Statistical Science
2018, Vol. 33, No. 1, 86–95
https://doi.org/10.1214/17-STS632

When Is a Sensitivity Parameter Exactly That?

Paul Gustafson and Lawrence C. McCandless

Abstract. Sensitivity analysis is used widely in statistical work. Yet the notion and properties of *sensitivity parameters* are often left quite vague and intuitive. Working in the Bayesian paradigm, we present a definition of when a sensitivity parameter is "pure," and we discuss the implications of a parameter meeting or not meeting this definition. We also present a diagnostic with which the extent of violations of purity can be visualized.

Key words and phrases: Bayesian inference, misclassification, missing data, selection bias, sensitivity analysis.

1. INTRODUCTION

In many statistical applications, the data we can collect are not enough. Fulsome insight into the scientific issue at hand may additionally require external information or inputs, and often this information is imprecise. When the external inputs are manifested as one or more numerical parameters (henceforth *sensitivity parameters*), with uncertainties about their true values, then the problem devolves to some sort of *sensitivity analysis*. A plausible setting of the sensitivity parameters, combined with the information from the data, yields a particular inference about a target parameter. Of course changing to an alternate but equally plausible setting of the sensitivity parameters leads to a quantitatively different inference. It is important to know whether the change in inference is slight or substantial.

To focus the discussion, consider the realm of observational studies, particularly in the health sciences domain. Often such studies are faced with one or more *threats to validity*, with the nature and extent of the threat governed by one or more sensitivity parameters. For instance, the sampling scheme may be systematically biased for/against potential study subjects having certain attributes, with sensitivity parameters describing this bias. Or a variable might be prone to missingness, with sensitivity parameters describing how the chance of missingness depends on the underlying value of the variable itself. Or a variable might be prone to measurement error, with sensitivity parameters describing the magnitude of this error. In all these circumstances, unless one additionally has access to some "threat-free" data (i.e., some validation data), we must resort to sensitivity analysis.

As an extremely simple example, say we wish we had threat-free data in the form of n i.i.d. realizations of Y, with mean $\psi = E(Y)$ being the target of inference. However, selection bias has a corrupting influence, forcing us to draw data from the distribution of $(Y \mid S = 1)$, for a binary selection variable S. In actuality then, our i.i.d. datapoints have mean $\phi = \psi - \lambda$, where $\lambda = E(Y) - E(Y \mid S = 1)$ governs the impact of the selection bias. We then think of our inference process as involving an amalgamation of what the data say about ϕ along with what justifiable sensitivity assumptions say about λ, in order to infer $\psi = \phi + \lambda$. Here the target parameter is a straight sum of a term informed by the data and a term not informed by the data, yielding intuitive headway about how inference works. In more complex situations, however, some definitions and theory can provide helpful guidance.

One common and general strategy for sensitivity analysis could be referred to as the "tabular" method. A discrete set of plausible values is specified for each sensitivity parameter involved in the problem. This generates a number of different scenarios, usually in a factorial manner presuming there is more than one sensitivity parameter at hand. The inferences arising under all these scenarios are then simply reported in a (typi-

Paul Gustafson is Professor, Department of Statistics, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada (e-mail: gustaf@stat.ubc.ca). Lawrence C. McCandless is Associate Professor, Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada (e-mail: lmccandl@sfu.ca).

cally large) table. Methods for tabular sensitivity analysis that have attracted considerable attention include Rosenbaum and Rubin (1983), Greenland (1996), Lin, Psaty and Kronmal (1998).

A rather different approach is the "how do I break this" strategy. That is, one has a baseline scenario in mind, often involving the presumption that there are no threats to validity. Then one seeks values of the sensitivity parameters that produce a qualitatively different inference than the baseline inference. For instance, if the baseline analysis yields a statistically significant association, then the task is to compute values of the sensitivity parameters at which the result "tips" from significant to nonsignificant. Then the question is whether or not these values are so extreme as to be implausible. For overviews of this approach, see Phillips (2003), Rosenbaum (2002, 2010).

A rather principled approach to sensitivity analysis is via Bayesian inference. The plausible set of sensitivity parameter values used in tabular analysis gives way to a joint prior distribution over the sensitivity parameters. Then Bayesian updating is applied as usual, to combine what this prior distribution says with what the data say. The advantage here is that the posterior distribution over a target parameter seamlessly amalgamates two sources of uncertainty: the usual statistical uncertainty due to sampling (i.e., the result of having a finite rather than infinite sample), plus the uncertainty about the values of the sensitivity parameters. Examples of Bayesian sensitivity analysis include Scharfstein, Daniels and Robins (2003), McCandless, Gustafson and Levy (2007), Daniels and Hogan (2008), Gustafson et al. (2010), Geneletti et al. (2013).

It is also worth noting that a "near-Bayesian" approach to sensitivity analysis has attracted a lot of attention, particularly in the epidemiology literature. Often referred to as *probabilistic sensitivity analysis* or *Monte Carlo sensitivity analysis*, this also starts with assigning a distribution over plausible ranges to the sensitivity parameters. However, the combination of this distribution with the information in the data is carried out in a "simpler-than-Bayesian" fashion. Rudimentary versions of this approach are somewhat like tabular sensitivity analysis, except a very large number of scenarios are constructed by sampling from the prior distribution, and a summary of the induced distribution of an inferential quantity (say a point estimate) is reported. More nuanced versions take sampling variability into account, that is, each random draw from the prior distribution of the sensitivity parameters is paired with a random draw of the nonsensitivity parameters, using a data-determined distribution that appropriately characterizes sampling variability. Some key references for these approaches generally are Greenland (2003, 2005), Lash, Fox and Fink (2009), while MacLehose and Gustafson (2012) offer some contrasts with a fully Bayesian approach.

Most problems where sensitivity analysis is required come with considerable baked-in intuition: it seems obvious which parameters are the sensitivity parameters and which are not. That is, strong intuitions are usually present about which parameters cannot be informed by the study data, and these are treated as the sensitivity parameters. In the present paper, and working in the Bayesian framework, we explore whether slightly more formality leads to clearer understanding. Is it worth thinking more formally rather than intuitively about what it means for a quantity to be a sensitivity parameter? Our consideration of this question leads to a definition of when a sensitivity parameter is *pure*, in terms of being completely uninformed by the observable data. We also offer a graphical summary of the extent to which an "impure" sensitivity parameter deviates from purity. The definition is useful, since if we know we are dealing with a pure sensitivity parameter, then we know there is equivalence between the quality of external information about the sensitivity parameter and the quality of the posterior inference on the target parameter. And the graphical summary is useful, since if it reveals only slight impurity in the sensitivity parameter, then we know a rough version of the aforementioned equivalence still applies.

