s; \beta_{\mathrm{R}}) w_+$ corrected for non-response only. Under the above prior, non-response is a much larger source of uncertainty than random error; for example it yields a 5% probability that the net response bias equals or exceeds the observed ω_{MH0} (i.e. that non-response alone could have moved the summary from less than or equal to 1 to ω_{MH0}). The third row gives percentiles of the estimates $\Sigma_s w_s\, \omega_{XY}(s)/B_{\mathrm{C}}(s; \beta_{\mathrm{C}}) w_+$ corrected for confounding only. Under the above prior, confounding uncertainty is similar to uncertainty due to random error; for example, there is only a 0.2% probability that the net confounding equals or exceeds ω_{MH0} (i.e. that confounding alone could have moved the summary from less than or equal to 1 to ω_{MH0}).

Table 2. Percentiles of corrected Mantel–Haenszel odds ratios from multiple-bias analyses of 250000 trials each, with different maximum upper bounds m_0 and m_1 for the false negative rates $\varepsilon_0 \equiv p(X=0|T=1, y, s)$ and false positive rates $\varepsilon_1 \equiv p(X=1|T=0, y, s)$

Factors corrected for	*2.5th percentile*	*50th percentile*	*97.5th percentile*	*Proportion <1*	*Proportion <1.27*
Random error only†	1.27	1.68	2.22	0.0001	0.025
Response bias only	0.94	1.45	2.28	0.05	0.29
Confounding only	1.32	1.69	2.33	0.002	0.019
$m_0 = 0.05$, $m_1 = 0.01$					
Classification only	1.55	2.07	7.81	<0.0005	<0.0005
All bias sources‡	1.01	1.90	7.32	0.023	0.12
All plus random error§	0.95	1.91	7.50	0.036	0.14
$m_0 = 0.05$, $m_1 = 0.05$					
Classification only	1.24	3.63	46.7	0.003	0.031
All bias sources‡	0.96	3.26	42.6	0.031	0.089
All plus random error§	0.92	3.27	43.2	0.037	0.095
$m_0 = 0.25$, $m_1 = 0.01$					
Classification only	1.72	2.04	6.95	<0.0005	<0.0005
All bias sources‡	1.06	1.90	6.59	0.015	0.10
All plus random error§	0.99	1.91	6.73	0.027	0.12
$m_0 = 0.25$, $m_1 = 0.05$					
Classification only	1.41	3.45	41.3	<0.0005	0.008
All bias sources‡	1.06	3.11	38.1	0.017	0.068
All plus random error§	1.01	3.14	38.6	0.023	0.077
m_0, m_1 random§§					
Classification only	1.42	2.92	35.1	0.001	0.011
All bias sources‡	1.04	2.67	32.0	0.019	0.077
All plus random error§	0.99	2.70	32.5	0.026	0.088

†Lower 95% limit, point estimate, upper 95% limit and lower P-value from a Mantel–Haenszel analysis.
‡Correcting for bias from misclassification, non-response and confounding.
§Adding estimated normal random error (from the first row) to the distribution with all biases.
§§m_0 and m_1 logit normal(0, 4) on (0.025,0.40) and (0.005,0.105) respectively (roughly uniform on their support).

The first rows of the next four blocks in Table 2 provide percentiles of the estimates $\Sigma_s w_s\, \omega_{TY}(s)/w_+$, corrected for misclassification only, under some reasonable pairs for the maximum bounds m_x of the ε_x across studies. With $m_0 = 0.05$ and $m_1 = 0.01$ (which forces $\varepsilon_0 < 0.05$ and $\varepsilon_1 < 0.01$ in all studies), the above specification results in a probability of less than 0.05% that the misclassification bias equalled or exceeded $\omega_{\mathrm{MH}0}$ (i.e. that misclassification alone could have shifted the summary estimate from less than or equal to 1 to $\omega_{\mathrm{MH}0}$). Increasing m_1 alone to 0.05 increases this probability to 0.3%, but then increasing m_0 to 0.25 reduces the probability to below 0.05% again. In all cases, however, it appears improbable that misclassification alone moved the summary from 1 or less to $\omega_{\mathrm{MH}0}$.

It may seem paradoxical that increasing classification error bounds can reduce the probability of bias exceeding $\omega_{\mathrm{MH}0}$. With 'nearly' non-differential misclassification, however, the bias that is produced by the misclassification is on average towards 1, in accord with the idea that non-differential exposure measurement error that is independent across units attenuates the observed association. As a result, the correction to the observed positive association must on average be upwards, and so (as can be seen in Table 2) the corrected estimates are distributed mostly above $\omega_{\mathrm{MH}0}$, regardless of the bounds. The behaviour of the lower tail of the distribution is more complex, however. First, note that when $m_0 = m_1 = 0$ there is no misclassification (all $\varepsilon_x = 0$), and so the probability that the classification correction exceeds $\omega_{\mathrm{MH}0}$ is 0. As the m_x increase from 0 the ε_x can vary more widely, and hence the dispersion of the corrected estimates initially expands as the location shifts upwards. The dispersion and location change in a highly non-linear fashion and have opposing effects on the lower tail percentiles. The dispersion can increase more rapidly than the location and thus increase the probability that the correction exceeds $\omega_{\mathrm{MH}0}$ (for example, compare the results for $m_0 = 0.01$ and $m_1 = 0.05$ with those for $m_0 = m_1 = 0.05$) but can also decline as the range of misclassification rates increases and thus reduce the probability that the correction exceeds $\omega_{\mathrm{MH}0}$ (for example, compare the results for $m_0 = 0.25$ and $m_1 = 0.05$ with those for $m_0 = m_1 = 0.05$). These phenomena can be further explained algebraically but for brevity I omit the details.

3.9. Combined corrections, and subsequent inferences

Let $\beta = (\beta_{\mathrm{C}}, \beta_{\mathrm{R}}, \beta_{\mathrm{M}})$. Table 2 provides the percentiles of the multiple-corrected estimates

$$\Omega_{TY}(\beta) = \sum_s w_s\, \omega_{TY}(s; \beta_{\mathrm{M}}) / B_{\mathrm{C}}(s; \beta_{\mathrm{C}})\, B_{\mathrm{R}}(s; \beta_{\mathrm{R}})\, w_+$$

for different m_x-pairs. It also gives percentiles after including log-normal random error at each draw of β, i.e. percentiles of $\Omega_{TY}(\beta) \exp(Z)$ where Z is normal$(0, \hat{s}_0^2)$. We can now look at features of the distribution of the corrected estimates without and with correction for random error and compare these results with those of the conventional analysis (which accounts only for random error). Uncertainty about the TY-effect due to uncertainty about classification error is sensitive to the m_x, especially to the false positive bound m_1 (which is unsurprising, given the low prevalence of exposure). For example, with random error included, the probability that the corrected estimate falls below 1 (i.e. that bias plus random error explain the observed association) is 3.6% for the first pair of m_x but 2.3% for the last pair. The results are also sensitive to the form of the ε_x-distributions within their support (which are not shown). These features should temper any conclusions about the TY-effect that might be drawn from the conventional results.

Uncertainty about appropriate bounds suggests adding the m_x to the model as hyperprior parameters. As an example, the final set of results in Table 2 comes from sampling m_0 and m_1 from logit–normal(0,4) distributions rescaled to (0.025,0.40) and (0.005,0.105) respectively, which are close to uniform on these intervals. The net result of this extension is an averaging of the fixed m_x results over the range of the m_x in the sensitivity analysis. Similar results can be obtained by making σ_{M} unknown with a prior.

Given the prior, the results in Table 2 might be taken as favouring the hypothesis of a leukaemogenic effect of magnetic fields or a close correlate for which they are a surrogate. None-the-less, no agreement about the existence of an effect (let alone its size) could be forced by the data without more precise knowledge of the classification errors. Classification error is the largest source of uncertainty because (unlike non-response and confounding) there are simply no relevant data or theory from which to develop a precise prior for η_{M}. Even if that information were available, uncertainty that is due to non-response is comparable with uncertainty that is due to random error, and so the overall uncertainty would remain high even if enormous studies with perfect measurements (which will never exist) were added to the analysis. The only positive note is that confounding alone seems to be of lesser importance than other biases, given the prior information that is used here.

Finally, before the earliest studies little credibility was given to the hypothesis that residential fields cause leukaemia. Thus, if we added a substantively justified prior for θ to the specification, the final distributions would be shifted towards the null because such a prior is concentrated near the null (Greenland, 2003b).

4. Discussion

4.1. Model forms

As in most conventional analyses, I have not addressed uncertainty about model forms; I used log-linear and logistic models with normal random effects only for tractability and to enforce range restrictions. This source of uncertainty could be included by adding parameters to index the model form (as is done in Bayesian model averaging), although that would greatly increase the complexity of the bias model and the priors.

In my experience, many users of statistics think that their results do not depend on the form of their model because they use categorized variables or purely tabular analyses. None-the-less, the justification and performance of categorical and tabular statistics depend on implicit regression models; for example, the tabular Cochran–Mantel trend test that is popular in epidemiology is the score test of the slope parameter in a logistic model with binomial errors, and it can be quite misleading when that model form poorly approximates reality (Maclure and Greenland, 1992). Such issues are ordinarily addressed by model diagnostics; on expansion to include bias parameters, however, the model forms (as well as their parameters) are not identified. Thus, in bias modelling the model form is an integral component of the prior specification, rather than a structure that is selected with guidance from the data, as most statistical research treats it.

4.2. Some problems in interpretation

Analysts sometimes conclude that the combination of bias and random error is sufficient to explain an elevated estimate if a plausible value or distribution of η could by itself produce a value that is as high as the conventional lower confidence limit; similarly, some analysts call inference about the null sensitive to hidden bias if a plausible value for η could make the two-sided $P=0.05$. These interpretations are misleading because they do not coherently integrate the uncertainties regarding bias and random error. In particular, they suggest that the bias prior makes the null more probable than it actually does. Consider Table 2: the correct probability (under the prior) that the combination of bias and random error equal or exceed the observed association (ω_{MH0}) is the ‘proportion < 1’ in the ‘all plus random error’ row. This probability is always much smaller than the corresponding probability that bias alone could have produced an elevation that is as high or higher than the conventional lower limit (the ‘proportion < 1.27’ in the ‘all bias sources’ row). Given a positive observed association, the use of the lower limit or $P=0.05$ as the criterion for evaluating bias sensitivity implicitly assumes that the random error is improbably positive (at least as positive as the difference between the point estimate and the lower limit, an event with only 2.5% probability). These criteria are thus biased in favour of the null hypothesis.

Other analysts report percentiles from MCSA as frequentist statistics; for example, in summarizing the distribution of corrected estimates, they may present the percentage below the null value as a one-sided lower P-value and refer to the 2.5th and 97.5th percentiles as 95% confidence limits. None-the-less, because the distributions of these summaries have a heavily subjective prior component $p(\eta)$, conventional frequentist interpretations are unjustified (Greenland, 2001a).

4.3. Metasensitivity and objections to bias modelling

Bayesian and MCSA outputs depend completely on the prior $p(\eta)$, which suggests that a metasensitivity analysis of the dependence is essential. Moving in this direction reintroduces the problem of basic sensitivity analysis, however: given the limitless possibilities for $p(\eta)$, a thorough metasensitivity analysis would only illustrate how various conclusions can be reached. A conclusion about the target θ, however, would require constraints on the $p(\eta)$. These constraints would constitute a subjective prior on priors (a hyperprior); incorporating them into the analysis would produce a subjective average of results over the hyperprior, as in the final block of Table 2. This result would itself be subject to concerns about sensitivity to the hyperprior, which would continue on into an infinite regress.

