A BAYESIAN ADJUSTMENT FOR COVARIATE MISCLASSIFICATION WITH CORRELATED BINARY OUTCOME DATA

by

Dianxu Ren

M.S. Statistics, University of Georgia, 2001

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Dianxu Ren

It was defended on

July 14, 2005

and approved by

Roslyn A. Stone, PhD, Associate Professor, Department of Biostatistics, Graduate School
of Public Health, University of Pittsburgh

Howard E. Rockette, PhD, Professor, Department of Biostatistics, Graduate School of
Public Health, University of Pittsburgh

Sati Mazumdar, PhD, Professor, Department of Biostatistics, Graduate School of Public
Health, University of Pittsburgh

Gong Tang, PhD, Assistant Professor, Department of Biostatistics, Graduate School of
Public Health, University of Pittsburgh

Michael J. Fine, MD, Professor, Department of Medicine, School of Medicine, University of
Pittsburgh

Dissertation Director: Roslyn A. Stone, PhD, Associate Professor, Department of
Biostatistics, Graduate School of Public Health, University of Pittsburgh

A BAYESIAN ADJUSTMENT FOR COVARIATE MISCLASSIFICATION WITH CORRELATED BINARY OUTCOME DATA

Dianxu Ren, PhD

University of Pittsburgh, 2005

Estimated associations between an outcome variable and misclassified covariates tend to be biased when the methods of estimation that ignore the classification error are applied. Available methods to account for misclassification often require the use of a validation sample (i.e, a gold standard). But in practice, such gold standard may be unavailable or impractical. We propose a Bayesian approach to adjust for misclassification in a binary covariate in fixed and random effect logistic models when a gold standard is not available. This Markov Chain Monte Carlo (MCMC) approach uses two imperfect measures of a dichotomous exposure under the assumptions of conditional independence and non-differential misclassification. This approach is validated with several simulation studies. We illustrate the proposed approach to adjust for misclassification with respect to oxygenation status in a multi-center trial of patients with pneumonia, where 16 per cent of patients are classified discordantly by two assessments. The estimated log odds of inpatient care and the corresponding standard deviation are much larger in our proposed method compared to the models ignoring misclassification. We also applied the proposed Bayesian method to the EDCAP trial to assess the intervention effect allowing for misclassification with respect to risk status. Ignoring misclassification produces downwardly biased estimates and underestimates uncertainty. The public health significance of this study is that the proposed approach can correct for the bias of an estimated association when a covariate is misclassified and no gold standard is available, which is common problem in epidemiology studies.

PREFACE

This dissertation is organized as follows. In Chapter 1, I give an introduction to the misclassfication problem and a motivating example. In Chapter 2, I review some statistical approaches that have been used to correct for measurement error in continuous variables and misclassification in categorical variables. The proposed Bayesian methods and results for the motivating example are described with greater detail in Chapter 3 and Chapter 4. In Chapter 5, the proposed Bayesian approach is applied to EDCAP trial. Finally, Chapter 6 is a conclusion chapter.

I wish to thank my advisor, Dr. Roslyn A. Stone for her earnest help and guidance during the course of my dissertation work. It is she who introduced the interesting misclassification problem from real data to me, which is the main topic of my dissertation. Moreover, Dr. Stone's continuous encouragement has greatly helped me build up my confidence in doing research and her insightful suggestions clarified lots of problems that had confused me. I truly am grateful to Dr. Stone's patient revision of my proposal and this dissertation.

I would like to thank Dr. Howard E. Rockette, Dr. Sati Mazumdar, Dr. Gong Tang and Dr. Michael J. Fine for serving as my committee members. I want to thank Dr. Gong Tang for valuable discussions for the likelihood identifiability issue. I want to thank Dr. Howard E. Rockette, Dr. Sati Mazumdar and Dr. Gong Tang, whose lectures on estimation theory, linear models, mixed models and missing data established indispensable background for my dissertation. I want to thank the Department, which provided financial support through my Ph.D study.

Special thanks to Dr. Michael J. Fine for providing me with excellent data to motivate the main idea for my dissertation.

Finally, I dedicate this dissertation to my wife and my parents for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Measurement error problems are common in many studies, due to the fact the measurement of many biological and enviromental risk factors is subject to error. For example, we might use inexact measures because of high cost of exact evaluation, or poorly recalled personal exposure history, or substitutes for latent variables that can not be measured at all. It has long been recognized that measurement error can cause considerably biased estimates of relationship between an outcome and covariates (Flegal, Brownie and Hass, 1986). The terms "measurement error", "mismeasurement", "misclassification" and "error in variables" refer to any discrepancy between a true variable and its measured value. "Misclassification" is often used with reference to categorical variables measured with error where "measurement error" refers to continuous variables.