2. DETAILS

In general, consider a statistical model for n observable datapoints D_n that is parameterized by a vector θ of p unknown parameters within a parameter space Θ. We say that a particular *context* for this model arises when we declare (i) a particular choice of prior distribution π for θ, and (ii) a particular choice of scalar target parameter $\psi = g(\theta)$. A putative q-dimensional sensitivity parameter is expressed as $\lambda = h(\theta)$. In what follows, a "dagger" notation is used, when helpful, to indicate true parameter values. That is, we use θ, ψ and λ to refer to parameters generically, and to discuss the structure of prior and posterior distributions. But when describing what happens when data are manifested, we use $\theta^\dagger$ to indicate the true value of θ that spawns these data. Commensurately, $\psi^\dagger = g(\theta^\dagger)$, and $\lambda^\dagger = h(\theta^\dagger)$.

Our starting point is a definition given, but not pursued, in Section 6.1 of Gustafson (2015). Stated slightly more formally now, this goes as follows.

DEFINITION 1. For a particular context (choice of prior π and scalar target ψ), $\lambda = h(\theta)$ is a *pure sensitivity parameter* provided there exists a reparameterization from θ to (ϕ, λ) satisfying the following four conditions:

C1. The distribution of the observable data D_n given (ϕ, λ) does not vary with λ.

C2. The distribution of D_n given ϕ constitutes an identified statistical model supporting consistent estimation of ϕ.

C3. A priori, ϕ and λ are independent of one another.

C4. Expressed as a function of (ϕ, λ), the target parameter ψ is strictly monotone in each component of λ.

As a first remark about the definition, a reparameterization satisfying C1 and C2 is referred to as a *transparent* reparameterization in Gustafson (2005). An immediate consequence of C1 is that $\pi(\lambda \mid \phi, d_n) = \pi(\lambda \mid \phi)$, that is, regardless of what data are observed, the posterior conditional distribution of $(\lambda \mid \phi)$ is the same as the prior conditional distribution. Provided that C1 and C2 both hold then, we have a very simple characterization of what happens to the posterior distribution as the sample size grows. The *limiting posterior distribution* (LPD) for (ϕ, λ) is formed by a point-mass distribution for ϕ at the true value $\phi^\dagger$, coupled with the prior conditional distribution of λ given $\phi = \phi^\dagger$. And this distribution induces the limiting posterior distribution of θ via transformation back to the original parameterization.

As a second remark, taken together C1 and C3 imply the prior and posterior marginal distributions of λ are identical—a very strong sense of the data having nothing to say about λ. Indeed, this alone seems like quite a "pure" sense in which λ plays the role of a sensitivity parameter, so the additional need for C4 is perhaps not so obvious. Before discussing this in detail, however, we make the following definition.

DEFINITION 2. Consider the case that $q = 1$, that is, λ is a scalar parameter. For a given context (choice of prior and scalar target of inference), let

$$p_{\text{pri}}(\lambda^\dagger) = \Pr_\pi\{\lambda < \lambda^\dagger\} \tag{1}$$

indicate where the truth about λ lies within the prior distribution. Similarly, let

$$p_{\text{pst}}(\theta^\dagger) = \lim_{n\to\infty} \Pr_\pi\{\psi < \psi^\dagger \mid D_n\} \tag{2}$$

indicate where the truth about the target parameter lies in the posterior distribution, in the large-sample limit. We say that λ is *quantile-preserving* for the context if, for every $\theta^\dagger \in \Theta$, either $p_{\text{pst}}(\theta^\dagger) = p_{\text{pri}}(\lambda^\dagger)$ or $p_{\text{pst}}(\theta^\dagger) = 1 - p_{\text{pri}}(\lambda^\dagger)$.

Stated intuitively, if λ is a quantile-preserving sensitivity parameter, then the extent to which the true value of the target parameter is in the tails (versus the center) of the posterior distribution is exactly the same as the extent to which the true value of the sensitivity parameter is in the tails (versus the center) of the prior distribution. For instance, the (limiting) 95% equal-tailed posterior credible interval for the target will contain the truth if and only if the 95% equal-tailed prior credible interval for the sensitivity parameter contains the truth. (Or replace 95% here with your own favorite coverage level.) More colloquially, if quantile-preservation holds, the veracity of the posterior inference on the target is governed completely by the veracity of the prior distribution for the sensitivity parameter.

Armed with definitions for when a sensitivity parameter is pure and when a sensitivity parameter is quantile-preserving, we have the following theorem.

THEOREM 1. *Consider a context* (*choice of prior* π *and scalar target* ψ) *for a statistical model. Assume that* π *is absolutely continuous. If a scalar parameter* λ *is a pure sensitivity parameter in this context, then* λ *is also quantile-preserving in this context.*

PROOF. Assuming λ is pure, let (ϕ, λ) be a parameterization satisfying C1 through C4. From C1 through C3, the limiting posterior distribution on ψ must be the distribution on $g(\theta(\phi^\dagger, \lambda))$ induced by the (marginal) prior distribution of λ. By C4 this transformation is strictly monotone. Therefore, if $\lambda^\dagger$ lies at the δth quantile of the prior distribution on λ, then $\psi^\dagger = g(\theta(\phi^\dagger, \lambda^\dagger))$ must lie at either the δth or $(1-\delta)$th quantile of the limiting posterior distribution on ψ. □

Taking stock, we now have a good handle on the characteristics that stem from Definition 1. A pure sensitivity parameter has exactly the same marginal posterior and prior distributions. Also, a scalar pure sensitivity parameter is such that the veracity of its prior distribution completely governs the veracity of the posterior distribution for the target parameter (in the large-sample limit as the role of random-sampling variation dissipates).

According to Definition 1, the "pure sensitivity parameter" moniker is bestowed in a very context-specific manner. Given only a parametric model and

a candidate sensitivity parameter λ, we cannot resolve the question of whether λ is pure. The question is only meaningful in the context of a specific prior distribution for θ and a particular scalar target parameter. Indeed, later in this paper we encounter a situation where a seemingly innocuous change to the prior distribution impacts whether λ meets the conditions for purity. Similarly, we encounter a situation where a seemingly innocuous change in the choice of target parameter has the same impact.