This regress is as unnecessary as it is impractical. Multiple-bias modelling can be treated as a project to discover and exhibit a prior that is arguably reasonable or defensible (in that it is consistent with known facts and established theory), and that leads to borderline conclusive results according to some operative criterion (e.g. a posterior probability for the null of 0.025) (Greenland, 2003a). Such a prior can help to show why the data cannot force agreement between all reasonable observers: defensible perturbations to such a prior can make the results appear moderately inconclusive or moderately conclusive, as the results in Table 2 do when evaluated against a two-sided 0.05-criterion (a criterion that is used in laws and precedents in the USA; see Greenland (2001a)). As an example, in the year following the publication of Greenland, Sheppard, Kaune, Poole and Kelsh (2000) one official of the California State Department of Health publicly asserted, with near certainty, that fields caused childhood leukaemia; this assertion fed demands on the Public Utilities Commission to impose very costly interventions to reduce field levels at schools and homes. Multiple-bias modelling can counterbalance such overconfident assessments of ambiguous evidence and provide more realistic inputs for decision makers (whose decisions will be guided by cost–benefit as well as evidential concerns).

The unlimited nature of metasensitivity may cause some to label bias modelling as a futile exercise. These objections correctly note that, for most topics in which bias modelling might be worthwhile, it would only show how all of an observed association can be plausibly attributed to bias and random error. This objection is no fault of bias modelling, however, but it instead reflects the weakness of available evidence. The demonstration of this weakness is worthwhile if not imperative in many cases, as above.

Metasensitivity has also led to charges that the quantification of uncertainty that is achieved under bias modelling is spurious. There is, however, nothing spurious about the quantification if the prior approximates the views of the analyst, for then the output gives the analyst an idea what his or her posterior bets about the value of θ should be. From a more broad perspective, charges of spurious precision embody a double standard relative to the *status quo*: the apparently precise quantification of uncertainty that is offered by conventional analysis is far more spurious than that from bias modelling. Within health sciences, at least, I believe that most researchers fail to grasp how poorly conventional analyses capture uncertainty, and they fail to compensate sufficiently for these deficits. Intuitive discussions of bias often rely on flawed heuristics, such as ‘non-differential misclassification introduces a bias toward the null in virtually every study’ (Rothman (1986), page 88). Such flawed heuristics ignore the effects of bias uncertainty, effects which are revealed by an exercise in bias modelling (see Section 3.8). A recent sample survey of the epidemiologic literature revealed that most papers do not even deploy flawed heuristics but instead dismiss biases as unlikely to be important, or else simply fail to mention the problems (Jurek *et al.*, 2004). In the rare case that sensitivity analysis is added, it is almost never coherently combined with the assessment of random error.

Another objection is that possible biases are always omitted from modelling. That is true, but the inevitable omission of some bias sources cannot justify the omission of all (which is what conventional analyses do) any more than the inevitable failure to apprehend all murderers can justify ignoring all the murderers who can be apprehended. The inevitability of omissions does suggest that no analysis can do more than to provide a lower bound on the uncertainty that we should have in light of the data and a prior. None-the-less, bias modelling can provide less misleadingly optimistic bounds than can conventional analysis.

4.4. Bias analysis versus better data?

Some recommend that formal bias analysis should be eschewed in favour of improving measurement, response and covariate data. This recommendation is a *non sequitur* and is often wildly impractical. Bias modelling and collection of better data are not mutually exclusive, although bias modelling is often the only feasible option. Exhortations 'just to collect better data' are especially empty when (as in the example) we can neither identify a gold standard measurement nor force subjects to participate or to submit to better measurements (which tax co-operation of subjects). Even when we can envision a way to collect better data, decisions must often be made immediately and so can only be based on *currently available* data; as in the example, it may be essential to model those data fully to counter naïve judgments.

'Collect better data' becomes a relevant slogan when it is feasible to do so. Multiple-bias modelling is then a useful ally in making clear that the added value of more observations of previous quality (e.g. case–control studies with unknown and possibly large amounts of bias) is much less than conventional statistical formulae convey (Eddy *et al.*, 1992). Conventional standard errors shrink to 0 as the number of observations increases, and total uncertainty approaches the combined bias uncertainty. At some point, mere replication or enlargement of observational studies is not cost effective, and innovations to reduce bias are essential. This point is passed when random error contributes a minority share to total uncertainty. In the example, three more studies of magnetic fields and childhood leukaemia have been published since completion of the pooling project, but none controlled the biases that are described above. Hence, adding these studies has little effect on the final uncertainty distributions; in fact, adding a study with no random error (infinite sample size) but the same bias uncertainty would have little effect.

Most would agree that proposals to confirm or test previous results should include effective safeguards to reduce sources of bias that were suspected in earlier findings, or at least should supply validation data that could provide usefully precise estimates of bias parameters. But the cost of such improvements may be prohibitive. Decisions about funding should also involve considerations of research value (Eddy *et al.*, 1992); the high cost of doing a very informative study may not justify the usual claim that 'more research is needed' (Phillips, 2001). When the cost of better data is prohibitive, multiple-bias analysis of existing data may become the best feasible option.

4.5. Concluding remarks

Extreme sensitivity of results to the priors is inevitable and unsurprising, given the many unidentified parameters in realistic bias models. It reflects an irreducible core of uncertainty about the mechanisms generating non-experimental observations (Leamer, 1978; Rubin, 1983). Unfortunately, this core uncertainty is hidden by adherence to identified models. It is more honest instead to bring uncertainty to the fore and to attempt to discover which parameters contribute most to the final uncertainty. Such discovery can help to guide research planning by focusing resources on reducing the largest sources of uncertainty. Those sources are not necessarily the largest sources of bias, but rather are the sources that are most poorly determined by prior information.

Compared with conventional analysis, multiple-bias modelling better captures uncertainty about effects but requires the specification of a much larger model and demands far more subject-matter knowledge. It also requires much more presentation space and more effort by the reader. Its key advantages may only make it more unappealing: if conducted and presented properly, it depicts how, in the absence of experimental evidence, effects of interest are identified by prior distributions for bias sources rather than by data. It thus belies methods that claim to 'let the data speak for themselves': without external inputs, observational data say nothing at all about causal effects. In many settings it also shows that only indefensibly precise (overconfident) priors can produce firm conclusions, and that conventional methods produce definitive looking results only because they assign probability 1 to the extremely implausible assumption of no bias ($\eta = 0$).

Multiple-bias modelling can be superfluous when conventional standard errors make clear that substantive inferences are unwarranted, as when only a few small studies are available. It may, however, be essential when an analysis purports to draw causal inferences from observational data, when bias uncertainty is comparable with random error or when decisions with costly consequences must be made on the basis of the available evidence (Eddy *et al.*, 1992). I thus argue that bias modelling should become part of the core training of scientists and statisticians who are entrusted with the analysis of observational data. For research planning and allocation, multiple-bias modelling can show when conventional approaches to reducing uncertainty, such as increasing the sample size or replicating studies, have become inefficient (in the magnetic field controversy, this point was reached with studies published in the mid-1990s). To be worthwhile after that point, further studies must give estimates that are more precise and unbiased than previous estimates or else must give precise estimates of biases. When improved studies are prohibitively expensive, multiple-bias modelling may be the best option for decision-making input.

Acknowledgements

The author is grateful to David Draper, W. Dana Flanders, Patrick Graham, Katherine J. Hoggatt, Tim Lash, George Maldonado, the Joint Editor and the referees for numerous helpful comments, and to Michinori Kabuto for translating results from his study. This research was supported by the Electric Power Research Institute and by grant 1R29-ES07986 from the National Institute of Environmental Health Sciences.

Appendix A

'Unconfounded' and 'confounding' have been formalized in various ways (Greenland *et al.*, 1999); for the present exposition the precise definition is unimportant as long as it implies that there is a U such that the true causal effect of T on Y can be identified from $P(T, Y|S, U)$. This U may be a compound of other variables. Existence can be shown under various causal models. For example, under a directed acyclic graphical model for causation, all confounding can be traced to common causes of T and Y, and hence such a U will exist if (as here) S is unaffected by T or Y (Pearl (2000), chapter 6). Existence is also guaranteed under a potential outcome model for the effect of T on Y, for U can then be taken as the potential outcome vector (Frangakis and Rubin, 2002). In the present analysis, with binary T, any sufficient multidimensional U can be reduced to a sufficient univariate summary; for example, the propensity score $P(T=1|S, U)$ is such a summary. This score is usually categorized and five levels are often deemed adequate (Rosenbaum, 2002); if the range of the score is very restricted or the relationship of (S, U) to T or Y is weak, fewer levels may be needed, although more may be needed if the relationship of (S, U) to T and Y is very strong. Note that, under a deterministic monotone effect model for a binary Y, the potential outcome vector $U=(Y_1, Y_0)$ has only three possible levels: (0,0), (1,1) and at most one of (1,0) or (0,1) (Angrist *et al.*, 1996).

References

Angrist, J., Imbens, G. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Ass.*, **91**, 444–472.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: Massachusetts Institute of Technology Press.

Bracken, M. B., Belanger, K., Hellebrand, K., Adesso, K., Patel, S., Trich, E. and Leaderer, B. (1998) Correlates of residential wiring configurations. *Am. J. Epidem.*, **148**, 467–474.

Brain, J. D., Kavet, R., McCormick, D. L., Poole, C., Silverman, L. B., Smith, T. J., Valberg, P. A., Van Etten, R. A. and Weaver, J. C. (2003) Childhood leukemia: electric and magnetic fields as possible risk factors. *Environ. Hlth Perspect.*, **111**, 962–970.

Breslow, N. E. (1981) Odds ratio estimators when the data are sparse. *Biometrika*, **68**, 73–84.

Coghill, R. W., Steward, J. and Philips, A. (1996) Extra low frequency electric and magnetic fields in the bedplace of children diagnosed with leukemia: a case-control study. *Eur. J. Cancer Prevn*, **5**, 153–158.

Copas, J. B. (1999) What works?: selectivity models and meta-analysis. *J. R. Statist. Soc.* A, **162**, 95–109.

Copas, J. B. and Li, H. G. (1997) Inference for non-random samples (with discussion). *J. R. Statist. Soc.* B, **59**, 55–95.

Cornfield, J., Haenszel, W., Hammond, W. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natn. Cancer Inst.*, **22**, 173–203.

Crouch, A. C., Lester, R. R., Lash, T. L., Armstrong, S. R. and Green, L. C. (1997) Health risk assessment prepared per the risk assessment reforms under consideration in the U.S. Congress. *Hum. Ecol. Risk Assessmnt*, **3**, 713–785.

Dockerty, J. D., Elwood, J. M., Skegg, D. C. G. and Herbison, G. P. (1998) Electromagnetic field exposures and childhood cancers in New Zealand. *Cancer Causes Contr.*, **9**, 299–309; erratum **10** (1999), 641.

Draper, D., Saltelli, A., Tarantola, S. and Prado, P. (2000) Scenario and parametric sensitivity and uncertainty analyses in nuclear waste disposal risk assessment: the case of GESAMAC. In *Mathematical and Statistical Methods for Sensitivity Analysis* (eds A. Saltelli, K. Chan and M. Scott), ch. 13, pp. 275–292. New York: Wiley.

Eddy, D. M., Hasselblad, V. and Schachter, R. (1992) *Meta-analysis by the Confidence Profile Method*. New York: Academic Press.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Electric Power Research Institute (2003) Selection bias in epidemiologic studies of EMF and childhood leukemia. *EPRI Report 1008149*. World Health Organization, Geneva.

Feychting, M. and Ahlbom, A. (1993) Magnetic fields and cancer in children residing near Swedish high-voltage power lines. *Am. J. Epidem.*, **138**, 467–481.

Flegal, K. M., Keyl, P. M. and Nieto, F. J. (1991) Differential misclassification arising from nondifferential errors in exposure measurement. *Am. J. Epidem.*, **134**, 1233–1244.

Frangakis, C. and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics*, **58**, 21–29.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd edn. New York: Chapman and Hall–CRC.

Good, I. J. (1983) *Good Thinking*. Minneapolis: University of Minnesota Press.