1.1 STATEMENT OF THE PROBLEM

Categorical covariates are often subject to misclassification, and this misclassification can distort the relationship to the outcome of interest (Newell, 1962; Gullen, Berman and Johnson, 1968; Walker and Blettner, 1985; Kristensen, 1992). Even small amounts of covariate misclassification can substantially bias the association (Copeland, Checkoway, McMichael and Holbrook, 1977). Misclassification can be either differential or non-differential, depending on whether or not the classification error varies with the level of the response or other covariates (Chen, 1989). Non-differential misclassification often arises in prospective studies, where the true and observed covariates are assessed before response is measured. Differential misclassification is more likely to occur in case-control studies, where the response is obtained

first and subsequently ascertained covariates may be subject to recall bias (Carroll, Rupert and Stefanski, 1995). When misclassification is ignored, non-differential misclassification generally biases the estimated association toward the null, while differential misclassification can bias the estimated association either toward or away from the null (Greenland, 1980).

Adjustment for potential bias due to misclassification requires information on the misclassification structure. In general, the misclassification is unknown and can be estimated from validation samples (i.e, a gold standard that can classify the covariate with complete accuracy). The sensitivity and specificity of imperfect measures can be estimated when a gold standard is available, and corrected estimates derived (Barron, 1977; Greenland, 1982; Greenland and Kleinbaum, 1983). But in practice the gold standard may be unavailable, due to high cost of exact evaluation or the fact that it is also measured with error (Sheps and Schechter, 1984; Wacholder, Armstrong and Hartge, 1993). When no gold standard but two imperfect measures of a covariate are available, one can correct for the bias due to misclassification using a maximum likelihood approach under 2×2 table and fixed effect logistic model settings (Hui and Walter, 1980; Flanders, Drews and Kosinski, 1993; Kosinski and Flanders, 1999). The likelihood function is integrated over the true but unobserved covariate. Often this integration can not be done explicitly in closed form, and the Expectation-Maxmization (EM) algorithm (Dempster, Laird and Rubin, 1977) is applied to iteratively maximize the likelihood function. In some measurement error problems, the expectation to be completed in the E-step may not have a closed form expression, so that some approximations must be used. Also, extra effort is required to estimate standard errors when maximum likelihood estimates are obtained via the EM algorithm (Louis, 1982).

This need for numerical integration is more problematic in the generalized linear mixed model (GLMM) setting. Due to the unobserved true covariates and random effects, this likelihood function contains many integrals that have to be evaluated. If no closed form expression exists, one has to opt for numerical integration, simulation or approximation (Rabe-Hesketh, Skrondal and Pickles, 2002; Booth and Hobert, 1999; Breslow and Clayton, 1993), which is either computationally intensive or inaccurate. Ko and Davidian (2000) used a Laplace approximation to correct for measurement error in a nonlinear mixed effect model.

The Bayesian alternatives using Markov Chain Monte Carto (MCMC) for measurement

error problems are attractive, especially in the GLMM framework. The integrals in the like-lihood function do not have to be evaluated, and the unobserved variables can be sampled along with the model parameters from their full posterior distribution. Bayesian inference does not require approximations, and uncertainty assessments are obtained directly. When a validation sample is available, a two-stage Bayesian model has been proposed to account for misclassification (Viana, 1994; Spiegelhalter, Thomas and Best, 1999). In my work, I will present a Bayesian approach to correct for misclassification when two imperfect measures, but no gold standard, are available in the fixed effect logistic model setting, then extend to random effect logistic model setting. This approach allows for adjustment other discrete or continuous covariates that are assumed to be measured without error.

