Assessing Sensitivity to Unmeasured Confounding in Observational Studies: A Bayesian Approach

by

Lawrence Cruikshank McCandless

B.Sc., University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

We accept this thesis as conforming
to the required standard

The University of British Columbia

August 2004

Library Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Lawrence M'Candless

Name of Author (please print)

20 August 2004

Date (dd/mm/yyyy)

Title of Thesis: Assessing Sensitivity to Unmeasured Confounding in Observational Studies: A Bayesian Approach

Degree: Master of Science   Year: 2004

Department of Statistics

The University of British Columbia
Vancouver, BC   Canada

Abstract

Systematic error due to possible unmeasured confounding may weaken the validity of findings from observational studies investigating the effects of exposures on disease. Because study subjects are assigned to exposure levels in a non-random way, hidden differences between exposure groups may bias effect estimates in a way which is difficult to predict. A solution is to conduct a Bayesian sensitivity analysis (BSA) which incorporates uncertainty about unmeasured confounding into the analysis as prior distributions on bias parameters. Markov chain Monte Carlo techniques can then be used to summarize the posterior distribution of the exposure effect given the data and prior beliefs about unmeasured confounding. We consider BSA in the context of logistic regression models for a binary exposure, binary outcome, binary unmeasured confounder and covariate vector. Because the resulting model is not identifiable, standard theory governing the large sample behaviour of posterior distributions cannot be applied, complicating an evaluation of the performance of BSA. However, using two simulation studies, we demonstrate that if the prior distribution for the analysis of datasets from a sequence of observational studies approximates the distribution from which study parameters arise, then the coverage probabilities of BSA 95% credible intervals will be approximately 95% when averaged over many studies. Moreover, we demonstrate that BSA credible intervals tends to yield greater coverages probabilities of the true exposure effect compared to methods which ignore unmeasured confouding. As an example, we investigate the effect of possible unmeasured confouding on risk of elevated triglyceride levels among HIV infected persons treated with protease inhibitors.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor Paul Gustafson, from the Department of Statistics, and my co-supervisor Adrian Levy, from the Department of Health Care and Epidemiology, for their thoughtful guidance and insight.

LAWRENCE CRUIKSHANK MCCANDLESS

The University of British Columbia
August 2004

Chapter 1

Introduction

Confounding can bias estimation of model parameters in observerational studies investigating the effect of exposures on disease in human populations (Szklo and and Nieto (1999)). Because study participants are not assigned to exposure levels using randomization, exposed subjects may be more predisposed to the outcome than no exposed subjects because of baseline differences in risk factors for the outcome. Consequently, the observed exposure effect will be a mixture of the effect of the exposure and the effect of the confounder (Rothman and Greenland (1999)).

For a variable to be a confounder, it must have three properties: It must be a risk factor for the disease; It must be associated with the exposure; and it must not be an intermediate variable between exposure and disease (Szklo and and Nieto (1999), Rothman and Greenland (2004)). An example presented by Gordis (2004) involves the effect of coffee drinking on cancer of the pancreas. In such an investigation, smoking might be a potential confounding variable. If we assume: firstly, smoking is a risk factor for pancreatic cancer; secondly, smoking is associated with coffee drinking; thirdly, smoking is not caused by coffee drinking. Then by comparing a group of coffee drinkers and non-coffee drinkers without taking smoking into consideration, the association between coffee drinking and risk of pancreatic cancer would be biased because coffee drinkers are more likely to smoke, and are therefore at greater risk of pancreatic cancer.

Control for confounding is accomplished using analysis techniques which involve strat-ification (Rosenbaum (2002). This proceedure is to compare exposed and unexposed subjects who appear similar with respect to measured convariates. If all confounding variables have been identified, then exposed and unexposed subjects in each strata can be treated as if they have been randomized into exposure levels. Multiple regression techniques including logistic regression are an example of stratified analysis (Rothman and Greenland (1998)). The imple-mentation of regression methods is straightforward and there is a variety of software available.

However, adjustment for measured confounders does not rule out unmeasured confound-ing. Because study subjects are assigned to exposure levels in a non-random way, hidden differences may also be induced between exposure groups which cannot be adjusted for in the analysis. If these differences affect the outcome under study, then effect estimates will be biased in a manner which is difficult to predict. A popular solution is to conduct a sensitivity analysis wherein the model of the relationship between exposure and disease is expanded to include bias parameters which reflect the investigator's assumptions about unmeasured confounding. If the exposure effect estimates are insensitive to a broad range of bias parameter values, then concerns about unmeasured confounding are ameliorated.

Sensitivity analysis was first proposed in the 1950's during the debate over the possible role of hidden confounders in the observed effect of tobacco smoking on the risk of lung cancer (see Cornfield, Haenszel, Hammond et al. (1959)). The method has seen later development by a numerous authors, including, Bross (1966), Schlesselman (1978), Rosenbaum and Rubin (1983), Yanagawa (1984), and Lin, Psaty, and Kronmal (1998). The books of Rothman and Greenland (1998) and Rosenbaum (2002) give in-depth coverage of the topic.

Most sensitivity analysis is characterized by the use of models of unmeasured confound-ing which are not identifiable, in the sense that bias parameters cannot be estimated consistently using the available data. For example, Schlessleman (1978) proposes a model relating the ob-served relative risk to the true relative risk acting under the influence of the confounding effect of an unmeasured dichotomous variable U. His model incorporates three bias parameters: the

2

relative risk of disease due to U, the prevalence of U in the exposed group and the prevalence of U in the unexposed group. But since U is never observed to begin with, estimating any of these three quantities is not possible.

Early pioneers in sensitivity analysis sought to overcome the challenge non-identifiability by using information external to the study to specify a range of possible values for the bias parameters. The investigator is then free to plug in parameter values, rendering the model identifiable. The remaining model parameters including the exposure effect can then be estimated using standard inference techniques. But such "externally adjusted" estimates suffer from the obvious limitation that different choices of bias parameters may result in different effect estimates depending on where information on bias parameters is obtained. Rosenbaum and Rubin (1983) use a more conservative approach which involves presenting effect estimates in tabular format using a range of possible bias parameters.

But this strategy also has limitations. The resulting tables are unwieldy and burden the presentation of results. Lash (2003) notes that since existing custom in the biomedical literature does not demand a quantitative assessment of systematic error due to unmeasured confounding, authors who struggle to meet space restrictions find the inclusion of such tables unappealing. One can also argue that specifying ranges of bias parameters may yield conclusions which are unnecessarily pessimistic. Investigators tend to focus on effect estimates derived from the most extreme bias parameters under consideration. What is lost is the fact that small biases are often more plausible than large biases.

Bayesian methods are well suited to sensitivity analysis because beliefs about the magnitude of unmeasured confounding can be incorporated into the analysis as prior distributions on bias parameters. The posterior distribution of the exposure effect yields a single summary which quantifies uncertainty due to unmeasured confounding in addition to random error. This approach makes effective use of all available information and more accurately captures the levels of bias that can be induced by unmeasured confounding.

A related method is Monte Carlo Sensitivity Analysis (MCSA) for quantifying uncer-

tainty due to systematic error in observational studies (Greenland (2001, 2003, 2005), Lash and Fink (2003) and Philips (2003)). The MCSA procedure is implemented as follows: First, the usual non-identifiable model is obtained which is indexed by a set of bias parameters. The investigator then specifies a distribution for the bias parameters and samples from this distribution repeatedly. Each random draw of bias parameters is substituted into the model to yield a corresponding effect estimate. The resulting distribution of effect estimates is conveniently summarized to yield an interval of values for the exposure effect which quantifies possible systematic error. Sampling variability can also be incorporated into the sensitivity analysis by using resampling techniques or methods based on the standard error of the effect estimates. (See Greenland (2001, 2005) for more details).

MCSA shares many of the benefits of Bayesian approaches to sensitivity analysis and has the intuitive appeal of attempting to capture prior beliefs about bias parameters in a distribution rather than a range of values. MCSA implementation is somewhat less technical than Bayesian methods and sidesteps the difficulties of Markov chain Monte Carlo sampler convergence. But the difficulty is that the approach is not fully Bayesian, in the sense of applying Bayes rule to incorporate prior beliefs about unmeasured confounding into the analysis. Greenland (2001) notes that MSCA intervals cannot be interpreted as frequentist confidence intervals, in the sense of proper 95% coverage over repeated sampling. A Bayesian interpretation of MCSA intervals may also be problematic because the MCSA distribution of effect estimates is not necessarily coherent with prior beliefs about systematic error. Consequently, interpreting MCSA intervals may be challenging because of difficulties in assessing their coverage and optimality properties.

Bayesian approaches to sensitivity analysis have been the topic of a number of recent articles including Robins, Rotnitzky and Scharfstein (1999), Greenland (2001, 2003, 2005), Scharfstein, Daniels and Robins (2003), Oakley and O'Hagan (2004) and Steenland and Greenland (2004). The approach of Greenland (2003) involves a dichotomous exposure X, outcome Y and unmeasured confounder U, where subjects are classified on X and Y in a standard 2 $\times$ 2 table. Greenland considers the unmeasured 2 $\times$ 2 $\times$ 2 table which is further stratified

4

over levels of U and models the expected cell counts using a log-linear model with parameters modeling the dependence and interaction between X, Y and U. His method might benefit further if measured covariates could be included in the analysis.

Accordingly, we consider Bayesian sensitivity analysis in the context of a binary exposure X, binary outcome Y, binary unmeasured confounder U and covariate vector Z. In Chapter 2, we present a model for unmeasured confounding using logistic regression models and include a discussion of prior specification for bias parameters and a posterior simulation using Markov Chain Monte Carlo methods. Chapter 3 discusses the results of two simulation studies which evaluate the performance of BSA relative to other methods of analysis which ignore unmeasured confounding. As an example of the method, Chapter 4 presents the implementation of BSA in a secondary analysis of a recent study which identified large increases in plasma triglycerides among HIV-infected persons treated with protease inhibitor. Chapter 5 concludes with a discussion and directions for further research.

Chapter 2

Bayesian Sensitivity Analysis

2.1 A model for unmeasured confounding

Let X, Y and U denote dichotomous variables taking values 1 or 0 to indicate the presence or absence of an exposure, disease or unmeasured confounder. Let Z denote a covariate vector with components which may be discrete or continuous.

If we assume that associations between X, Y and U are not causal effects of X or Y on U, then U will be a confounder if (1) U and Y are conditionally dependent given X, (2) U and X are conditionally dependent given Y (Rothman and Greenland (1995)). To model these dependencies, we specify the joint distribution distribution of X, Y, U and Z as

$$P(Y, X, U, Z) = P(Y|X, U, Z)P(U|X, Z)P(X, Z). \tag{2.1}$$

Previous models of unmeasured confounding with a dichotomous outcome, exposure, and confounder employ logistic regression models (Rosenbaum (2002), Lin, Psaty, and Kronmal (1998)). We follow in the same spirit letting

$$\text{logit}[Pr(Y = 1|X, U, Z)] = \alpha_0 + \alpha_1 X + \lambda U + Z\beta, \tag{2.2}$$

$$\text{logit}[Pr(U = 1|X, Z)] = \gamma_0 + \gamma_1 X \tag{2.3}$$

A few comments are in order. First, we assume that X and U do not interact in their

influence on Y. Extending equation (2.2) to include interaction terms is straightforward, but creates confusion about the targets of inference. Adding a term such as "$..+\alpha_2 XU+..$" to equation (2.2) would imply that both α_1 and α_2 govern the effect of X on Y and therefore a study of the exposure effect would need to consider each parameter separately, or some defensible average of the two. Interaction terms also requires the elicitation of additional prior distributions for non-identifiable parameters which may be challenging. Greenland (2003) discusses these issues in a similar setting and argues that including interaction terms has only a modest effect on inferences obtained from sensitivity analysis.

Additionally, note that Equation (2.3) assumes that U and Z are conditionally independent given X. This assumption is technically incorrect because if U and Z are both confounders of the association between X and Y, then this implies that U and Z must be conditionally dependent given X (Hernan and Robins (1999)). However, provided that Z is only weakly associated with X, equation (2.3) may be considered an adequate approximation.

