

Interval Estimation for Messy Observational Data

Paul Gustafson
Department of Statistics
University of British Columbia
Vancouver, BC V6T 1Z2, Canada
gustaf@stat.ubc.ca

Sander Greenland
Departments of Epidemiology and Statistics
University of California
Los Angeles CA 90095-1772, USA
lesdomes@ucla.edu

March 25, 2009

SUMMARY. We review some aspects of Bayesian and frequentist interval estimation, focusing first on their relative strengths and weaknesses when used in ‘clean’ or ‘textbook’ contexts. We then turn attention to observational-data situations which are ‘messy,’ where modeling that acknowledges the limitations of study design and data collection leads to nonidentifiability. We argue, via a series of examples, that Bayesian interval estimation is an attractive way to proceed in this context even for frequentists, because it can be supplied with a diagnostic in the form of a calibration-sensitivity simulation analysis. We illustrate the basis for this approach in a series of theoretical considerations, simulations, and an application to a study of silica exposure and lung cancer.

KEY WORDS: Bayesian analysis; Bias; Confounding; Epidemiology; Hierarchical prior; Identifiability; Interval coverage; Observational studies.

1. Introduction

The conventional approach to observational-data analysis is to apply statistical methods that assume a designed experiment or survey has been conducted. In other words, they assume that all unmodeled sources of variation are randomized under the design. In most

settings, deviations of the reality from this ideal are dealt with informally in post-analysis discussion of study problems. Unfortunately, such informal discussion seldom appreciates the potential size and interaction of sources of bias, and as a consequence the conventional approach encourages far too much certainty in inference (Eddy, Hasselblad and Schachter, 1992; Greenland, 2005; Greenland and Lash, 2008; Molitor, Jackson, Best and Richardson, 2008).

The entrenchment of the conventional approach derives in part from the fact that realistic models for observational studies are not identified by the data, a fact which renders conventional methods and software useless (except perhaps as part of a larger fitting cycle). The most commonly proposed mode of addressing this problem is sensitivity analysis, which however leads to problems of dimensionality and summarization. The latter problems have in turn been addressed by Bayesian and related informal simulation methods for examining nonidentified models (which are often dealt with under the topic of nonignorability). These methods include hierarchical (multilevel) modeling of biases (Greenland, 2003, 2005), which is intertwined with the theme of the present paper.

We start in Section 2 by reviewing some notions of interval estimator performance, with emphasis on coverage averaged over *different* parameter values. Section 3 then extends this discussion to include intervals arising from *hierarchical* Bayesian analysis when data from multiple studies are at hand. These two sections reframe existing theory and results in a manner suited for our present needs. We emphasize a well-known tradeoff: To the extent the selected prior distribution is biased relative to reality, the coverage of a Bayesian posterior interval will be off, but perhaps not by much; and in return the intervals can deliver substantial gains in precision and reduced false-discovery rates compared to frequentist confidence intervals. In addition, hierarchical priors provide a means to reduce prior misspecification as studies unfold.

In section 4 we turn to the more novel aspect of our work, by studying the case which we believe better captures observational-study reality, in which priors are essential for identification. Here the usual order of robustness of frequentist vs. Bayesian procedures reverses: Confidence intervals become only extreme posterior intervals, obtained under degenerate priors, with coverage that rapidly deteriorates as reality moves away from

these point priors. In contrast, the general Bayesian framework with proper priors offers some protection against catastrophic undercoverage, with good coverage guaranteed under a spectrum of conditions specified by the investigator and transparent to the consumer. Section 5 summarizes the lessons we take away from our observations and makes a recommendation concerning the practical assessment of interval estimator performance. We conclude that Bayesian interval estimation is an attractive way to proceed even for frequentists, because its relevant calibration properties can be checked in each application via simulation analysis. We close with an illustration of our proposed practical approach in an application to a study of silica exposure and lung cancer in which an unmeasured confounder (smoking) renders the target parameter nonidentified.

2. The Well-Calibrated Lab

Let θ denote the parameter vector, and D the observable data, for a study that is to be carried out. Assume for now that the distribution of $(D|\theta)$ (i.e., ‘the model’) is known correctly. Say that $\phi=g(\theta)$ is the scalar parameter of interest, and that $I(D)$ is an interval estimator for this target. We define the *lab-wise coverage* (LWC) of I with respect to a *parameter-generating distribution* (PGD) P as

$$C(I,P) = \Pr\{\phi \in I(D)\}. \quad (1)$$

Here the probability is taken with respect to the distribution of (θ, D) jointly, with $\theta \sim P$ and $(D | \theta)$ following the model distribution.

Interval coverage with respect to a joint distribution on parameters and data, as in (1), has been considered by many authors, but not with a consistent terminology. While it might be tempting to refer to (1) as ‘Bayesian’ coverage, we find this confusing since (1) can be evaluated for Bayesian or non-Bayesian interval estimators. We choose to call it lab-wise coverage since $C(I,P)$ is the proportion of right answers reported by a lab or research team applying estimator I in a long series of studies of *different* phenomena (different exposure-disease relationships, say) within a research domain. The role of the PGD P is then to describe the corresponding across-phenomena variation in the underlying parameter values. Interest in lab-wise coverage might be very direct in some contexts, in that estimator operating characteristics in a long sequence of actual studies

really are the primary consideration. Or interest may be more oblique, in that performance on the ‘next’ study is of interest, and this performance is being measured conceptually by regarding the next study as a random draw from the population of ‘potential’ or ‘future’ studies.

If I is a frequentist confidence interval (abbreviated FCI), then it will attain nominal coverage exactly for any PGD. That is, if $\Pr\{\phi \in I(D) \mid \theta\} = 1 - \alpha$ for every value of θ , then $C(I,P)=1 - \alpha$ for any P . Thus correct coverage for a hypothetical sequence of studies with the same parameter values implies correct coverage in the more realistic setting of repeatedly applying a procedure in a sequence of differing real problems. While this fact is often viewed as a robustness property of an FCI, Bayarri and Berger (2004), citing Neyman (1977), emphasize that it is the lab-wise coverage that is relevant for practice. Put another way, if a lab is well-calibrated in the LWC sense of producing 95% intervals that capture the true parameter for 95% of studies, and the cost of failing to capture is the same across studies (as might be the case in some genome studies or screening projects), there is little obvious benefit if the intervals happen to also have correct frequentist coverage.

Bayesian Intervals under PGDs

For a given choice of prior distribution Π on the parameter vector θ , a $1 - \alpha$ Bayesian posterior credible interval (BPCI) for the target parameter ϕ would be any interval having Bayesian probability $1 - \alpha$ of containing ϕ given the observed data D . The most common choices of BPCI are the *equal-tailed* BPCI (i.e, the interval formed by the $\alpha/2$ and $1-\alpha/2$ posterior quantiles of the target parameter), and the *highest-posterior-density* (HPD) BPCI. Though HPD intervals are optimally short, we consider only equal-tailed intervals here, given their simple interpretation and widespread use.

If the prior Π and the PGD P coincide, then a BPCI is guaranteed to have correct lab-wise coverage. This strikes us as a fundamental property of BPCIs, though it is surprisingly unemphasized in most introductions to Bayesian techniques. Henceforth we refer to a BPCI arising from a prior distribution set equal to the PGD as an *omniscient* or “oracular” BPCI (abbreviated OBPCI), in the sense that the investigator is omniscient in knowing the actual PGD giving rise to future studies. It is indeed a fanciful assumption to think that the PGD would be known exactly, so throughout this paper we pay much

attention to non-omniscient BPCIs (abbreviated NBPCI). That is, we will evaluate lab-wise coverage when the investigator's prior distribution Π differs from the PGD P .