Graham, P. (2000) Bayesian inference for a generalized population attributable fraction. *Statist. Med.*, **19**, 937–956.

Green, L. M., Miller, A. B., Villeneuve, P. J., Agnew, D. A., Greenberg, M. L., Li, J. and Donnelly, K. E. (1999) A case-control study of childhood leukemia in southern Ontario, Canada, and exposure to magnetic fields in residences. *Int. J. Cancer*, **82**, 161–170.

Greenland, S. (1996) Basic methods for sensitivity analysis of bias. *Int. J. Epidem.*, **25**, 1107–1116.

Greenland, S. (1998) The sensitivity of a sensitivity analysis. *Proc. Biometr. Sect. Am. Statist. Ass.*, 19–21.

Greenland, S. (2001a) Sensitivity analysis, Monte-Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Anal.*, **21**, 579–583.

Greenland, S. (2001b) Putting background information about relative risks into conjugate priors. *Biometrics*, **57**, 663–670.

Greenland, S. (2003a) The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia. *J. Am. Statist. Ass.*, **98**, 47–54.

Greenland, S. (2003b) Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics*, **59**, 92–99.

Greenland, S. (2004a) Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int. J. Epidem.*, **33**, in the press.

Greenland, S. (2004b) Smoothing epidemiologic data. In *Encyclopedia of Biostatistics*, 2nd edn (eds P. Armitage and T. Colton). New York: Wiley.

Greenland, S., Pearl, J. and Robins, J. M. (1999) Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.

Greenland, S. and Robins, J. M. (1985) Confounding and misclassification. *Am. J. Epidem.*, **122**, 495–506.

Greenland, S., Schwartzbaum, J. A. and Finkle, W. D. (2000) Problems from small samples and sparse data in conditional logistic regression analysis. *Am. J. Epidem.*, **151**, 531–539.

Greenland, S., Sheppard, A. R., Kaune, W. T., Poole, C. and Kelsh, M. A. (2000) A pooled analysis of magnetic fields, wire codes, and childhood leukemia. *Epidemiology*, **11**, 624–634.

Gustafson, P. (2003) *Measurement Error and Misclassification in Statistics and Epidemiology*. New York: Chapman and Hall.

Hatch, E. E., Kleinerman, R. A., Linet, M. S., Tarone, R. E., Kaune, W. T., Auvinen, A., Baris, D., Robison, L. L. and Wacholder, S. (2000) Do confounding or selection factors of residential wire codes and magnetic fields distort findings of electromagnetic fields studies? *Epidemiology*, **11**, 189–198.

Jurek, A. M., Maldonado, G., Greenland, S. and Church, T. R. (2004) Exposure-measurement error is frequently ignored when interpreting epidemiologic study results (abstract). *Am. J. Epidem.*, **159**, S72.

Kabuto, M. (2003) A study on environmental EMF and children's health: final report of a grant-in-aid for scientific research project, 1999-2001 (in Japanese). *Report*. Japanese Ministry of Education, Culture, Sports, Science and Technology, Tokyo.

Kavet, R. and Zaffanella, L. E. (2002) Contact voltage measured in residences: implications for the association between magnetic fields and childhood leukemia. *Bioelectromagnetics*, **23**, 464–474.

Langholz, B. (2001) Factors that explain the power line configuration wiring code–childhood leukemia association: what would they look like (with discussion)? *Bioelectromagn. Suppl.*, **5**, S19–S31.

Lash, T. L. and Fink, A. K. (2003) Semi-automated sensitivity analysis to assess systematic errors in observational epidemiologic data. *Epidemiology*, **14**, 451–458.

Lash, T. L. and Silliman, R. A. (2000) A sensitivity analysis to separate bias due to confounding from bias due to predicting misclassification by a variable that does both. *Epidemiology*, **11**, 544–549.

Leamer, E. E. (1974) False models and post-data model construction. *J. Am. Statist. Ass.*, **69**, 122–131.

Leamer, E. E. (1978) *Specification Searches*. New York: Wiley.

Linet, M. S., Hatch, E. E., Kleinermann, R. A., Robison, L. C., Kaune, W. T., Friedman, D. R., Severson, R. K., Hainer, C. M., Hartsoak, C. T., Niwa, S., Wacholder, S. and Tarone, R. E. (1997) Residential exposure to magnetic fields and acute lymphoblastic leukemia in children. *New Engl. J. Med.*, **337**, 1–7.

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

London, S. J., Thomas, D. C., Bowman, J. D., Sobel, E., Cheng, T.-C. and Peters, J. M. (1991) Exposure to residential electric and magnetic fields and risk of childhood leukemia. *Am. J. Epidem.*, **134**, 923–937.

Maclure, M. and Greenland, S. (1992) Tests for trend and dose-response: misinterpretations and alternatives. *Am. J. Epidem.*, **135**, 96–104.

Maclure, M. and Schneeweiss, S. (2001) Causation of bias: the episcope. *Epidemiology*, **12**, 114–122.

McBride, M. L., Gallagher, R. P., Theriault, H. G., Armstrong, B. G., Tamaro, S., Spinelli, J. J., Deadman, J. E., Fincham, S., Robson, D. and Choi, W. (1999) Power-frequency electric and magnetic fields and risk of childhood cancer. *Am. J. Epidem.*, **149**, 831–842.

Michaelis, J., Schüz, J., Meinert, R., Semann, E., Grigat, J. P., Kaatsch, P., Kaletsh, U., Miesner, A., Brinkmann, K., Kalkner, W. and Kärner, H. (1998) Combined risk estimates for two German population-based case-control studies on residential magnetic fields and childhood leukemia. *Epidemiology*, **9**, 92–94.

Morgan, M. G. and Henrion, M. (1990) *Uncertainty*. New York: Cambridge University Press.

Mosteller, F. and Tukey, J. W. (1977) *Data Analysis and Regression*. New York: Addison-Wesley.

Olsen, J. H., Nielsen, A. and Schulgen, G. (1993) Residence near high voltage facilities and risk of cancer in children. *Br. Med. J.*, **307**, 891–895.

Pearl, J. (2000) *Causality*. New York: Cambridge University Press.

Phillips, C. V. (2001) The economics of "more research is needed". *Int. J. Epidem.*, **30**, 771–776.

Phillips, C. V. (2003) Quantifying and reporting uncertainty from systematic errors. *Epidemiology*, **14**, 459–466.

Poole, C. and Greenland, S. (1997) How a court accepted a possible explanation. *Am. Statistn*, **51**, 112–114.

Powell, M., Ebel, E. and Schlossel, W. (2001) Considering uncertainty in comparing the burden of illness due to foodborne microbial pathogens. *Int. J. Food Microbiol.*, **69**, 209–215.

Robins, J. M., Rotnitzky, A. and Scharfstein, D. O. (1999) Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology* (eds M. E. Halloran and D. A. Berry), pp. 1–92. New York: Springer.

Rosenbaum, P. (2002) *Observational Studies*, 2nd edn. New York: Springer.

Rothman, K. J. (1986) *Modern Epidemiology*. Boston: Little, Brown.

Rothman, K. J. and Greenland, S. (1998) *Modern Epidemiology*, 2nd edn. Philadelphia: Lippincott.

Rubin, D. B. (1983) A case study of the robustness of Bayesian methods of inference. In *Scientific Inference, Data Analysis, and Robustness* (eds G. E. P. Box, T. Leonard and C. F. Wu), pp. 213–244. New York: Academic Press.

Savitz, D. A., Wachtel, H., Barnes, F. A., John, E. M. and Tvrdik, J. G. (1988) Case-control study of childhood cancer and exposure to 60-Hz magnetic fields. *Am. J. Epidem.*, **128**, 21–38.

Schüz, J., Grigat, J. P., Brinkmann, K. and Michaelis, J. (2001) Residential magnetic fields as a risk factor for acute childhood leukemia: results from a German population-based case-control study. *Int. J. Cancer*, **91**, 728–735.
Steenland, K. and Greenland, S. (2004) Monte-Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *Am. J. Epidem.*, **160**, 384–392.
Stigler, S. M. (1986) *The History of Statistics*. London: Belknap.
Tomenius, L. (1986) 50-Hz electromagnetic environment and the incidence of childhood tumors in Stockholm County. *Bioelectromagnetics*, **7**, 191–207.
Tynes, T. and Haldorsen, T. (1997) Electromagnetic fields and cancer in children residing near Norwegian high-voltage power lines. *Am. J. Epidem.*, **145**, 219–226.
UK Childhood Cancer Study Investigators (1999) Exposure to power-frequency magnetic fields and the risk of childhood cancer. *Lancet*, **354**, 1925–1931.
Verkasalo, P. K., Pukkala, E., Hongisto, M. Y., Valjus, J. E., Järvinen, P. J., Heikkilä, K. K. and Koskenvuo, M. (1993) Risk of cancer in Finnish children living close to power lines. *Br. Med. J.*, **307**, 895–899.
Vose, D. (2000) *Risk Analysis*. New York: Wiley.
Wacholder, S., Armstrong, B. and Hartge, P. (1993) Validation studies using an alloyed gold standard. *Am. J. Epidem.*, **137**, 1251–1258.
Yanagawa, T. (1984) Case-control studies: assessing the effect of a confounding factor. *Biometrika*, **71**, 191–194.

Discussion on the paper by Greenland

John Copas (*University of Warwick, Coventry*)
It is a pleasure to welcome Professor Greenland to the Society and to propose the vote of thanks for his interesting paper. Most of the work that is reviewed in the paper is published in the epidemiological literature, which is not widely read by statisticians working in other areas. But the central problems of response bias, measurement error and confounding rear their ugly heads in many and probably most applications of statistics, not just in the areas that are usually associated with epidemiology. This is therefore an important topic for all of us.

We are so used to using conventional statistical methods which convert information about a sample S into a conclusion C about the population that we all too easily forget the essential impossibility of arguing from the particular to the general. Such induction is only possible if we make assumptions A, i.e. statistical inference is $(S, A) \rightarrow C$ and not $S \rightarrow C$. Fisher was the first to show that, if we can choose how to obtain S, then we can do so in such a way that A is self-evident, in the sense that the randomization that we have actually used to obtain S also gives us the probability space from which we can obtain C. If everyone agrees on the truth of A then there is no need to mention it explicitly. But in all other cases honesty requires us to emphasize that C depends on A as well as on S, and failure to do so is an abuse of our subject. Every day the media invite us to believe claims like 'studies show that if you eat Corn Flakes for breakfast you are twice as likely to . . .'. No doubt S is observational, and A is some absurd assumption of randomization. If A is not mentioned how can we assess it? If the conclusion is unbelievable, then what is discredited is not A, as it should be, but statistics, and hence statisticians. I welcome Professor Greenland's paper for his clear discussion of these issues.

The late George Barnard, in one of these discussion meetings, once remarked 'We statisticians spend too much time trying to find sensible answers to silly questions'. What question can we ask from Table 2? We see that the upper ends of the intervals for the odds ratio θ vary very widely across the different settings for (m_0, m_1), but from Professor Greenland's discussion there seems no very clear reason for preferring any one setting over any other. So, if the question is 'how big is the risk', then perhaps we should follow Barnard and answer 'the quality of the data is not sufficiently good for us to make any sensible estimate'. But, if the question is 'do we have clear evidence that there is a risk', we note the remarkable finding that the lower ends of the intervals are all near the null value, even including the setting $m_0 = m_1 = 0$ if we take this from the 'response-bias-only' figures. If we can show that this happens for all reasonable attempts to model these biases, then we have a non-silly question, and the answer is no. This would be an important and perhaps the only convincing analysis of these data.

Professor Greenland argues that sources of bias should be modelled simultaneously and not one by one, as is usually done. Approximating this in terms of sequential transformations on the expected cell counts is an attractive simplification, both conceptually and computationally. So it is a pity that the example does not demonstrate the force of his argument, at least as far as bias is concerned. If we think of each source of bias as adjusting the estimate up or down by a certain factor, and assume that these factors are additive on the log-scale, then we can use the 50% points in the single bias rows of Table 2 to predict the 50% points for the combined bias rows. This gives adjusted values that are very close to the 50% points calculated for the multiple-bias models.