As with other models for unmeasured confounding, our model is non-identifiable and indexed by a set of bias parameters γ_0, γ_1 and λ which reflect assumptions about the confounding effect of U. The quantity $\exp(\lambda)$ is the conditional odds ratio of the association between U and Y given X and Z, and the quantity $\exp(\gamma_1)$ is the conditional odds ratios of the association between U and X given Y and Z. Similarly, the prevalences of U in the unexposed and exposed groups are given by $\frac{\exp(\gamma_0)}{1+\exp(\gamma_0)}$ and $\frac{\exp(\gamma_0+\gamma_1)}{1+\exp(\gamma_0+\gamma_1)}$. Non-identfiability is apparent from the fact that any two pairs of bias parameters $(\lambda,\gamma_0,\gamma_1)^o=(0,a,b)$ or $(\lambda,\gamma_0,\gamma_1)'=(0,-a,-b)$ yield identical likelihoods for the observed data. To illustrate, first notice that among subjects with exposure X and covariate Z, the risk of disease is an average of the unobservable U-specific risks of

7

disease. This average is given by

$$
\begin{aligned}
P(Y = 1|X, Z) &= P(Y = 1|X, U = 1, Z)P(U = 1|X, Z) + \\
&\quad P(Y = 1|X, U = 0, Z)P(U = 0|X, Z) \\
&= \left(\frac{\exp(\alpha_0 + \alpha_1 X + \lambda + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + \lambda + Z'\beta)} \right) \left(\frac{\exp(\gamma_0 + \gamma_1)}{1 + \exp(\gamma_0 + \gamma_1)} \right) + \\
&\quad \left(\frac{\exp(\alpha_0 + \alpha_1 X + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + Z'\beta)} \right) \left(\frac{1}{1 + \exp(\gamma_0 + \gamma_1)} \right).
\end{aligned}
$$

But now we see that substitution of either $(\lambda, \gamma_0, \gamma_1)^o$ or $(\lambda, \gamma_0, \gamma_1)'$ into the above expression yields identical expressions for the marginal risk of Y. For $(\lambda, \gamma_0, \gamma_1)^o = (0, a, b)$, we have

$$
\begin{aligned}
P(Y = 1|X, Z) &= \left(\frac{\exp(\alpha_0 + \alpha_1 X + 0U + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + 0U + Z'\beta)} \right) \left(\frac{\exp(a + b)}{1 + \exp(a + b)} \right) + \\
&\quad \left(\frac{\exp(\alpha_0 + \alpha_1 X + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + Z'\beta)} \right) \left(\frac{1}{1 + \exp(a + b)} \right)
\end{aligned}
$$

and for $(\lambda, \gamma_0, \gamma_1)' = (0, -a, -b)$ we have

$$
\begin{aligned}
P(Y = 1|X, Z) &= \left(\frac{\exp(\alpha_0 + \alpha_1 X + 0U + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + 0U + Z'\beta)} \right) \left(\frac{\exp(-a + -b)}{1 + \exp(-a + -b)} \right) + \\
&\quad \left(\frac{\exp(\alpha_0 + \alpha_1 X + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + Z'\beta)} \right) \left(\frac{1}{1 + \exp(-a + -b)} \right).
\end{aligned}
$$

These two expressions are identical. Hence two different sets of bias parameters yield the same distribution for the observed data. Distinguishing between either set of parameters is not possible implying that the model is not identifiable.

2.2 Prior specification

Since $\alpha_0, \alpha_1, \beta$ are the primary parameters of interest and do not govern the confounding be-haviour of U, each is assigned a non-informative prior distribution which is independent normal with mean zero and variance equal to ten. Such a distributions asserts that, a priori, each parameter lies between $\pm 1.96\sqrt{10}$ on the log odds scale with 95% probability or equivalently, between $\exp(-1.96\sqrt{10}) \approx 1/500$ and $\exp(1.96\sqrt{10}) \approx 500$ on the odds scale.

For suitable values of c_1, c_2 and c_3, we assign informative prior distributions for the bias parameters $\lambda, \gamma_0,$ and γ_1 as follows. Let

$$f(\lambda) \propto \exp\left\{-\frac{\lambda^2}{2c_1}\right\}$$

represent the prior belief about the association between U and Y. The density $f(\lambda)$ is sym-metric and centered at zero to reflect the fact that nothing is known about the direction of the association between U and Y. The variable U is not simply unmeasured but is also *unknown* and therefore little can be said about whether λ is positive or negative. But this fact does not preclude a priori assertions about the possible magnitude of λ. Large biases are unlikely since covariate information on strong risk factors for Y is usually collected at the design stage. Con-sequently, we can specify the parameter c_1 to restrict prior beliefs about the possible magnitude of λ.

A prior distribution for γ_0 and γ_1 is assigned on the truncated bivariate normal distri-bution

$$f(\gamma_0, \gamma_1) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix}'\begin{pmatrix} c_2 & -c_3/2 \\ -c_3/2 & c_3 \end{pmatrix}^{-1}\begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix}\right\} \qquad \text{for } \gamma_1 \geq 0. \qquad (2.4)$$

To understand this choice of prior, ignore the restriction $\gamma_1 \geq 0$ for the moment and consider a prior distribution on the entire bivariate normal density. We then have $\text{Var}(\gamma_0) = c_2$, $\text{Var}(\gamma_1) = c_3$ and $\text{Var}(\gamma_0 + \gamma_1) = \text{Var}(\gamma_0) + \text{Var}(\gamma_1) + 2\text{Cov}(\gamma_0, \gamma_1) = c_2 + c_3 - 2c_3/2 = c_2$. Therefore, c_3 limits the strength of the possible association between U and X. The parameter c_2 restricts the possible

magnitude of γ_0 and $\gamma_0 + \gamma_1$ and consequently, the exposure-specific prevalences of U which are given by $\frac{\exp(\gamma_0)}{1+\exp(\gamma_0)}$ and $\frac{\exp(\gamma_0+\gamma_1)}{1+\exp(\gamma_0+\gamma_1)}$. Note that this prior distribution asserts that γ_0 and γ_1 are dependent a priori.

Imposing $\gamma_1 \geq 0$ changes the prior variances of γ_0, γ_1, and $\gamma_0 + \gamma_1$, but does not affect on the resulting posterior distributions for the exposure effect α_1. The reason for this is that U is unobserved and hence the labeling $U = 1$ or $U = 0$ is somewhat arbitrary. Without the requirement that $\gamma_1 \geq 0$, we have an identifiability problem in which different sets of bias parameters result in an identical distribution for Y given X and Z except for a switching of the labels $U = 1$ and $U = 0$. To see why this is so, recall that given exposure X and covariates Z, the distribution of Y can be written as

$$
\begin{aligned}
P(Y = 1|X, Z) ={}& P(Y = 1|X, U = 1, Z)P(U = 1|X, Z) + \\
& P(Y = 1|X, U = 0, Z)P(U = 0|X, Z) \\
={}& \left(\frac{\exp(\alpha_0 + \alpha_1 X + \lambda + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + \lambda + Z'\beta)} \right) \left(\frac{\exp(\gamma_0 + \gamma_1)}{1 + \exp(\gamma_0 + \gamma_1)} \right) + \\
& \left(\frac{\exp(\alpha_0 + \alpha_1 X + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + Z'\beta)} \right) \left(\frac{1}{1 + \exp(\gamma_0 + \gamma_1)} \right). \quad (2.5)
\end{aligned}
$$

But suppose that we instead define the unmeasured confounder as $U^* = 1 - U$, by switching the labels of U. If we used the choice of bias parameters $(\lambda, -\gamma_0, -\gamma_1)$, then the distribution of Y, averaged over U^*, would be written as

$$
\begin{aligned}
P(Y = 1|X, Z) ={}& P(Y = 1|X, U^* = 1, Z)P(U^* = 1|X, Z) + \\
& P(Y = 1|X, U^* = 0, Z)P(U^* = 0|X, Z) \\
={}& P(Y = 1|X, U = 0, Z)P(U = 0|X, Z) + \\
& P(Y = 1|X, U = 1, Z)P(U = 1|X, Z) \\
={}& \frac{\exp(\alpha_0 + \alpha_1 X + \lambda + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + \lambda + Z'\beta)} \frac{1}{1 + \exp(-\gamma_0 + -\gamma_1)} + \\
& \frac{\exp(\alpha_0 + \alpha_1 X + Z'\beta)}{1 + \exp(\alpha_0 + \alpha_1 X + Z'\beta)} \frac{\exp(-\gamma_0 + -\gamma_1)}{1 + \exp(-\gamma_0 + -\gamma_1)}.
\end{aligned}
$$

But this distribution is identical to the distribution in equation (2.5).

Hence requiring that γ_1 be greater than or equal to zero implies the a priori assumption that U is more prevalent in the exposed group than the unexposed group. The result is gains in efficiency of posterior simulation. The posterior distribution for the exposure effect α_1 remains unchanged by this restriction. Therefore, the prior distribution in equation (2.4) can be seen as equivalent to the same distribution without the restriction $\gamma_1 \geq 0$. This logic permits the convenient interpretations of the parameters c_2 and c_3 described above. For a related discussion on non-identifiability due to switching of parameter labels, see Shaffer and Chinchilli (2004).

It should also be noted that there are at least two approaches to specifying prior distributions for bias parameters. Philips (2003) describes *Bias-level* sensitivity analysis and *Target-adjustment* sensitivity analysis. The former, proceeds by specifying a prior distribution for bias parameters, a priori, to yield a posterior distribution for the exposure effect which accounts for bias uncertainty. Target-adjustment sensitivity analysis operates in reverse. First, a posterior distribution is obtained which reflects only the usual random error. Then a family of prior distributions for bias parameters is identified, such that the respective posterior distributions for members of this family yield no clear decision about the effect of the exposure. If any member of this family is deemed a reasonable representation of prior beliefs about bias parameters, then conclusions about the existence of an exposure effect in the original analysis can be more easily disputed. An example of Target-adjustment sensitivity analysis is presented by Greenland (2003). Although our method is conducive to either approach, we favor Bias-level sensitivity analysis in this investigation. Philips (2003) notes that Target-adjustment sensitivity analysis is primarily concerned with assessing the plausibility of competing hypotheses. But this strategy may be less useful if investigator is interested in determining plausible ranges of the exposure effect.

2.3 Posterior simulation

Let y, x and u denote vectors of length n of the observed responses, exposures and unmeasured confounders in n study subjects and and let z denote an $n \times p$ matrix of measured covariates. Our objective is to summarize the posterior distribution of the exposure effect given y, x and z and prior assumptions about unmeasured confounding described in Section 2.2. This can be accomplished via posterior simulation using Markov Chain Monte Carlo (MCMC) techniques that treat U as a latent variable which is integrated out of the joint posterior distribution of model parameters and U:

$$f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1 | y) = \int f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, u | y) du. \tag{2.6}$$

Simulating from either $f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1 | y)$ or $f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, u | y)$ directly is challenging. Instead, we sample sequentially from the conditional distributions

$$f(\alpha_0, \alpha_1, \lambda, \beta | \gamma_0, \gamma_1, u, y), \quad f(\gamma_0, \gamma_1 | \alpha_0, \alpha_1, \lambda, \beta, u, y), \quad \text{and} \quad f(u | \alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, y). \tag{2.7}$$

This approach follows the general procedure of data augmentation which was first formalized by Tanner and Wong (1987), and can be shown to ultimately yield a sample from the desired posterior distribution $f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, u | y)$.

To sample from each of the conditional densities in equation (2.7), we begin with an expression for the joint distribution of y, u and model parameters

$$
\begin{aligned}
f(y, \alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, u) &= f(y | \alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, u) f(u | \gamma_0, \gamma_1) f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1) \\
&= \prod_{i=1}^{n} \left[\left(\frac{\exp(y_i(\alpha_0 + \alpha_1 x_i + \lambda u_i + z_i \beta))}{1 + \exp(\alpha_0 + \alpha_1 x_i + \lambda u_i + z_i \beta)} \right) \left(\frac{\exp(u_i(\gamma_0 + \gamma_1 x_i))}{1 + \exp(\gamma_0 + \gamma_1 x_i)} \right) \right] \times \\
&\quad f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1,)
\end{aligned} \tag{2.8}
$$

where y_i and u_i denote the response and unmeasured confounder in the ith study subject, $i \in 1, \ldots, n$.

By conditioning appropriately, equation (2.8) yields unormalized expressions for each of

the conditional distributions in (2.7),

$$f(\alpha_0, \alpha_1, \lambda, \beta | \gamma_0, \gamma_1, u, y) \quad \propto \quad \prod_{i=1}^{n} \left[\frac{\exp(y_i(\alpha_0 + \alpha_1 x_i + \lambda u_i + z_i \beta))}{1 + \exp(\alpha_0 + \alpha_1 x_i + \lambda u_i + z_i \beta)} \right] f(\alpha_0, \alpha_1, \lambda, \beta), \quad (2.9)$$

$$f(\gamma_0, \gamma_1 | \alpha_0, \alpha_1, \lambda, \beta, u, y) \quad \propto \quad \prod_{i=1}^{n} \left[\frac{\exp(u_i(\gamma_0 + \gamma_1 x_i))}{1 + \exp(\gamma_0 + \gamma_1 x_i)} \right] f(\gamma_0, \gamma_1), \quad (2.10)$$

$$f(u_i | \alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, y_i) \quad \propto \quad \frac{\exp(y_i \lambda u_i + u_i \gamma_0 + u_i \gamma_1 x_i)}{1 + \exp(\alpha_0 + \alpha_1 + \lambda u_i + z_i \beta)}. \quad (2.11)$$

In equations (2.9) and (2.10), $f(\alpha_0, \alpha_1, \lambda, \beta)$ and $f(\gamma_0, \gamma_1)$ denote the prior distributions for $\alpha_0, \alpha_1, \beta, \lambda, \gamma_0$ and γ_1 which are specified in Section 2.2.

Simulation from the desired posterior distribution, $f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, u | y)$, can now be accomplished by simulating sequentially from each of the conditional densities (2.9), (2.10) and (2.11). Sampling from $f(u | \alpha_0, \alpha_1, \lambda, \gamma_0, \gamma_1, y)$ is the most straightforward. Since u_i are dichotomous, the expression in equation (2.11) is easily normalized to yield the distribution of u_i given $\alpha_0, \alpha_1, \beta$ and y_i.

$$P(u_i = 1 | \alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1, y_i) = \frac{\frac{\exp(y_i \lambda + \gamma_0 + \gamma_1 x_i)}{1 + \exp(\alpha_0 + \alpha_1 + \lambda + z_i \beta)}}{\left[\left(\frac{\exp(0)}{1 + \exp(\alpha_0 + \alpha_1 + z_i \beta)} \right) + \left(\frac{\exp(y_i \lambda + \gamma_0 + \gamma_1 x_i)}{1 + \exp(\alpha_0 + \alpha_1 + \lambda + z_i \beta)} \right) \right]}.$$

The u_i are conditionally independent given x_i and model parameters, therefore sampling from the conditional density in equation (2.11) can be accomplished directly using a Bernoulli random variable.

To sample from the conditional distributions $f(\alpha_0, \alpha_1, \lambda, \beta | \gamma_0, \gamma_1, u, y)$ and $f(\gamma_0, \gamma_1 | \alpha_0, \alpha_1, \lambda, \beta, u, y)$, note that both densities are simply the usual posterior distributions for logistic regression. The expression in equation (2.9) is derived from the posterior distribution of a logistic regression of y on x, z and u. Equation (2.10) is obtained from a similar regression of u on x.