It is worth noting that BPCIs have desirable properties from a decision-theoretic point of view. The situation is complicated in that both coverage and length must be reflected in the loss function. Hence this function must be bivariate, or be a univariate combination of coverage and length terms (which would necessitate some weighting of the two). Robert (1994) gives some general discussion of this point. Despite this complication, there are still results which link, and come close to equating, BPCIs and admissible interval estimators (see, for instance, Meeden and Vardeman 1985). Thus the common argument for Bayesian point estimators having desirable frequentist properties does extend, albeit with complications, to the case of interval estimators.

Additionally, there are large-sample results saying that in "regular" modeling situations with large sample sizes and priors with unrestricted support, BPCIs will have frequentist coverage that converges to nominal coverage, at every possible set of parameter values. These results are based on obtaining a likelihood that dominates the prior given enough data; as such they are not very useful for our purposes, because later we turn to problems in which no such domination occurs. We will however find use for a variant of this result in which information is accumulated over a sequence of studies. First, however, we illustrate the operating characteristics of some interval estimators in a simple but relevant situation.

Example: Mixture of Near-Null and Important Effects

Say that θ represents the strength of a putative exposure-disease relationship (which may indeed be one of a sequence of such exposure-disease combinations to be investigated). For instance, θ might be a risk difference or a log odds-ratio relating binary exposure and disease variables. Supposed that D is a univariate sufficient statistic such that $D|\theta \sim N(\theta, \sigma^2)$ where σ^2 is known. Then $(D \pm q_{\alpha/2}\sigma)$ can be reported as a $100 \times (1 - \alpha)\%$ frequentist confidence interval (FCI) for θ , where $q_{\alpha/2}$ is the $1 - \alpha/2$ standard normal quantile.

In the context of observational epidemiology, null or minimal effects are common, and large effects are rare. Thus the PGD giving rise to a sequence of studies

might have most of its mass at or near zero. For instance, say the PGD is a mixture of two normal distributions: $N(0, \varepsilon^2)$ with weight p and $N(0, k^2 \varepsilon^2)$ with weight $1-p$, for a “small” ε and $k > 1$. This is interpreted as the first component generating minimal or near-null associations, while the second gives rise to important as well as near-null associations, e.g., $|\theta_i| < 2\varepsilon$ and $|\theta_i| > k\varepsilon$ might reasonably be described as near-null and important respectively.

We simulate 500,000 parameter-data ensembles with $\varepsilon=0.05$, $p=0.85$, $k=8$, and $\sigma^2=0.025$. If θ is a log odds-ratio then these values have $\exp(\theta_i)$ within $(0.91, 1.1)$ as near-null, and $\exp(\theta_i)$ outside $(0.67, 1.5)$ as important. The choice of $\sigma^2 = 2/((500)(0.2)(0.8))$ approximates the amount of information for the log odds ratio when comparing two independent groups (as in an unmatched case-control study) with 500 subjects per group and exposure prevalences around 20%.

The first two rows of Table 1 give operating characteristics of the FCI and the (equal-tailed) OBPCI as interval estimators for θ (both at the nominal 95% level). Note that when used as the prior distribution for θ the mixture distribution is conjugate, so that computation of the OBPCI is straightforward. As is consistent with theory, the lab-wise coverages of both procedures are within simulation error of the nominal 95%. On average, though, the OBPCI is considerably shorter than the FCI, by almost a factor of two. This results from the infusion of prior information.

Motivated by taking $|\theta| < 2\varepsilon$ as a minimal effect, we also define the *total discovery rate* (TDR), *false discovery rate* (FDR) and *false non-discovery rate* (FNR) for interval estimation as follows: The TDR is simply the proportion of reported intervals that exclude the minimal range, i.e., give confidence that the effect is not minimal. The FDR is then the proportion of these “discoveries” that are false, i.e., in which the parameter actually does lie in the minimal range. Similarly, amongst intervals intersecting the minimal range, the FNR is the proportion for which the target is actually outside this range. We can then describe how the OBPCI is more conservative than the FCI: The OBPCI attains a lower FDR at the cost of a higher FNR, as evidenced in the first two rows of Table 1.

Investigators are not omniscient. To illustrate consequences of defective prior information, we examine results in which the prior distribution deviates from the PGD. Our example is far from a comprehensive study of prior misspecification, and we doubt that such a study could be done given all the contextual elements involved. Rather, we wish to illustrate some qualitative points that will be relevant later, regarding potential consequences of such misspecification.

Two sets of NBPCI results are given in Table 1. The first set correspond to an investigator using the same form of a mixture-normal prior with the correct value of ε (which defines the notion of a minimal effect and so is contextually established), but with misspecified values of p and k (choosing $p=0.50$ or $p=0.95$, and $k=4$ or $k=12$). The second set corresponds to an investigator who doesn't elucidate a mixture structure for the prior, but rather simply applies a mean-zero normal prior. The case where the prior variance τ^2 equals the PGD variance of $v^2 = p\varepsilon^2 + (1-p)k^2\varepsilon^2$ is considered, as are the cases where the prior variance is half/double the PGD variance. The results in Table 1 underscore the disadvantage of the NBPCI relative to the FCI and the unattainable OBPCI: The lab-wise coverage now deviates from nominal. Arguably, however, these deviations are modest. Moreover, the NBPCI tend to maintain the other attractive features seen with the OBPCI, namely the much shorter average length and lower FDR compared to the FCI. Note that the deviations from nominal coverage are less pronounced and tend toward conservatism when the prior is more spread out than the PGD ($p=0.50$ in the first set of results, $\tau^2=2v^2$ in the second). This is not surprising, since NBPCIs will resemble FCIs more and more in the “flat” prior limit. In contrast, the deviations can be markedly anticonservative when the prior is more concentrated ($p=0.95$ in the first set, $\tau^2=0.5v^2$ in the second). Thus, by using very dispersed priors, we can improve the precision and reduce the FDR of our intervals without incurring objectionable deviations from nominal coverage. If however we “get greedy” and attempt to improve performance by using overconfident priors, we risk unacceptable deterioration of coverage.

In this and subsequent examples, we have used equal-tailed BPCIs because these are intuitive and commonly reported. It is well known, however, that for a given dataset the HPD interval (or possibly region) is the shortest interval with the specified Bayesian

probability content. In fact, Uno, Tian and Wei (2005) prove an interesting result about lab-wise coverage of HPD intervals, in the case that the prior and PGD coincide. They show that the HPD interval with coverage $1-\alpha$ does not always minimize *average* width subject to obtaining lab-wise coverage $1-\alpha$. Rather, the minimizing procedure in general involves the HPD interval with coverage $1-\alpha(D)$, such that $E\{\alpha(D)\}=\alpha$, where the data-dependent coverage level $\alpha(D)$ arises by thresholding the posterior densities for all studies at the same cutoff value. Thus in cases where the width of the posterior density varies across studies, HPD intervals with higher (lower) coverage levels will be reported for studies with narrower (wider) posterior densities. While we do not pursue this further here, it is worth emphasizing that the simple and intuitive interpretations associated with using a BPCI of fixed coverage level can be sacrificed in order to obtain intervals which are narrower on average.

Given the performance issues illustrated in Table 1, we find it overly simplistic to argue against the use of Bayesian interval estimation simply because prior specification is required, and because it is impossible to get this specification exactly right in terms of matching the PGD. Furthermore, as we discuss in the next section, by using a hierarchically structured prior one can effectively move the prior closer to the PGD as a sequence of studies unfolds.

3. Hierarchical Prior Distributions

As we have emphasized, consideration of a sequence of studies is a conceptual device capturing the reality facing most investigators. Studies in medicine and public health are nowhere near identical in design, conduct, and population studied, and hence there is no basis for asserting parameter equality across these studies. This fact is the rationale for random-effects models in meta-analysis, which typically employ very simple models for the PGD. When a new study is performed, however, data from m previous studies can be used to improve the prior distribution for θ by using a hierarchically structured prior distribution, which in turn will make the lab-wise coverage closer to nominal (as m increases, particularly). This further strengthens the frequentist appeal of Bayesian intervals.