For a sensitivity analysis to be useful, it is surely necessary that the assumptions which drive the different conclusions are sufficiently transparent that they can be communicated. Even to a statistical audience, Professor Greenland's bias models have taken several pages to explain and involve various approximations whose validity is not always clear, at least to me. Any attempt to model these bias processes explicitly is bound to be complicated, contentious and arbitrary. Is there a simpler approach?

My suggestion is to turn the problem back to front and first to ask whether we can define a parameter η which is interpretable in the context of the data, and which, in some sense, reflects the scope of the biasing process. For example, in assessing publication bias, we could define η to be the unknown number of unpublished papers. If we can formulate a suitable η then, at least in principle, we can replace the explicit modelling by a 'worst case' argument. Subject-matter knowledge is still essential, as Professor Greenland emphasizes. But, instead of high dimensional priors on nuisance parameters, we could rely on informed judgment about plausible values of η: hence the need for η to be interpretable.

The naïve analysis is tantamount to fitting a model $f(x, y, s, z; \theta)$, where z is an additional random variable defined by the kind of bias that we are considering. For the three sources of bias discussed, z would be a binary response indicator, a true exposure, or a latent confounder. What we observe is a sample from the derived distribution $f_{\text{OBS}}(x, y, s; \theta)$, the naïve ignorability assumption in f ensuring that θ is identifiable. For response bias f_{OBS} is the conditional distribution given $z = 1$ or is simply the marginal distribution of the data in the other two cases.

The distribution f_{OBS} defines a score function $u(x, y, s; \theta)$. However, the ignorability assumptions behind f_{OBS} are almost certainly false, so we imagine that the real distribution is $g(x, y, s)$, say. Our chosen η is now a functional of $g, \eta(g)$, say. Then the bias bounds are

$$\theta^-_\eta = \inf_g [\theta : E_g\{u(x, y, s; \theta)\} = 0, \eta(g) = \eta], \tag{14}$$

$$\theta^+_\eta = \sup_g [\theta : E_g\{u(x, y, s; \theta)\} = 0, \eta(g) = \eta]. \tag{15}$$

Geometrically, we are holding $\eta(g)$ fixed and finding the extreme projections of g onto f. For the sensitivity analysis we just estimate θ from the crude analysis and plot these quantities against η.

This approach is explored in two recent references: Copas and Jackson (2004) for publication bias and Copas and Eguchi (2001) for missing data and confounding. It would be interesting to see whether this approach can be extended to cover the more complicated set-up which we have here.

This paper has raised some very challenging questions, with an impressive discussion of a major issue in public health. It gives me great pleasure to propose the vote of thanks.

David R. Jones (*University of Leicester*)
This is a very stimulating paper about the very important (and relatively neglected) issue of dealing with bias in analysis and interpretation of observational data. This problem of bias in observational studies, and almost inadequate analytical responses to it, has a long history. With some imagination, Shakespeare can be held to have made several arguably relevant comments (although admittedly in different contexts), including '... with assays of bias, by indirections find directions out', and 'Bias and thwart, not answering the aim' (Crystal and Crystal, 2002). Much more directly, this paper draws on Eddy's confidence profile ideas of bias modelling (Eddy *et al.*, 1992), seeking to account explicitly for all major sources of uncertainty, as in risk assessment and comprehensive decision analysis. The exposition of the problem is clear here, and the two-step characterization of the conventional (epidemiological) approach to analysis, and the selection bias that is often inherent in implementation of Section 1.1, point (b), will be a skewer through many epidemiologists' hearts.

The author argues that multiple-bias modelling should be part of the core training of analysts of observational data. I agree of course that major improvements in approach are needed, but questions about the feasibility of such training are focused for me by the introductory Masters course in statistical methods for epidemiology that I shall be teaching next week. Our students should (I think!) be able to cope with the more technical statistical aspects. However, although this paper—in particular the example—demonstrates the potential for fuller, quantified analyses of bias in capable hands, it does not give a general, operational specification. Not all epidemiological hands are as capable across the wide range of statistical confidence and understanding of the application area that is required by the approach.

Perhaps in recognition of this, at the end of Section 1, readers who are 'uninterested in details' are encouraged to skim the theory (Section 2) and the magnetic fields and leukaemia example (Section 3). I suspect that many epidemiologists would indeed skip (not just skim) the theory and start with the example.

So, can we move in any way towards wider implementation? In particular, how can the example be generalized in practice? The specification of sensitivity analyses and/or priors is bespoke; how to decide what is 'major' in 'accounting for all major sources of uncertainty' is demonstrated in the example but not in general. In Section 2.2 the potential which sensitivity analyses offer for nihilism, through finding an η that yields any preselected value for θ, is indicated. As manipulation may indeed not be obvious when there are multiple-bias sources, the approach is not sufficiently transparent to allay such concerns in all applications.

I suggest that we could start to extend the paper by seeking to draw up guidelines for *minimal* good practice. These might include attempts to parallel the use of sceptical and enthusiastic priors in trial contexts (Spiegelhalter *et al.*, 1994). Although their derivation in an observational study context will be less straightforward, it could usefully draw on another possible feature of the guidelines: presentation of a systematic review of evidence relating to distribution of η in other studies as the basis of a more transparent indication of the source of the analyst's prior, so as partially to prevent the covert selectivity of approach that is outlined in Section 1.1, point (b). This systematic review will not necessarily include the same studies as a review of outcomes. An adequate analysis in any particular example will need to move beyond the inevitably arbitrary guideline specifications; the guidelines are proposed simply as a starting-point for relatively inexperienced analysts.

As the author indicates, adoption of his approach will require greater resources for both analysis and presentation, though the latter should be relatively easily accommodated by supplementary use of the Web. There are also hints towards expected value of information approaches in determining the next step in research in the area of interest (Claxton, 1999). The message that ignoring bias uncertainty in conventional power calculations overestimates benefits of study replication is anyway important for research prioritization decisions even if less formal methods are used. Multiple-bias modelling, or performing another conventional epidemiological study, will not always prove to be the best buy; sometimes more radical choices, such as invoking Mendelian randomization, using genetic markers of exposure propensity to 'eliminate' confounding (Davey Smith and Ebrahim, 2003), will be possible and preferable.

In summary, this paper does not of course solve all the problems of bias in the analysis and inference of observational studies; in particular it is not easy to see how to apply the approach generally. However, it does provide a stimulating starting-point for a very necessary development of better future practice. With further apologies to Shakespeare I could say that I come not to 'bury' Sander but to praise him, and with great pleasure I second the vote of thanks.

The vote of thanks was passed by acclamation.

David Spiegelhalter (*Medical Research Council Biostatistics Unit, Cambridge*)

This paper reveals an admirable effort to broaden sensitivity analysis beyond its usual unrealistically limited boundaries. The author appears keen to avoid a full likelihood or Bayesian approach, presumably with the feeling that epidemiologists will feel more comfortable with a sequence of adjustments to the standard Mantel–Haenszel point estimates. But I wonder whether perhaps a 'full probability model' would be both conceptually and computationally simpler.

Fig. 1 shows a graphical model constructed in the 'causal' (in a loose sense) direction, revealing how an underlying 'true' odds ratio of primary interest is steadily modified by problems of measurement and design, until finally giving rise to the observed counts.

Each of the model assumptions that are assumed by the author can be placed within this framework: the only additional component is a random-study effect instead of the conditional Mantel–Haenszel analysis.

Such a model appears conceptually straightforward and moreover is readily implementable in freely available software such as WinBUGS: my colleague Ken Rice will report the considerable insights that are revealed by using such a model to reanalyse the author's example. A full probability model allows access to the whole range of likelihood-based techniques for model criticism and comparison, implementation of proper prior distributions and appropriate interval estimates with full allowance for parameter uncertainty.

In many areas of statistics there are arguments between those who prefer full probability models (perhaps labelled 'generative' or 'causal'), with those who prefer statistical procedures applied directly to data (e.g. generalized estimating equations, robust methods and classification and regression trees). In technological problems such as machine learning, with defined objectives of prediction or classification, procedural approaches may outperform models and be preferable: see, for example, Breiman (2001) for an excellent

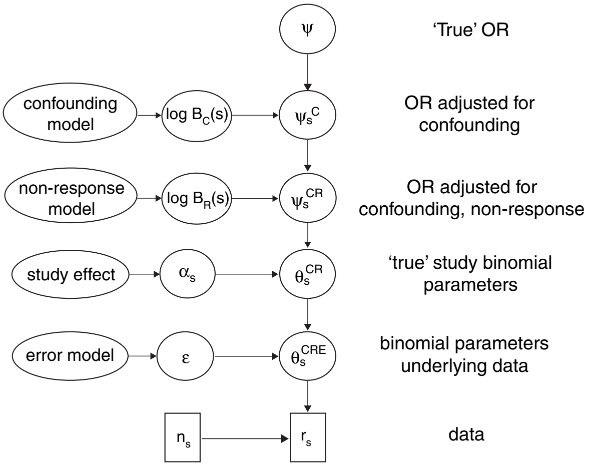
ψ
'True' OR
confounding model
log B_C(s)
ψ_s^C
OR adjusted for confounding
non-response model
log B_R(s)
ψ_s^CR
OR adjusted for confounding, non-response
study effect
α_s
θ_s^CR
'true' study binomial parameters
error model
ε
θ_s^CRE
binomial parameters underlying data
n_s
r_s
data

Fig. 1. Graphical model showing how an underlying true odds ratio is modified by the processes that were identified by Greenland

discussion of these issues. But I would question whether this necessarily holds in epidemiology, where we are presumably trying to gain understanding of the effects of whatever underlying process generated the observed data. I would therefore welcome the author's opinion on the role of full probability models in epidemiology.

Kenneth Rice (*Medical Research Council Biostatistics Unit, Cambridge*)
The author expressly aims to approximate a Bayesian analysis optimally. However, by carefully carrying out a full Bayesian analysis, pertinent analytical issues can be identified and addressed explicitly. My comments concern clarification of how the paper's method implicitly deals with such issues.

Suppose, for simplicity, that we assume only classification error, at fixed rates ε_0 and ε_1. Then a reasonable-looking Bayesian model for the author's example is

$$X_{ys} \sim \text{binomial}(n_{ys}, \theta_{ys}), \qquad y=0,1,\ s=1,\ldots,14,$$

$$\theta_{ys} = \varepsilon_0\theta^*_{ys} + (1-\varepsilon_1)(1-\theta^*_{ys}), \qquad y=0,1,\ s=1,\ldots,14,$$

$$\text{logit}(\theta^*_{ys}) = \mu_s + y\log(\psi), \qquad y=0,1,\ s=1,\ldots,14,$$

$$\text{expit}(\mu_s) \sim U(0,1), \qquad s=1,\ldots,14,$$

$$\log(\psi) \sim U(-3,3).$$

The only extra assumptions are of a flat prior on $\log(\psi)$, covering reasonable values, and independent flat priors on the 'true' exposure probability for controls.

This choice of 'base-line' seems innocuous, but one might reasonably have used flat priors on the case exposure probabilities, alternatively defining

$$\text{logit}(\theta^*_{ys}) = \mu_s + (y-1)\log(\psi), \qquad y=0,1,\ s=1,\ldots,14,$$

above. In Fig. 2 we see that this trivial change can have a massive effect on the posterior. An explanation follows, but the choice of prior clearly matters. What choice of prior is the author aiming to approximate?

(The imputed priors on 'non-base-line' exposure probabilities have peaks near 0 and 1. Also, for large negative values of μ_s, optimal support for $\theta^*_{0s} = \theta^*_{1s} = 0$ comes with large negative value of $\log(\psi)$. The two parameterizations bias the analysis towards this extreme by different amounts. We could instead find a compromise between them (Smith *et al.*, 1995), but more theoretically appealling resolutions exist (Rice, 2004).)