Posterior simulation for logistic regression can be accomplished using the Metropolis Hastings algorithm with an independence sampler. (See Chib and Greenberg (1995) for an accessible introduction). The algorithm is an iterative procedure for sampling from a target density $f(\theta) = \frac{\pi(\theta)}{k}$, where k is an unknown constant of normalization. The implementation proceeds as follows: At iteration i, given a current sampled parameter value $\theta^{(i)}$, a candidate

13

value θ^* is generated from a candidate generating density $Q(\theta^*)$ which possibly depends on $\theta^{(i)}$. We assign $\theta^{(i+1)} \leftarrow \theta^*$ with probability

$$\min \left[\frac{f(\theta^*)Q(\theta^{(i)})}{f(\theta^{(i)})Q(\theta^*)}, 1 \right] \qquad (2.12)$$

or assign $\theta^{(i+1)} \leftarrow \theta^{(i)}$ otherwise. For a sufficiently large burn-in size k and simulation size M, $\theta^{(k+1)}, \theta^{(k+2)}, \dots, \theta^{(k+M)}$ constitute a dependent sample from the target density $f(\theta)$.

The choice of candidate generating density $Q(\theta^*)$ impacts the performance of the Metropolis Hastings algorithm. For posterior simulation with logistic regression models, a common choice is an a multivariate normal density with mean equal to the maximum likelihood estimate of the model parameters and covariance matrix given by the inverse of the observed information. This candidate distribution is known to well approximate the posterior distribution in large samples and yields high acceptance rates for candidate values. This choice of $Q(\theta^*)$ is an example of an *independence sampler* Metropolis Hastings algorithm because the candidate density does not depend on the current parameter values $\theta^{(i)}$. Another popular class of candidate distributions admits the so called *random walk* Metropolis Hastings algorithm. An example is $Q(\theta^*)$ equal to a multivariate normal density, centered at $\theta^{(i)}$, with a suitable choice of covariance matrix. At the ith iteration, a candidate θ^* is generated consisting of $\theta^{(i)}$ plus noise. However, random walk Metropolis Hastings algorithms tend to perform poorly in logistic regression models, particularly for parameters of high dimension. Finding a suitable covariance matrix for the candidate density can be onerous with a poor choice leading to either high rates of rejection or slow movement movement through the target density. It is for this reason that independence sampler candidate distributions are popular for logistic regression.

Adapting this idea in the present context, we use the Metropolis Hastings algorithm to sample from the conditional densities $f(\alpha_0, \alpha_1, \lambda, \beta | \gamma_0, \gamma_1, u, y)$ and $f(\gamma_0, \gamma_1 | \alpha_0, \alpha_1, \lambda, \beta, u, y)$. In both instances, a candidate density is obtained using a multivariate normal density with mean and covariance matrix obtained from fitting logistic regression models of y on x, z and u for the conditional distribution (2.9) and u on x for the conditional density in (2.11).

To summarize, first let $\alpha_0^{(i)}$, $\alpha_1^{(i)}$, $\lambda^{(i)}$, $\beta^{(i)}$, $\gamma_0^{(i)}$, $\gamma_1^{(i)}$ denote the sampled parameter values and let $u^{(i)}$ denote the simulated unmeasured covariate vector for n subjects at the ith iteration. Simulation from the desired posterior distribution $f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1 | y)$ proceeds through the following five steps:

1. Obtain starting values for the unknown parameters $\alpha_0^{(0)}$, $\alpha_1^{(0)}$, $\lambda^{(0)}$, $\beta^{(0)}$, $\gamma_0^{(0)}$, $\gamma_1^{(0)}$ and $u^{(0)}$. Our method is to simulate the components of $u^{(0)}$ as Bernoulli(1/2) and then choose values for $\alpha_0^{(0)}$, $\alpha_1^{(0)}$, $\lambda^{(0)}$ and $\beta^{(0)}$ by regressing y on x, z, and $u^{(0)}$ and choose values for $\gamma_0^{(0)}$ and $\gamma_1^{(0)}$ by regressing $u^{(0)}$ on x.

2. Generate $\alpha_0^{(i+1)}, \alpha_1^{(i+1)}, \lambda^{(i+1)}, \beta^{(i+1)}$ from $f(\alpha_0, \alpha_1, \lambda, \beta | \gamma_0^{(i)}, \gamma_1^{(i)}, u^{(i)}, y)$ using the Metropolis Hastings algorithm and a candidate generating density with mean and covariance obtained fitting logistic regression of y on x, z and $u^{(i)}$.

3. Generate $\gamma_0^{(i+1)}, \gamma_1^{(i+1)}$ from $f(\gamma_0, \gamma_1 | \alpha_0^{(i+1)}, \alpha_1^{(i+1)}, \lambda^{(i+1)}, \beta^{(i+1)}, u^{(i)}, y)$ using the Metropolis Hastings algorithm and a candidate generating density with mean and covariance obtained by fitting logistic regression of $u^{(i)}$ on x.

4. Draw $u^{(i+1)}$ from the conditional distribution $f(u | \alpha_0^{(i+1)}, \alpha_1^{(i+1)}, \lambda^{(i+1)}, \beta^{(i+1)}, \gamma_0^{(i+1)}, \gamma_1^{(i+1)}, y)$ which can be determined exactly as outlined above.

5. Repeat steps 2 through 4 a total of $M + k$ times where M is the desired number of posterior simulations and k is the number of burn-in iterations.

6. Discard $\{u^{(i)} | i \in 1, \ldots, M\}$ to obtain $\{\alpha_0^{(i)}, \alpha_1^{(i)}, \lambda^{(i)}, \beta^{(i)}, \gamma_0^{(i)}, \gamma_1^{(i)}, | i \in 1, .., M\}$, a posterior sample from the desired posterior distribution $f(\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1 | y)$.

The sample $\alpha_1^{(1)}, \alpha_1^{(2)}, \ldots, \alpha_1^{(M)}$ is then a sample from the posterior distribution of the exposure effect given the data and prior beliefs about possible unmeasured confounding. We can see that the general procedure involves treating u as a latent variable which is simulated in Step 4 and then used to repeatedly fit logistic regression of y on x, u and z in step 2 to obtain a sample

15

from the posterior distribution of α_1 as u is averaged over the prior distribution on the bias parameters λ, γ_0 and γ_1 according to Bayes rule.

The high acceptance rates of our choice of candidate generating distribution are advantageous, but the scheme also has limitations. Iterative simulation through steps 2 to 4 requires fitting logistic regression models for each sampled value of u. This approach is very computationally intensive, particularly if standard regression packages are used. By writing a function in the R language which maximized the log likelihood using the Newton Raphson method with the ith sampled parameters as starting values for the maximization, substantial gains in speed were obtained compared to the glm() function. While slow posterior simulation is unlikely to be a problem for Bayesian sensitivity analysis of a single dataset, it is a major obstacles in the repeated analysis of simulated datasets. This issue is discussed in greater detail in Chapter 3.

A more serious challenge is that the latent $u^{(i)}$ can periodically take on values which prevent fitting logistic regression on $u^{(i)}$. By the very nature of the prior distributions specified in Section 2.2, the bias parameters λ, γ_0 and γ_1 can occasionally take on extreme values depending on the choice of c_1, c_2 and c_3 in equations (2.2) and (2.4). Extreme values of λ or γ_1 will yield values of $u^{(i)}$ which are highly correlated with the response or exposure respectively. Similarly, extreme values of γ_0 can result in values of $u^{(i)}$ with zero variation. Each of these instances can prevent fitting the regressions in steps 2 and 3 because the likelihood function cannot be easily maximized. Moreover, these events do not occur infrequently. Using a typical dataset of sample size 1000 with reasonable choices for c_1, c_2, and c_3, regression model fitting fails about once out of every 50000 posterior simulations.

Two possible solutions are implemented with good results. The simplest solution is to restrict the values to c_1, c_2 and c_3 thereby decreasing the prior probability of bias parameters which are large in magnitude. This method clearly has limitations because it prevents a sensitivity analysis involving large possible biases due to unmeasured confounding. A more complicated solution is to create an alternative procedure for assigning candidate distribution in steps 2 and 3. The Newton Raphson algorithm is an iterative procedure which maximizes

the log likelihood by attempting to obtain a sequence of parameter values which converge to a maximum. But this procedure will tend to fail if the second derivative of the log likelihood is small over the parameter space, as is typically the case if the variation of a covariate is small or if two covariates are highly correlated. In such instances the sequence of parameter values in the Newton Raphson algorithm may behave unpredictably and move rapidly away from the maximum to regions of the parameter space where the second derivative of the log likelihood tends to zero. A strategy for assigning a candidate distribution in steps 2 and 3 is carry out the Newton Raphson algorithm as usual, but condition on the values of the second derivative (or determinant of second derivative in the multivariate case). If the magnitude of the second derivative passes below some tolerance level, then fitting the regression is assumed to be not possible and an alternative scheme for assigning a candidate distribution is chosen. Our preference is for a candidate distribution based on a random walk which samples each component of the regression model one at a time using a normal distribution and Metropolis Hastings steps. Regardless of the choice of candidate distribution at each MCMC iteration, we can compensate appropriately by suitably modifying the Metropolis Hastings acceptance ratio. Hence the Markov property of the algorithm is preserved and convergence to the target distribution is guaranteed. Of course, the candidate distribution for the random walk component must be independently tuned to ensure a suitable acceptance rate.

Chapter 3

Two simulation studies

3.1 A simulation study using fixed parameter values

To contrast the performance of Bayesian sensitivity analysis to other statistical methods which ignore unmeasured confounding, we consider a comparison based on the analysis of simulated datasets. Datasets are simulated for three sample sizes, $n = 250, 1000, 4000$, and two choices of fixed parameters

$$
\begin{aligned}
\theta^+ &= (\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1)^+ \\
&= \left(\log\left(\frac{2}{3}\right), \log\left(\frac{9}{4}\right), \log\left(\frac{9}{4}\right), (\log(2), -\log(2))', \log\left(\frac{2}{3}\right), \log\left(\frac{9}{4}\right) \right)
\end{aligned}
$$

and

$$
\begin{aligned}
\theta^- &= (\alpha_0, \alpha_1, \lambda, \beta, \gamma_0, \gamma_1)^- \\
&= \left(\log\left(\frac{2}{3}\right), \log\left(\frac{9}{4}\right), -\log\left(\frac{9}{4}\right), (\log(2), -\log(2))', \log\left(\frac{2}{3}\right), \log\left(\frac{9}{4}\right) \right).
\end{aligned}
$$

In this setting, Z is a continuous vector of length 2 and the parameters θ^+ and θ^- are chosen to yield reasonable prevalences of Y and U and dependencies between Y, X, U and Z. In particular, the following properties hold:

- For $U = 0$ and $Z = (0,0)'$ the prevalence of the response Y in the exposed and unexposed groups are $\exp(\alpha_0 + \alpha_1)/(1 + \exp(\alpha_0 + \alpha_1)) = 60\%$ and $\exp(\alpha_0)/(1 + \exp(\alpha_0)) = 40\%$

respectively.

- The prevalences of the unmeasured confounder U in the exposed and unexposed groups are $\exp(\gamma_0 + \gamma_1)/(1 + \exp(\gamma_0 + \gamma_1)) = 60\%$ and $\exp(\gamma_0)/(1 + \exp(\gamma_0)) = 40\%$ respectively.

- The conditional odds ratio of the association between X and Y, given U and Z, is $\exp(\alpha_1) = 9/4$.

- The conditional odds ratios of the associations between the components of Z and Y, given X and U, is $\exp(\beta) = (2, 1/2)'$.

- The conditional odds ratio of association between U and X, given Z and Y, is $\exp(\gamma_1) = 9/4$.

- The conditional odds ratio of association between U and Y, given X and Z, is $\exp(\lambda) = 9/4$ for θ^+ and $\exp(\lambda) = 4/9$ for θ^-.

To understand the basis for selecting θ^+ and θ^-, first note that the true odds ratio for the exposure effect of X on Y is $\exp(\alpha_1) = 9/4$ for either parameter. But since U and X have a positive association, and since U is either positively (using θ^+) or negatively (using θ^-) associated with Y, this means that the effect of X on Y is confounded by U. Analyses which do not adjust for U will yield estimates of α_1 which are either biased positively (using θ^+) or biased negatively (using θ^-).

Using θ^+ and θ^-, we seek to simulate a collection of datasets for analysis with different methods. This provides a basis for comparing methods. To simulate a dataset of sample size n using the fixed parameters θ^+ or θ^-, we first simulate the exposure X as Bernoulli(1/2) and the components of Z as Uniform(-2,2). We then simulate the unmeasured confounder U for each subject using γ_0, γ_1 and the simulated X according equation (2.3). The responses Y for each subject are simulated with α_0, α_1, λ, β and the covariates X, U and Z using equation (2.2).

To study the performance of Bayesian sensitivity analysis on simulated datasets, denoted henceforth as BSA, we consider a comparison with two other inference techniques which we denote as NAIVE and GOLD. NAIVE is standard logistic regression of the observed data Y on X and Z, and is the usual technique for analyzing data with a binary outcome while ignoring unmeasured confounding. Since NAIVE does not adjust for U, the posterior distribution for the exposure effect α_1 will tend to be biased. GOLD is logistic regression of Y on the complete data X, U and Z under the assumption that U is measured. While GOLD cannot be implemented in practice, it is a useful benchmark for comparison with BSA and NAIVE.