2.1 Impurity Plots

The preceding discussion suggests a way to visualize the characteristics of a putative scalar sensitivity parameter. Particularly, the quantities (1) and (2) suggest the use of a probability–probability plot to summarize the nature of a putative sensitivity parameter λ in a given context (choice of prior distribution and target parameter). Specifically, for an ensemble of $\theta^\dagger$ values, we plot points $\{p_{\text{pri}}(\lambda^\dagger), p_{\text{pst}}(\theta^\dagger)\}$ in the form of a scatterplot. (The most obvious choice of ensemble is a suite of values drawn from the joint prior distribution of θ.) The aggregate extent to which the plotted points deviate from the diagonal reference lines ($p_{\text{pst}} = p_{\text{pri}}$ and $p_{\text{pst}} = 1 - p_{\text{pri}}$) then reflects the extent to which λ is not quantile-preserving (and therefore not pure) as a sensitivity parameter. In particular, a point that is far from the diagonals must correspond to the centrality of $\lambda^\dagger$ in the prior for λ differing greatly from the centrality of $\psi^\dagger$ in the limiting posterior distribution for ψ. Such a point corresponds to "more going on" than the quality of posterior inference on the target being driven directly by the quality of prior assertion on the putative sensitivity parameter. Henceforth, we refer to a scatterplot constructed as above as an *impurity plot*.

The impurity plot generalizes readily to the situation that λ is comprised of two or more parameters, at least in situations where ϕ and λ are variation-independent of one another (i.e., the parameter space is the product space of the possible values for ϕ crossed with the possible values for λ). We generalize (1), which reflects the a priori extremity of a scalar λ, to

$$\begin{aligned} p^*_{\text{pri}}(\theta^\dagger) &= \Pr_\pi\{\psi(\phi^\dagger, \lambda) < \psi(\phi^\dagger, \lambda^\dagger)\} \\ &= \Pr_\pi\{\psi(\phi^\dagger, \lambda) < \psi^\dagger\}, \end{aligned} \tag{3}$$

which reflect the a priori extremity of a vector λ *in the direction that determines the target parameter*. So the generalized procedure is to plot $\{p^*_{\text{pri}}(\theta^\dagger), p_{\text{pst}}(\theta^\dagger)\}$ pairs for an ensemble of $\theta^\dagger$ values, where again a natural choice of ensemble is a sample from the prior distribution of θ. This version of the impurity plot retains the property that if λ is a pure bias parameter then all points fall on the diagonal reference lines.

3. EXAMPLES

3.1 Example A: Estimating Exposure-Disease Association in the Face of Misclassified Exposure

Say that interest lies in estimating the association between binary exposure X and binary disease status Y, however the observable data are of the form (X^*, Y), where X^* is a possibly misclassified surrogate for X. For simplicity, we assume nondifferential misclassification, so that X^* and Y are conditionally independent given X. We also assume data arise via case-control sampling, ergo the data directly inform us about the $(X^* \mid Y)$ distribution.

We let γ describe the exposure classification scheme according to $\gamma_x = \Pr(X^* = x \mid X = x)$, for $x = 0, 1$. That is, γ_0 and γ_1 are respectively the *specificity* and *sensitivity* of X^* as a surrogate for X. For the first version of this problem, we assume that the classification scheme is known to never produce false positives, that is, $\gamma_0 = 1$ is known a priori. However, γ_1 is taken as an unknown parameter, as the extent to which X^* has imperfect sensitivity is unknown. Thus, (r_0, r_1, γ_1) is an initial parameterization for this problem, where $r_y = \Pr(X = 1 \mid Y = y)$. And $\psi = \log \text{OR}(Y, X) = \text{logit}(r_1) - \text{logit}(r_0)$ can be taken as the target parameter.

For this problem, a reasonable prior specification might take the three parameters as independent, with $r_y \sim U(0, 1)$ for $y = 0, 1$, and $\gamma_1 \sim \text{Beta}(a_1, b_1)$. (For a technical reason soon to be apparent, we presume $a_1 > 2$.) This corresponds to being noninformative concerning the (X, Y) association, but drawing upon expert knowledge to arrive at a defensible choice of hyperparameters (a_1, b_1), thereby committing to plausible prior information about the extent of the exposure misclassification. Clearly the classification parameter γ_1 is the putative sensitivity parameter in this setting.

For $y = 0, 1$, let $r^*_y = \Pr(X^* = 1 \mid Y = y) = r_y\gamma_1$. One can quickly verify that $\phi = (r^*_0, r^*_1)$, $\lambda = \gamma_1$ comprises a transparent parameterization, with the target parameter expressed as

$$\psi = \log\left(\frac{r^*_1}{\gamma_1 - r^*_1}\right) - \log\left(\frac{r^*_0}{\gamma_1 - r^*_0}\right). \tag{4}$$

Letting $g(\cdot; a, b)$ denotes the Beta(a, b) density function, the prior density in the initial parameterization is

$$\pi(\gamma_1, r_0, r_1) = g(\gamma_1; a_1, b_1) I_{(0,1)}(r_0) I_{(0,1)}(r_1).$$

Upon transforming to obtain the prior density of (γ_1, r_0^*, r_1^*), we can read off the conditional prior density of γ_1 given $r^* = (r_0^*, r_1^*)$ to be

$$\pi(\gamma_1 \mid r^*) \propto g(\gamma_1; a_1 - 2, b_1) I_{A_1(r^*)}(\gamma_1), \tag{5}$$

where the interval $A_1(r^*) = (\max\{r_0^*, r_1^*\}, 1)$ is the support of this distribution. This characterizes the LPD. So we see the update from the prior to limiting posterior distribution of γ_1 involves both a change in the Beta hyperparameters and truncation. Moreover, (5) clearly reveals that C3 fails, since γ_1 and r^* are a priori dependent (and this dependence is "structural," being driven by the fact that γ_1 and r^* are variation-dependent). In the present context then, γ_1 is not a pure sensitivity parameter.