1.2 THE EDCAP STUDY

The Emergency Department Community Acquired Pneumonia (EDCAP) Trial (Yealy, Auble, Stone et al., 2004) motivated the proposed method. Community-acquired Pneumonia (CAP) causes over four million episodes of illness each year and has high morbidity, mortality and total cost of care. The initial site of treatment (home or hospital) is often identified as the single most important clinical decision for patients with CAP. The identification of low risk patients with CAP may reduce unnecessary hospitalizations and could significantly reduce health care related expenses. The 32-site EDCAP study was designed to evaluate three guideline implementation strategies of increasing intensity (low intensity, moderate intensity and high intensity) to guide the admission decision when combined with a validated mea-sure of pneumonia severity, the Pneumonia Severity Index (PSI) (Fine, Auble, Yealy et al., 1997). Outpatient care was recommended for low risk patients, defined as non-hypoxemic patients in PSI risk classes I–III. Inpatient care was recommended for higher risk patients (hypoxemic and/or PSI risk classes IV and V). Hypoxemia (evidence of clinically significant arterial oxygen desaturation) was defined as pulse oximetry less than 90% or PO_2 less than 60mm Hg. In my dissertation, the EDCAP data are used in two ways: (1) the data from

control arm (8 sites) serve as a motivating example for proposed Bayesian method, and (2) the proposed Bayesian methods are applied to full EDCAP study data (32 sites).

In our motivating example, we estimate the odds of inpatient status as a function of true oxygenation status and PSI risk class in the 8 control sites, accounting for site as (i) a fixed effect and (ii) a random effect. The true initial oxygenation status (x) is unobserved and was ascertained by two imperfect assessments: prospectively ascertained oxygenation (x_1) in the Emergency Department and retrospectively ascertained oxygenation (x_2) documented in the medical chart. We assume the two assessments are independent conditional on the true unobserved status. The PSI risk class was assumed to be measured without error, and class IV and V are combined in our analysis because of low frequencies. Table 1.1 displays the data for 740 patients enrolled from 8 control sites in western Pennsylvania on inpatient status, PSI risk class and the two imperfect assessments of oxygenation. Although the prospective assessment was used in the primary analysis of the EDCAP trial, subsequent analysis has raised some question about the accuracy of the prospective assessment although the agreement with the retrospective assessment is substantial (Aujesky, Stone, Obrosky et al, 2005). Due to time lags in data recording, oxygenation status may be obtained retrospectively from ED medical chart review using information that was not in fact available in real time. Among these 740 patients, 84% were classified concordantly (70% non-hypoxemic and 14% hypoxemic) on the two assessments and 16% were classified disconcordantly, with 11% classified as non-hypoxemic by the prospective but not the retrospective assessment, and 5% classified as non-hypoxemic by the retrospective but not the prospective assessment. The proposed methods were developed to assess the association between true oxygenation status and inpatient status adjusting for Pneumonia Severity Index (PSI) and accounting for site as random effect in this motivating example.

Then we apply the proposed Bayesian approach to the entire EDCAP trial, to evaluate the intervention effect in the presence of risk misclassification in the random effect logistic model. To be consistent with the published paper for EDCAP study, the outcome variable here is outpatient status, instead of inpatient status in our motivating example. We will give more details for the application to full EDCAP data in chapter 5.

4

Table 1.1: Frequencies of observed patterns of covariates and inpatient status for the EDCAP control data

Covariate pattern		Oxygenation status assessment		
Inpatient status	PSI	Prospective	Retrospective	Frequency
0	1	0	0	87
0	1	0	1	1
0	1	1	0	1
0	1	1	1	0
0	2	0	0	54
0	2	0	1	3
0	2	1	0	1
0	2	1	1	1
0	3	0	0	21
0	3	0	1	1
0	3	1	0	0
0	3	1	1	0
0	4	0	0	3
0	4	0	1	1
0	4	1	0	0
0	4	1	1	0
1	1	0	0	43
1	1	0	1	6
1	1	1	0	0
1	1	1	1	8
1	2	0	0	103
1	2	0	1	17
1	2	1	0	4
1	2	1	1	9
1	3	0	0	89
1	3	0	1	20
1	3	1	0	8
1	3	1	1	20
1	4	0	0	116
1	4	0	1	31
1	4	1	0	23
1	4	1	1	69
			Total	740

Inpatient status (1=inpatient, 0=outpatient)
PSI (Pneumonia Severity Index from 1 to 4 (low to high))
Prospective oxygenation (1=abnormal, 0=normal)
Retrospective oxygenation (1=abnormal, 0=normal)

2.0 REVIEW OF KEY LITERATURE

2.1 TERMINOLOGY AND NOTATION

There are two classes of measurement error: random or systematic error. For random error, the average value of many repeated measures will approach the true value. It can be quantified with a single replicate measure for a subject. With systematic error, which is not distributed randomly around the true value, the mean of repeated determinations differs from the true value. It requires a second "gold standard" measure of exposure, i.e a validation study.