For BSA, NAIVE and GOLD analyses of simulated datasets, we assign uninformative prior distributions for α_0, α_1 and β which are independent normal with mean zero and variance equal to 10. (See Section 2.1 for a discussion of this choice of prior distribution). For BSA, we specify prior distributions for bias parameters λ, γ_0 and γ_1 according to Section 2.2 with c_1, c_2 and c_3 chosen to model prior beliefs about the possible magnitude of unmeasured confounding. We assign $c_1 = (\log(6)/1.96)^2$ and $c_3 = (\log(6)/1.96)^2$ to model the prior belief that the log odds ratio of the associations between U and Y, and between U and X are normally distributed with 95% of the prior distribution lying between $\log(1/6)$ and $\log(6)$. We assign $c_2 = (\log(6)/1.96)^2$ to represent the prior belief that the exposure specific logit prevalences of U are normally distributed with 95% of the distribution lying between $\text{logit}(1/(6+1))$ and $\text{logit}(6/(1+6))$.

For each of the fixed parameters θ^+ and θ^-, three datasets of sample sizes $n = 250, 1000$ and 4000 are simulated and then analyzed using BSA, NAIVE and GOLD. Posterior samples of size 50000 (1000 burn-in) are used to construct posterior histograms of the exposure effect and are displayed in Figures 3.1 and 3.2. Note that by construction, the true exposure effect for each of the datasets is known to be $\alpha_1 = \log(9/4)$. In Figures 3.1 and 3.2, this is indicated with a vertical bar in each of the nine graphs. Examining the figures, it is clear that the NAIVE histograms are biased positively in Figure 3.1 or negatively in Figure 3.2 relative to GOLD and will likely yield poor coverage of the true exposure effect for large sample sizes. BSA posterior histograms are wider than NAIVE histograms, reflecting additional uncertainty about

Figure 3.1: Posterior histograms of the exposure effect α_1 for BSA, NAIVE and GOLD analyses of datasets with sample sizes $n =250$, 1000 and 4000 using the parameter θ^+.
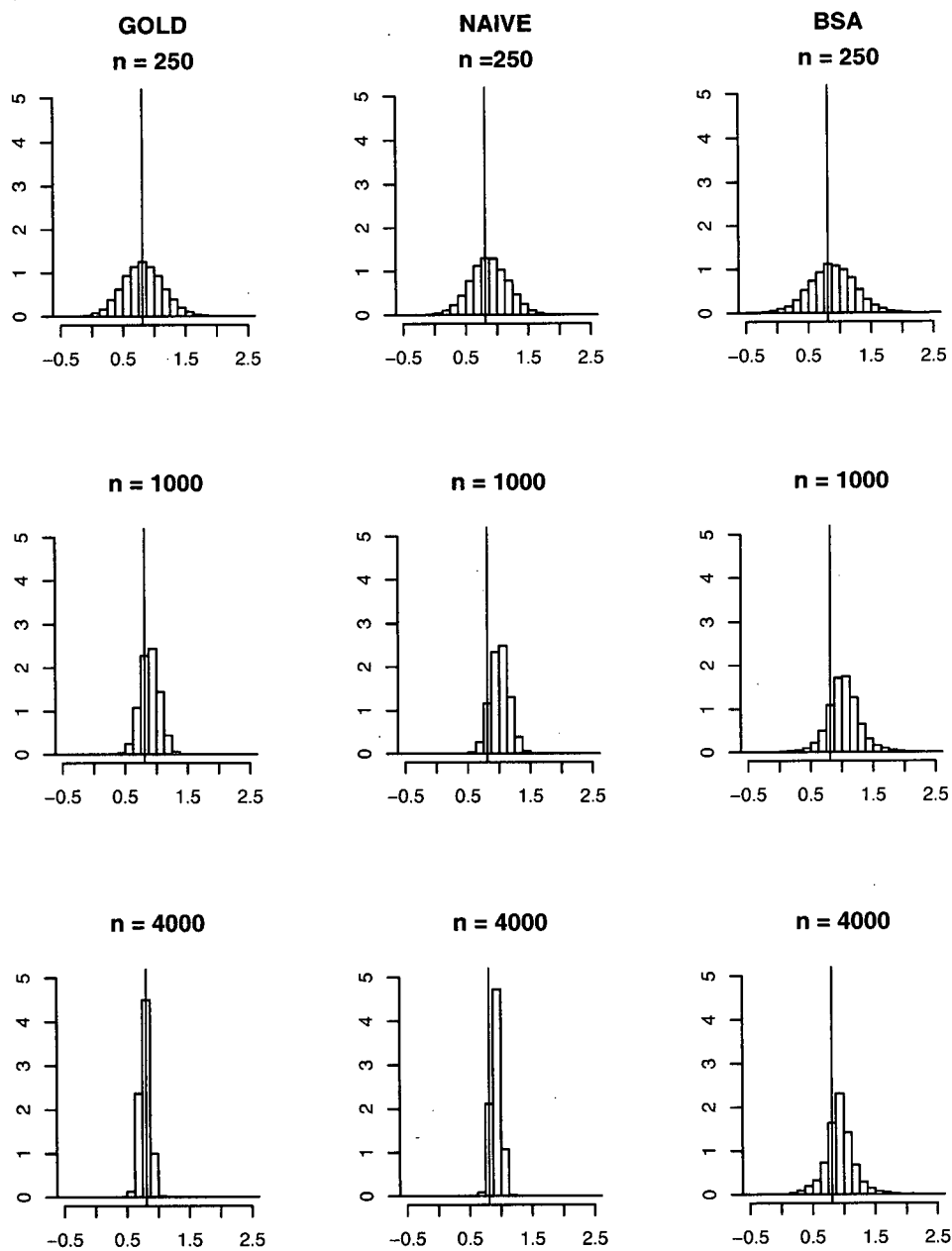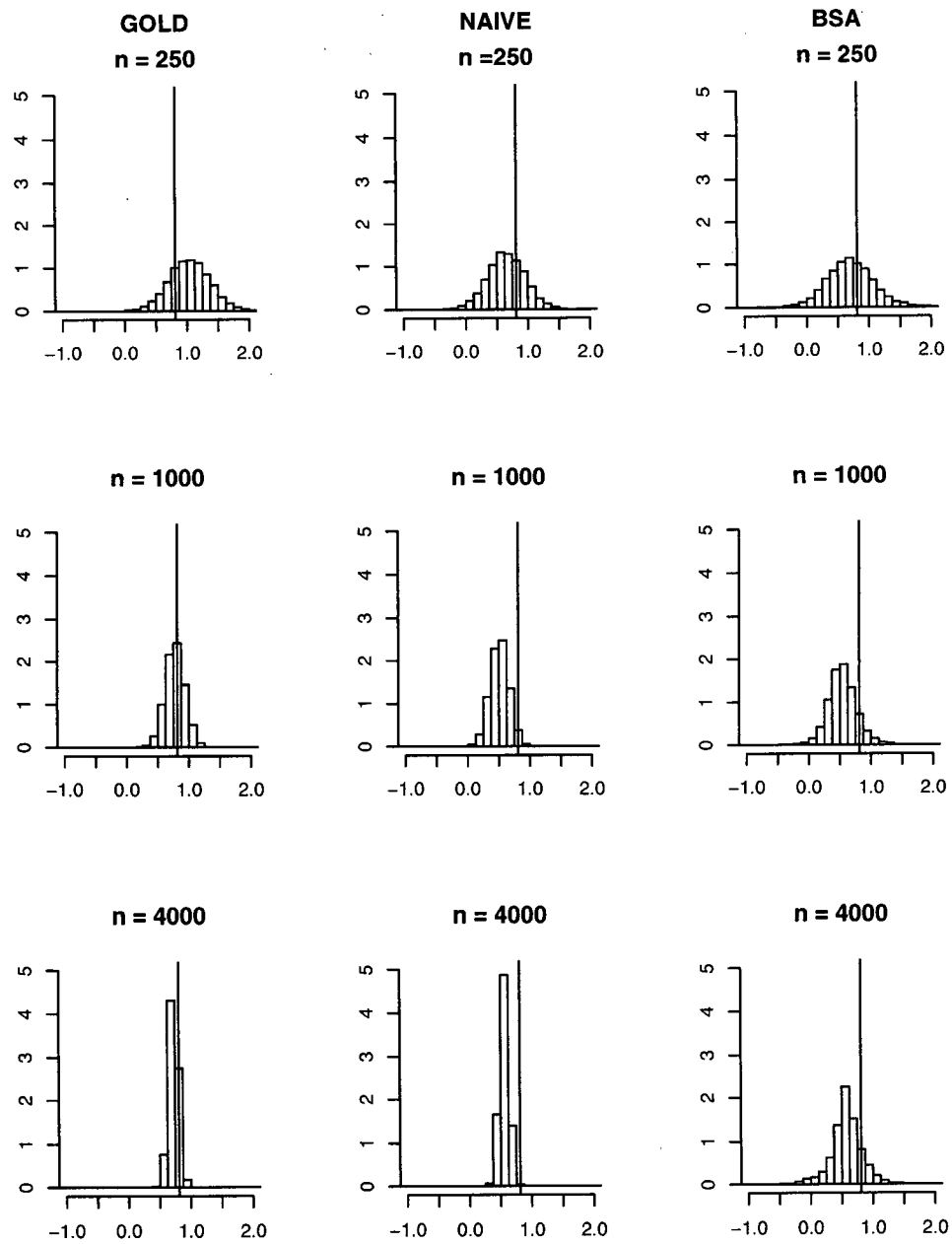
Figure 3.2: Posterior histograms of the exposure effect α_1 for BSA, NAIVE and GOLD analyses of datasets with sample sizes $n =$ 250, 1000 and 4000 using the parameter θ^-.

unmeasured confounding that is modelled by the prior distributions on the bias parameters λ, γ_0 and γ_1. Figures 3.1 and 3.2 suggest that BSA credible intervals may yield better coverage of the true exposure effect than NAIVE intervals for these choices of fixed parameters.

We note in passing that the MCMC mixing behaviour of BSA is poor. Figures 3.3 and 3.4 depict MCMC chains for model parameters based on 1000 posterior simulations (after burn-in) from a simulated dataset with parameter θ^+ and a sample size of 1000. For the parameters $\alpha_0, \alpha_1, \beta$, mixing appears to be adequate, but for the bias parameters $\lambda, \gamma_0, \gamma_1$, many iterations are required for the Markov chain to move thoroughly through the support of the target distribution. Moreover, the slow mixing appears to worsen as sample size increases. This is demonstrated in Figure 3.5, which displays MCMC chains for the bias parameter λ, based on three datasets with sample sizes n=250, 1000, 4000 with fixed parameter θ^+. While 50000 posterior simulations may be adequate in the context of Figures 3.1 and 3.2, poor mixing presents a much more serious obstacle in the *repeated* analysis of simulated datasets. Unfortunately, there does not appear to be a straightforward solution to this problem because of current limitations in computing resources. Modifying the MCMC algorithm might yield modest improvement and will be investigated at a later time. Poor mixing in non-identifiable models has also be observed by Gustafson (2005) and Gelfand and Sahu (1999) who discuss the problem in greater detail.

Naturally, the posterior histograms in Figure 3.1 and 3.2 are specific to the choice of simulated datasets. Hence the performance of BSA relative to NAIVE and GOLD is compared under repeated simulations. For sample sizes $n =250, 1000, 4000$ and parameters θ^+ and θ^-, 400 datasets are simulated and then analysed using BSA, NAIVE and GOLD. Posterior samples of size 2000 (500 burn-in) are obtained for each dataset and used to construct highest posterior density (HPD) 80% credible intervals. We choose the α-level of 80% rather than the usual 95% to more easily highlight the differences in coverages between methods.

A posterior sample size of 2000 in clearly inadequate given the poor mixing behaviour of BSA, but this brings to bear the computational challenges highlighted at the end of Section 2.2. BSA requires fitting two logistic regression models for each MCMC iteration, and thus repeated