To gauge the lack of purity of γ_1, we produce an impurity plot. Since (4) is monotone in γ_1 over $A_1(r^*)$, computing $p_{\text{pst}}(\theta^\dagger)$ reduces to computing the probability of a subinterval of $A_1(r^{*\dagger})$ under the truncated Beta prior distribution (5) (the subinterval below $\gamma_1^\dagger$ if $r_0^{*\dagger} < r_1^{*\dagger}$, the subinterval above otherwise). As an example, consider hyperparameters $(a_1, b_1) = (17, 3)$, a situation where a priori the misclassification is assumed to be mild. For an ensemble of $\theta^\dagger$ values drawn from the prior distribution, the impurity plot appears in Figure 1. While many points do lie close to the reference lines, we see it is possible for $\gamma_1^\dagger$ to be near the middle of the prior distribution on γ_1, while at the same time $\psi^\dagger$ lies in the tail of the limiting posterior distribution. This happens in circumstances where the truncation takes a particularly large "bite" in moving from the prior to limiting posterior distribution of γ_1. For instance, if $r_1^* > r_0^*$, then (4) is decreasing in γ_1. Should

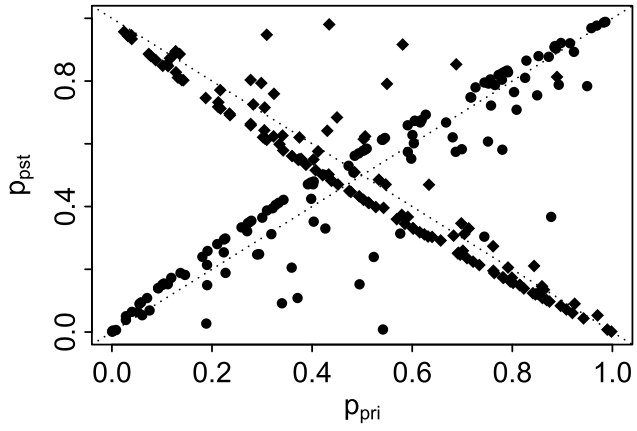

FIG. 1. *Impurity plot for Example* A. *The points correspond to* 250 *draws from the prior distribution, with hyperparameters* $(a_1, b_1) = (17, 3)$. *Note that points for which* r_1 *is greater* (*less*) *than* r_0 *are plotted as diamonds* (*circles*).

the true value of γ_1 be in the middle of its prior distribution but also happen to lie only a bit above the truncation point $\max\{r_0^*, r_1^*\} = r_1^*$, then (4) will map this a posteriori left-tail value for γ_1 into an a posteriori right-tail value for ψ. Overall then, γ_1 is not particularly pure as a sensitivity parameter.

An obvious extension to this example arises when both classification parameters, sensitivity and specificity, are unknown. This proves somewhat challenging, in that λ has two components, but is *not* variation-independent of ϕ. So neither (1) nor (3) applies. However, it turns out we can customize (3) to deal with this situation. We take this up later in Section 3.4.

3.2 Example B: Estimating Prevalence in the Face of Nonignorable Missingness

Say that interest lies in estimating the prevalence of a binary trait Y, that is, $\psi = E(Y) = \Pr(Y = 1)$. However, random sampling from the population results in Y being unobserved for some sampled subjects. In contrast, an auxiliary binary variable X can be obtained for all those sampled. Taking a "pattern mixture model" approach to the missing data problem, and letting $R = 1$ and $R = 0$ indicate observation and missingness of Y respectively, the available data can be regarded as i.i.d. realizations of (X, R, YR).

Let $s = \Pr(X = 1)$, let $t_x = \Pr(R = 1 \mid X = x)$, and let $u_{rx} = \Pr(Y = 1 \mid R = r, X = x)$. Then the seven parameters $\theta = (s, t, u)$ clearly characterize the joint distribution of (X, R, Y) and can be taken as an initial parameterization for this problem.

Prior specification (i): A first prior specification arises by cleaving the elements of θ into $\phi^{(1)} = (s, t_0, t_1, u_{10}, u_{11})$ and $\lambda^{(1)} = (u_{00}, u_{01})$. This satisfies C1 and C2. Moreover, since $\phi^{(1)}$ and $\lambda^{(1)}$ are variation independent, one can legitimately choose a prior distribution for θ under which $\phi^{(1)}$ and $\lambda^{(1)}$ are independent, that is, one can opt to satisfy C3. Finally, the target parameter

$$\psi = \sum_{x=0}^{1} \sum_{r=0}^{1} s^x (1-s)^{1-x} t_x^r (1-t_x)^{1-r} u_{rx}$$

is clearly coordinate-wise monotonic in $\lambda^{(1)}$ for fixed $\phi^{(1)}$. Therefore C4 holds, and $\lambda^{(1)}$ meets the criteria for being a pure sensitivity parameter.

With a priori independence of $\phi^{(1)}$ and $\lambda^{(1)}$, it is not possible to express a probabilistic bound on the extent to which the *missing at random* (MAR) assumption is violated. That is, we do not have direct control in the prior distribution over the extent to which (u_{00}, u_{01})

might deviate from (u_{10}, u_{11}). This leads us to the following alternative.

Prior specification (ii): Let $\phi^{(2)} = (s, t_0, t_1, u_{10}, u_{11})$ (exactly the same as $\phi^{(1)}$), but now consider $\lambda^{(2)} = (\delta_0, \delta_1)$, where

$$\delta_x = \log \mathrm{OR}(Y, R \mid X = x).$$

Straightforward exercises verify that the map from $(\phi^{(1)}, \lambda^{(1)})$ to $(\phi^{(2)}, \lambda^{(2)})$ is invertible, and that $\phi^{(2)}$ and $\lambda^{(2)}$ are variation independent. The latter property allows us to assert a priori independence of $\phi^{(2)}$ and $\lambda^{(2)}$ if we wish. Moreover, expressed as a function of $(\phi^{(2)}, \lambda^{(2)})$, ψ is readily seen to be coordinate-wise monotonic in $\lambda^{(2)}$. Thus, $\lambda^{(2)}$ is a pure sensitivity parameter under this specification. In addition, this prior lends itself to probabilistically bounding the extent of deviation from the MAR assumption of $\lambda^{(2)} = (0, 0)$. For instance, upon assigning a mean zero bivariate normal distribution to $\lambda^{(2)}$, the choice of covariance matrix describes the potential magnitude of departure from MAR.

One perhaps troubling aspect of prior specification (ii) is a lack of symmetry. The distribution of $(Y \mid R = 1, X)$ appears explicitly, while the distribution of $(Y \mid R = 0, X)$ is entirely implicit. A more symmetric prior construction is as follows.