Say that the study to be carried out has parameter-data ensemble (D, θ) and is preceded by m earlier studies with ensembles $(D_1^*, \theta_1^*), \dots, (D_m^*, \theta_m^*)$. If the $m+1$ ensembles are independent and identically distributed (i.e., each according to the PGD and the data model), it makes sense to allow the interval estimator for θ to depend on the earlier data as well as the current data. The lab-wise coverage (1) then generalizes to

$$C(I, P) = \Pr\{\phi \in I(D; D^*)\}, \quad (2)$$

where $D^* = (D_1^*, \dots, D_m^*)$ and the probability is taken with respect to the joint distribution of the $m+1$ parameter-data ensembles.

The standard Bayesian approach to borrowing strength across studies involves a hierarchical prior. That is, the prior Π asserts that the $m+1$ components of (θ^*, θ) are independent and identically distributed given a further parameter vector λ . Then λ itself is assigned a prior distribution. Application of Bayes theorem to form the posterior distribution on θ involves the likelihood contribution of $D|\theta$, with $\Pi(\theta|D^*) = \int \Pi(\theta|\lambda)\Pi(\lambda|D^*)d\lambda$ playing the role of the prior. That is, the earlier studies inform the value of λ , which in turn informs θ , in advance of observing D . If the PGD is well approximated by the posited $(\theta|\lambda)$ prior for some value of λ (say λ_0), and if the number of previous studies m is large, then $\Pi(\lambda|D^*)$ should be concentrated near λ_0 . Thus the ‘effective prior’ being applied to θ will be close to the PGD, which should result in lab-wise coverage for BPCIs that is close to nominal.

The use of hierarchical priors to ‘borrow strength’ across studies and the evaluation of coverage along the lines of (2) originates under the rubric of ‘empirical Bayes’ procedures (see, for instance, Morris 1983), which typically involve a non-Bayesian approximation to $\Pi(\lambda|D^*)$. With the advent of better algorithms and machines for Bayesian computation, however, fully Bayesian ‘hierarchical modeling’ is now commonplace. It should also be noted that treating the parameter values for the $m+1$ studies as exchangeable as described is a modeling assumption that will sometimes be inappropriate. Notably, in a situation where all studies focus on the same relationship at different calendar times, the assumption is dubious but may be weakened to allow for trends. For instance, the “Ty Cobb” example in Morris (1983) involves explicit modeling of a time trend for parameter values corresponding to consecutive calendar years. The exchangeability inherent in assuming the study parameters are conditionally *iid* may be

more appropriate when the studies address *different* relationships in a research domain, with no information conveyed by the time-ordering of the studies.

To give a simple illustration, suppose again that $D|\theta \sim N(\theta, \sigma^2)$ with σ^2 known. For computational ease we first consider a simpler PGD than previously, namely a $N(\lambda_0, \tau^2)$ distribution. Consider a partially omniscient investigator who knows the variance of the PGD, but not the mean, and in hierarchical fashion assigns the prior $\theta|\lambda \sim N(\lambda, \tau^2)$, $\lambda \sim N(\delta, \omega^2)$. The marginal prior distribution is then $\theta \sim N(\delta, \tau^2 + \omega^2)$. In the absence of previous studies, or with an *iid* prior assuming independence of θ and θ^* , the posterior on θ would arise from combining this prior with D , and the discrepancy between this prior and the PGD would induce some degree of non-nominal lab-wise coverage. With a correct hierarchical prior, however, the previous data will pull $\Pi(\lambda|D^*)$ toward λ_0 , and hence pull $\Pi(\theta|D^*)$ toward the PGD.

The present setting is sufficiently simple that the LWC given in (2) can be computed directly (one-dimensional numerical integration is required, but repeated simulation of data is not). As an example, suppose $\sigma=1$, and the PGD has $\lambda_0 = 3$, $\tau = 1$. Consider four prior distributions for λ , with $\omega=1$, and $\delta = 0, 1, 2, 3$. Note that these priors range from very bad (mean of the PGD lies three prior standard deviations away from the prior mean) to unrealistically good (mean of the PGD coincides with the prior mean). Figure 1 illustrates the coverages (2) for the resulting 95% BPCIs, as the number of previous studies m increases. When $m=0$ (thought of as either no previous studies, or as an *iid* prior across studies), the coverage ranges from less than 80% for the ‘worst’ prior to somewhat above 95% for the ‘best’ prior. As Figure 1 illustrates, however, for all the priors the coverage converges quite quickly to the nominal 95% as m increases, to match the OBPCI coverage. The figure also displays the interval width as a function of m . In this simple setting the width is governed by

$$\text{Var}(\theta|D, D^*) = \sigma^2 \tau^2 (\sigma^2 + \tau^2)^{-1} [1 + \sigma^2 \omega^2 \tau^{-2} \{(m+1)\omega^2 + \sigma^2 + \tau^2\}^{-1}], \quad (3)$$

which depends on neither the observed data nor the hyperparameter δ . Note that by $m=30$ most of the potential reduction in width has been realized, i.e., the width is close to the OBPCI width which corresponds to the $m \rightarrow \infty$ limit of (3). While the convergence of coverage and length to match the OBPCI represents a well-known calibration feature

of Bayes and empirical-Bayes estimation, it is interesting to see how rapidly it can proceed in simple settings.

Of course the above illustration is very simplistic, particularly as the variance components involved in the prior are taken to be known, and only the mean of the PGD must be learned via previous data. At the other end of the spectrum, one anticipates that complex PGDs, such as those involving a mixture of near-null and important effects, will require a larger number of studies before they are estimated well. To investigate this we re-consider the example of the previous section, involving a mixture of near-null and important effects. Now, however, we treat (p,k) as unknown parameters with prior distributions $p \sim \text{Uniform}(0,1)$ and $k \sim \text{Uniform}(4,20)$. As before the PGD is based on $p=0.85$ and $k=8$. Empirical results on coverage and average length appear in Table 2. These are based on only 1,000 simulated parameter-data meta-ensembles (with each meta-ensemble encompassing $m+1$ parameter-data ensembles), since the posterior computation is burdensome. In particular, a simple Markov chain Monte Carlo algorithm (with random walk proposals) is applied to sample from $(p,k|D,D^*)$, while $(\theta|p,k,D,D^*)=(\theta|p,k,D)$ can be sampled from directly. Results for coverage are quite appealing, in that the lab-wise coverage (2) modestly exceeds nominal when the number of previous studies m is small, presumably because $\Pi(\theta|D^*)$ is very flat, and also modestly exceeds nominal when m is large, presumably because $\Pi(\theta|D^*)$ is close to the PGD. However, the very slow convergence of $\Pi(\theta|D^*)$ to the PGD is manifested by the average interval width. Even with $m=100$ previous studies, the average width is still 44% larger than that of the OBCI. Nonetheless, it is 23% narrower than the FCI, a worthwhile gain paid for by a minor conservatism.

There are many examples of hierarchical modeling in the literature, where unknown means and variances are themselves modeled via prior distributions or estimated via marginal likelihood. These methods have performed quite well in large-scale simulations and in applications that provide subsequent validation (Brown 2008). Special methodology for inference about the distribution of a sequence of effects has expanded apace, driven by work on multiple comparisons (and particularly false discovery rates) in genome studies (see, for instance, Efron *et. al.* 2001, Newton and Kendziorowski 2003).

We have emphasized that *formal* inclusion of previous studies on various phenomena within a research team's domain of study can have positive benefits for subsequent studies within this domain, in terms of both lab-wise coverage and average width. Consequently, a formal scheme to obtain a prior which is close to the PGD for a given domain seems desirable when practical. In other circumstances, however, it should be possible to *informally* use previous studies in constructing a reasonable prior distribution. As alluded to earlier, for instance, in many sub-fields of epidemiology investigators do have well-grounded notions concerning the across-study prevalence of near-null effects and magnitude of important effects. One anticipates that a prior formed from direct elicitation of the investigators' views should not deviate greatly from a prior formed from formally updating a 'flat' prior based on previous studies. Regardless of which route is taken, construction of a prior which is reasonably close to the PGD for future studies in the domain seems to be a realistic and worthwhile goal. With this encouraging message in hand, we now turn to examining the use of Bayesian interval estimators in nonidentified model settings.