Figure 3.3: MCMC mixing for model parameters $\alpha_0, \alpha_1, \lambda, \beta_1, \beta_2$ based on 1000 posterior simulations (after burn-in) from a dataset with fixed parameter θ^+ with n=1000.
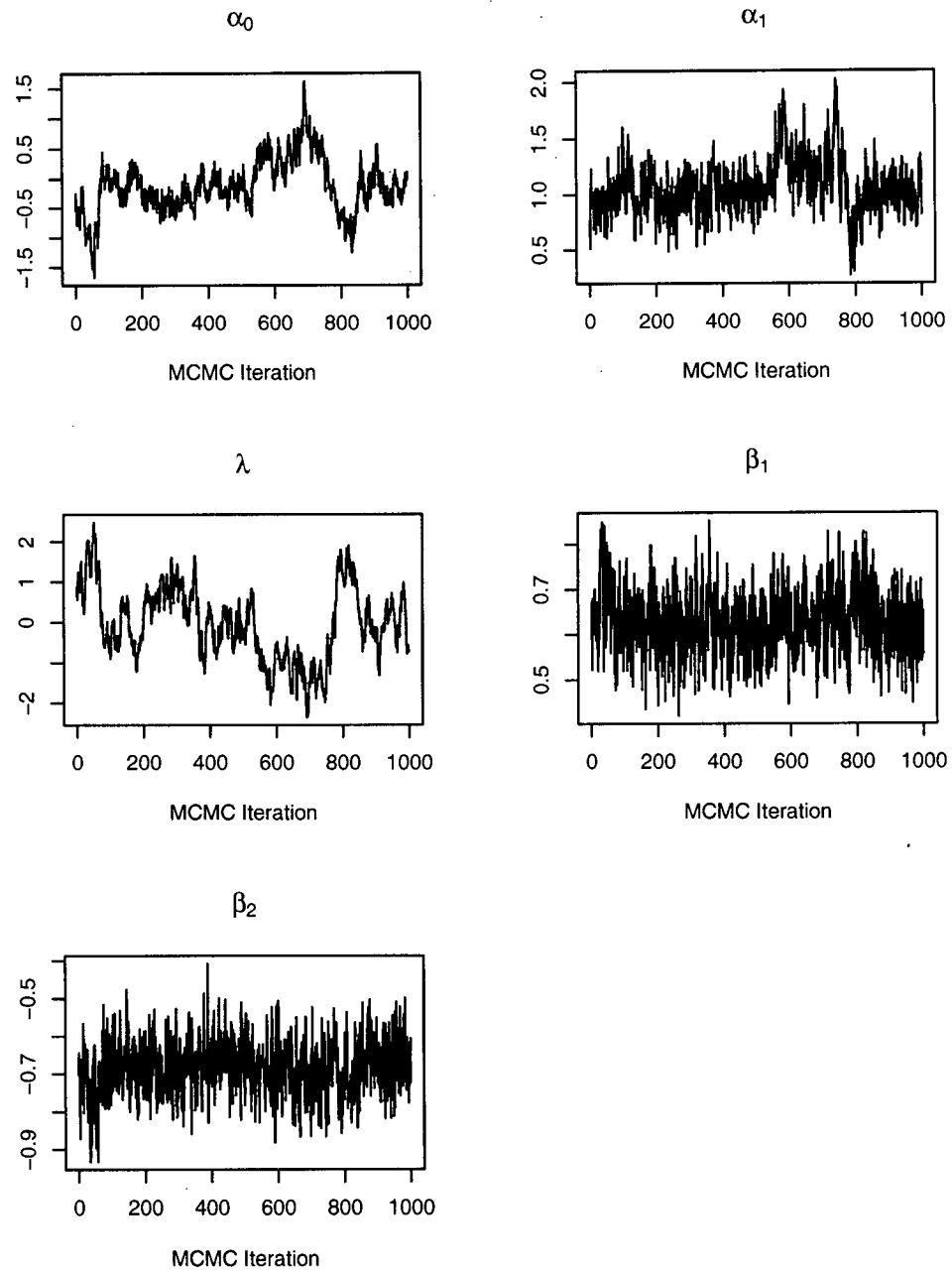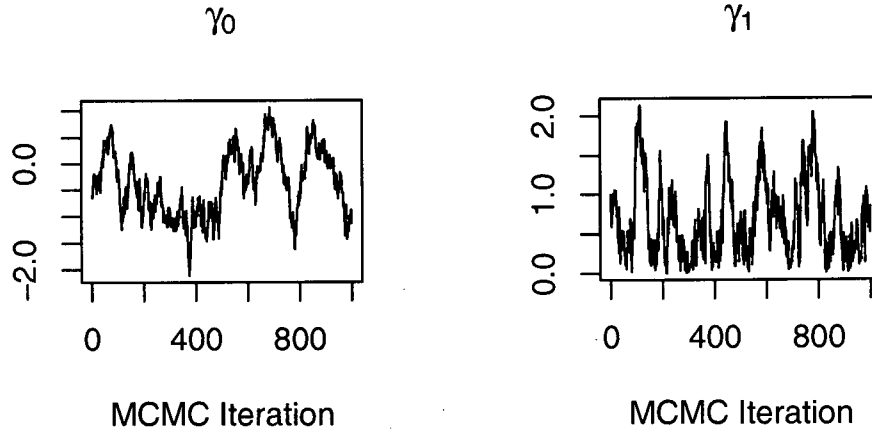
Figure 3.4: MCMC mixing for model parameters γ_0, γ_1 based on 1000 posterior simulations (after burn-in) from a dataset with fixed parameter θ^+ with n=1000.

posterior simulation from simulated datasets can be time consuming. Using a personal computer with dual 1.7GHz processors and 1Gb of memory, the simulation results presented in Chapter 3 alone require more than one hundred hours of computing time.

Table 3.1: Empirical coverages of the exposure effect α_1 and average length of 80% HPD credible intervals for BSA, GOLD and NAIVE analyses of datasets simulated with parameter θ^+.

Sample sizes		GOLD	NAIVE	BSA
250	Coverage	77.5%	67.5%	81%
	Length	0.791	0.787	0.911
1000	Coverage	77.5%	56.2%	80.8%
	Length	0.39	0.377	0.596
4000	Coverage	80%	27.3%	71%
	Length	0.194	0.189	0.447

Tables 3.1 and 3.2 display the results of the simulation study including the empirical coverage probabilities of the true exposure effect, and the average credible interval length for BSA, NAIVE and GOLD credible intervals. In both tables, the empirical coverage of GOLD

Figure 3.5: MCMC mixing for the bias parameter λ based on 1000 posterior simulations (after burn-in) from three datasets with fixed parameter θ^+ and increasing sample sizes n=250, 1000, 4000.

n = 250

MCMC Iteration

n = 1000

MCMC Iteration
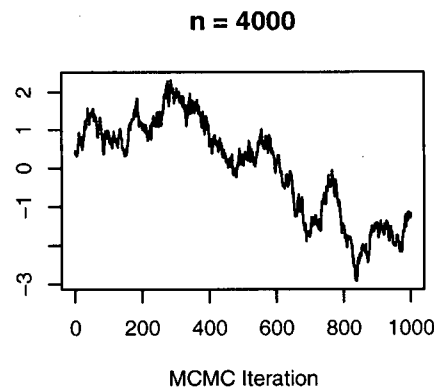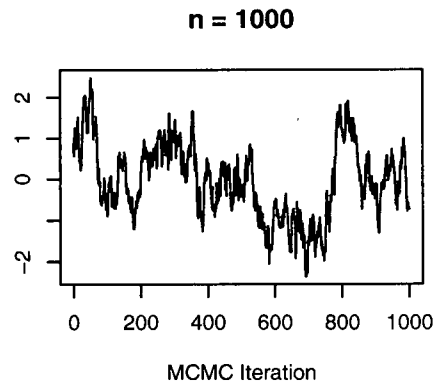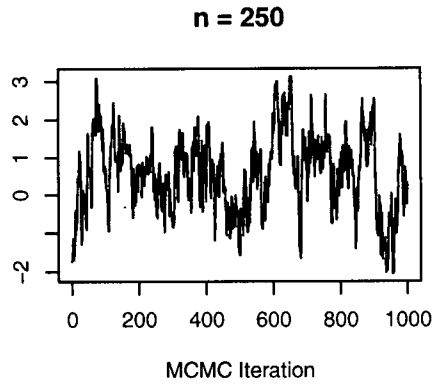
n = 4000

MCMC Iteration

Table 3.2: Empirical coverages of the exposure effect α_1 and average length of 80% HPD credible intervals for BSA, GOLD and NAIVE analyses of datasets simulated with parameter θ^-.

Sample sizes		GOLD	NAIVE	BSA
250	Coverage	81.8%	71%	81.2%
	Length	0.796	0.749	0.912
1000	Coverage	79.5%	48.2%	80.5%
	Length	0.395	0.37	0.585
4000	Coverage	76.5%	11.5%	69.2%
	Length	0.194	0.185	0.458

intervals approaches 80% as the sample sizes increases. This result is not unexpected because GOLD correctly adjust for U. As the sample size increases, large sample theory dictates that the GOLD posterior distribution is asymptotically normal with mean equal to the true value of α_1. Credible intervals obtained from GOLD approximately match the usual frequentist confidence intervals and therefore, the GOLD coverage approaches 80%. In contrast, NAIVE coverage tends to degrade as sample size increases because of bias due to U. In Figures 3.1 and 3.2, we see that the NAIVE posterior distribution tends to be asymptotically biased suggesting that NAIVE coverages tends to zero with increasing sample size. BSA credible intervals provide better coverage of the true exposure effect relative to NAIVE and this appears to be true for our six choices of sample size and fixed parameter values.

Comparing the lengths of BSA and NAIVE intervals, we see that improved coverage comes at the expense of greater uncertainty. As expected, the length of GOLD and NAIVE credible intervals decreases at the usual $\sqrt{n}$ rate, but BSA intervals shrink more slowly. In particular, the asymptotic distribution of the BSA interval is non-degenerate, since the model is non-identifiable, and therefore BSA intervals may be considerably longer than their NAIVE counterparts. For example, with sample size 1000 and the selected prior distributions and fixed parameters, BSA credible intervals appear to be roughly fifty percent longer than NAIVE intervals reflecting a substantial increase in uncertainty.

Our conclusions are limited to the specific choices of fixed parameters and prior distributions and it is clearly desirable to explore the performance of BSA in greater generality. It is this idea that provides the motivation for the simulation study in Section 3.2.

3.2 A simulations study using sampled parameter values

Simulations using fixed parameters may not be the best method to evaluate the performance of BSA. As suggested in Tables 3.1 and 3.2, the improvement in coverage for BSA relative to NAIVE depends on the choice of parameters used to simulate datasets. Parameters which reflect large biases (i.e. values of λ, γ_0, γ_1 which lie far from the regions of highest density in the prior distributions (2.2) and (2.4)) yield BSA credible intervals with poor coverage of the true exposure effect, particularly for large sample sizes. Coverage improves when simulating datasets using fixed parameters which reflect smaller levels of bias. Fixing $\lambda = 0$ or $\gamma_1 = 0$ to reflect no unmeasured confounding yields BSA 95% credible intervals with coverage of the exposure effect which *exceed* 95%.

Indeed since the posterior distribution for BSA does not obey the usual large sample rules governing identifiable models, it may be unreasonable to study it's performance using standard simulation methods. In identifiable models, large sample theory dictates that under the usual regularity conditions, the posterior mean is asymptotically normal, with mean equal to the true parameter value and that credible intervals have approximately matching frequentist coverage. Hence simulations with fixed parameter values can be used to demonstrate acceptable coverage in large samples. In non-identifiable models, posterior distributions have different large sample properties. The posterior mean is again asymptotically normal but may be biased in a manner which is difficult predict and dependent on the choice of fixed parameters and prior distributions (Gustafson (2005)). Hence there is no reason to believe that in non-identifiable models, empirical coverages of credible intervals will tend to 95% in large samples and consequently, simulation studies using fixed parameters may not be informative about the overall performance of BSA. This is not a deficiency of Bayesian methods. Without identifiability, no statistical method will afford interval estimates which have correct frequentist coverage levels.

An alternative is to simulate datasets using a range of different parameters values. One might specify a distribution function over the parameter space and use it to sample parameters

for the simulation study. Instead of simulating many datasets for a fixed parameter, a single dataset could be simulated for each of many sampled parameters. Assessing BSA using these simulated datasets might be more representative of overall performance of the method as a whole.

This approach has interesting implications if we consider the following fact: If the distribution used to simulate parameters is also used as a prior distribution for Bayesian analysis of simulated datasets, then the empirical coverages of simulated parameters will be exactly 95%, regardless of whether the model is identifiable or not. Specifically, consider a parameter θ and model $f(y|\theta)$ for the data y. Let $f(\theta)$ denote a density function for θ. Denoting $\theta_1, \ldots, \theta_k$ as a sample of size k from $f(\theta)$, we obtain k simulated datasets $y_1, \ldots, y_k$ where the ith dataset is simulated from $f(y|\theta_i)$. Using $f(\theta)$ as a prior distribution for Bayesian analysis, we obtain k posterior distributions $f(\theta|y_1), \ldots, f(\theta|y_k)$ where $f(\theta|y_i) \propto f(y_i|\theta)f(\theta)$. Since the distribution used to simulate the $\theta_i's$ is identical to that used for analysis of simulated datasets, this implies that $1 - \alpha$ posterior credible intervals obtained from $f(\theta|y_i)$ will cover the respective θ_i with probability equal to $1 - \alpha$ exactly. Moreover, this is not a large sample property. Proper $1 - \alpha$ coverage of credible intervals is achieved regardless of sample size. Rubin (1984) discusses this idea and refers to such a sequence of intervals as *well calibrated*.

Why is this result useful in the current context? Tables 1 and 2 suggest that coverage of the true exposure effect for BSA may be low in some instances or high in others. But one can argue that coverages may equal 95% when averaged over many different studies. Let us suppose that model parameters in a sequence of observational studies arise from some unknown density which we refer to as "Nature's prior". Data are collected in each study and posterior distributions are obtained using a separate prior distribution which we denote as "Investigator's prior". To obtain a sequence of credible intervals which cover the true exposure effect with probability $1 - \alpha$, it is clearly desirable to set the Investigator's prior equal to Nature's prior.

Now practically speaking, Nature's prior cannot be known with certainty by the investigator and it would seem that obtaining such a sequence of credible intervals would be

difficult if not impossible. But it is possible that small deviations between the Investigator's prior and Nature's prior might not greatly affect the calibration. Using a reasonable choice for the Investigator's prior, the empirical coverages of 95% credible intervals over a sequence of different studies may remain close to 95% on average. Moreover, even if the two distributions are different, BSA may still yield intervals which outperform intervals obtained using methods which ignore unmeasured confounding.

To explore this idea, we conduct a simulation using simulated parameters where the Investigator's prior for Bayesian analysis of simulated datasets differs from Nature's prior used to sample parameters. If we assume that both Nature's prior and the Investigators prior have the parametric form of equations (2.2) and (2.4), then it suffices to assess the effects of differences between Investigator's prior and Nature's prior exclusively for the parameters λ, γ_0 and γ_1. We can reasonably assume that in a sequence of studies, values of α_0, α_1 and β arise from a distribution which is relatively uninformative (diffuse independent normal) and that an identical uninformative prior distribution is used to obtain posterior distributions for α_0, α_1 and β. For the bias parameters λ, γ_0 and γ_1, we model Nature's prior and the Investigator's prior following equations (2.2) and (2.4) using different combinations of the variance parameters c_1, c_2 and c_3 to reflect differences between the two distributions.