The following notation is used for illustrating different statistical methods to adjust for both types of measurement error (Carroll, Rupert and Stefanski, 1995):

Y: response, could be continuous or binary (1=yes 0=no)

X: vector of true predictor variables of primary interest, usually unobserved

W: vector of observed predictor variables, erroneous measurement of X

Z: vector of other covariates (assumed to be measured without error)

β: parameter vector of the outcome model, usually of interest

λ: parameter of the measurement model

The outcome model: $P(Y|X, Z) = f(X, Z, \beta)$ or $P(Y = 1|X, Z) = f(X, Z, \beta)$.

The classical measurement error model: $W \sim f(X, \lambda)$

The Berkson measurement error model: $X \sim f(W, \lambda)$, also called the Regression Calibration model

The marginal distribution of the true predictor X in the population $P(X)$, i.e, the exposure model

It is important to make a distinction between differential and nondifferential measure-

6

ment error. Nondifferential measurement error occurs when W has no information about Y other than what is available in X, which is equivalent to the assumption of conditional independence of W and Y given X. i.e $f(Y|X, W) = f(Y|X)$. In this case, W is said to be a surrogate, i.e. W does not contain any information about Y that is not provided by X. The measurement error is differential otherwise.

Many problems can plausibly be classified as having nondifferential measurement error. With nondifferential measurement error, one can typically estimate parameters in models for response given covariates even when true covariates are not observable. But with differential measurement error, one must observe the true covariate on some study subjects (Carroll, Ruppert and Stefanski, 1995).

2.2   SOURCE OF DATA

To perform a measurement error analysis, one needs information about the error structure. The data sources can be broken up into two main categories: an internal subset of primary data and an external set of independent studies. Within each of those broad categories, there are three types of data, all of which might be available only in a random subsample of the study population:

1. Validation data: in which X is observable directly.

2. Replication data: in which replicates of W are available.

3. Instrumental variable: in which another variable T is observable in addition to W. The instrumental variable is a second surrogate measurement of X obtained by an independent method.

An internal validation dataset is ideal, because it can be used with all known analytical techniques and typically leads to much greater precision of estimation and inference. With external validation data, one must assume that the same error structure appears in the primary data. i.e that the models and parameters are transportable.

2.3 DESCRIPTION OF STATISTICAL METHODS

The statistical approaches for adjusting for measurement error and misclassification can be classified in the following groups: Regression calibration method, SIMEX algorithm, Maximum-Likelihood methods, Semiparametric approaches, Bayesian methods, and Bootstrap methods.

2.3.1 Regression calibration method

This regression calibration algorithm was suggested as a general approach for correcting measurement error in a continuous variable in linear models by Carroll and Stefanski (1990) and Gleser (1990). Armstrong (1985) suggests regression calibration for generalized linear models. Rosner et al. (1989) developed the idea for logistic regression into a workable and popular methodology based on the model

$$ln\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)} = \alpha^* + \beta^* X$$

$$X = \lambda_0 + \lambda_1 W + \varepsilon$$

where

$$\varepsilon \sim N(0, \sigma^2)$$

Rosner first estimate the parameters in the second equation from validation study by Ordinary Least Squares (OLS). Second, calcuate an expected value of X

$$E(X|W) = \lambda_0 + \lambda_1 W$$

for each person in the main study population based on the result in the first step. Third, substitute the expected value $E(X \mid W)$ for X for each person in the main study population and estimate β^*. Regression calibration produces a simple imputed value for each X. Fourth, adjust the resulting standard error to account for the additional variability caused by the estimation of the measurement error using either the bootstrap or asymptotic methods.

Another linear approximation approach yields the same point estimate as the regression calibration method (single imputation method). One can use first order Taylor series expansion and the Delta method to obtain a variance estimator that incorporates the measurement error into the variance for the estimated true exposure-disease relationship.

The advantage of single imputation approach is that it can be extended readily to the case in which the true response is derived from multiple surrogate measures, some of which may be related through a multivariable model. But the standard error has to be adjusted to account for the estimation of the measurement model. The Taylor series approximation for the standard error for the linear approximation approach will always result in a larger variance. Kuha (1994) extended the linear approximation by improving the approxmation of the logistic model using a second-order Taylor series instead of the first-order series.

2.3.2   The SIMEX method

Simulation extrapolation (SIMEX) is a simulation-based method of estimating and reducing bias due to measurement error in a continuous variable. SIMEX estimates are obtained by adding additional measurement error to the data in a resampling-like stage, establishing a trend of measurement error-induced bias versus the variance of the added measurement error, and extrapolating this trend back to the case of no measurement error. This technique was proposed by Cook and Stefanski (1995) and further developed by Carroll et al. (1996) and Stefanski and Cook (1995). The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation.