4. Interval Estimation in Nonidentified Models

The case for BPCIs versus FCIs seems mixed thus far, particularly as FCIs are guaranteed to have correct lab-wise coverage, without requiring any knowledge of the PGD. But in a large class of statistical problems, construction of valid FCIs is not possible. Recall that in general a model is nonidentified if there are multiple sets of parameter values giving rise to the same distribution of observables. We have argued that this class of models is the only realistic choice in most observational studies of human health and society (Greenland 2005; Gustafson 2006). This is particularly true in disciplines such as epidemiology where honest appraisal of what modeling assumptions are justified, and what limitations are inherent in the available data, *ought* to lead investigators to nonidentified models routinely.

Identifiable models are desirable when they can supply root-n consistent estimators of target parameters, as in classic industrial and laboratory experiments. With

study problems such as measurement error, missing data, selection bias, and unmeasured confounders, however, extremely strong assumptions may be required to attain an identified model. Most statistical methods assume absence of such problems, and the remainder assume that the form of the problems is known up to a few identifiable parameters. Either way, there is a strong possibility that the resulting model is grossly misspecified, with the resulting FCIs exhibiting excessive precision and severe undercoverage for the inferential target.

Put another way, using an overly simplified model for the sake of identifiability results in root- n consistent inference *for the wrong parameter* (e.g., an unconditional association, when the desired inferential target is an association conditional on an unmeasured covariate) (Greenland 2003, 2005; Gustafson 2006). If as usual the parameter being estimated does not equal the target parameter, the interval coverage for the latter will tend to zero as the sample size increases.

Backing away from untenable assumptions may result in a model that is better specified (closer to reality, or at least better representing the true inferential target), but which lacks identifiability. There is extreme hesitance amongst statisticians regarding the use of nonidentified models, because they don't give rise to estimators with familiar statistical properties, such as root- n shrinkage of interval estimators *to some value*. But for Bayesian analysis there is no conceptual or computational difference in how inferences are obtained from a nonidentified model compared to an identified model. In fact, from a radical subjective Bayesian perspective, identification is a matter of a degree and always a function of the full prior (including the prior for the data given the parameters).

In summary, in nonidentified problems there is no route to FCIs achieving exactly nominal coverage for any set of underlying parameter values. If in these settings we simplify the model to the point of identifiability, then FCIs are readily obtained via standard methods, but are likely to have grossly incorrect coverage probabilities due to misspecification. Without simplification, models are nonidentified, which precludes construction of FCIs having the nominal coverage probability at every point in the parameter space.

Some frequentist approaches to problems of this sort involve (i) specifying bounds (rather than prior distributions) on key parameters, and (ii) constructing interval estimators having *at least* nominal coverage at every point in the parameter space, with the consequence that the coverage will be higher than nominal at most parameter values. Some recent suggestions along these lines include Imbens and Manski (2004), Vansteelandt et al. (2006), and Zhang (2009); we illustrate such an approach in the first of the two examples below. Conversely, the use of Bayesian or approximately Bayesian inferences from nonidentified models was suggested at least as far back as Leamer (1974), and has long been discussed under special topics such as nonignorable missingness (Little and Rubin, 2002). It has also attracted considerable attention in recent literature; see, for instance, Dendukuri and Joseph (2001); Greenland (2003, 2005); Gustafson (2005b); Gustafson and Greenland (2006a,b); Hanson, Johnson, and Gardner (2003); Joseph, Gyorkos, and Coupal (1995); McCandless, Gustafson, and Levy (2007, 2008); Scharfstein, Daniels and Robins (2003).

For Bayesian procedures, the exact attainment of nominal lab-wise coverage by an OBPCI still holds under nonidentified models. The result in general (for any kind of model) is known, but surprisingly unemphasized in the literature (see Rubin 1984; Rubin and Schenker 1986 for exceptions). Yet it seems to be a useful reference point, as it provides a clear calibration, or “anchor,” for an interval estimation procedure in a nonidentified model. On the other hand, we generally expect the choice of prior to be far more influential on the posterior distribution when the model is nonidentified, so that lab-wise coverage may deviate rapidly from nominal as the prior distribution deviates from the PGD. We investigate this phenomenon in the two examples below.

Example: Prevalence Survey with Nonresponse

Vansteelandt *et al.* (2006) illustrate some frequentist techniques for sensitivity analysis in nonidentified models in the following setting. A binary outcome Y may be observed ($R=1$) or missing ($R=0$, nonresponse) for each study unit, so that the available data consist of n *iid* realizations of (RY, R) . The inferential target is the outcome prevalence, $\pi = \Pr(Y=1)$, while the missingness may be informative, i.e., Y and R may be

associated. One parameterization for this situation is $p = \Pr(R=1)$, $s = \Pr(Y=1|R=1)$, and $\gamma = \text{logit}\{\Pr(Y=1|R=0)\} - \theta$ where $\theta = \text{logit}(s)$. Then the inferential target is $\pi = (1-p)\text{expit}(\theta+\gamma) + ps$. This is a nonidentified inference problem because the likelihood for the observed data depends only on p and s , while the inferential target also depends on γ .

We consider the coverage and average length of three interval estimators for π . The first is the naïve interval estimator obtained by assuming $\gamma = 0$, i.e., assuming the missingness is completely at random, and estimating π as the sample proportion of the observed outcomes. The second is an interval estimator suggested by Vansteelandt *et al.* (2006), designed to have at least nominal frequentist coverage (approximately) under every fixed value of γ in a specified interval I ; we take $I = (-2,2)$ in the present example. Let $\hat{\pi}_l$ and $\hat{\pi}_u$ be the estimates of π when fixing the value of γ at the lower and upper endpoints of I respectively. Then the interval estimator with target level $1-\alpha$ is of the form $(\hat{\pi}_l - q_{\alpha^*/2} \text{se}(\hat{\pi}_l), \hat{\pi}_u + q_{\alpha^*/2} \text{se}(\hat{\pi}_u))$, where α^* is chosen to make the minimum coverage as γ varies in I equal to $1-\alpha$ (with the minimum attained at one of the endpoints). We refer to this interval as a conservative frequentist confidence interval (CFCI). The relationship between α^* and α depends on the unknown parameters, hence estimates are plugged-in and the coverage properties become approximate rather than exact. Vansteelandt *et al.* (2006) call interval estimators of this form “pointwise estimated uncertainty regions,” since the coverage claim applies to the true value of the target parameter. These authors also propose “weak” and “strong” estimators with coverage claims pertaining to the set of all target parameter values consistent with the observed data law (i.e., interval estimation of an interval). For more details see Vansteelandt *et al.* (2006).

The third interval estimator is the equal-tailed Bayesian credible interval arising from a uniform prior distribution for γ on the same interval I , along with uniform(0,1) priors for both p and s . Under this specification the parameters p, s , and γ remain independent of one another *a posteriori*, with beta distributions for p and s arising from binomial updating, and a uniform posterior distribution on I for γ ; that is, no updating of γ occurs.