For simplicity and ease of presentation, we assume that $c_1 = c_2 = c_3$ for both distributions. Recall that c_1 and c_3 model prior beliefs about the association between U and Y, and between U and X respectively. The constant c_2 models prior beliefs about the prevalence of U for unexposed subjects. Letting $c_I = c_1 = c_2 = c_3$ for the Investigator's prior and $c_N = c_1 = c_2 = c_3$ for Nature's prior, we contrast c_I and c_N for combinations of the values $(\log(3)/1.96)^2, (\log(6)/1.96)^2$ and $(\log(9)/1.96)^2$. These values are chosen in the same spirit as detailed in Section 3.1 to reflect realistic levels of confounding one might encounter in practice. For the parameters $\alpha_0, \alpha_1, \beta$, we assign identical distributions for Nature's prior and Investigators prior which are independent normal with mean zero and with variance equal to $(\log(15)/1.96)^2$. While this distribution is not strictly uninformative, it is chosen based on

31

the practical realities of conducting this type of simulation. Using a more diffuse distribution can result in sampled parameters values which are very large, thereby preventing regression analyses from working correctly during posterior simulation. Moreover, in some sense, these prior distributions are uninformative because they are diffuse relative to the typical effect sizes in epidemiological investigations.

Table 3.3: Empirical coverages of the exposure effect α_1 and average length of 80% HPD credible interval for various c_I and c_N.

	Investigator	Nature		
		$c_N = \left(\frac{\log(3)}{1.96}\right)^2$	$c_N = \left(\frac{\log(6)}{1.96}\right)^2$	$c_N = \left(\frac{\log(9)}{1.96}\right)^2$
	$c_I = \left(\frac{\log(3)}{1.96}\right)^2$	78%, 0.540	73%, 0.525	67%, 0.533
BSA	$c_I = \left(\frac{\log(6)}{1.96}\right)^2$	90%, 0.654	77%, 0.651	74%, 0.653
	$c_I = \left(\frac{\log(9)}{1.96}\right)^2$	91%, 0.736	85%, 0.741	82%, 0.740
NAIVE		74%, 0.503	70%, 0.490	65%, 0.512

We simulate 400 datasets of sample size 1000 using parameters which are sampled from Nature's prior with c_N equal to $(\log(3)/1.96)^2, (\log(6)/1.96)^2$ or $(\log(9)/1.96)^2$. Each dataset is then analysed using BSA with three choices for the Investigator's prior corresponding to $c_I = (\log(3)/1.96)^2, (\log(6)/1.96)^2$ or $(\log(9)/1.96)^2$. For comparison, NAIVE analyses are also conducted. 2000 posterior simulations (500 burn-in) are obtained for each dataset and method of analysis and choice of c_N, to obtain 80% credible intervals for the exposure effect. As in Section 3.1, we choose 80% credible intervals to more easily highlight differences in coverages. We are also faced with the same limitations of using small posterior samples of size 2000.

Table 3.3 presents the results of the simulation. The upper half of the table shows a comparison of the empirical coverage probabilities and average length of 80% credible intervals for the exposure effect using BSA and different combinations of c_N and c_I. As expected, the

diagonal elements of the table are consistent with 80% reflecting proper coverage of the interval procedure because of equality between Nature's prior and the Investigator's prior. Indeed, standard errors for the coverage estimates, obtained using the usual formula $1.96\sqrt{\hat{p}(1-\hat{p})/400}$, are all less than 2.5% indicating agreement with 80%. (2.5% corresponds to the worst case scenario with an empirical coverage percentage of 50%.) Correct 80% coverages indicate that simulation error due small posterior samples is not terribly important.

On the off-diagonals, we see overcoverage or undercoverage of the true exposure effect depending on whether c_I is greater or less than c_N. For example, a choice of $c_I = (\log(9)/1.96)^2$ and $c_N = (\log(3)/1.96)^2$ models the instance where the investigator believes that large biases are present when in fact only small biases arise due to nature. The result is incorrect calibration. BSA credible intervals tend to cover the true exposure effect too often (91% of the time in our simulation), because they are too great in length. For NAIVE analyses, we assume that bias due to unmeasured confounding is absent and use uninformative prior distributions to obtain credible intervals for α_0, α_1 and β. Consequently, we need only evaluate the performance of NAIVE for different values of c_N. The results of this comparison are shown at the bottom of Table 3.3 and indicate that for large values of c_N, coverage of NAIVE intervals is poor but tends to improve as c_N approaches zero.

Here we see the merit of BSA relative to NAIVE. For NAIVE analyses, proper 80% coverage is achieved only in the absence of unmeasured confounding. For $c_N = 0$, corresponding to the instance of no unmeasured confounding, the NAIVE posterior distribution is asymptotically normal with mean equal to the true exposure effect. Hence in large samples, NAIVE credible intervals have correct coverage probabilities irrespective of the value of the true exposure. But for most observational studies, the possibility of unmeasured confounding is everpresent and therefore c_N should be non-zero. In this case, BSA credible intervals will achieve correct 80% coverage when the Investigator's prior approximates Nature's prior.

Critics may argue that since Nature's prior is unknown, there can be no guarantee that BSA credible intervals will achieve proper coverage. But Table 3.3 illustrates that sizeable

deviations between Nature's prior and Investigator's prior may still result in improved coverage of BSA relative to NAIVE. For example, a choice of $c_N = (\log(9)/1.96)^2$ and $c_I = (\log(6)/1.96)^2$, reflecting sizeable differences between distributions, results in an empirical coverage of 73.8% indicating a sizeable improvement over the NAIVE coverage of 64.8% using the same choice for Nature's prior. Since the risk of unmeasured confounding is commonplace in observational studies, Table 3.3 suggests that an educated choice for the Investigator's prior may yield credible intervals which cover the true exposure effect more frequently on average than NAIVE intervals when averaged over many different studies.

Our investigation is not without limitations. While BSA appears to outperform NAIVE in these simulations, we have by no means performed an exhaustive comparison of the effects of differences between the two distributions. The parameter c_2, which reflects prior belief about the prevalence of the unmeasured confounder, may have a different impact on overall calibration compared to c_1 or c_3. Exploring a comparison without the assumption that $c_1 = c_2 = c_3$ for Nature's prior and the Investigator's prior would likely provide insight into such behaviour. Additionally, we base our comparisons on the assumption that the Nature's prior and the Investigator's prior follow the parametric models in equations (2.2) and (2.4). An investigation of more complicated models might reveal subtleties that are lost in this simulation.

Chapter 4

An application in the study of treatment for HIV/AIDS

4.1 Protease inhibitors and risk of elevated triglycerides

We consider a secondary analysis of a recent study by Levy, McCandless, Harrigan et al. (2005) which identified increases in plasma triglyceride levels among HIV-infected persons treated with triple drug combination therapy including a protease inhibitor (PI). Unscheduled blood lipoprotein measurements were taken from 709 HIV infected patients who were treated through the British Columbia Centre for Excellence in HIV/AIDS Drug Treatment Program between August 1996 and January 2002. A record of patient data including demographic characteristics, use of antiretroviral medication, and CD4 counts was maintained for all subjects. In regression analyses adjusting for age, sex, adherence, time-dependent CD4 count, time-dependent concurrent use of other therapy and correlation due to repeated measures on study subjects, Levy, McCandless, Harrigan et al. (2005) found that uninterrupted use of PI was associated with an annual increase of 22.8% (7.9%, 39.8%) in triglyceride levels. This estimate was obtained by restricting the analysis to subjects who were ever dispensed a PI over the course of therapy and by stratifying triglyceride measurements according the magnitude of cumulative exposure

to PI at the time each measurement was taken.

We mimic this analysis in the context of a dichotomous exposure and dichotomous response by estimating impact of PI exposure on risk of elevated triglyceride levels (≥ 5.6mmol/L). (National Cholesterol Education Program Adult Treatment Panel III (2001) guidelines). We conduct a retrospective study of antiretroviral naive patients who initiated triple drug combination therapy between August 1996 and January 2002 and were dispensed a PI at least once over the course of follow up. For the ith subject, with measurements $j \in 1 \ldots T_i$, we observe the binary response Y_{ij} taking values 1 or 0 to indicated presence or absence of elevated triglycerides. We let x_{ij} denote a binary indicator variable taking values 1 or 0 to denote whether the ith subject was dispensed a PI at the beginning of the month in which the jth measurement was taken. Fixed covariates including age, sex, baseline CD4 count and adherence to prescribed therapy are denoted with the vector z_i. Adherence is calculated as the ratio the number of months in which antiretroviral medications were dispensed to the number of months of follow-up in the first year after therapy initiation. Incomplete adherence represents the gap between the time that the previous medication supply ran out until the next refill date. For details on this method for quantifying adherence see Wood, Hogg, Yip et al. (2003,2004).

Classifying subjects as "exposed" or "unexposed" to PI is challenging because all subjects included in the analysis are dispensed a PI at least once over the course of follow-up. If we assume that the number and times of triglyceride measurements occur at random and in a way which does not depend on the data, then we can attempt to classify exposure status based on whether the subject was receiving a PI at the times when measurements were taken. Stratifying the n subjects into groups A, B or C where

$$
\begin{aligned}
A &= \{i \in 1, \ldots, n | x_{ij} = 1 \text{ for } \forall j \in 1, \ldots, T_i\}, \\
B &= \{i \in 1, \ldots, n | x_{ij} = 0 \text{ for } \forall j \in 1, \ldots, T_i\}, \\
C &= \{i \in 1, \ldots, n | x_{ia} = 0, \ x_{ib} = 1 \text{ for some } a, b \in 1, \ldots, T_i\},
\end{aligned}
$$

we assign exposure status x_i for the ith subject according to the rule

$$x_i = 1 \quad \text{if } i \in A \quad \text{(exposed)},$$ (4.1)

$$x_i = 0 \quad \text{if } i \in B \quad \text{(unexposed)},$$

where x_i is a binary indicator variable taking values 1 or 0 to indicate exposed or unexposed to PI. Therefore, subjects in group A (exposed) are those individuals who were receiving PI therapy at all of the times when measurements were taken. Group B (unexposed) are subjects who were not receiving therapy at all of the times when measurements were taken. Note that group B has non-zero membership because the measurement times for the ith subject need not correspond to months receiving PI therapy.

Of course, this leaves an ambiguity for subjects in group C, the individuals who were receiving treatment at some measurement times but not others. But any attempt to dichotomize exposure status for these individuals will be unsatisfactory. For subjects in groups A and B, the classification scheme in equation (4.1) is reasonable if we assume that there are no carry-over effects from changes in therapy. Subjects in groups A or B are observed as either always exposed or always unexposed at measurement times $1, \ldots, T_i$ and we may attempt to classify them as such. For subjects in group C, we choose to move these individuals into group B and classify them as unexposed by discarding all "exposed" data pairs (Y_{ij}, x_{ij}) where $x_{ij} = 1$. This classification decision has certain consequences which become apparent when we consider the classification of disease status.

For each subject, assign disease status Y_i according to

$$Y_i = \max\{Y_{ij} | j \in 1 \ldots T_i\}.$$ (4.2)

Hence individuals are classified as diseased if they are ever observed to have elevated triglycerides over the course of follow-up. Triglyceride measurements prior to therapy initiation are not available for most of the sample, but since triglyceride levels exceeding 5.6mmol/L are rare among men and women of the same age in the general population (MacLean, Petrasovits,

Connelly et al. (1999)), we assume that all subjects had triglyceride levels less than 5.6mmol/L prior to initiation of therapy. Based on this assumption, Y_i is an indicator of incident cases of elevated triglycerides which occurred after PI exposure.

A consequence of this disease classification is that subjects with many measurements over the course of follow-up are more likely to be classified as diseased. If the distribution of the number of measurements in exposed and unexposed groups were identical, this would not present a problem. But unfortunately, this is not the case. Subjects in group A are likely to have more measurements than subjects in group C by virtue of the fact that all subjects are exposed to PI at some point over the course of therapy. Hence, our method of assessing disease status has the consequence of inflating the observed risk of disease in the exposed group relative to the unexposed group. There does not appear to be a straightforward way of correcting this problem because the indicators of disease status Y_{ij} are time-varying. The allocation of subjects from group C to group B as described above will offset this bias somewhat.

4.2 Bayesian sensitivity analysis

We identify 709 eligible subjects (91% male) with a median age of 38 years (IQR 33-44) and median CD4 count of 210 cells/μL(IQR 70-380) at the time of initiation of therapy. Based on records of PI dispensation at the times of triglyceride measurements, 333 subjects are classified as exposed and 376 as unexposed to PI. A total of 2216 measurements are used to identify 93 incident cases of elevated triglycerides.

Cross tabulating subjects by exposure and disease status

	Diseased	Not Diseased
Exposed	62	271
Not exposed	31	345

yields a crude log odds ratio of 0.93 with 95% confidence interval (0.48, 1.39) suggesting that PI exposure is associated with a increased risk of elevated tryglycerides. The corresponding odds ratio is 2.53 (1.62, 4.01).

Treating Y as disease status, X as exposure status and Z as the vector of covariates age, sex, adherence and baseline CD4, we contrast NAIVE and BSA analyses of the data according to the definitions of Section 3.1. Age is coded in years, sex as binary with the value one for males and zero for females, adherence as a percentage between one and one hundred, and CD4 as cells per micro liter. Prior to analysis, all covariates are rescaled to mean zero and unit variance to improve MCMC posterior simulation. For the NAIVE analysis, we use uninformative priors for the regression coefficients (independent normal with mean 0 and variance 10). For BSA, we use the model and prior distributions of Sections 2.1 and 2.2 with three choices of prior distributions for the bias parameters λ, γ_0, γ_1. Setting $c = c_1 = c_2 = c_3$, we choose prior distributions with values of c equal to $(\log(3)/1.96)^2$, $(\log(6)/1.96)^2$ or $(\log(9)/1.96)^2$ to reflect three choices of increasing degrees of uncertainty about the possible magnitude of unmeasured confounding.