We assume that

$$Y = \beta_1 + \beta_x X + \varepsilon$$

with additive measurement error

$$W = X + U$$

where

$$U \sim N(0, \sigma_u^2)$$

The OLS estimate of β_x, denoted $\widehat{\beta_{x,naive}}$, consistently estimates not β_x but rather $\frac{\beta_x \sigma_x^2}{\sigma_x^2 + \sigma_u^2}$. Suppose that in addition to the original data used to calculate $\widehat{\beta_{x,naive}}$, there are M-1 additional data sets available, each with successively larger measurement error variances, say $(1 + \lambda_m)\sigma_u^2$ where $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_M$. The least squares estimate of the slope from the mth data set, $\widehat{\beta_{x,m}}$, consistently estimates $\frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1+\lambda_m)\sigma_u^2}$. We can think of this problem as a nonlinear regression model, with dependent variable $\widehat{\beta_{x,m}}$ and independent variable λ_m, having a mean function of the form

$$\mathcal{G}(\lambda) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1 + \lambda)\sigma_u^2}$$

where $\lambda \geq 0$.

The parameter of interest, β_x, is obtained from $\mathcal{G}(\lambda)$ by extropolation to $\lambda = -1$. Inference for SIMEX estimators can be based on either the bootstrap or the theory of M-estimators, in particular by means of the sandwich estimator.

The SIMEX algorithm works only for measurement error in a continuous variable and is sensitive to the choice of functional form of the measurement error variance. Also it is always dangerous to extropolate the function beyond the range in which values are observed.

2.3.3   Maximum-likelihood (ML) method

A more formal approach for conducting likelihood based inference is to consider the overall likelihood in which the true unobserved variable X is integrated out. Likelihood methods require stronger distributional assumptions, but they can be applied to more general problems, including those with discrete covariates subject to misclassification. See, for instance, Satten and Kupper (1993), Breslow and Holubkov (1997) and Spiegelman et al. (2000).

Suppose X is not observable for all subjects, but there are sufficient data to characterize the distribution of W given (X, Z). The likelihood function of Y given (Z, X) will be called $f_{Y|Z,X}(y|z, x, \beta)$ and interest lies in estimating β. To perform a likelihood analysis, one must specify a model for every component of the data. There are three components required for

the likelihood function.

1. A model relating the response to the "true" covariate.

For example, if Y is normally distributed with mean $\beta_0 + \beta_x x + \beta_z z$ and variance σ^2 then $\beta = (\beta_0, \beta_x, \beta_z, \sigma^2)$ and

$$f_{Y|Z,X}(y|z, x, \beta) = \sigma^{-1}\phi(\frac{y - \beta_0 - \beta_x x - \beta_z z}{\sigma})$$

where $\phi(v) = 2\pi^{-\frac{1}{2}}exp(-0.5v^2)$ is the standard normal density function.

2. An error model, here called $f_{W|Z,X}(w|z, x, \tilde{\lambda})$.

In many applications, the error model does not depend on Z. For example, in the classsical additive measurement error W=X+U, $U \sim N(0, \sigma_u^2)$, and σ_u^2 is the only component of $\tilde{\lambda}_1$. The error model density is $\sigma_u^2\phi(\frac{w-x}{\sigma_u})$ where $\phi(.)$is the standard normal density function.

3. A model for the distribution of the latent variable, here called $f_{X|Z}(x|z, \tilde{\alpha})$.

The parameter $\tilde{\alpha}$ formulate the relationship between latent variable X and other perfectly measured covariates Z.

As discussed by Carroll, Rupert and Stefanski (1995), the likelihood of the observed data under nondifferential measurement error is

$$f_{Y,W|Z}(y, w|z, \beta, \tilde{\lambda}_1, \tilde{\lambda}_2) = \prod_{i=1}^{n} \int f_{Y|Z,X}(y|z, x, \beta)f_{W|Z,X}(w|z, x, \tilde{\lambda}_1)f_{X|Z}(x|z, \tilde{\lambda}_2)dx$$

If we use the Berkson error model rather than the classical additive error model, the likelihood function becomes

$$f_{Y|Z,W}(y|z, w, \beta, \tilde{\lambda}) = \prod_{i=1}^{n} \int f_{Y|Z,X}(y|z, x, \beta)f_{X|Z,W}(x|z, w, \tilde{\lambda})dx$$