The prior on ψ is a related concern. Given such rare exposures, if we entertain misclassification rates of 5% (suggested by Sections 3.5 and 3.8) then the observed data give non-trivial support to underlying true

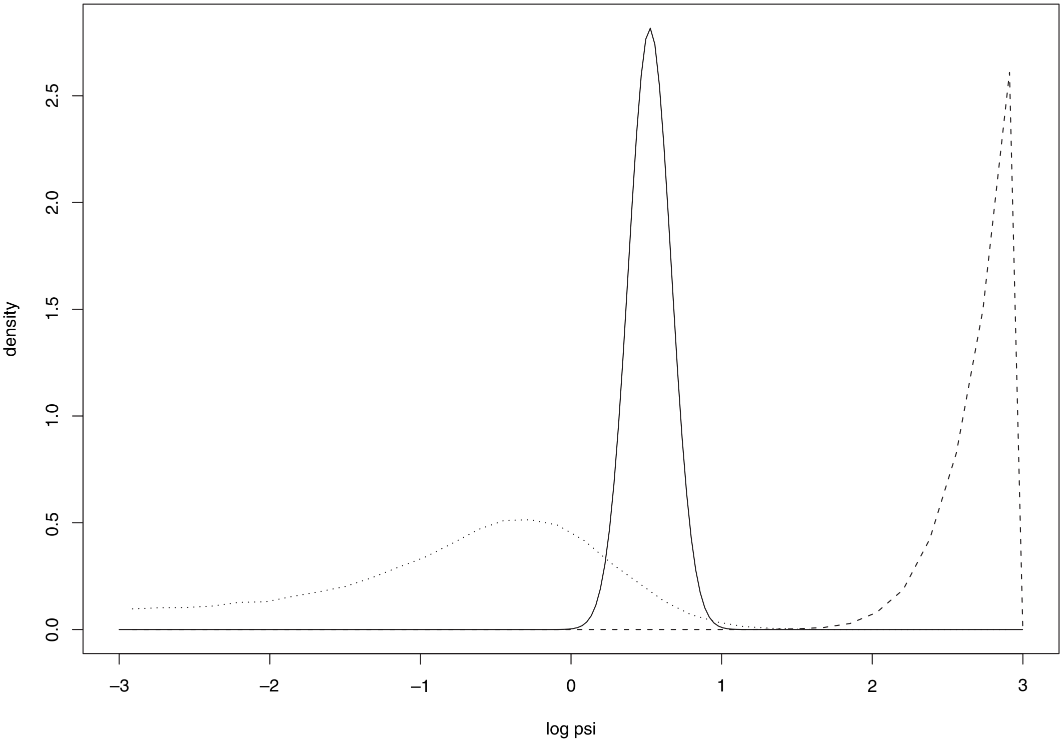

Fig. 2. Posterior distributions for various nuisance parameter priors, when $\varepsilon_0 = \varepsilon_1 = 0.05$ (for comparison with $\varepsilon_0 = \varepsilon_1 = 0$, the normal distribution that is suggested by the ψ-interval from Table 1 is superimposed (———)): · · · · · · ·, P(control exposure) ∼ $U(0,1)$; - - - - - - -, P(case exposure) ∼ $U(0,1)$

data with many zero counts—including where no controls are exposed. Hence for all values out to $\psi = 0$ or $\psi = \infty$ the likelihood remains above 0, and this forces the prior and posterior to have identically shaped 'tails', which in turn makes the posterior's moments and quantiles very sensitive to the ψ-prior. This is well known for completely unidentifiable models (Paulino *et al.*, 2003) but also impacts in this rather less special case (Rice, 2003). The problem is particularly acute if the likelihood's maximum lies at an extreme value, which is quite possible with realistic data sets.

The author's method needs no prior for ψ, but his 'semi-Bayes' smoothing technique effectively does the same job, stopping corrected counts E_η from becoming close to or below 0—indeed the smoothing parameter σ is justified by considering the variance of prior beliefs. I would therefore welcome an 'epidemiologist-friendly' interpretation of this parameter directly in terms of the exposure probabilities, and some idea of whether and when the analysis will be sensitive to its value.

Ben Armstrong (*London School of Hygiene and Tropical Medicine*)

The paper presented by Professor Greenland proposes methods to bring more formal and more numerical reasoning to bear on a problem—bias in observational studies—which has to date typically been addressed by using less formal, less quantitative approaches. The potential benefits of such a move—improved clarity in assumptions and in conclusions—are easily seen by us statisticians. However, the explicit models that clarify for us can obscure for non-statisticians. We should avoid discounting important insights on bias from such scientists. I have some suggestions that might help to avoid such a schism between statisticians and non-statisticians developing if, as I hope, methods such as these become more popular among statisticians.

(a) The formal method should be presented as an extension of a conventional sensitivity analysis, with presentation of the posterior distribution of the effect measure preceded by graphs of the dependence of the effect measure on a range of values of the 'bias' parameters. This clarifies the role and influence of the prior distribution of the bias parameters and provides an aid to understanding for people who are unfamiliar with the approach.

(b) The method should run in parallel with rather than replace less formal discussions of potential effects of bias.

(c) Priority should be given to (the minority of?) situations where less formal arguments are clearly insufficient. In particular, I suspect that this will be where no one source of bias can be argued informally to predominate. Where there is one predominant source of bias, informal reasoning sometimes accompanied by conventional sensitivity analysis will often be adequate.

Professor Greenland chose his example well, as one in which potential biases dominate uncertainty, and it is not obvious that one source of bias predominates. However, even here parallel informal reasoning can reach quite far, motivate the modelling and reassure model sceptics.

Finally, I wonder whether formal modelling can ever incorporate all the issues that are important for a conclusion. My prime candidate for consideration in Professor Greenland's example is his choice of focus on exposures above one cut point—3 mG. Professor Greenland appears to acknowledge that part of the reason for the focus is that this dichotomy yields the strongest evidence for association. This weighs quite heavily—but informally—on my personal post-Greenland probability that magnetic fields cause leukaemia.

Stephen Senn (*University of Glasgow*)
I think that there may be more that can be said for conventional analyses than Professor Greenland is willing to admit. Suppose that a meta-analyst is ambitious to carry out an analysis along the lines that are suggested in this stimulating paper and suppose that he or she is given a choice: either to be provided with conventional frequentist analyses of the original data, study by study, or to be provided with multiple-bias-adjusted analyses of these individual studies. It seems to me that the task of our would-be summarizer will be much easier if given the former rather than the latter. In fact, it is an irony, which George Barnard for one was wont to point out, that the last thing you want as raw input as a Bayesian is Bayesian posteriors unless they happen to be your own.

In fact, if they are not, you get in a terrible mess. You need to date-stamp all analyses and to make it clear what has and has not been included in any report of them. To make an analogy, sometimes, when you ask for the accounts you just want them item by item, even if the entries are not entirely reliable, rather than having some accountant's assesssment of current liquidity (Senn, 2000). Of course, if we follow this thought to the bitter end we come to the depressing conclusion that there is no such thing as public statistical analysis, the only things worth communicating are data but maybe frequentist point estimates and standard errors provide a useful compromise between unusable posteriors and overwhelming detail.

This is not to say that we should not welcome Professor Greenland's proposals. It is an all-too-frequent mistake to imagine that a conventional statistical analysis delivers with it an automatic recommendation for action. The paper here shows that much more may be needed before decisions can be taken.

James Carpenter and Mike Kenward (*London School of Hygiene and Tropical Medicine*)
It occurs to us that perhaps the problem that is tackled in this stimulating paper is equivalent to data coarsening (Heitjan and Rubin, 1991), a framework which includes missing observations, transcription errors, measurement errors and so forth. We agree with the author that, in such cases, it is always advisable to assess the sensitivity of the conclusions to the modelling assumptions, especially the ignorability of the coarsening mechanism.

Greenland presents an elegant framework to tackle this problem when the data form a multiway contingency table. In effect, a series of imputations are performed under a non-ignorable data coarsening mechanism.

However, we conjecture that a more general approach is to recast the problem as 'missing not at random multiple imputation', and to approximate draws as follows. Write the data Y, consisting of coarsened and uncoarsened (correctly observed) observations, as $Y=(Y_{\mathrm{C}}, Y_{\mathrm{O}})$. Our aim is assess the sensitivity of estimates of $\mathbf{E}_{Y_{\mathrm{C}}|Y_{\mathrm{O}}}\, \theta(Y_{\mathrm{C}}, Y_{\mathrm{O}})$ to non-ignorable coarsening mechanisms. Approximate draws from the posterior of this distribution can be obtained as follows:

(a) draw $Y_{\mathrm{C}}|Y_{\mathrm{O}}$ under an ignorable data mechanism and

(b) calculate the probability of seeing these data under the coarsening mechanism (which depends on Y_{C}), and then draw from a uniform distribution in $[0, 1]$ to decide whether to accept this draw.

This is closely related to using a weighted bootstrap (Smith and Gelfand, 1992) and has a couple of potential advantages:

(a) the modelling framework is no longer limited to count data and
(b) inference uses Rubin's rules for combining multiply imputed data sets.

Further, if the main analysis uses an ignorable data model, then a 'non-ignorable' weight can be calculated for each unit. A weighted version of the statistic of interest could then be calculated, with a sandwich estimate of variance. This approach has the attraction that

(i) weighted analyses can be done in most statistical packages,
(ii) ignorable imputations can be done semiautomatically in a variety of packages and
(iii) the model for coarsening can often be readily written down.

Alternatively, if data are categorical and only a small proportion are coarsened, a more direct assessment of sensitivity can be obtained from looking directly at the likelihood, which can in some settings reduce to examining worst–best scenarios. See Raab and Donnelly (1999), Molenberghs *et al.* (2001) and Verzilli and Carpenter (2002).

Prior information for non-ignorable models is often contentious. Useful progress can sometimes be made by eliciting priors from experts, or consumers, of research. For a discussion of this approach in the context of clinical trials, see White *et al.* (2004).

Bianca De Stavola and Dorothea Nitsch (*London School of Hygiene and Tropical Medicine*)
Professor Greenland's drive for an explicit modelling of the biases affecting epidemiological studies is extremely welcome. As practitioners, however, we have concerns about our ability to apply and interpret multibias modelling.

The study that is discussed in the paper is a pooled analysis of 14 case–control studies, where Y is a binary variable identifying the cases, X a binary indicator of magnetic field exposure and S the study identifiers (Fig. 3). A Mantel–Haenszel statistic could be used to obtain a weighted summary of the study-specific effects and therefore control for the confounding effect of S. Alternatively, as suggested by Professor Greenland, a more realistic model would recognize that X might be only a proxy for the true exposure T, that the study participants are likely to be a biased subset of the population of interest and that S is not the only confounder. In other words it would include submodels to represent measurement error, selection (and response) bias and the unaccounted confounders (Fig. 4).

The forms that are taken by each of these submodels, plus the range of values that are allowed for their parameters, are crucial to the results and should be guided by subject-matter considerations. However, there are aspects of the Monte Carlo sensitivity analysis strategy that is suggested by Professor Greenland that should be generalizable. This is where we wish to have some clarifications.

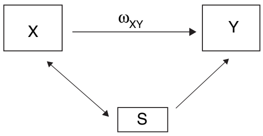

Fig. 3. Standard model for an outcome Y, an exposure X and a confounder S

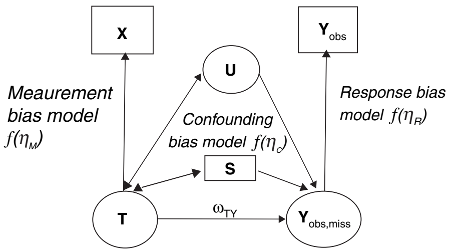
X
Y_obs
U
Meaurement bias model f(η_M)
Confounding bias model f(η_C)
Response bias model f(η_R)
S
T
ω_TY
Y_obs,miss

Fig. 4. Path diagram for Greenland's multiple-bias model

(a) The example focuses on a pooled analysis. Could it be equally applied to data from a single study? Or should we adopt such complex sensitivity analyses only when information is available from several studies so that assumptions are more believable?