Table 4.1: Posterior means (95% HPD credible intervals) for covariate effects using NAIVE and BSA analyses. Values of c reflect three choices of prior distributions for bias parameters.

Variable	NAIVE		BSA $c = \left(\frac{\log(3)}{1.96}\right)^2$		$c = \left(\frac{\log(6)}{1.96}\right)^2$		$c = \left(\frac{\log(9)}{1.96}\right)^2$	
PI Exposure	0.84	(0.36,1.31)	0.85	(0.34,1.35)	0.86	(0.27,1.50)	0.88	(0.15,1.62)
Gender	0.1	(-0.13,0.33)	0.1	(-0.13,0.33)	0.1	(-0.14,0.33)	0.1	(-0.13,0.34)
Baseline Age	0.02	(-0.21,0.25)	0.02	(-0.21,0.25)	0.02	(-0.22,0.25)	0.02	(-0.22,0.26)
Baseline CD4	0.55	(0.17,0.96)	0.55	(0.17,0.95)	0.55	(0.18,0.97)	0.56	(0.16,0.96)
Adherence	0.88	(-0.17,1.96)	0.88	(-0.16,1.99)	0.89	(-0.12,2.05)	0.91	(-0.18,2.01)

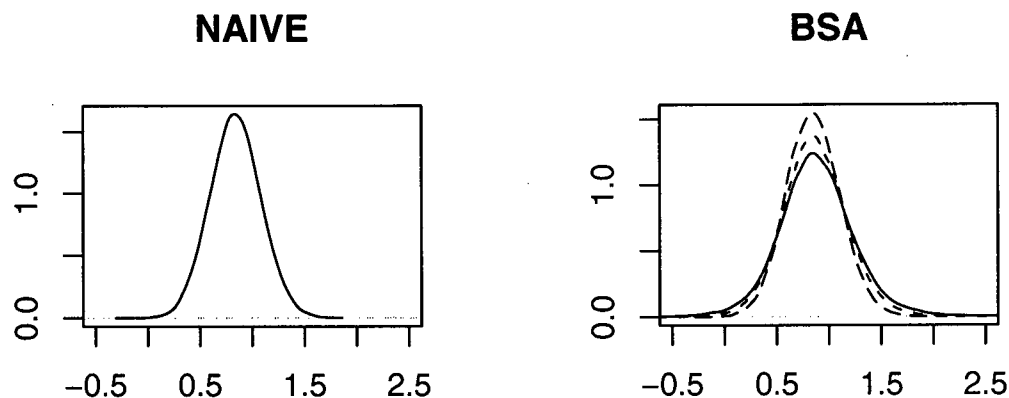Figure 4.1: Posterior distribution of the exposure effect for NAIVE and BSA.

Table 4.1 gives posterior means and 95% HPD credible intervals for the log odds ratio of the PI exposure effect and other covariate effects using NAIVE and BSA methods with 100 000 posterior simulations (after burn-in). A graphical comparison of NAIVE and BSA based on kernel density estimation of the posterior distribution of the exposure effect is provided in Figure 4.1. The results indicate that PI exposure is associated with increased risk of ele-

vated triglycerides even after adjusting for measured covariates and assessing sensitivity to a unmeasured confounder. Moreover, the strength of the association is robust to a variety of assumptions about the magnitude of possible unmeasured confounding. A choice of c equal to $(\log(9)/1.96)^2$ reflecting the prior belief of strong confounding due to U yields a 95% HPD credible interval of $(0.13, 1.58)$ which excludes zero. For most observational studies, BSA using such a large choice of c is unnecessary because modest rather than large levels of unmeasured confounding are expected.

Table 4.1 and Figure 4.1 reveal also some interesting properties of BSA. As expected, an increase in c results in a posterior distribution for the exposure effect which is more diffuse. This finding is consistent with our intuition that greater uncertainty about unmeasured confounding results in greater uncertainty about the exposure effect. But interestingly, the posterior mean for BSA analyses are shifted slightly relative to NAIVE and the magnitude of this shifting appears to depend on c. Using the R function effectiveSize() (coda library), we estimate the effective sample size for 100000 MCMC posterior simulations of parameter α_1 to be at least 10000 for the chosen values of c. The largest standard deviation among the posterior samples for α_1 is 0.369. Consequently, simulation standard errors of the posterior means are less than $0.0037 = 0.369/\sqrt{10000}$. Therefore, 95% confidence intervals for the posterior means do not overlap. This indicates that the BSA credible intervals are shifted relative to NAIVE intervals, although the shifting is small relative to the posterior standard deviations.

It is important to emphasize that this shifting of the posterior means is not necessarily indicative of a better estimation α_1. As stated in Section 3.2, posterior means in non-identifiable models may be asymptotically biased in a manner which cannot be predicted without knowledge of the true exposure effect. In a recent investigation, Gustafson (2005) studies the factors which influence the posterior distribution and asymptotic bias in non-identifiable models involving misclassification. Several important variables are identified including: the location, orientation and concentration of the prior distribution; the relation between prior distribution and true parameters values; sample size. We do not consider such a thorough analysis in this

investigation, but note simply that the posterior mean appears to be shifted away from zero and that the degree of shifting depends on the choice of c. It is important to emphasize that the true exposure effect remains unknown and may be either greater than or less than the NAIVE estimate.

The analysis would appear to present us with two unsatisfactory alternatives. On one hand, we have BSA which makes full use of available prior information but yields estimates which may be asymptotically biased in a manner which cannot be predicted. On the other, we have a NAIVE analysis which assumes artificially that unmeasured confounding is absent and may also yield biased estimates. The advantage of BSA becomes clear if we can guarantee that 95% credible intervals have approximately 95% coverage when averaged over many different observational studies. In the study of the effect of PI exposure on risk of elevated triglycerides, the magnitude of unmeasured confounding remains unknown. Bayesian sensitivity analysis provides credible intervals which may yield large or small coverage probability of the true exposure effect depending on the magnitude of unmeasured confounding. However, by choosing a suitable value of c from Table 4.1, the coverage probability of the true exposure effect may be approximately 95% *when averaged over many different studies.* Moreover, even if we misspecify c, the average coverage of BSA intervals may exceed the coverage of the usual NAIVE analysis.

To achieve correct 95% coverage, we must select credible intervals from Table 4.1 using the choice of c which most closely approximates the distribution from which study parameters can be expected to arise over repeated observational studies. Or equivalently, using the terminology of Section 3.2, we must ensure that the Investigator's prior approximates Nature's prior. A reasonable choice for c is $(\log(3)/1.96)^2$. The logic of this choice of c rests on the following assumptions: In a typical sequence of unrelated observational studies involving logistic regression models, the magnitude of bias due to unmeasured confounding is small. The log odds ratio of the association between confounder and disease, and between confounder and exposure is usually close to zero on the log-odds scale, with magnitude less than $\log(3)$ about 19 times out of 20. The exposure-specific logit prevalences of the unmeasured confounder are typically

between logit(25%) and logit(75%) about 19 times out of 20. Strengthening this argument would require an appropriate literature review.

Accordingly, our final interval estimate for the log odds ratio of the effect of PI on risk of elevated triglycerides is 0.85 (0.34,1.35) with corresponding odds ratio 2.3 (1.4, 3.9). The choice of c may only crudely approximate the distribution from which study parameters arise, but the simulations of Section 3.2 suggest that even a poor choice of c may yield intervals with better average coverage probabilities than NAIVE intervals.

It is also of interest to consider the posterior distributions of the bias parameters $\lambda, \gamma_0, \gamma_1$ themselves. Little is known about the possible values of bias parameters a priori and this is represented in the prior distributions of Section 2.2. But interestingly, the posterior distribution of bias parameters, given the data, may differ slightly from the prior distributions. Consider Figure 4.2 which superimposes the prior and posterior distributions for each bias parameter based on BSA with $c = (\log(6)/1.96)^2$. The differences between distributions is quite subtle. Table 4.2 highlights the differences in greater detail and includes means (95% CI) and medians (IQR) for the two distributions.

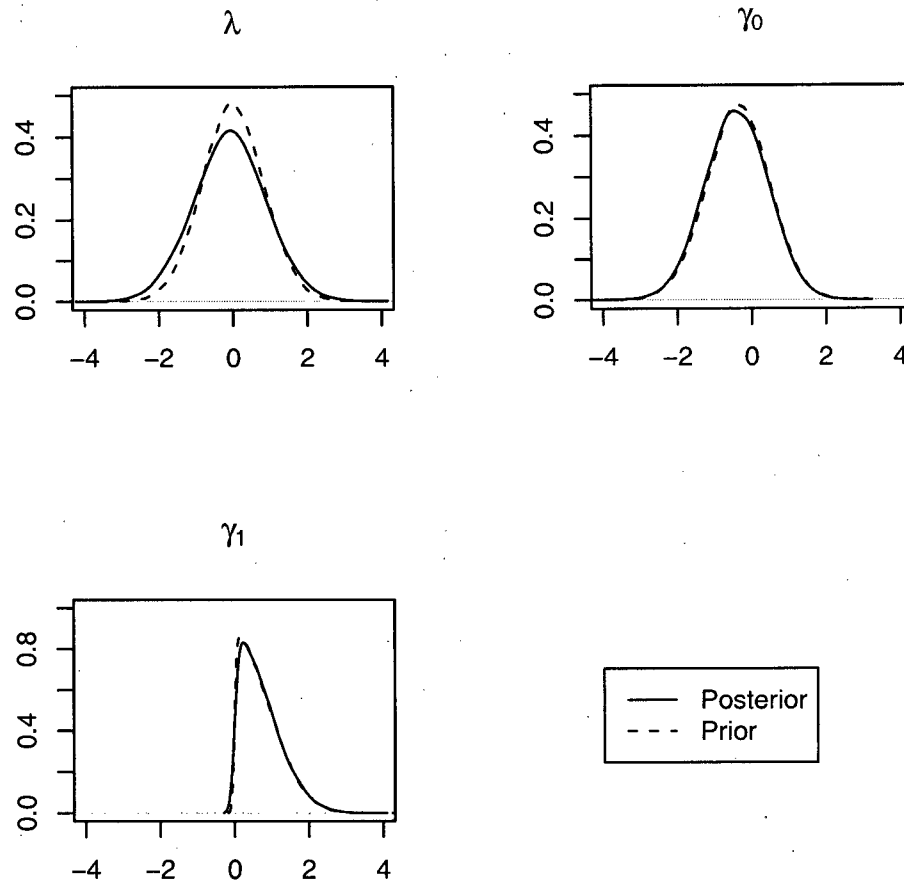Table 4.2: Prior and posterior distributions for bias parameters $\lambda, \gamma_0, \gamma_1$.

	Prior			Posterior			
	Median	(IQR)	Mean	Median	(IQR)	Mean	(95% CI)*
λ	0	(-0.56, 0.56)	0	-0.07	(-0.72, 0.57)	-0.08	(-0.12,-0.04)
γ_0	-0.35	(-0.93, 0.21)	-0.36	-0.40	(-0.97, 0.17)	-0.41	(-0.44,-0.37)
γ_1	0.61	(0.29, 1.00)	0.73	0.62	(0.30, 1.10)	0.73	(0.71,0.76)

IQR = Interquartile range, CI = Confidence interval.
* Confidence intervals are based on an effective sample size of 2000 from 100000 posterior simulations.

Because of poor mixing, these results should be interpreted with caution. The effective sample size for 100000 posterior simulations is approximately 2000 for the bias parameters $\lambda, \gamma_0, \gamma_1$ and therefore, the observed differences between prior and posterior may be an artifact

Figure 4.2: Prior and posterior distributions for bias parameters $\lambda, \gamma_0, \gamma_1$

of poor mixing. Regardless, the prior and posterior distributions do appear to differ, particularly for the bias parameter λ. Thus it would seem that while non-identifiability precludes consistent estimation of the bias parameters, there may be still may be something to be learned about them from the data. In the context of models for misclassification, Gustafson (2005) describes this phenomenon as "indirect learning" about non-identifiable parameters. In future, it would be informative to explore this behaviour in more detail. The use of MCMC thinning techniques or convergence diagnostics with multiple chains would permit a more valid comparison of prior

and posterior distributions.

An important implication of indirect learning is that BSA and MCSA may yield interval estimates for the exposure effect which differ slightly. Greenland (2005), describes a set of conditions in which MCSA intervals will approximate Bayesian credible intervals. One condition is that there can be no substantial indirect learning about bias parameters. While this condition appears to be satisfied in the current context, it would be informative to compare MCSA and BSA intervals in detail via simulation to explore the conditions in which the approximation hold correctly.

Chapter 5

Discussion and directions for further research

5.1 Discussion

Philips (2003) and Greenland (2005) argue that the reporting of results from observational studies should include a quantitative assessment of systematic error. The usual confidence intervals obtained using standard statistical packages quantify uncertainty due to random error only. But this is often dwarfed by uncertainty arising from systematic error due to unmeasured confounding, selection bias and misclassification (Greenland (2005)). In present context, we introduce a method for assessing sensitivity to unmeasured confounding in observational studies with a binary exposure X, binary response Y, binary unmeasured confounder U and covariate vector Z. Prior beliefs about unmeasured confounding are included in a Bayesian analysis as prior distributions over bias parameters which govern the confounding behaviour of U. The posterior distribution for the exposure effect is not only a tool for sensitivity analysis, but also a single summary which reflects uncertainty due to systematic error from unmeasured confounding in addition to the usual random error. We argue that credible intervals from BSA may yield more valid inference compared to standard methods because they better approximate

the total uncertainty in observational studies. This perspective is consistent with the recent initiatives of Greenland (2005) to quantify multiple sources of uncertainty due to systematic error using Bayesian methods.