11

The likelihood functions are very different in the differential and nondifferential cases. In general, and dropping parameters, the likelihood of the observed data is

$$
\begin{aligned}
f_{Y,W|Z}(y, w|z) &= \int f_{Y,W,X|Z}(y, w, x|z)dx \\
&= \int f_{Y|Z,X}(y|z, x)f_{W|Y,Z,X}(w|y, z, x)f_{X|Z}(x|z)dx
\end{aligned}
$$

(2.1)

Under nondifferential measurement error, W and Y are independent given X, so that

$$
f_{W|Y,Z,X}(w|y, z, x) = f_{W|Z,X}(w|z, x)
$$

Under differential measurement error, we must ascertain the distribution of W given the other covariates and the response Y. This is essentially impossible to do in practice unless one has a subset of validation data in which all of (Y, Z, X, W) are observed.

This general approach can be applied to either measurement error in a continuous variable or misclassification in a categorical variable. The specification of the measurement error model and the model for the distribution of latent variable will be different.

Typically one maximizes the logarithm of the overall likelihood in the unknown parameters. There are two ways one can maximize the likelihood function. The most direct is to compute the likelihood function itself, and then use numerical optimization techniques to maximize the likelihood. Likelihoods in which X has some continuous components can be computed using a number of different approaches, such as Gaussian quadrature approximation and Monte Carlo techniques. The second general approach is to view the problem as a missing data problem, and then use missing data techniques such as data augmentation (Little et al. 2002 and Tanner, 1996).

2.3.4   Semiparametric approaches

The ML methods are based on structural models. They make parametric assumptions about the distribution of the measurement error, e.g about $f(W|X)$, and therefore run the risk of model misspecification. The semiparametric methods consider functional models. They avoid parametric formulation of the distribution of X, and avoid possible non-robustness of the estimated parameters due to model misspecification.

One approach is nonparametric estimation of the marginal or conditional distribution of X or W. Pepe and Fleming (1991) proposed an "estimated likelihood method" for studies with measurement error in categorical covariates. The unknown probability of X conditional on W is estimated empirically from validation data. From this, the estimated conditional probability of W given the gold standard X is derived for the main study units, and the parameter estimates are calculated numerically by use of standard ML estimation.

A mean score or weighted regression method was proposed by Reilly and Pepe (1995) for the case where (Y, Z, W) are all discrete. First a nonparametric estimate of the density of X conditional on W and Y is calculated from the validation data. Then the ML estimate of the parameter vector β is derived iteratively via the EM-algorithm.

The second approach is non-parametric estimation of the conditional moments of X or W. Carroll and Stefanski (1990) presented quasilikelihood and variance function (QVF) models. No assumptions are made regarding the distribution of X given W, but the first two moments of the model given W. The pseudolikelihood algorithm estimates the quantities of mean and variance functions nonparametrically, but otherwise employs the standard estimation scheme, i.e. maximizing likelihood and solving QVF estimating equations. Whittemore and Keller (1988) and Sepanski et al. (1994) presented a similar idea.

2.3.5   Bayesian approaches

Bayesians and frequentists differ in their concepts of probability. As a consequence, they handle uncertainty in model parameters differently. Bayesians think of model parameters as random variables, and interpret the probability distribution of a model parameter in terms

of a degree of belief about the value of that parameter. Frequentists think of probabilities as frequencies observed in a long run of repeated experiments, and they view model parameters as fixed (non-random) quantities that, therfore, do not have probability distributions.

In the Bayesian approach, information available at the start of the study leads to specification of the prior distribution of the parameters. Once data have been gathered, inference is made on the basis of the posterior distribution of the parameters given the data, which, by Bayes theorem, is proportional to the product of the likelihood and the prior distribution. From this posterior distribution, point and interval estimates of the parameters can be computed.

Richardson and Gilks (1993) proposed a general Bayesian framework for the analysis of measurement error models. A key assumption in building the structure of such a model is that of conditional independence. To state that Y_i and W_i are conditionally independent given X_i is equivalent to making the classical assumption that surrogate measures W_i do not provide any information on the disease status Y_i if true values of risk factors X_i are known (i.e, measurement error is nondifferential). The three components of the Bayesian model specification are:

1. The disease model, which expresses the relationship between the risk factors Z and X and the disease status Y, i.e $[Y_i|X_i, Z_i, \beta]$.

2. The measurement model, which expresses the relationship between the surrogate measures W and the true unknown risk factor X, i.e $[W_i|X_i, \lambda]$.

3. The exposure model, which specifies the distribution of the unknown risk factor X in the general population, i.e $[X_i|Z_i, \pi]$.