Prior specification (iii): Let $v_x = \Pr(Y = 1 \mid X = x)$. Then one could work with $\theta^{(3)} = (s, t_0, t_1, v_0, v_1, \delta_0, \delta_1)$ as a parameterization, and it would be natural to assert a priori independence between (s, t_0, t_1, v_0, v_1) and (δ_0, δ_1). From here, a transparent parameterization takes the form

$$\phi^{(3)} = \{s, t_0, t_1, u_{10}(t_0, v_0, \delta_0), u_{11}(t_1, v_1, \delta_1)\},$$

along with $\lambda^{(3)} = (\delta_0, \delta_1)$, where $u_{1x}(t_x, v_x, \delta_x)$ is defined implicitly as the solution to

$$v_x = (1 - t_x)\,\mathrm{expit}\{\mathrm{logit}(u_{1x}) - \delta_x\} + t_x u_{1x},$$

for $x = 0, 1$. Applying the change of variables, the prior density of $(\phi^{(3)}, \lambda^{(3)})$ takes the form

$$\begin{aligned} &\pi\big(\phi^{(3)}, \lambda^{(3)}\big) \\ &\quad = \pi_{s,t,v}\big(s, t, v(t, u, \delta)\big)\pi_\delta(\delta) \\ &\quad\quad \times \prod_{x=0}^{1} \left\{ t_x \right. \\ &\quad\quad \left. + (1 - t_x)\frac{w[\mathrm{expit}\{\mathrm{logit}(u_x) - \delta_x\}]}{w(u_x)} \right\}, \end{aligned} \tag{6}$$

where $w(\,)$ is simply defined as $w(p) = p(1 - p)$.

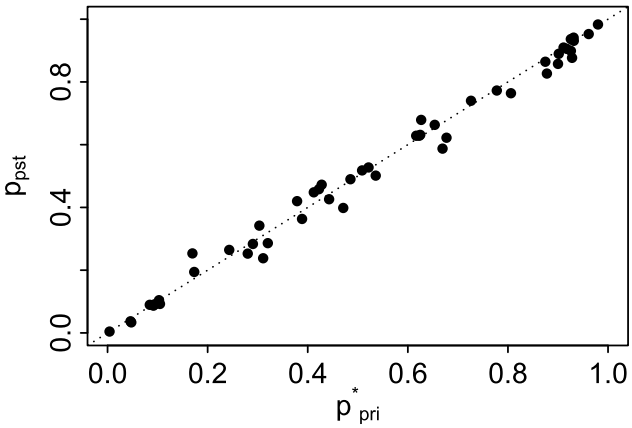

FIG. 2. *Impurity plot for Prior* (iii) *in Example* B. *The points correspond to* 50 *draws from the prior distribution, with hyperparameter* $\sigma_d = 0.35$.

Clearly C3 is violated in this situation, that is, $\phi^{(3)}$ and $\lambda^{(3)}$ are not independent of one another in the prior distribution (6). But the extent and impact of the dependence is not intuitively clear from the mathematical form of (6). So we proceed to an impurity plot to examine the extent to which $\lambda^{(3)}$ is not a pure bias parameter. A simple example ensues when all seven components of $\theta^{(3)}$ are assumed mutually independent in the prior, with $U(0, 1)$ distributions for the five probabilities (s, t_0, t_1, v_0, v_1), and $N(0, \sigma_\delta^2)$ priors for the two association parameters (δ_0, δ_1). We can "read off" the difference between the limiting posterior and prior densities for (δ_0, δ_1) from (6), that is,

$$\begin{aligned} &\frac{\pi(\delta \mid \phi^\dagger)}{\pi(\delta)} \\ &\quad \propto \prod_{x=0}^{1} \left\{ t_x^\dagger + (1 - t_x^\dagger)\frac{w[\mathrm{expit}\{\mathrm{logit}(u_{1x}^\dagger) - \delta_x\}]}{w(u_x^\dagger)} \right\}. \end{aligned}$$

Given this form, for a given $\theta^\dagger$ it is a simple exercise in two-dimensional numerical integration to compute both $p^*_{\mathrm{pri}}(\theta^\dagger)$ as per (3) and $p_{\mathrm{pst}}(\theta^\dagger)$ as per (2). Taking the hyperparameter value $\sigma_\delta = 0.35$, this is done for an ensemble of $\theta^\dagger$ values drawn from the prior distribution, with the resulting impurity plot given in Figure 2. Here the points are quite clustered around the reference line, indicating that $\lambda^{(3)}$ is only "mildly impure" as a sensitivity parameter.

3.3 Example C: Estimation in the Face of a Hidden Subpopulation

Here we elaborate on a problem studied by Xia and Gustafson (2012), Xia and Gustafson (2014), as motivated by the use of venue sampling to make inferences

about hard-to-reach populations. Say data are available in the form of i.i.d. realizations of (X, Y, Q), where X and Y are binary traits of interest, while Q is positive and continuous. In fact, each realization arises from weighted sampling of the population, where a unit's selection probability is proportional to Q. In the venue sampling context, for example, Q can be observed by asking sampled individuals about their proclivity to attend venues where sampling occurs. The plot thickens, however, in situations where some population members are hidden. That is, an unknown proportion p of population members have $Q = 0$, hence will never be sampled. In the venue sampling context, for instance, a proportion p of the population are never-attenders at all venues where sampling is undertaken. Intuitively, at least, p feels like a sensitivity parameter. Nothing in the observable data speaks directly to whether the hidden proportion is small or large.

To organize inference for this problem, we use $f(x, y, q)$ to denote the joint density *from which the observed data realizations arise*, noting that by construction this density puts all its mass on $Q > 0$. Against this, the actual distribution of (X, Y, Q) across the population of interest is given by

$$\begin{aligned}&\tilde{f}(x, y, q)\\ &\quad = \Big\{ p\delta_0(q)\\ &\qquad + (1-p)\frac{q^{-1} f(q) I_{(0,\infty)}(q)}{E_f(Q^{-1})} \Big\} f(x, y|q),\end{aligned} \tag{7}$$

where $\delta_0(\,)$ is the Dirac delta function, that is, the "density" of a point-mass distribution at zero. Importantly, here the conditional distribution of $(X, Y \mid Q = q)$ is assumed to be continuous in q at $q = 0$, so that the law of $(X, Y \mid Q = 0)$ can be learned from the observable data. In the venue sampling context, this continuity assumption corresponds to never attenders "looking like" the limiting case of ever-more-seldom attenders.