(b) Is the sequential correction that is recommended in the paper only applicable to discrete data (i.e. binary Y and X and categorical S)? The general Monte Carlo sensitivity analysis that is described in Section 2.3 seems impractical for metric exposures (and/or outcomes), unless we adopt a fully Bayesian approach. However should one be advised to attempt multibias modelling only after dichotomizing a metric exposure? But would that not lead itself to misclassification bias (Flegal *et al.*, 1991)?

(c) The assumption that all the intercepts in the bias models have mean 0 may be too restrictive, especially for the logit of the latent confounder among the non-exposed non-cases (equation (12)): has Professor Greenland considered sensitivity analyses for the means of their distributions?

D. Nitsch and B. De Stavola (*London School of Hygiene and Tropical Medicine*)

As applied statisticians we fully support Greenland's drive for a structured quantification of bias due to systematic errors in epidemiologic studies, such as confounding, selection bias and measurement error or misclassification. In the current paper it might be worth mentioning that before even considering doing such multiple-bias modelling a satisfactory directed acyclic graph should be drawn (an example is given in Fig. 5 (Greenland *et al.*, 1999)). This helps to understand the data generation process and leads to the identification of possible sources of bias. Owing to the case–control design, the sampling of the study subjects is according to the full arrows, whereas the usual argument with respect to causality or confounding is in the direction of the broken arrows. S, D and V are all related by separate functions to the latent variables T (the true measurement), $Y_{\text{miss,obs}}$ (observed and missing cases or controls) and U (the confounder) which are denoted by ovals. For simplicity additional arrows between S, D and V and these latent quantities have been omitted.

The current application is designed to deal with systematic errors that might affect a pooled analysis of case–control studies using only binary variables. In other situations the problem of misclassification of exposure might be solved by prior information, in contrast with non-response where a sensitivity analysis might be more important. Is it right to assume that the proposed data unwrapping process changes for cohort studies (adjust for differential non-response; then misclassification; then confounding) and that then the sensitivity analysis can be carried out in a similar way, assuming a differential non-response? How might multiple imputation solve the problem with respect to this question? Would this extension in the latter setting allow for additionally dealing with continuous variables?

Is our understanding of the 'bias in favour of the null hypothesis' (near the end of the first paragraph of Section 4.2) right in the sense that by using sensitivity analysis we try to reject findings that favour the alternative hypothesis? Would an *a priori* formalization of the effect sizes that we wish to confirm or reject help? Computation of such prespecified equivalence 'bands' (the proportion of an effect of a given range) might yield information of interest and avoids such a bias in favour of the null hypothesis.

Using a weighted Mantel–Haenszel average to summarize study results uses uncorrected weights. As noted in Section 2.6, the more the bias parameters move away from zero, the less efficient the $W_{0\eta}$-weighted bias-corrected estimator is. Did we understand correctly that this was the reason for using the original weights that are derived from the smoothed table frequencies? If there is an assumed high rate of misclassification, must these weights be amended to the assumed relationship between the true exposure T and outcome Y after adjusting for misclassification of study results?

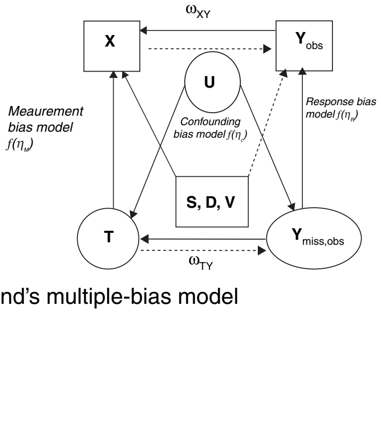
ωXY
X
Yobs
U
Meaurement bias model f(ηM)
Confounding bias model f(ηC)
Response bias model f(ηR)
S, D, V
T
Ymiss,obs
ωTY

Fig. 5. Path diagram for Greenland's multiple-bias model

Colin R. Muirhead (*National Radiological Protection Board, Chilton*)
I would like to echo the comments made by Dr Armstrong, namely that epidemiologists often attempt to assess the effect of various potential sources of bias, albeit not using the formal modelling approach that is described in this paper. The conclusions from the magnetic fields and childhood leukaemia example that are presented here appear to be similar to those reached in a report by an Advisory Group on Non-ionising Radiation (2001), chaired by Sir Richard Doll, which used a more traditional epidemiological approach. It would be interesting to know of practical situations in which multiple-bias modelling and the usual epidemiological approaches to assessing the effect of bias produce substantially different conclusions and—in such instances—to understand why this is the case. (Publication of these comments is copyright of the National Radiological Protection Board).

Jim Hodges (*University of Minnesota, Minneapolis*)
Professor Greenland has belled a nasty cat with a sober, concrete paper. Not so long ago, a paper like this would have elicited a strong majority vote—from statisticians, at least—that the author was evil, stupid or both. Such a view was often rationalized by using a distinction, pungently expressed by Fisher, between restrained, disinterested utterances appropriate to science (i.e. along the lines of a conventional analysis) and the grubbier utterances of mere decision makers. This paper refutes that old rationale with two arguments. First, we technicians can hardly wash our hands of biases and trust decision makers to weigh them subtly, but nor can we dispose of them with our customary bit of ‘arm waving’. A second and less familiar argument is that bias analysis has important implications for allocating research resources, and it is often not sufficiently good simply to do more of what we can easily sell to our statistical colleagues.

Another way to view this paper is that it realizes a theory of statistics in application, as distinct from the idealized theories of statistics that are considered in the defunct branch of our field called ‘Foundations of statistics’. A theory of statistics in application acknowledges that statistical reality is messy but asserts that we have some powerful tools to deal with the mess and that we are obligated to use them for urgent substantive questions. In this sense Greenland’s paper is a fitting companion to the observational study work of Don Rubin, Paul Holland, Jamie Robins, Susan Murphy and others.

N. T. Longford (*SNTL, Leicester*)
The paper is a well-aimed indictment on the ubiquitous practice of applying text-book methods in settings that deviate substantially from the text-book assumptions of good representation, perfect measurement, subjects’ full co-operation and the like. I am surprised that the paper makes no reference to the literature on nuisance parameters, and a new term, bias parameters, is used instead. The problem that is addressed is estimation of a target parameter vector $\boldsymbol{\theta}$ in the presence of a nuisance parameter vector $\boldsymbol{\xi}$; the current practice sets $\boldsymbol{\xi}$ to a default value, and Greenland’s proposal uses a prior for $\boldsymbol{\xi}$.

Notwithstanding the integrity of the approach, a less sympathetic reader may feel short changed by the paper because information about the bias parameters is contained solely in the prior distributions, and their specification is usually not particularly scientific—we cannot capture the relevant information with much precision or credibility. Ignoring the uncertainty about the prior leads to an error that cannot be practically taken care of by hyperpriors. Priors cannot replace information or overcome the difficulties of eliciting it, especially in several dimensions.

An alternative can be motivated by Longford (2001). A small number of extreme scenarios, values of $\boldsymbol{\xi}$ that are judged to be on the borderline of what could occur in reality, are specified. The efficiency of the estimator that is optimal for the default setting is then assessed for each of these extreme scenarios as a form of sensitivity analysis (the author’s objections granted), or the estimator that is optimal for an extreme setting is assessed for the default and the other extreme settings. In this way, the low mean-squared error at the default setting is traded off for improved estimation at (some) extreme settings.

The paper focuses on asymptotics, where model selection is not a problem. For finite samples, submodels of the ‘true’ model may yield more efficient estimators because their bias may be small in relation to the variance reduction. This is a poorly explored area that may offer the current practice a weak justification in small samples.

Finally, I want to point out that the studies in the meta-analysis should themselves be regarded as a sample from a population of settings (time and country in particular). Given no formal sampling from these settings, the pooled estimator has an unknown bias that should also be associated with one or several nuisance parameters.

Andrew Gelman (*Columbia University, New York*)
Accounting for bias in causal inference is hard work, and Greenland's paper provides a theoretical structure and an application to epidemiology. I would like to add a small point regarding sensitivity analysis, which typically involves extra parameters that cannot be estimated from data and thus must be swept over some range of possibilities.

Bias models are similar to non-ignorable missing data models in that by their nature they are commonly non-identifiable from the observed data likelihood. When bias or non-ignorable missingness parameters are identified, it is usually a weak identification that is highly sensitive to distributional assumptions or selection mechanisms that are themselves not identifiable from data (e.g. Heckman (1979)).

Although the extra parameters cannot really be estimated, we can sometimes establish how much they can reasonably vary by examining their implication for an unbiased experiment. We illustrate with a simple example from Abayomi *et al.* (2004).

Fig. 6 shows hypothetical data of heights of boys in a school, measured when the basketball team (representing 20% of the population) is away. The distribution of the observed data is skewed, and various models can be imagined to impute the missing data. Missingness can be at random or can depend on unobserved data (Rubin, 1976). The graphs in Fig. 6 show distributions that were estimated from three models:

(a) reproducing the skewed distribution with a model with missingness completely at random,
(b) filling in a normal distribution with a model with missingness not at random in which taller students are less likely to be sampled and
(c) estimating a more extreme model with missingness not at random in which the underlying complete-data distribution turns out to be bimodal.

The data alone do not distinguish between the three models of Fig. 6. And, in fact, one could imagine model (a) being appropriate (if basketball players had the same height distribution as the other boys) or model (b) being appropriate (with an approximate normal distribution for all the boys). Model (c) seems less plausible, as it would correspond to a peculiar distribution for the *complete data* that would be seen without any selection bias. Thus it seems unnecessary, in considering a sensitivity analysis, to consider models as extreme as (c). This idea differs from the 'device of imaginary results' (Good, 1950) in that we are considering completed data—i.e. imputed data combined with observed data—rather than imaginary data simulated from the model alone.

This example illustrates that parameters that cannot be estimated can still be bounded by examining implied complete-data distributions. Such bounds could be considered as a form of informally specified prior information allowing sensitivity analysis to remain within a plausible range.

The following contributions were received in writing after the meeting.

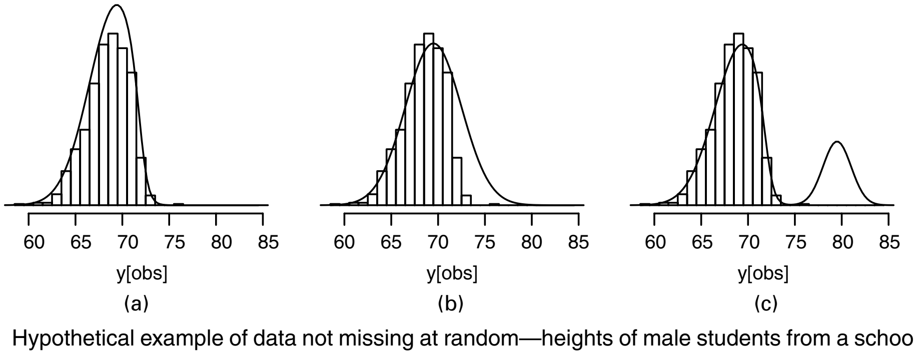

Fig. 6. Hypothetical example of data not missing at random—heights of male students from a school, measured when the basketball team was away: estimates of the distribution of the entire population of male students fitted under the three models (a) data missing completely at random, (b) data not missing at random (model 1) and (c) data not missing at random (model 2), with models 1 and 2 allowing taller students to be more likely to be missing; the distribution in (b) looks reasonable whereas the distribution in (c) is implausible—it is difficult to believe that the complete data are actually bimodal; this example illustrates that some aspects of missing data models, although theoretically untestable, can be bounded

David Draper (*University of California, Santa Cruz*)
I am in great sympathy with the aims of this work. However difficult it is to imagine that analyses of this type will become routine, it should be even more difficult for us to imagine that the right way forward is to perpetuate the *status quo*, in which sources of uncertainty other than random error are quantitatively ignored. In Draper (1995) (also see Holland (1989)) I referred to the need to estimate judgmentally 'a variance component for nonexchangeability' between the observed units in an observational study and units in the population of real scientific interest, and a variance component for the effects of unmeasured confounders; here Greenland has essentially provided a method for judgmentally quantifying these variance components.