Empirical lab-wise coverage and average length for nominal 95% intervals are given in Table 3. The PGDs used have normal distributions for $\beta = \text{logit}(p)$ and $\theta = \text{logit}(s)$ with $\mu_\beta = \text{logit } 0.67$, $\sigma_\beta = (\text{logit } 0.89 - \text{logit } 0.67)/2$, $\mu_\theta = \text{logit } 0.5$, and $\sigma_\theta = (\text{logit } 0.8 - \text{logit } 0.5)/2$. Thus the PGD for (p,s) concentrates around more typical-use scenarios than does the prior for these parameters. The PGD is completed by $\gamma \sim \text{uniform}(J)$ for various specifications of interval J . Note that one specification is the single-point interval $J=[2,2]$, which corresponds to fixing γ at the endpoint of I , and hence corresponds to a partially frequentist evaluation of coverage. Note also that the average interval lengths do not depend on the specification of J for this problem, since the distribution of the observed data (under the joint distribution of parameters and data) does not depend on J . Thus the average lengths of 0.11 for the naïve interval, 0.33 for the CFCI, and 0.28 for the BPCI apply for any J .

Table 3 verifies that when J in the PGD and I in the prior coincide, the Bayesian intervals have LWC within simulation error of nominal, despite the discrepancy between the uniform priors for (p,s) and the logit-normal PGDs for (p,s) . In contrast, the CFCI approach is indeed quite conservative when I and J coincide, with lab-wise coverage of 99% and average length 17% greater than the BPCI. As expected, the lab-wise coverage of both the CFCI and the BPCI is highly affected by any discrepancy between I and J . As advertised, the CFCI achieves conservative coverage in all cases, except for a slight dip below nominal in the case that J is wider than (and contains) I . Note in particular that the CFCI achieves nominal coverage when γ is fixed at an endpoint of I , whereas the BPCI coverage drops to 71% in this setting.

The differences between lab-wise coverage of BPCIs and CFCIs are somewhat hidden in Table 3, since nominal 95% intervals do not have much 'room' to obtain higher than nominal coverage. Thus we also report results for nominal 80% intervals (Table 4). Admittedly such intervals are seldom reported in practice (though see Greenland *et al.* 2000 for an exception), but they are useful for gauging the extent to which a given interval estimator is conservative. The average lengths of these intervals are 0.069 (naïve), 0.29 (CFCI) and 0.22 (BPCI). When I and J match, we now see very substantial over-coverage (96%) for the CFCI, with an average width 30% greater than for the

OBPCI. We also see more clearly the over-coverage that results for both CFCI and BPCI when J is narrower than I .

The BPCI and the CFCI are constructed to satisfy different criteria, and we are not attempting to argue that one is better than the other. In particular, note the tradeoff exhibited in Tables 3 and 4. If the investigator has an interval of values I in mind for γ , then the CFCI has a conservatism which may be appealing: at least nominal coverage can be obtained with respect to any averaging across values in I , including the selection of single points. On the other-hand, if lab-wise coverage with respect to the Uniform(I) distribution is at issue, then the BPCI will be shorter on average, and have correct coverage. We do emphasize that this correct coverage constitutes a calibration property of the BPCI which the CFCI does not possess. That is, without doing simulation, we do not know to what extent the CFCI based on interval I will exhibit higher than nominal lab-wise coverage when the PGD is based on I . But we do know automatically that the BPCI using I in the prior will exhibit correct lab-wise coverage when the PGD is based on I . Thus the BPCI is anchored via the investigator's knowledge that *exactly* nominal coverage *would* be obtained in a sequence of studies with PGD equal to the prior, and presumably *at least* nominal coverage *would* eventually be attained in a sequence of studies in which the support of the prior contains the PGD. In this sense, posterior coverage is conservative precisely when the prior is conservative relative to the PGD. The CFCI lab-wise coverage has a more murky connection to the PGD, which is the price it pays for obtaining correct frequentist coverage at the endpoints of the prior interval I .

Example: Case-Control Study with Misclassification

Consider an unmatched case-control study of the association of a disease indicator Z and a binary exposure indicator X , with X subject to independent nondifferential misclassification. Let $r_0 = \Pr(X=1|Z=0)$ and $r_1 = \Pr(X=1|Z=1)$ be the prevalences of actual exposure among nondiseased and diseased source population members, and let $SN = \Pr(X^*=1|X=1)$ and $SP = \Pr(X^*=0|X=0)$ be the sensitivity and specificity of the exposure classification in the study. The numbers *apparently* exposed amongst the n_0 nondiseased

controls and n_1 diseased cases in the study are modeled as $Y_i \sim \text{Bin}(n_i, \theta_i)$ for $i=0$ and $i=1$ respectively, with $\theta_i = r_i SN + (1-r_i)(1-SP) = \Pr(X^*=1|Z=i)$. If all four parameters (r_0, r_1, SN, SP) are unknown, then this model is not identified by the observed counts $(y_1, y_0, n_1-y_1, n_0-y_0)$. Bayesian inference under this model is considered by Gustafson, Le, and Saskin (2001), Gustafson (2003), Greenland (2005), Chu *et al.* (2006) and Gustafson and Greenland (2006a), among others.

We consider prior distributions and PGDs of the following form: A bivariate normal distribution for the logit prevalences $(\text{logit } r_0, \text{logit } r_1)$, with correlation ρ and identical marginals (mean μ and variance τ^2). The log-odds ratio, $\beta = \text{logit}(r_1) - \text{logit}(r_0)$, is then distributed as $N\{0, (1-\rho)2\tau^2\}$. The correlation is essential to reflect the fact that information about the exposure prevalence in one group would alter bets about the prevalence in the other group, due to prior information about β (Greenland, 2001). SN and SP are here taken as independent of the exposure prevalences and each other, however, with $SN \sim \text{Beta}(a_N, b_N)$ and $SP \sim \text{Beta}(a_P, b_P)$; more realistic priors might allow dependent SN and SP (Chu *et al.*, 2006; Greenland and Lash, 2008).

Bayesian computation is readily implemented via the efficient algorithm of Gustafson, Le and Saskin (2001). While this algorithm takes advantage of structure imbued by assigning uniform priors on prevalences, we can use importance sampling to adapt the algorithm output to the present prior specification. As an example, $m = 10,000$ parameter-data ensembles with $n_1 = n_2 = 500$ are drawn from the PGD based on $\mu = -2.3$, $\tau = 1.17$, $\rho = 0.76$, $a_N = a_P = 18$, $a_N = a_P = 4$. These choices produce 95% logit-symmetric interval for each r_i of $(0.01, 0.50)$ and a 95% log-symmetric interval for e^β of $(0.2, 5.0)$. Also, the modes of the SN and SP distributions are 0.85, with 95% logit-symmetric intervals of $(0.637, 0.946)$.

For each dataset, seven interval estimates for β are constructed:

- (i) the standard FCI assuming no misclassification;
- (ii) an FCI derived by taking $SN=0.85$ and $SP=0.85$ as known values;
- (iii) the omniscient BPCI arising when the prior distribution coincides with the PGD;

non-omniscient BPCIs with priors based on correct specification of (μ, τ, ρ) but:

- (iv) $a_N = a_P = 9.5$, $b_N = b_P = 2.5$ (keeping the prior modes on SN and SP at 0.85 but making the distribution more diffuse);
- (v) $a_N = a_P = 26.5$, $b_N = b_P = 5.5$ (modes at 0.85 but overly concentrated);
- (vi) $a_N = a_P = 23.5$, $b_N = b_P = 8.5$ (still overly concentrated and modes shifted down to 0.75);
- (vii) $a_N = a_P = 28.5$, $b_N = b_P = 3.5$ (still overly concentrated and modes shifted up to 0.95).

Empirical properties of the interval estimators (at the nominal 95% level) are described in Table 5.

In the previous example, the joint posterior density was a product of the marginal posterior density for the two parameters appearing in the likelihood function and the marginal posterior density (equal to the prior density) for the one parameter not in the likelihood. This factorization simplified the mathematics of how the prior influences the posterior distribution of the target parameter. In the present example, however, the structure of the problem is more nuanced. As emphasized by Gustafson, Le and Saskin (2001), the support of the two parameters not in the likelihood, (SN, SP) , depends on the values of the two parameters in the likelihood, (θ_0, θ_1) , since by construction $1-SP$ and SN must straddle both θ_i values. To some extent then, the posterior distribution of (SN, SP) can depend on the data, even though these parameters do not appear in the likelihood function. Gustafson (2005a) discusses such *indirect learning* about parameters in nonidentified models in more general terms.