Directly letting ϕ parameterize $f(x, y, q)$ while setting $\lambda = p$, we immediately have a transparent parameterization, that is, C1 and C2 hold. Moreover, in light of the interpretation of p as the proportion of the population that is hidden, it seems quite reasonable to make independent prior assertions about ϕ and p, so we can plausibly choose a prior distribution for which C3 holds. (In fact, it seems hard to imagine any intuitively sensible way in which a priori dependence between p and ϕ could be introduced.) Thus, whether or not p is a pure sensitivity parameter for a given target parameter of interest comes down to condition C4.

If the target parameter of interest is a population mean with respect to $\tilde{f}$, then the monotonicity of (7) with respect to p for fixed $f(\,)$ immediately ensures C4 is satisfied. So for target parameters such as $E_{\tilde{f}}(X)$ or $E_{\tilde{f}}(Y)$, the population prevalence of X or of Y, p is a pure sensitivity parameter.

On the other hand, say the association between X and Y in the target population is of interest. For instance, the log odds-ratio,

$$\psi = \log \frac{E_{\tilde{f}}\{(1-X)(1-Y)\} E_{\tilde{f}}\{XY\}}{E_{\tilde{f}}\{(1-X)Y\} E_{\tilde{f}}\{X(1-Y)\}}, \tag{8}$$

might be targeted. While each of the four constituent terms on the right-hand side of (8) is monotone in p for fixed $f(\,)$, this does not guarantee monotonicity of (8) as a whole.

To explore this situation, say that $f(\,)$, the observable distribution of (X, Y, Q), is parameterized by $\phi = (a, b, \gamma^{(0)}, \gamma^{(1)})$ according to $Q \sim \text{Beta}(a, b)$ [where necessarily $a > 1$, to ensure (7) is well defined], while

$$\Pr(X = x, Y = y \mid Q) = (1-Q)\gamma_{xy}^{(0)} + Q\gamma_{xy}^{(1)}.$$

Under this specification, the terms in (8) have simple structure. For instance,

$$E_{\tilde{f}}(XY) = p\gamma_{11}^{(0)} + (1-p)\frac{k_{a,b}(-1, 1)\gamma_{11}^{(0)} + \gamma_{11}^{(1)}}{k_{a,b}(-1, 0)},$$

where $k_{a,b}(c, d) = E\{B^c(1-B)^d\}$ for $B \sim \text{Beta}(a, b)$.

For illustration, say the available prior information about the relative magnitude of the hidden population is $p \sim \text{Beta}(2, 18)$. This corresponds to a prior mean of 0.1, a prior mode of 0.056, and a prior 95th percentile of 0.226. We complete the prior specification with $a \sim \text{Unif}(1, 3)$, $b \sim \text{Unif}(0, 2)$, and $\gamma^{(i)} \sim \text{Dirichlet}(2, 2, 2, 2)$, for $i = 0, 1$. Generating an ensemble of parameter values from this prior distribution yields the impurity plot in Figure 3. The resulting behaviour is rather curious. The vast majority of points lie on a reference line, but there are a few exceptions. This arises as for most, but not all, values of $\phi^\dagger = (a^\dagger, b^\dagger, \gamma^{(0)\dagger}, \gamma^{(1)\dagger})$, ψ is monotone in p.

To elaborate, Figure 4 explores the situation for three "cherry-picked" exceptional values for $\phi^\dagger$ under which ψ is indeed not monotone in p. These underscore the possibility of a true value of p in the middle of the prior distribution producing a true value of the target ψ in the tail of the posterior. Conversely, the plots also show it possible that a value of p in the tail of the prior gives rise to a value of ψ in the mid-range of the posterior. The intuition that the quality of the inference on the target is driven exclusively by the quality of the prior information about the hidden proportion, while good in most instances, can fail.

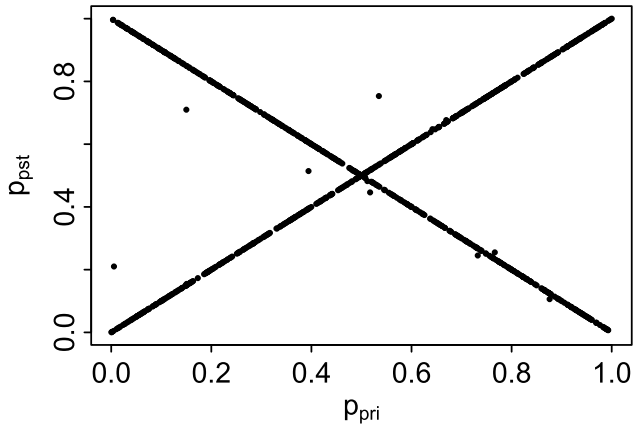

FIG. 3. *Impurity plot for Example* C. *The* 1000 *parameter values are generated from the prior distribution described in the text.*

3.4 Example A, Continued

We now return to Example A from Section 3.1. Now, however, we extend to both exposure classification parameters being unknown. Recall these parameters are the sensitivity $\gamma_1 = \Pr(X^* = 1 \mid X = 1)$ and the specificity $\gamma_0 = \Pr(X^* = 0 \mid X = 0)$. We extend the earlier prior distribution as

$$\begin{aligned}&\pi(r_0, r_1, \gamma_1, \gamma_0)\\ &\quad \propto g(r_0; 1, 1)g(r_1; 1, 1)\\ &\qquad \times g(\gamma_1; a_1, b_1)g(\gamma_0; b_0, b_0)I\{\gamma_1 + \gamma_0 > 1\},\end{aligned}$$

where the truncation is to the region where the classification scheme is better than random. Following Gustafson, Le and Saskin (2001), the LPD is then characterized by

$$\begin{aligned}&\pi(\gamma_1, \gamma_0 \mid r^*)\\ &\quad \propto \frac{g(\gamma_1; a_1, b_1)g(\gamma_0; b_0, b_0)I_{A_1(r^*)}(\gamma_1)I_{A_0(r^*)}(\gamma_0)}{(\gamma_1 + \gamma_0 - 1)^2},\end{aligned} \tag{9}$$

where $A_0(r^*) = (1 - \min\{r_0^*, r_1^*\}, 1)$, while, as before, $A_1(r^*) = (\max\{r_0^*, r_1^*\}, 1)$. Thus again updating from the marginal prior to limiting posterior of (γ_1, γ_0) involves both a change in shape of density and truncation. Computations associated with the LPD (9) can be handled with two-dimensional numerical integration.