I have two additional remarks on the paper.

(a) The author is not clear on what, precisely, he would report to policy makers by way of conclusions on the effects of electromagnetic field exposure on childhood leukaemia from his Table 2. We have 18 different interval estimates, with median odds ratios ranging from 1.45 to 3.63; what does Greenland suggest we should conclude?
(b) Since the techniques of this paper are based on prior distributions that are unmodified by data in the analytic journey from prior to posterior, it might appear that the only evaluation possible of results such as those of Table 2 is process based: do the assumptions seem reasonable?

But here are two outcome-based ideas for evaluation of multiple-bias modelling on which Greenland might wish to comment:

(i) if the fields of medicine and epidemiology were to keep a public database of observed biases, by comparing the results of randomized controlled trials and observational studies designed to answer the same questions, then the prior distributions in Greenland's approach could in principle be based on past data instead of being entirely judgmental;
(ii) one way (albeit slow) to validate the methods of this paper externally is to regard results such as Table 2 as attempts to predict accepted scientific truth k years in the future (for some value of k like 5 or 10) when better data are available (see Leonhardt (2001), based on Pocock and Spiegelhalter (1992), for a nice example of this); we can then wait k years and see, for example, whether the currently accepted odds ratio for electromagnetic field exposure in the year $2005 + k$ is closer to 1.68 (the median in the first row in Table 2) or 2.70 (the median in the last row).

Paul Gustafson and Lawrence McCandless (*University of British Columbia, Vancouver*)
We congratulate Professor Greenland for a compelling discussion of a difficult problem. Modelling bias from misclassification, non-response and unobserved confounding simultaneously is a daunting task. Yet such efforts seem vital for the credible analysis of observational data.

We like Professor Greenland's interpretation of Monte Carlo sensitivity analysis (MCSA) as an approximation to a Bayesian analysis, and we agree that often, but not always, this approximation will be good. Echoing his comments about misclassification, Gustafson (2005) considers misclassification and measurement error scenarios where the posterior marginal distribution of bias parameters can differ from the prior distribution, so the MCSA answer may differ from the Bayesian answer.

One potential objection to the overall approach is that the actual and nominal frequentist coverage of interval estimates cannot agree in the face of non-identified models. Those with such objections may be comforted by a reminder that the Bayesian coverage of intervals will still be satisfactory, i.e. actual and nominal coverage of Bayesian intervals (perhaps approximated by MCSA) will agree for a hypothetical sequence of studies in which nature generates study data by first sampling parameters from the prior distribution and then sampling data given parameters. Neither identifiability nor large samples are required for this agreement, so long as nature and the investigator use the same prior distribution. Thus one way to shed light on the 'sturdiness' of multiple-bias modelling (again either Bayesian or the MCSA approximation thereof) is to study how badly the agreement breaks down when nature and the investigator use different priors. In the context of unobserved confounding, this is taken up by McCandless (2004).

The comment in Section 4.4 that at some point replication and enlargement of observational studies cease to be cost effective is interesting, and should be eye opening to epidemiologists. We envision an approach for quantifying this idea in a design context. One tool may be the decomposition of mean-squared error for a posterior mean in a non-identified model that is considered by Gustafson (2005). This involves a squared bias term which does not fall off with sample size and a variance term which does. One could choose a sample size which is just sufficiently large to make the variance small compared with the squared bias (in an *a priori* expected sense), thereby conserving resources for further studies of different exposure–disease relationships. Also related is the interesting work of Rahme *et al.* (2000) and Dendukuri *et al.* (2004), who studied sample size criteria when model non-identifiability due to misclassification is acknowledged.

Donald B. Rubin (*Harvard University, Cambridge*)
Professor Greenland's paper is a welcome addition to the literature on causal inference in observational studies. The history of using the Bayesian paradigm to incorporate non-sampling sources of uncertainty has several relatively early examples, including Mosteller and Wallace (1954), Box and Tiao (1973) and Rubin (1977). Although none of these specifically addressed causal inference, Rubin (1978), section 4.2, does address the sensitivity of Bayesian causal inference when faced with non-ignorable data assignment mechanisms.

Greenland emphasizes the multiple sources of bias that are typically present in real world studies and the consequential futility of conducting simple sensitivity analyses to all sources simultaneously. This criticism also applies to analyses that provide large sample bounds for estimates. Although the technical work can be interesting (e.g. Manski (l990)) and the idea has a long history in statistics, going back at least to Cochran (1953) in the context of survey non-response, Cochran's conclusion seems equally true in the context of observational studies:

> 'With a continuous variate, the only bounds that can be assigned with certainty are often so wide as to be useless'.

This last paragraph relates to, what I believe is, a somewhat misguided focus in much of current statistical education on methods for the analysis of existing data, whether Bayesian or frequentist, to the exclusion of education on the use of creative designs for the collection of relevant data. I think that Greenland's wisdom in Section 4 may be misread by some as supporting this tendency, even though one of the crucial products of such analyses must be the exposure of which sources of uncertainty are crucial, with accompanying suggestions for what sorts of cost-effective data could reduce this uncertainty. As Box (1976) noted, science progresses through a series of iterations, with the analysis of one data set leading to the design of the next data collection effort, and the field of statistics should serve both enterprises. For example, we should consider not only what sort of data could be relevant to the science (e.g. from other species) but also what sort of data could inform us about the non-ignorable data assignment mechanisms that are operating in the existing studies (e.g. research on the reasons why people move to certain locations in Greenland's example).

I agree with Greenland that the formal consideration of sources of inferential uncertainty, beyond sampling uncertainty, should become part of the training of those who deal with the analysis of existing data or the collection of future data, especially when used for causal inference.

The **author** replied later, in writing, as follows.

I thank the discussants for their comments, and regret that I do not have space to address them all.

Bias analysis versus informal bias assessment
No-one disagrees that bias problems are vital, yet some question formal approaches. Bias analysis has a long history (e.g. Berkson (1946)) but has never taken root in basic statistics teaching and hence is uncommon in health science reports. In the USA, at least, appeals to reform epidemiologic statistics have failed because many health scientists cannot stomach quantitative reasoning beyond a 2^3-table, and because of arguments that conventional methods are adequate, that biases should be handled informally (even as random error is handled with absurd precision) and that miniscule and simplistic 'validation' studies fully address biases.

In large scale decisions, statistics serves as an input to informal judgments. The key question is, what inputs are needed? Muirhead mentions that an illustrious panel used only conventional inputs to reach conclusions that were in accord with my formal analysis. Unfortunately, some health officials in the USA were not in accord. Illustrious panels are unavailable for most issues, and informal bias assessments sometimes fail spectacularly. Formal analyses can check informal assessments and reveal complexities that are otherwise overlooked. And, unlike with informal assessments, one can derive formal conditions under which formal assessments will yield valid answers.

Draper asks how we should evaluate our assessment methods in practice. Randomized trials are often taken as a gold standard, but they are usually infeasible (as in my example). Furthermore, their treatments and subjects rarely approximate natural exposures or populations, and hence trial results are easily dismissed, as Lawlor *et al.* (2004a, b) lament in studies of nutrients and of oestrogen replacement. Regardless, trials are costly and tardy tests of informal assessments based on conventional analyses; recent history suggests that more economical and rapid correctives are needed.

By 1981, leading researchers focused on β-carotene as responsible for the lower cancer incidence that is seen among people with high fruit and vegetable consumption (Peto *et al.*, 1981). Their judgments did not formally account for the stupendous correlations between dietary intakes and their measurement errors, and the compound error induced by computing nutrient intakes from dietary measures. Application of multiple-bias analysis to the observational evidence favouring such nutrients would have rendered the evidence far less compelling. Even simple sensitivity analyses revealed the trial designs as unjustifiable, banking too heavily on β-carotene as the key preventive nutrient (I raised this point as an observer at a trial planning meeting in 1984; it was politely dismissed).

In 1990, informal assessments from certain authorities combined with overwhelmingly significant conventional results (Stampfer and Colditz, 1991) won US Food and Drug Administration approval to label oestrogens preventive for heart disease, despite pleas from others for randomized trial evidence (Lawlor *et al.*, 2004b). Subsequent trials found no benefit and possible harm, suggesting that people died as a result of the prevailing informal assessments. Some authorities still maintain that the observational results are correct, but the trials at least call into question the certitude of the original judgments (Petitti, 2004). If the Food and Drug Administration had demanded formal bias analysis as it demands statistical significance, the label approval could not have been justified.

As long as statistical methods for randomized studies remain the standard that is taught and accepted for observational studies, with 'statistical significance' (a randomization P under 0.05) the centre-piece of inference, the field of statistics should bear partial blame for the consequences of incorrect informal assessments. If bias analysis is not incorporated into basic teaching and regulatory evaluation, its use will remain patchwork 'ad hockeries' carried out by a brave few or, worse, it will continue to be neglected in favour of muddled intuitive arguments and ever more elaborate treatments of random error (whose distribution in epidemiologic studies is as hypothetical as any prior). To paraphrase Draper: although difficult to imagine the methods that I discussed are suitable for mass consumption (I doubt they are), it should be more difficult to imagine that the right way forward is to continue conventional practice.

Bias analyses developed because in many contexts (including those of greatest policy importance) bias can be more important than random error. The distributions that they produce are their message. In the example those are too dispersed ever to be contradicted, which tells policy makers that no definitive inference should be drawn from such studies. In other settings they may show their worth by encompassing trial results, something that conventional intervals did not do in the nutrient and oestrogen controversies. Whether they encompass future opinions seems less relevant, especially if those opinions are formed to defend past opinions.

Sensitivity analysis versus Bayesian analysis

I agree with Copas's description of the logical problem of inference, although we diverge on preferred solutions. He and Longford suggest examining extreme scenarios, wishing to avoid the complication, contention and arbitrariness of high dimensional priors. My arguments against the adequacy of these sensitivity analyses apply unchanged. Sensitivity analysis is computationally simpler but does not avoid arbitrariness and contentious subjectivity; after all, who decides what are 'worst cases', 'extremes' or 'plausible values' of bias parameters? Projections collapse over high dimensional phenomena; they do not account for uncertainties or information about these phenomena.

Still, sensitivity analysis is a vast improvement over conventional analysis. Why then should we go further?: because, to paraphrase Rubin's quote from Cochran, with multiple-bias parameters, the only bounds that can be assigned with certainty will be so wide as to be useless. With unlimited sensitivity, using boundary points for extremes only shows that anything is possible. Less extreme worst case analysis amounts to using a prior that is concentrated at two implausible points, with nothing between. This prior is held by no-one and is unduly pessimistic in situations that are already plagued by uncertainties. Sensitivity analysis says nothing without context to explain what constitutes sensitivity. It helps by forcing construction of a bias model, but that model must be coupled with priors to make any inference. Given that inferences will be drawn, we either use explicit priors or else follow frequentist tradition and pretend that the priors are not there (which results in the use of implicit absurd priors).

Sensitivity to priors

Rice rightly emphasizes the extreme sensitivity of results to priors, as is inevitable when target parameters are not identified. We must become accustomed to and constantly underscore these facts: absent randomization, every causal inference depends entirely on priors, and what seems reasonable to one may seem unreasonable to another. Consider Rice's comment, with its implied prior case–control symmetry: 'one might reasonably have used flat priors on the case probabilities' instead of for the control probabilities. The controls stand in for the exposure experience of the entire population before the occurrence of disease (Rothman and Greenland (1998), chapter 7), whereas the cases arise from this experience and the highly selective force of leukaemia incidence. Scientific priors must account for causal (and response and measurement) ordering; thus, absent case series data, priors for cases should be induced by population and effect priors, rendering prior symmetry unreasonable (Greenland, 2001).