Given that the data can provide some information about (SN, SP) , one might anticipate that the NBPCI coverage would be less sensitive to the choice of prior than in a situation without any indirect learning. The results in Table 5 bear this out, with the coverage of nominal 95% NBPIs ranging from 87% to 95% across the priors considered. In accord with theory, the OBPCI coverage is within simulation error of nominal, which can be regarded as a check that our scheme for posterior computation is working adequately (see Cook, Gelman and Rubin 2006 for elaboration).

While the link between (SN, SP) and (θ_0, θ_1) is exploited to advantage under a Bayesian analysis, it is problematic for the FCI based on taking SN and SP as fixed values less than one. In particular, the FCI is not defined for datasets with one or both $\hat{\theta}_i$ falling outside the interval $(1-SP, SN)$. Moreover, this can happen via sampling variation even if the postulated values of (SN, SP) happen to be correct. Tu, Litvak and Pagano (1994, 1995) discussed this problem, and offered some mitigating strategies when exposure prevalence (say in a single population) is the inferential target of interest. Such strategies yield interval estimates for prevalence with an endpoint at zero or one, which limits their utility for odds-ratio inference. In our case, the results for estimator (ii) in Table 5 are based on only the 81% of sampled parameter-data ensembles not giving rise to the aforementioned problem. Perversely, this method is failing in situations where the data are most suggestive that the guessed values of (SN, SP) might be wrong. Put another way, the FCI fails on datasets where Bayesian intervals may do particularly well via more prior-to-posterior updating of (SN, SP) .

As a final point concerning this example, we recognize that is quite reasonable to also study the frequentist properties of the Bayesian interval estimator. This can become quite computationally burdensome, however, if evaluation of frequentist coverage at many points in the parameter space is desired: each point necessitates simulation of many datasets, and each dataset may require many MCMC iterations in order to compute the interval estimate. Rather than pursuing this course, we note that the simulation of parameter-data ensembles as used to evaluate lab-wise coverage also yields information about frequentist coverage.

Thus, say that the frequentist coverage for parameter vector θ^* is of interest. If m parameter-data ensembles are simulated, then we might consider the proportion α of ensembles for which θ is closest to θ^* in some sense. Then the empirical coverage for this subset of ensembles approximates the frequentist coverage at θ^* , with the approximation improving as $\alpha \rightarrow 0$ and $m\alpha \rightarrow \infty$. Admittedly, it may be computationally prohibitive to make the approximation error very small, so we refer to the reported coverage as ‘near-frequentist coverage’ around θ^* . Notwithstanding its approximate nature, this can still reveal trends in frequentist coverage across the parameter space.

To apply this to the present example, we extend the simulation size to $m=100,000$ parameter-data ensembles, and set $\alpha=0.01$. Various points θ^* in the parameter space are considered, by fixing $r_0=0.10$, $r_1=0.15$, and then setting SN and SP at values corresponding to specific prior quantiles. Thus we investigate how the frequentist coverage depends on the compatibility between the prior and the true SN and SP values. Results appear in Table 6. We see under-coverage for SP values which are low in relation to the prior, and over-coverage when SP or SN is high in relation to the prior. Generally, however, the variation in frequentist coverage as SN and SP values move around the region supported by the prior distribution seems quite modest.

5. Recommendations

The above arguments and illustrations are intended to summarize and explain in simple form several practical recommendations that we and others have reached in the course of numerous theoretical studies, simulations, and real applications. Like others before us, we first recommend forming prior distributions and then reporting Bayesian interval estimates for parameters of interest, particularly in nonidentified model contexts. Based on our investigations, however, we further suggest that a special form of sensitivity analysis be carried out as well.

Sensitivity analysis is conducted in much applied work; typically this involves reporting multiple inferences corresponding to multiple models and (for Bayesians) multiple prior distributions. While these analyses are often better than standard reports of results from just one model, the resulting collection of interval estimates leads to problems of summarization and interpretation of the collection. Thus we recommend instead that one start with a single, relatively inclusive “covering” prior distribution that subsumes the diversity of opinions and possibilities for the parameters. Then, as a safeguard, we would evaluate the lab-wise coverage of Bayesian intervals arising from this prior, for a variety of PGDs differing somewhat from the prior. If the coverage does not fall much below nominal as the PGD deviates from the prior, then we may argue that our statistical procedure is probably (in the subjective judgmental sense) at least roughly

calibrated, in the across-study sense of lab-wise coverage. Otherwise we may consider ourselves alerted to a potentially serious miscalibration.

Table 3, in the context of prevalence surveys with nonresponse, provides one example of studying the sensitivity of lab-wise coverage as the PGD deviates from the prior distribution. We close with a further example from a specific and well-developed scientific context.

Example: Silica exposure and lung cancer

We revisit the investigation of Steenland and Greenland (2004) on the relationship of silica exposure to lung cancer. In a cohort of 4,626 industrial sand workers with high silica exposure, 109 lung-cancer deaths were observed, compared to an expected count of 68.1 under the null hypothesis of no association between silica exposure and lung cancer. This comparison of the cohort to U.S. population data is adjusted for age, race, calendar time and sex. It is not adjusted for smoking status though, because smoking histories were not collected for this cohort.

Steenland and Greenland used prior information derived from other studies in order to remedy this situation using both Monte-Carlo sensitivity analysis (MCSA) and Bayesian analysis. To describe this analysis, let β_1 be the log relative risk of lung-cancer death for silica exposure versus no exposure, within strata defined by smoking behavior, and let β_2 and β_3 be log relative risks for current smokers compared to never smokers, and former smokers compared to never smokers. Assuming a log-linear model without products between silica exposure and smoking effects, the observed death count can be regarded as a Poisson realization with log-mean λ , where

$$\lambda = c + \beta_1 + \log(p_1 + p_2 e^{\beta_2} + p_3 e^{\beta_3}) - \log(q_1 + q_2 e^{\beta_2} + q_3 e^{\beta_3}).$$

Here c is a known offset obtained from population data ($c = \log 68.1$ in the present example), while (p_1, p_2, p_3) and (q_1, q_2, q_3) are probability distributions over (never, current, former) smokers, in the exposed and unexposed populations respectively. This is a highly nonidentified model, with nine unknown parameters involved in the mean function. Identification of the target parameter β_1 can only be obtained via a strong assumption, e.g., that smoking behavior and occupational silica exposure are

unassociated, i.e., $(p_1, p_2, p_3) = (q_1, q_2, q_3)$, which is known to be false. Thus a far more principled analysis combines the Poisson model for data along with prior distributions for $(\beta_1, \beta_2, \beta_3)$, (p_1, p_2, p_3) , and (q_1, q_2, q_3) .

Based on data from a large cohort study of smoking and lung cancer, Steenland and Greenland took $\beta_2 \sim N(\log(23.6), 0.094^2)$ and independently $\beta_3 \sim N(\log(8.7), 0.094^2)$. They used smoking data on a small sample of 199 workers inform the prior $p \sim \text{Dirichlet}(199 \times (0.26, 0.40, 0.34))$, and used a large national survey to inform the prior on q . This survey involved 56,000 subjects, but to account for various uncertainties it was discounted by a factor of four to yield the prior $q \sim \text{Dirichlet}(14,000 \times (0.34, 0.35, 0.31))$. Steenland and Greenland used a very diffuse prior on β_1 . This is not appropriate for investigating lab-wise coverage, however, as some datasets simulated from parameters generated under this prior will have implausibly low (i.e., zero) death counts, while other will have implausibly large counts. Thus, for present purposes we take the prior $\beta_1 \sim N[0, \{\ln(5)/2\}^2]$, which puts most of its weight on relative risks between 1/5 and 5. This completes specification of the prior distribution.