In the present situation the putative sensitivity parameter $\gamma = (\gamma_1, \gamma_0)$ is bivariate, and is *not* variation-independent of $\phi = r^*$. Thus, neither (1) nor (3) can form the basis of an impurity plot. In particular, (3) is not well defined, since $\psi(\phi^\dagger, \lambda)$ is not well defined for

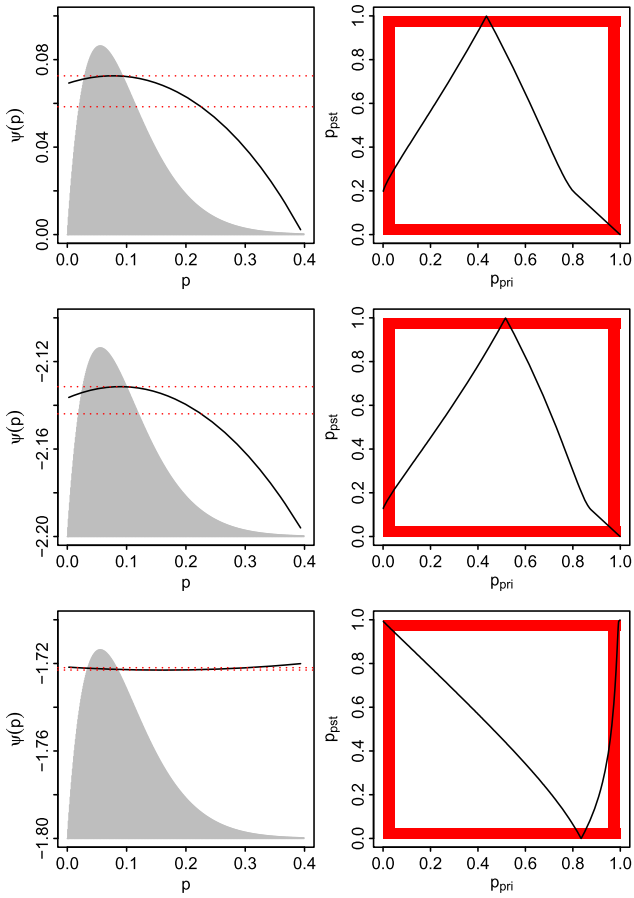

FIG. 4. *Target log-odds ratio* ψ *as a function of hidden proportion* p *with* $\phi^\dagger$ *fixed (left panels), and the corresponding restricted impurity plots (right panels). In the left panels, the prior distribution of* p *and the* 90% *equal-tailed limiting posterior credible interval for* ψ *(dotted horizontal lines) are superimposed. The right panels are impurity plots which are restricted in the sense that only* p *varies, with* $\phi^\dagger = (a^\dagger, b^\dagger, \gamma^{(0)\dagger}, \gamma^{(1)\dagger})$ *fixed. The red areas correspond to lowest* 5% *and highest* 5% *tails. Throughout,* $a^\dagger = 2$ *and* $b^\dagger = 1$. *In the first row,* $\gamma^{(0)\dagger}$ *and* $\gamma^{(1)\dagger}$ *are* $(0.53, 0.01, 0.36, 0.1)$ *and* $(0.07, 0.4, 0.22, 0.31)$, *respectively [using the order* $\gamma = (\gamma_{00}, \gamma_{11}, \gamma_{10}, \gamma_{01})$*]. In the second panel, these values are* $(0.02, 0.28, 0.33, 0.37)$ *and* $(0.19, 0.03, 0.29, 0.49)$. *In the third panel, these values are* $(0.04, 0.38, 0.22, 0.37)$ *and* $(0.1, 0.16, 0.61, 0.13)$.

all a priori plausible values of λ. However, there is a natural adaptation. Rather than (3) we now work with

$$p_{\text{pri}}^{**}(\theta^\dagger) = \Pr_\pi\{\psi_e(\phi^\dagger, \lambda) < \psi^\dagger\}, \tag{10}$$

for a suitable extension $\psi_e(\cdot, \cdot)$ of $\psi(\cdot, \cdot)$, to the product space of the prior support of ϕ crossed with the prior support of λ. The particular extension is motivated by the following observation. Say r^* is fixed, with $r_0^* < r_1^*$. Then the target ψ increases to positive infinity as γ_0 or γ_1 decreases to the lower boundary of the rectangle $A_0(r^*) \times A_1(r^*)$. That is, the a priori

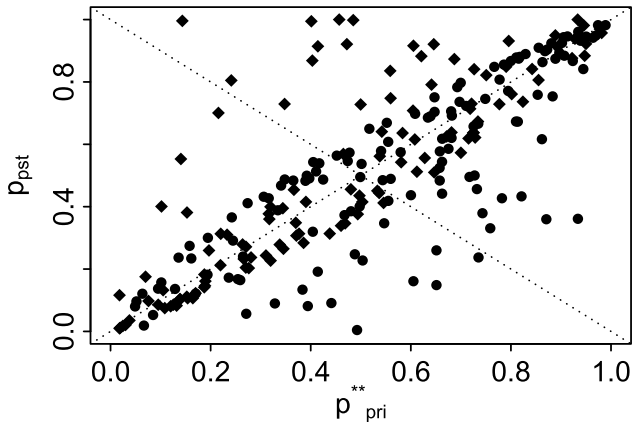

FIG. 5. *Impurity plot for the extended version of Example* A. *The points correspond to* 250 *draws from the prior distribution, with hyperparameters* $(a_1, b_1) = (b_0, b_0) = (17, 3)$. *Note that points for which* r_1 *is greater* (*less*) *than* r_0 *are plotted as diamonds* (*circles*).

plausible values of γ that are excluded a posteriori can be viewed as mapping to "beyond" the right tail of the posterior distribution. Similarly, if $r_1^* < r_0^*$, then ψ decreases to negative infinity as γ_1 and/or γ_0 decrease to the same lower boundary. This motivates

$$\psi_e(r^*, \gamma) = \begin{cases} \psi(r^*, \gamma) & \text{if } \gamma \in A(r^*), \\ +\infty & \text{if } \gamma \notin A(r^*), r_0^* < r_1^*, \\ -\infty & \text{if } \gamma \notin A(r^*), r_0^* > r_1^*, \end{cases} \tag{11}$$

where $A(r^*) = A_0(r^*) \times A_1(r^*)$.