With unlimited sensitivity, the best that we can do is to display the priors we tried, facilitating criticism. If the priors are compatible with existing data or accepted theory, the most that we can say is 'here are posteriors from some currently reasonable priors'. Reasonable priors may conflict; hence so may reasonable posteriors. We can, however, reject unreasonable priors and their results. If randomization and perfect measurement are unreasonable priors for the data, then conventional analyses are unreasonable because they use these priors.

Of course, the status of 'reasonable' is subjective and may change given new results or new criticisms. Hence priors play the role of theories in falsificationist philosophy always tentative and highly disposable. This view violates rigid Bayesian philosophy; instead, it treats Bayesian analysis as a device to organize current information coherently, facilitating deductions from that information (Good, 1983).

Other issues

Several misunderstandings occurred. Armstrong thought that the 3 mG cut point was chosen for 'giving the strongest evidence'. That is incorrect; it was not picked to maximize some statistic, but rather because global summaries are fairly insensitive to the choice and there is no evidence of an association below that point. Rice thought that I aimed 'to approximate a Bayesian analysis optimally' but I was only aiming to make Monte Carlo sensitivity analysis (MCSA) better approximate Bayesian analysis. Like Spiegelhalter I prefer fully Bayesian computations over MCSA. None-the-less, MCSA has greater intuitive appeal for non-statistician epidemiologists and so has taken users away from Bayes in epidemiology and risk assessment.

Longford commented that bias parameters are nuisance parameters. The converse is not true, however; in the example, the $Y=1$ probabilities are nuisance parameters but not bias parameters. Draper and Longford implied that the prior distributions that I used were 'entirely judgmental', yet in Sections 3.6 and 3.7 I described how priors for confounder–field associations and selection bias were based on actual surveys. More generally, there are many epidemiologic studies that could supply information on bias parameters. None-the-less, there will always be gaps in data that must be filled by speculation, as with the example confounder–disease and misclassification priors. Multiple-bias analysis can help to identify gaps in most urgent need of filling. In the example the gaps are so large that further studies like those reviewed are a waste of funds; such studies continue to be done, however, and there will be no trial to staunch them.

De Stavola and Nitsch ask many important questions. Whether one study or several, bias analysis seems unnecessary when confidence intervals exclude only implausible values, and it seems most needed to evaluate claims that the data mandate action. Correction factors are not limited to discrete data; any correction method that is based on second-stage ('validation') data can be used by sampling validation data from a prior. The bias–intercept means are prior means when all prior covariates are 0—they require subject-matter choice. In cohort studies, the bias that is created by base-line non-response *is* confounding (indeed, in econometrics 'selection bias' means confounding). The cohort analogue of differential case–control response is differential drop-out; measurements come before drop-out but may precede or follow the response (e.g. records *versus* interviews). Diagrams clarify bias ordering and thus help to determine correction ordering. Output distributions estimate posterior probabilities of any interval (decision) hypothesis; conversely, we can see what sort of priors yield a fixed posterior probability (e.g. 0.95) for a given interval hypothesis such as '$\theta > 0$', as in conventional Bayesian analyses (Matthews, 2001). The uncorrected weights are used in MCSA to approximate Bayes analysis better; if misclassification is very high, however, fully Bayesian analysis is advisable (Gustafson, 2003).

Carpenter and Kenward note that MCSA could be subsumed under multiple imputation. I agree and elsewhere have encouraged imputation approaches (Cole *et al.*, 2005; Fox *et al.*, 2005); there are, however, problems with 'Rubin's rules' for combining multiple imputations (Robins and Wang, 2000). Rice questioned the smoothing parameter σ^2. It is a prior variance on the log-odds of $X=1$; the example results are insensitive to σ^2 because the prior is dominated by large counts, but sensitivity would arise when all counts are small (small samples imply sensitivity even for identified parameters). Finally, I admit to Senn that conventional estimates are helpful if viewed as quick data summaries *and nothing more*. I always provide them, but when allowed I also provide multiway data tables. Anyone can do their own analysis using the data presented in my paper; those are all the data that I used and are superior to conventional statistics as input for future analyses.

References in the discussion

Abayomi, K., Gelman, A. and Levy, M. (2004) Diagnostics for multivariate imputations. *Technical Report*. Department of Statistics, Columbia University, New York.

Advisory Group on Non-ionising Radiation (2001) ELF electromagnetic fields and the risk of cancer. *Document NRPB 12(1)*. National Radiological Protection Board, Chilton. (Available from `http://www.nrpb.org/publications/documents of nrpb/abstracts/absd12-1.htm`.)

Berkson, J. (1946) Limitations of the application of fourfold tables to hospital data. *Biometr. Bull.*, **2**, 47–53.

Box, G. E. P. (1976) Science and statistics. *J. Am. Statist Ass.*, **71**, 791–799.

Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. Reading: Addison-Wesley.

Breiman, L. (2001) Statistical modeling: the two cultures (with discussion). *Statist. Sci.*, **16**, 199–231.

Claxton, K. (1999) Bayesian approaches to the value of information: implications for the regulation of new pharmaceuticals. *Hlth Econ.*, **8**, 269–274.

Cochran, W. G. (1953) *Sampling Techniques*, 1st edn. New York: Wiley.

Cole, S. R., Chu, H. and Greenland, S. (2005) A simulation study of multiple-imputation for measurement error correction. *Am. J. Epidem.*, to be published.

Copas, J. and Eguchi, S. (2001) Local sensitivity approximations for selectivity bias. *J. R. Statist. Soc.* B, **63**, 871–895.

Copas, J. B. and Jackson, D. (2004) A bound for publication bias based on the fraction of unpublished studies. *Biometrics*, **60**, 146–153.

Crystal, D. and Crystal, B. (2002) *Shakespeare's Words*. London: Penguin.

Davey Smith, G. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidem.*, **32**, 1–22.

Dendukuri, N., Rahme, E., Belisle, P. and Joseph, L. (2004) Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics*, **60**, 388–397.

Draper, D. (1995) Inference and hierarchical modeling in the social sciences (with discussion). *J. Educ. Behav. Statist.*, **20**, 115–147, 233–239.

Eddy, D. M., Hasselblad, V. and Shachter, R. (1992) *Meta-analysis by the Confidence Profile Method*. Boston: Academic Press.

Flegal, K. M., Keyl, P. M. and Nieto, F. J. (1991) Differential misclassification arising from nondifferential errors in exposure measurement. *Am. J. Epidem.*, **134**, 1233–1244.

Fox, M. P., Lash, T. L. and Greenland, S. (2005) A SAS macro to automate probabilistic sensitivity analyses of misclassified binary variables.

Good, I. J. (1950) *Probability and the Weighing of Evidence*. London: Griffin.

Good, I. J. (1983) *Good Thinking*. Minneapolis: University of Minnesota Press.

Greenland, S. (2001) Sensitivity analysis, Monte-Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Anal.*, **21**, 579–583.

Greenland, S., Pearl, J. and Robins, J. M. (1999) Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.

Gustafson, P. (2003) *Measurement Error and Misclassification in Statistics and Epidemiology*. New York: Chapman and Hall.

Gustafson, P. (2005) On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured data (with discussion). *Statist. Sci.*, to be published.

Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and coarse data. *Ann. Statist.*, **19**, 2244–2253.

Holland, P. (1989) Discussion of "Fisher scoring algorithm for variance component analysis of data with multilevel structure," by Longford NT. In *Multilevel Analysis of Educational Data* (ed. R. D. Block), pp. 311–317. San Diego: Academic Press.

Lawlor, D. A., Davey Smith, G., Bruckdorfer, K. R., Kundu, D. and Ebrahim, S. (2004a) Those confounded vitamins: what can we learn from the differences between observational versus randomized trial evidence? *Lancet*, **363**, 1724–1727.

Lawlor, D. A., Davey Smith, G. and Ebrahim, S. (2004b) The hormone replacement–coronary heart disease conundrum: is this the death of observational epidemiology? *Int. J. Epidem.*, **33**, 464–467.

Leonhardt, D. (2001) Adding art to the rigor of statistical science. *The New York Times*, Apr. 28th. (Available from `www.nytimes.com`.)

Longford, N. T. (2001) Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited. *Statist. Med.*, **20**, 3189–3203.

Manski, C. F. (1990) Nonparametric bounds on treatment effects. *Am. Econ. Rev. Pap.*, **80**, 319–323.

Matthews, R. A. J. (2001) Methods for assessing the credibility of clinical trial outcomes. *Drug Inform. J.*, **35**, 1469–1478.

McCandless, L. (2004) Assessing sensitivity to unmeasured confounding in observational studies—a Bayesian approach. *MSc Thesis*. Department of Statistics, University of British Columbia, Vancouver.

Molenberghs, G., Kenward, M. G. and Goetghebeur, E. (2001) Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Appl. Statist.*, **50**, 15–29.

Mosteller, F. and Wallace, D. (1954) *Inference and Disputed Authorship: the Federalist*. Reading: Addison-Wesley.

Paulino, C. D., Soares, P. and Neuhaus, J. (2003) Binomial regression with misclassification. *Biometrics*, **59**, 670–675.

Petitti, D. B. (2004) Hormone replacement therapy and coronary heart disease: four lessons. *Int. J. Epidem.*, **33**, 461–463.

Peto, R., Doll, R., Buckley, J. D. and Sporn, M. B. (1981) Can dietary beta-carotene materially reduce human cancer rates? *Nature*, **290**, 201–208.

Pocock, S. J. and Spiegelhalter, D. J. (1992) Domiciliary thrombolysis by general practitioners. *Br. Med. J.*, **305**, 1015.

Raab, G. M. and Donnelly, C. A. (1999) Information on sexual behaviour when some data are missing. *Appl. Statist.*, **48**, 117–133.

Rahme, E., Joseph, L. and Gyorkos, T. W. (2000) Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Appl. Statist.*, **49**, 119–128.

Rice, K. M. (2003) Full-likelihood techniques for misclassification of exposure in matched case control studies. *Statist. Med.*, **22**, 3177–3194.

Rice, K. M. (2004) Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *J. Am. Statist. Ass.*, **99**, 510–522.

Robins, J. M. and Wang, N. S. (2000) Inference for imputation estimators. *Biometrika*, **87**, 113–124.

Rothman, K. J. and Greenland, S. (1998) *Modern Epidemiology*, 2nd edn. Philadelphia: Lippincott.

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

Rubin, D. B. (1977) Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Statist. Ass.*, **72**, 538–543.

Rubin, D. B. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.

Senn, S. J. (2000) Consensus and controversy in pharmaceutical statistics (with discussion). *Statistician*, **49**, 135–176.

Smith, A. F. M. and Gelfand, A. E. (1992) Bayesian statistics without tears: a sampling-resampling perspective. *Am. Statistn*, **46**, 84–88.

Smith, T. C., Spiegelhalter, D. J. and Thomas, A. (1995) Bayesian approaches to random-effect meta-analysis: a comparative study. *Statist. Med.*, **14**, 2685–2699.

Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. (1994) Bayesian approaches to randomized trials (with discussion). *J. R. Statist. Soc.* A, **157**, 357–416.

Stampfer, M. J. and Colditz, G. A. (1991) Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev. Med.*, **20**, 47–63.

Verzilli, C. and Carpenter, J. R. (2002) Estimating uncertainty in parameter estimates with incomplete data: an application to repeated ordinal measurements. *Technical Report*. Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London. (Available from `www.missingdata.org.uk`.)

White, I., Carpenter, J., Evans, S. and Schroter, S. (2004) Eliciting and using expert opinions about non-response bias in randomised controlled trials. Submitted to *Clin. Trials*. (Available from `www.missingdata.org.uk`.)