Bayesian computation for the present situation is readily implemented in a two-stage manner. First, an approximate posterior sample is simulated by drawing λ values ‘as if λ had a flat prior, and independently drawing $(\beta_2, \beta_3, p_1, p_2, p_3, q_1, q_2, q_3)$ values from their prior distribution. Second, this posterior sample is ‘made exact’ via importance sampling, which recognizes the actual prior distribution in the $(\lambda, \beta_2, \beta_3, p_1, p_2, p_3, q_1, q_2, q_3)$ parameterization. Note that omitting the second step corresponds to the MCSA in Steenland and Greenland. In the present example this second step has negligible impact, though in general importance sampling can be used to convert MCSA inferences to fully Bayesian inferences in situations where the two do not agree so closely.

Applied to the cohort data, a 95% equal-tailed BPCI for $\exp(\beta_1)$ is (1.12, 1.73), which is very similar to the interval reported by Steenland and Greenland using their slightly different prior. For comparison, the analysis which ignores the confounding effect of smoking gives the interval (1.31, 1.91). This result is based on the same prior for β_1 as above, with the presumption that $(p_1, p_2, p_3) = (q_1, q_2, q_3)$. Thus the impact of acknowledging smoking as a confounder is to push the interval estimate for β_1 toward (but not across) the null, and to widen the interval by about 15%. This widening is

somewhat modest, since there is relatively good prior data about smoking effects and smoking behaviour in the two populations and the association of smoking with silica exposure in these data appears to be small.

We know that BPCIs based on this prior will have correct lab-wise coverage for a PGD equal to the prior. We wish to see how far the coverage deviates from nominal as the PGD deviates from the prior. We thus examine eight PGDs, starting with the prior and considering all possible combinations of:

- (i) shifting the prior mean for β_2 left or right by one prior standard deviation;
- (ii) shifting the prior mean for β_3 left or right by one prior standard deviation;
- (iii) discounting the prior on (p_1, p_2, p_3) by a factor of two or (further) discounting the prior on (q_1, q_2, q_3) by a factor of two.

Table 7 gives coverage results using 95% equal-tailed BPIs for β_1 . When the PGD equals the prior, the lab-wise coverage is within simulation error of nominal, as theory dictates. As the model is highly nonidentified, we are not surprised to see lower than nominal coverage for most of the PGDs considered. We are pleasantly surprised, however, to see that the loss of coverage is very mild. This adds credence to the Bayesian results given by Steenland and Greenland (2004).

Based on examples as well as theoretical and simulation studies, we recommend that PGD sensitivity analysis be used when inference using nonidentified models is required. No important sensitivity was seen in the preceding example. Nonetheless, high sensitivity to plausible PGD specifications would have suggested that the full model (including those for the prior distribution and data-generating mechanism) had inadequately captured posterior uncertainty given the actual prior uncertainty of the analysts, and that intervals estimates from the model could be seriously miscalibrated. Hence, as with failed regression diagnostics, we would find ourselves advised to revise our model rather than rely on it.

Of course, this advice raises classic issues of the impact of post-data model revision based on diagnostics, long recognized as a challenge for applied Bayesians as well as for applied frequentists (Box, 1980). We thus regard these issues as an important direction for further research in our proposed approach.

References.

- Bayarri, M.J., and Berger, J.O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science* **19**, 58-80.
- Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**, 383–430.
- Brown, L.D. (2008). In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Annals of Applied Statistics* **2**, 113-152.
- Cook, S., Gelman, A., and Rubin, D.B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics* **15**, 675-692.
- Chu, H., Wang, Z., Cole, S.R. and Greenland, S. (2006). Illustration of a graphical and a Bayesian approach to sensitivity analysis of misclassification. *Annals of Epidemiology*, **16**, 834-841.
- Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* **57**, 158-167.
- Eddy, D. M., Hasselblad, V. and Schachter, R. (1992). *Meta-analysis by the Confidence Profile Method*. New York:Academic Press.
- Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151-1160.
- Greenland, S. (2001). Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Analysis* **21**, 579-583.

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. *Journal of the American Statistical Association* **98**, 47-54.

Greenland, S. (2005). Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A* **168**, 267-308.

Greenland, S., and Lash, T.L. (2008), Bias analysis. Ch. 19 in: Rothman, K.J., Greenland, S., and Lash, T.L, eds.. *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott-Williams-Wilkins, 345-380.

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: impacts and Bayesian adjustments*. Boca Raton: Chapman and Hall.

Gustafson, P. (2005a). On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with discussion). *Statistical Science* **20**, 111-140.

Gustafson, P. (2005b). The utility of prior information and stratification for parameter estimation with two screening tests but no gold-standard. *Statistics in Medicine* **24**, 1203-1217.

Gustafson, P. (2006). Sample size implications when biases are modelled rather than ignored. *Journal of the Royal Statistical Society, Series A* **169**, 883-902.

Gustafson, P. and Greenland, S. (2006a). Curious phenomena in adjusting for exposure misclassification. *Statistics in Medicine* **25**, 87-103.

Gustafson, P. and Greenland, S. (2006b). The performance of random coefficient regression in accounting for residual confounding. *Biometrics* **62**, 760-768.

Gustafson, P., Le, N., and Saskin R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* **57**, 598-609.

Hanson, T.E., Johnson, W.O., and Gardner, I.A. (2003). Hierarchical models for the estimation of disease prevalence and the sensitivity and specificity of dependent tests in the absence of a gold-standard. *Journal of Agricultural, Biological, and Environmental Statistics* **8**, 223-239.

Imbens, G.W. and Manski, C.F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72**, 1845-1857.

Joseph, L., Gyorkos, T.W., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* **141**, 263-272.

Leamer, E.E. (1974). False models and post-data model construction. *Journal of the American Statistical Association*, **69**, 122–131.

Little, R.J.A., Rubin, D.B. (2002). *Statistical analysis with missing data*, 2nd ed. New York: Wiley.

McCandless, L.C., Gustafson, P., and Levy, A.R. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine* **26**, 2331-2347.

McCandless, L.C., Gustafson, P., and Levy, A.R. (2008). A sensitivity analysis using information about measured confounders yielded improved uncertainty assessments for unmeasured confounding. *Journal of Clinical Epidemiology* **61**, 247-255.

Meeden, G. and Vardeman, S. (1985). Bayes and admissible set estimation. *Journal of the American Statistical Association* **80**, 465-471.

Molitor, J., Jackson, C., Best, N. B., and Richardson, S. (2008). Using Bayesian graphical models to model biases in observational studies and to combine multiple data sources: Application to low birthweight and water disinfection by-products. *Journal of the Royal Statistical Society, Series A*, in press.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Association* **78**, 47-55.

Newton, M.A. and Kendziorski, C.M. (2003). Parametric empirical Bayes methods for microarrays. In *The Analysis of Gene Expression Data Methods and Software* (G. Parmigiani, E.S. Garrett, R. Irizarry and S.L. Zeger, eds.), Springer-Verlag: New York, pp. 254-271.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36**, 97-131.

Robert, C.P. (1994). *The Bayesian Choice: A Decision-Theoretic Motivation*. Springer-Verlag: New York.

Rubin, D.B.(1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**, 1151-1172.

Rubin, D.B. and Schenker N. (1986). Efficiently simulating the coverage properties of interval estimates. *Applied Statistics* **2**, 159-167.

Scharfstein, D.O., Daniels, M., and Robins, J.M. (2003). Incorporating prior beliefs into the analysis of randomized trials with missing outcomes. *Biostatistics* **4**, 495-512.

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. *American Journal of Epidemiology* **160**, 384-392.

Tu, X., Litvak, E., and Pagano, M. (1994). Studies of AIDS and HIV surveillance screening tests: can we get more by doing less? *Statistics in Medicine* **13**, 1905-1919.

Tu, X., Litvak, E., and Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application in HIV screening. *Biometrika* **82**, 287-297.