Since we are in a Bayesian framework, prior distributions for β, λ, π are also required (denoted respectively by $[\beta]$, $[\lambda]$ and $[\pi]$). The description of the structure is completed by specifying that the joint distribution of all the variables be written as the product of all the model conditions

$$[\beta][\lambda][\pi] \prod_i [X_i|Z_i, \pi] \prod_i [W_i|X_i, \lambda] \prod_i [Y_i|X_i, Z_i, \beta]$$

Bayesian estimation of parameter is based on the posterior distribution of the parameters given the data. Thus, our interest is really in the marginal posterior distribution of β given

the data. Computing this marginal posterior distribution leads to a very high dimensional integral that is totally intractable. Gibbs sampling is a Markov Chain Monte Carlo method for generating samples from the joint posterior distribution of the model parameters. It was originally proposed by Hastings (1970) and rediscovered by Geman and Geman (1984). The wide applicablity of the algorithm to general statistical modelling was recognized by Gelfand and Smith (1990) and Gelfand et al. (1990), and has since been demonstrated by many authors (Smith and Roberts, 1993; Gilks et al. 1993; Gilks and Wild, 1992). At first, arbitrary starting values for each parameter are chosen. Then, in turn, one parameter at a time is updated by a new value for that parameter from its conditional distribution given the data and the current values of all other parameters in the model, refered to as the current full conditional distributions. A cycle of the Gibbs sampler is completed when all the unknown variables in the model have been updated once. The updating cycle is repeated a large number of times. It has been shown under weak regularity conditions that this process generates a Markov Chain whose equibrium distribution is the distribution of interest, i.e. that the samples generated can be considered after a while as samples from the joint posterior distribution of all the parameters.

Dellaportas and Stephens (1995) also proposed a method that can be used generally under Bayes theory. The full conditional posterior distributions for each of the unknown parameters and covariates that can be specified using Bayes theorem and the conditional priors of the unknown parameters are needed. The full conditionals are then used to derive the Gibbs sampler from the joint distribution of the unknown parameters. If the full conditional densities are not of closed from, algorithms such as rejection methods (Carlin and Gelfand, 1991; Zeger and Karim, 1991), or the Metropolis algorithm (Tierney, 1994) have to be applied.

Kuha (1997) use a special case of Gibbs sampling, namely data augmentation, to estimate the parameters of a regression models, taking into account measurement error in the covariates. The key idea of the approach is to introduce the true values of the covariates measured with error explicitly into the estimation process as missing data. Thus it involves two type of unknown quantities, the parameters to be estimated and the missing data, unobserved because of measurement error. The algorithm works by alternating between two

main steps. In the first step values are imputed (filled in) for the unobserved covariates to generate a data set with no measurement error. In the second step this data set is analysed using standard methods to obtain new values for the parameters. In the terminology of Rubin (1987), regression calibration thus uses "single imputation" of the missing data, while data augmentation is based on "multiple imputation". This approach is a fully parametric structural model and the distributional assumptions are difficult to check as X is missing in the main data set.

Muller and Roeder (1997) and Mallick and Gelfand (1996) proposed "nonparametric or semiparametric Bayes" method for the measurement error model. They incorporate a nonparametric model for the exposure model or the disease model under the Bayesian framework and obtain robust estimation for the effects.

2.3.6   Bootstrap methods

The bootstrap is a very general approach for calculating standard errors or confidence limits of parameter estimates, or bias for sample statistics (Efron and Tibshirani, 1993). The basic idea is quite simple. One creates simulated data sets with replacement called bootstrap data sets, whose distribution is equal to an estimate of the probability distribution of the actual data. Any statistical method that is applied to the actual data can also be applied to the bootstrap data sets. Thus, the empirical distribution of an estimator or test statistic across the bootstrap data sets can be used to estimate the actual sampling distribution of that statistic.

Haukka (1995) proposed a two-stage bootstrap method for the correction of covariate measurement error in the generalized linear model framework. The validation data set and a gold standard are required. The algorithm is as follows:

1. Take a bootstrap sample from the validation data (X_v^*,W_v^* and Z_v^*. The subscript v denotes for validation data and superscript * denotes a bootstrap sample).

2. Estimate the values for α_1^* and α_2^* from the bootstrap sample of the validation data according to the measurement model. e.g: $X = W\alpha_1 + Z\alpha_2 + \varepsilon$

3. Take a bootstrap sample from the main data (Y^*,W^*,Z^*). Estimate β^* of disease

model from the GLM model where X is replaced with its expectation $E(X^*|W^*, Z^*)$.