BSA provides a number of advantages compared to standard methods for sensitivity analysis. As illustrated in the example of Chapter 4, the investigator can explore the sensitivity of effect estimates over a range of possible biases and present the results in a manner which requires less printed area compared to standard methods of sensitivity analysis which use tables. Credible intervals obtained from BSA may also be narrower than intervals obtained using standard methods. When reporting tables of effect estimates over a range of possible bias parameters, there is a tendency to focus attention on the most extreme effect estimates corresponding to the most extreme bias parameters. BSA incorporates the prior belief that small biases are often more likely than large biases and can therefore yield conclusions which are less pessimistic.

Like other models of unmeasured confounding, our model is non-identifiable and indexed by a set bias parameters which cannot be estimated consistently using the available data. Rather than achieve identifiability by plugging in specified values for bias parameters, we place informative prior distributions on bias parameters. This approach treats the unmeasured confounder U as a latent variable which is averaged over the prior distributions and later integrated out of the joint posterior distribution of model parameters. Our perspective is that while this model is non-identifiable, it is a better representation of observational data since the latent U and bias parameters λ, γ_0, and γ_1 have have substantive meaning. Models of this type of are increasingly commonplace in the biostatistical literature and do not preclude a proper Bayesian analysis. (See for example Gustafson (2005) or Gelman (2004)) If the prior distribution is an accurate predata summary of knowledge, then the posterior is equally credible even if the model is not identifiable.

Critics may tend to favor the NAIVE analysis for estimating exposure effects in observational studies. This method, described in Section 3.1, is standard logistic regression of Y on X

and Z while ignoring possible unmeasured confounding. Since the resultant models are identifiable, the usual large sample theory for posterior distributions can be applied, simplifying the analysis. Additionally, the NAIVE approach requires no subjective prior distributions. What must be emphasized is that the NAIVE method is inherently subjective because it requires the subjective assumption that unmeasured confounding is absent. (Philips (2003)) While some investigators may be reluctant to incorporate prior beliefs into the analysis of observational data, the fact remains that the use of standard identifiable models may tend to oversimplify the mechanisms from which observational data arise, yielding biased estimates and intervals which are falsely precise.

Non-identifiability complicates the implementation of Bayesian sensitivity analysis because the usual large sample theory governing the behaviour of posterior distributions is not applicable.(Gustafson (2005)) The posterior distribution does not converge to a single point at the true model parameter, but instead converges to a non-degenerate distribution which is strongly governed by the choice of prior. Thus, without identifiability, we have the paradoxical situation where the prior tends to dominate the posterior distribution as the sample size goes to infinity. The posterior mean is asymptotically normal, but may be biased in a manner which cannot be predicted without knowledge of the true model parameters. Consequently, as illustrated in Section 3.1, the coverage probability of 95% credible intervals does not tend to 95% as the sample size increases. These facts are discouraging for some and may explain the limited use of non-identifiable models in the biostatistical literature.

We demonstrate that, under certain conditions, the coverage probability of 95% BSA credible intervals may be approximately 95% *when averaged over many different studies.* Suppose that the true model parameters from repeated observational studies arise from arise from a distribution with density function $g(\theta)$ (Nature's prior) and the corresponding data are analysed with a prior distribution $f(\theta)$ (Investigator's prior). In Section 3.2, we illustrate that if Nature's prior and the Investigator's prior are approximately equal ($g(\theta) \approx f(\theta)$), then the expected coverage probability of $1 - \alpha$ BSA credible intervals may be approximately $1 - \alpha$.

How can this be interpreted? Suppose that an investigator conducts a sequence of observational studies and obtains 95% credible interval for the exposure effect in each study using Bayesian sensitivity analysis. In normal identifiable models, if the sample were large, then each interval would cover the true exposure effect with very high probability, regardless of the true value of the exposure effect. But since the model is not identifiable, some intervals may cover the true exposure effect with low probability, and others with higher probability. If the investigator chooses a prior distribution for data analysis which is approximately the same as the distribution from which study parameters arise, then the sequence of credible intervals will cover the true exposure effect with probability 95% on average.

5.2 Limitations

While we illustrate the conditions for proper coverage when averaging over many different studies, we provide no guidelines for determining the properties of Nature's prior, the distribution from which study parameters arise. Indeed, these properties cannot be determined from the available data, and the investigator must once again look outside the study to obtain the needed information to ensure that that credible intervals have correct 95% coverage over repeated studies.

The simulation studies of Section 3.2 suggest that small differences between Nature's prior and the Investigator's prior may maintain approximately $1 - \alpha$ coverage over repeated studies, but our comparison of the two distributions was not exhaustive. A more thorough investigation might reveal a specific individual role of the parameters c_1, c_2, c_3 which was lost in the comparison. The conclusions of Section 3.2 are intended to be primarily descriptive in nature and motivate further research in the area. It remains possible that a casual choice for the Investigator's prior may yield coverages which are far from 95% on average.

Technical challenges are a major hurdle in this investigation. The simulations of Sections 3.1 and 3.2 are based on the analyses of 400 simulated datasets with only 2000 posterior

simulations. As illustrated in Figures 3, 4 and 5, many iterations are required to ensure that MCMC chains moves thoroughly through the support of the posterior distribution. By using only 2000 simulations, intervals obtained from posterior samples may be poorly representative of the posterior distribution and the empirical coverages of intervals may be prone to error. While the use of small posterior samples may be adequate for exploratory purposes, longer posterior simulations and improved algorithms are needed to better estimate coverage probabilities via simulation. However, it should also be emphasized that while the mixing of bias parameters is often very slow, this problem may be less serious than it seems. Different combinations of bias parameters may yield approximately equivalent level of bias in the exposure effect estimate. For example, if x is strongly associated with a weak risk factor for Y, then this may be approximately equivalent to the case where X is weakly associated with a strong risk factor for Y.

5.3 Further research

While the distributions presented in Section 2.2 are in a convenient parametric form for the Investigator's prior, it would be informative to consider a more general distribution for Nature's prior. In particular, to more rigorously identify the conditions for approximately correct $1 - \alpha$ coverage of credible intervals. Additionally, since the prior distributions on bias parameters tend to dominate the posterior asymptotically, it would be informative to explore the performance and coverages of Bayesian sensitivity analysis using distributions other than normal densities presented in equation (2.2) and (2.4).

A natural extension of BSA is to adapt the method to include other sources of systematic error in observational studies including selection bias and misclassification. Unmeasured confounding is only one possible source of bias and consequently, BSA will only poorly quantify the total uncertainty in observational studies. Greenland (2005) formalizes this idea in the context of 'multiple bias modeling' which involves the use of a non-identifiable model indexed by a collection of bias parameters which model all sources of bias simultaneously . His discussion is

primarily framed in the context of the MCSA, and consequently our approach could be adapted along similar lines.

An interesting avenue of research is to adapt BSA to instances in which unmeasured confounding is thought to be particularly likely. The prior specification in Sections 2.2 is very general and assumes that little is known about the possible confounding behaviour of U. But in fact, there are many instances in epidemiological research where more information is available about U, despite the fact that U itself is unknown and unmeasured. One example builds on the following citation by Joffe (2000):

"The argument is sometimes advanced that if adjustment for measured covariates fails to change the measure of effect, then there must be little unmeasured confounding".

The logic of this statement rests on the assumption that the measured covariates Z are somehow informative about the confounding behaviour of U. How is this reasonable? In conducting observational studies, health researchers seek to collect information on as many confounding variables as conveniently possible. Presumably they will be successful in identifying at least a few of them. Therefore, if no measured covariates are determined to be confounders, then this may indicated that there is no unmeasured confounding (Joffe (2000)). This idea can be formalized into a Bayesian sensitivity analysis by modifying the model of Section 2.1 in such a way that the parameters β and λ are treated as exchangeable, meaning that they are indistinguishable a priori with joint prior density $f(\lambda, \beta)$ which is invariant to permutations. The joint distribution

$$P(U, X, Z) = P(U|X, Z)P(X, Z)$$

can also be suitably modified to something of the form

$$\text{logit}[Pr(U = 1|X, Z)] = \gamma_0 + \gamma_1 X,$$

$$\text{logit}[Pr(X = 1|Z)] = \eta_0 + \eta_1 Z.$$

The parameters η_1 and γ_1 can then be treated as exchangeable to permit sharing of information

between the observed association of X and Z with the unmeasured association between X and U.

A similar example where we have additional information possible unmeasured confounding, is in assessments of confounding by indication in the study of the intended effects of therapeutic interventions. The "intended effect" of a therapeutic intervention is defined as the efficacy of the intervention in curing or preventing the disease for which it was prescribed. "Unintended effects" refer to side-effects and the safety of the intervention. It is well known that observational studies investigating the intended effects of therapeutic interventions are particularly prone to unmeasured confounding because the risk factors of disease are often closely related to indications for prescription of treatment. (McMahon (2003) and Joffe (2000)). This confounding by indication for treatment can cause new treatments to appear non-efficacious or harmful and it is particularly difficult to adjust for using standard methods.

Our approach to Bayesian sensitivity analysis could be tailored according to instances in which confounding by indication is thought to be likely. Existing methods for adjusting for confounding by indication are diverse, but suffer from of the limitations of the NAIVE method of analysis of Section 3.1 because they typically involve adjusting for measured covariates (Joffe (200)). Adapting Bayesian sensitivity analysis might involve adjusting the prior distributions in Section 2.2, for example, by specifying c_3 to account for beliefs about the association between physician prescribing practices and latent risk factors for the intended outcome. Similarly, small values of c_3 could be used in the evaluation of drug safety where confounding by indication is thought to be less likely.

Bibliography

[1] Bross, I. D. J. Spurious effects from an extraneous variable. *Journal of Chronic Diseases.* *19* (1966), 637–47.

[2] Chib, S., and Greenberg, E. Understanding the metropolis-hastings algorithm. *The American Statistician.* *49* (1995), 327–35.

[3] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute.* *22* (1959), 173–203.

[4] Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP). *Journal of the American Medical Association.* *2001* (2001), 2486–2497.

[5] Gelfand, A. E., and Sahu, S. K. Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association.* *94* (1999), 247–53.

[6] Gelman, A. Parametrization and Bayesian modeling. *Journal of the American Statistical Association.* *99* (2004), 537–45.

[7] Gordis, L. *Epidemiology.* Saunders, Philadelphia, 2004.

[8] Greenland, S. Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Analysis.* *21* (2001), 579–83.

[9] Greenland, S. 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* (2003), 47–54.

[10] Greenland, S. Multiple bias modeling for analysis of epidemiologic data (with discussion). *Journal of the Royal Statistical Society, Series A. (in press)* (2005).

[11] Gustafson, P. On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science. (in press)* (2005).

[12] Gustafson, P. The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. *Statistics in Medicine. (in press)* (2005).

[13] Hernan, M. A., and Robins, J. M. Correspondence. *Biometrics. 55* (1999), 1316–7.

[14] Joffe, M. M. Confounding by indication; the case of calcium channel blockers. *Pharmacoepidemiology and Drug Safety. 9* (2000), 37–41.

[15] Lash, T. L., and Fink, A. K. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology. 14* (2003), 451–8.

[16] Levy, A. R., McCandless, L. M., Harrigan, R. S., Hogg, R. S., Bondy, G., Drummond, A., Iloeje, U., Mukherjee, J., and Montaner, J. S. Lipid profiles amond persons infected with HIV/AIDS starting highly active antiretroviral therapy. *Antiviral therapy. (under review)*.

[17] Lin, D. Y., Psaty, B. M., and Kronmal, R. A. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics. 54* (1998), 948–63.

[18] Maclean, D. R., Petrasovits, A., Connelly, P., Joffres, M., O'Connor, B., and Little, J. A. Plasma lipids and lipoprotein reference values, and prevalence of dyslipoproteinemia in canadian adults. *Canadian Journal of Cardiology. 15* (1999), 434–44.

[19] McMahon, A. D. Approaches to combat confounding by indication in observational studies of indended effects. *Pharmacoepidemiology and Drug Safety. 12* (2003), 551–8.

[20] Oakley, J. E., and O'Hagan, A. Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society, Series B. 66* (2004), 751–69.

[21] Philips, C. Quantifying and reporting uncertainty from systematic errors. *Epidemiology. 14* (2003), 459–66.

[22] Robins, J. M., Rotnitzky, A., and Scharfstein, D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In. *Statistical Models in Epidemiology*, Ed. E. Halloran and D. Berry, New York, Springer. (1999), 1–94.

[23] Rosenbaum, P., and Rubin, D. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. 45* (1983), 212–8.

[24] Rosenbaum, P. R. *Observational Studies, 2nd ed.* Springer, New York, 1999.

[25] Rothman, K. J., and Greenland, S. *Modern Epidemiology, 2nd ed.* Lippincott, Philadelphia, 1998.

[26] Rubin, D. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics. 12* (1984), 1151–72.

[27] Schaffer, M. L., and Chinchilli, V. M. Bayesian inference for randomized clinical trials with treatment failures. *Statistics in Medicine. 23* (2004), 1215–28.

[28] Scharfstein, D., Daniels, M., and Robins, J. Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing data. *Biostatistics. 4* (2003), 495–512.

[29] Schlesselman, J. J. Assessing effects of confounding variables. *American Journal of Epidemiology. 108* (1978), 3–8.

[30] Steenland, K., and Greenland, S. 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* (2004), 384–92.

[31] Szklo, M., and Nieto, F. J. *Epidemiology : Beyond the Basics.* Aspen, Gaithersburg, 1999.

[32] Tanner, M. A., and Wong, W. H. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association. 82* (1987), 528–40.

[33] Yanagawa, T. Case-control studies: Assessing the effect of a confounding factor. *Biometrika. 71* (1984), 191–4.