Uno, H., Tian, L., and Wei, L.J. (2005). The optimal confidence region for a random parameter. *Biometrika* **92**, 957-964.

Vansteelandt, S., Goetghebeur, E., Kenward, M.G., and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica* **16**, 953-980.

Zhang, Z. (2009). Likelihood-based confidence sets for partially identified parameters. *Journal of Statistical Planning and Inference*, to appear.

Table 1: Frequency properties of interval estimators based on 500,000 simulated parameter-data pairs from PGD with $\varepsilon = 0.05$, $p = 0.85$, $k = 8$. The “omniscient” posterior (OBPI) uses these parameter values; the “non-omniscient” posterior (NBPI) uses the values of p and k shown.*

	Coverage %	Avg. Length	TDR %	FDR %	FNR %
FCI	95.0	0.62	6.3	17.0	11.5
OBPI	95.0	0.33	2.2	0.6	14.1
NBPI:					
$p=0.50, k=4$	95.8	0.41	2.1	0.5	14.2
$p=0.50, k=12$	97.3	0.46	3.2	2.5	13.3
$p=0.95, k=4$	89.8	0.24	0.9	0.0	15.2
$p=0.95, k=12$	91.7	0.26	1.8	0.2	14.4
$N(0, v^2)$	94.8	0.44	2.1	0.6	14.1
$N(0, 0.5v^2)$	92.1	0.36	0.8	0.0	15.2
$N(0, 2v^2)$	96.5	0.51	3.5	3.4	13.0

*Simulation standard errors for coverage, $TDR \approx 0.04\%$. The simulation standard errors for FDR are considerably larger and variable, since only a small portion (the TDR) of the simulated pairs contribute to the estimated proportion.

Table 2: Empirical properties of interval estimators with a prior distribution on (p, k) and m previous studies. Results based on 1000 simulated parameter-data meta-ensembles.

m	Coverage %	Avg. Length
0	96.4	0.515
10	95.9	0.491
20	95.8	0.487
100	95.4	0.476

Table 3: Empirical coverage probabilities and average lengths for nominal 95% interval estimators of a prevalence π . Results are given for naïve estimator, the CFCI, and the BPI. For each choice of PGD (i.e., choice of interval J), results are based on 10,000 simulated parameter-data ensembles with a sample size of $n=500$. Simulation standard errors for coverages are 0.5% or less. Both the CFCI and the BPI assume an interval range $I=(-2,2)$ for γ .

J in PGD:	Naïve	CFCI	Bayes
$J=(-2,2)$	42%	99%	95%
$J=(-3,3)$	30%	93%	80%
$J=(-1,1)$	67%	100%	100%
$J=(-1,3)$	40%	95%	84%
$J=(2,2)$	9%	95%	71%
Average length	0.11	0.33	0.28

Table 4: Empirical coverage probabilities and average lengths for nominal 80% interval estimators of a prevalence π . Both the CFCI and the BPI assume an interval range $I=(-2,2)$ for γ . The table entries are as per Table 2.

J in PGD:	Naïve	CFCI	Bayes
$(-2,2)$	27%	96%	80%
$(-3,3)$	19%	83%	59%
$(-1,1)$	47%	100%	98%
$(-1,3)$	26%	87%	69%
$(2,2)$	4%	80%	31%
Average length	0.069	0.29	0.22

Table 5: Empirical lab-wise properties of nominal 95% interval estimators for a log odds ratio β based on 10,000 simulated parameter-data ensembles. The simulation standard errors for coverage are less than 0.5%. Results for estimator (ii) are based only on the 81% of ensembles for which the method works.

	Coverage	Avg. Length
(i)-FCI	44%	0.60
(ii)-FCI	81%	2.20
(iii)-OBPI	95%	2.02
(iv)-NBPI	95%	2.12
(v)-NBPI	94%	1.94
(vi)-NBPI	95%	2.32
(vii)-NBPI	87%	1.56

TABLE 6: Near-frequentist coverage in the case-control study with misclassification example.

	$SP^*=0.63$	$SP^*=0.77$	$SP^*=0.83$	$SP^*=0.88$	$SP^*=0.95$
$SN^*=0.63$	90%	95%	97%	97%	98%
$SN^*=0.77$	92%	98%	99%	99%	99%
$SN^*=0.83$	93%	99%	99%	99%	99%
$SN^*=0.88$	92%	99%	99%	99%	99%
$SN^*=0.95$	93%	98%	99%	99%	98%

NOTE: Evaluation is for parameter values θ^* given by $r^*=(0.10, 0.15)$ and the indicated values of (SN^*, SP^*) , using $\alpha=0.01$ of the $m=100,000$ simulated parameter-data ensembles in each instance. The chosen values for (SN^*, SP^*) correspond to 2.5th, 25th, 50th, 75th and 97.5th percentiles of the prior distribution.

Table 7: Lab-wise coverage of 95% Bayesian intervals for β_1 as the PGD varies, in the silica and lung cancer example. The first row gives coverage when the PGD equals the prior. The remaining eight rows give coverage when the PGD is an alteration of the prior. The three-character code describes the alteration. The first character (+ or -) indicates whether the mean of β_2 is increased or decreased, the second character does the same for the mean of β_3 , and the third character (p or q) indicates whether the prior on p or the prior on q is discounted. Results are based on 100,000 realizations, giving simulation error for coverage less than 0.1%.

PGD	Coverage %
Prior	94.8
- - p	92.6
- - q	94.8
- + p	93.1
- + q	95.2
+ - p	92.0
+ - q	94.5
+ + p	92.4
+ + q	94.8

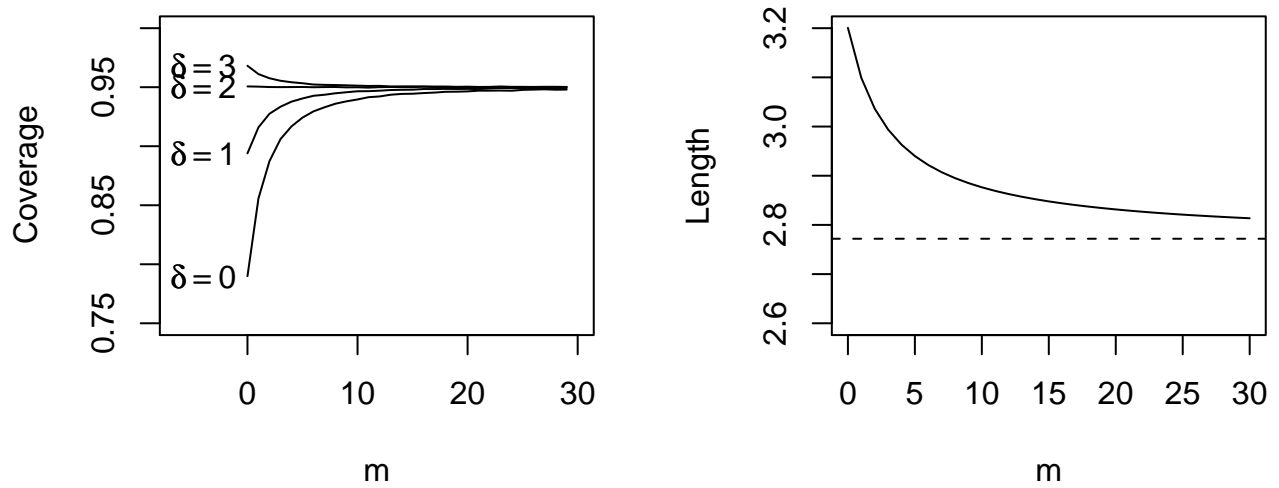


Figure 1: Labwise coverage and interval length for the nominal 95% BPCI, as a function of the number of previous studies m . Coverage is given for four choices of hyperparameter δ , whereas the length does not depend on δ . The dashed horizontal line in the second panel corresponds to the length of the OBCPI.