Using the extended target (11) leads to the impurity plot in Figure 5. The wide scatter of points reflects the fact that γ is actually quite impure as a sensitivity parameter. In particular, much as was seen with the simpler version of this problem, the plot reflects the possibility that a priori nonextreme values of $\gamma^\dagger$ can give rise to a posteriori extreme values of $\psi^\dagger$. This can occur precisely because of the truncation inherent in the prior-to-posterior updating.

4. DISCUSSION

We are not the first to attempt to define what a sensitivity parameter is. For instance, Daniels and Hogan (2008) give a definition of a sensitivity parameter that is well adapted to missing data problems. While their definition has some elements in common with ours, it is not sufficiently detailed to distinguish between what we are calling "pure" and "impure" situations. And in fact much of the literature on sensitivity analysis really does rely on intuition concerning what constitutes a sensitivity parameter. It is common to see, for instance, either an explicit or implicit suggestion that a sensitivity parameter is simply something that, if fixed, renders all the other parameters identified. [A more explicit example of such a suggestion appears in Greenland (2005), whereas both Greenland (2003) and Lash, Fox and Fink (2009) are more implicit.]

One thought to take away from the present work is that *if* one can reasonably work with a pure sensitivity parameter, then there are advantages to doing so. One can proceed knowing that final inferences are drawn together from the data and prior in a simple, intuitive way: the prior on the pure sensitivity parameter will be crucial, while the prior on other parameters will get swamped by enough data, in the usual manner. And there can't be any "funny business" in terms of the quality of the prior information on the pure sensitivity parameter being upgraded or downgraded in its impact on the quality of the posterior distribution on the target parameter. However, the "if one can reasonably work" qualifier above cannot be ignored. For instance, the first prior specification in Example B yields a pure sensitivity parameter. However, while pure, it's not a very *useful* sensitivity parameter. Assigning a prior to this λ and asserting prior independence between λ and ϕ just does not work in a practical sense. It precludes the investigator from specifying a prior that sensibly and probabilistically limits the extent of departure from MAR. So going for purity at any cost is not a sensible strategy.

Another rationale for wanting a definition of a pure sensitivity parameter lies in the reality that intuition alone can be unreliable. And this applies notwithstanding one's level of statistical expertise. As is emphasized in Kahneman (2011), even experts can flunk out when asked to solve statistical problems by intuition alone (specifically see the introduction to this book for a description of empirical tasks on which it was shown that "even statisticians were not good intuitive statisticians"). Along these lines, we think that relying on intuition alone to determine the nature of a sensitivity parameter does present some dangers. For example, one might easily intuit in Example C that the proportion of the population that is hidden would be "informally" very pure as a sensitivity parameter. Without having a definition, and without delving into its applicability, one could easily be misled. That is, without a mathematical investigation of whether the target parameter is monotone in the sensitivity parameter, an important nuance could be missed. In fact, the prior information about the hidden proportion propagates through to the posterior distribution on the target in either a simple or

complicated way, depending on which target parameter is under consideration. In subtle situations such as these, having a definition, and knowing the implications of whether the putative sensitivity parameter meets the test of purity, seems like a step forward.

ACKNOWLEDGEMENTS

Research supported by a grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN 183772-13).

REFERENCES

DANIELS, M. J. and HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies*: *Strategies for Bayesian Modeling and Sensitivity Analysis*. *Monographs on Statistics and Applied Probability* **109**. Chapman & Hall/CRC, Boca Raton, FL. MR2459796

GENELETTI, S., BEST, N., TOLEDANO, M., ELLIOTT, P. and RICHARDSON, S. (2013). Uncovering selection bias in case–control studies using Bayesian post-stratification. *Stat. Med.* **32** 2555–2570. MR3067407

GREENLAND, S. (1996). Basic methods for sensitivity analysis of biases. *Int. J. Epidemiol.* **25** 1107–1116.

GREENLAND, S. (2003). 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. Amer. Statist. Assoc.* **98** 47–54. MR1977199

GREENLAND, S. (2005). Multiple-bias modelling for analysis of observational data. *J. Roy. Statist. Soc. Ser. A* **168** 267–306. MR2119402

GUSTAFSON, P. (2005). On model expansion, model contraction, identifiability, and prior information: Two illustrative scenarios involving mismeasured variables (with discussion). *Statist. Sci.* **20** 111–140. MR2490546

GUSTAFSON, P. (2015). *Bayesian Inference for Partially Identified Models*: *Exploring the Limits of Limited Data*. Chapman & Hall/CRC Press, Boca Raton, FL. MR3642458

GUSTAFSON, P., LE, N. D. and SASKIN, R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* **57** 598–609. MR1855698

GUSTAFSON, P., MCCANDLESS, L. C., LEVY, A. R. and RICHARDSON, S. (2010). Simplified Bayesian sensitivity analysis for mismeasured and unobserved confounders. *Biometrics* **66** 1129–1137. MR2758500

KAHNEMAN, D. (2011). *Thinking*, *Fast and Slow*. Macmillan, New York.

LASH, T. L., FOX, M. P. and FINK, A. K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer, New York.

LIN, D. Y., PSATY, B. M. and KRONMAL, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54** 948–963.

MACLEHOSE, R. F. and GUSTAFSON, P. (2012). Is probabilistic bias analysis approximately Bayesian? *Epidemiology* **23** 151–158.

MCCANDLESS, L. C., GUSTAFSON, P. and LEVY, A. R. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat. Med.* **26** 2331–2347. MR2368419

PHILLIPS, C. V. (2003). Quantifying uncertainty in epidemiologic studies. *Epidemiology* **14** 459–466.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

ROSENBAUM, P. R. and RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.

SCHARFSTEIN, D. O., DANIELS, M. J. and ROBINS, J. M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics* **4** 495–512.

XIA, M. and GUSTAFSON, P. (2012). A Bayesian method for estimating prevalence in the presence of a hidden sub-population. *Stat. Med.* **31** 2386–2398. MR2967761

XIA, M. and GUSTAFSON, P. (2014). Bayesian sensitivity analyses for hidden sub-populations in weighted sampling. *Canad. J. Statist.* **42** 436–450. MR3254036