4. The sampling steps are repeated and the distribution of β^* is used to approximate the true distribution of β.

The sampling variation in the measurement model and in the disease model are accounted for in the first step and third step respectively. The results compare well with other methods are shown to be equally good in general. The bootstrap method is applicable to all link and error function in GLMs when a gold statndard is available, and is thus a good general method for measurement error correction.

3.0 PROPOSED BAYESIAN METHODS

In the Bayesian framework, it is relatively easy to specify a measurement error model for measurement error in continuous variable (we can assume a standard normal distribution for the measurement error model). It is much harder to specify such a model for misclassification in a categorical variable because there is no standard distribution available.

3.1 LIKELIHOOD FORMULATION

We describe our methods in the context of our motivating example. Suppose each patient $(i = 1, \cdots, n)$ has a known inpatient status y_i (1=inpatient, 0=outpatient), and oxygenation results x_{1i} and x_{2i} (1=hypoxemia, 0=non-hypoxemia) obtained by two ascertainments of the underlying latent variable, true oxygenation status, x_i. Let p_1, q_1, p_2, q_2 denote the sensitivity and specificity, respectively, of the two assessments x_1 and x_2 given the true status x. The vector z includes other covariates such as the PSI risk class, which are assumed to be measured perfectly. Our goal is to estimate the odds of inpatient status associated with a true oxygenation status x adjusting for PSI risk class. We also estimate the sensitivities and specificities of the two oxygenation assessments.

We first ignore site and consider a fixed-effect logistic model. We extend the proposed method to incorporate random effects for site in section 3.5. The likelihood function of the data can be formalized by writing three submodels (Richardson and Gilks, 1993a and 1993b): The measurement model:

$$(3.1) \qquad\qquad f(x_1, x_2|y, x, z, \alpha)$$

The outcome model:

$$(3.2) \qquad\qquad\qquad f(y|x, z, \beta)$$

The exposure model:

$$(3.3) \qquad\qquad\qquad f(x|z, \gamma)$$

Where α', β', γ' are three vectors of parameters. In standard logistic regression for this EDCAP example, the outcome and exposure models would be

$$(3.4) \qquad\qquad logit(Pr(y = 1|x, z)) = \beta_0 + \beta_1 x + \widetilde{\beta}_2 z$$

$$(3.5) \qquad\qquad logit(Pr(x = 1|z)) = \gamma_0 + \widetilde{\gamma}_1 z$$

Where x denotes true oxygenation status and z would include 3 indicator variables denoting PSI risk classes 2, 3 and 4/5 combined.

Under the nondifferential misclassification assumption, the measurement model can be simplified as following:

$$(3.6) \qquad\qquad f(x_1, x_2|y, x, z, \alpha) = f(x_1, x_2|x, \alpha)$$

Under the conditional independence assumption of the two imperfect oxygenation assessments, we can specify two conditional distributions:

$$(3.7) \qquad\qquad p(x_{1i}, x_{2i}|x_i = 1) = p_1^{x_{1i}}(1 - p_1)^{1-x_{1i}} p_2^{x_{2i}}(1 - p_2)^{1-x_{2i}}$$

and

$$(3.8) \qquad\qquad p(x_{1i}, x_{2i}|x_i = 0) = q_1^{1-x_{1i}}(1 - q_1)^{x_{1i}} q_2^{1-x_{2i}}(1 - q_2)^{x_{2i}}$$

Where $p_1 = P(x_1 = 1 \mid x = 1)$, $p_2 = P(x_2 = 1 \mid x = 1)$, $q_1 = P(x_1 = 0 \mid x = 0)$, $q_2 = P(x_2 = 0 \mid x = 0)$. So p_1 and q_1 represent the sensitivity and specificity respectively of the retrospective assessment, p_2 and q_2 represent the sensitivity and specificity respectively of the prospective assessment.

Under these two assumptions, we can combine the likelihood function of the measurement model by an indicator function:

$$(3.9) \qquad f(x_{1i}, x_{2i}|x_i) = \prod_{i=1}^{n} p(x_{1i}, x_{2i}|x_i = 1)^{x_i} p(x_{1i}, x_{2i}|x_i = 0)^{1-x_i}$$

For the observed data likelihood, we need to integrate out the unobserved quantity x. From the idea of Data Augmentation in Bayesian statistics, we augment the observed data with the latent variable oxygen status (x) and then define an augmented data likelihood to be used in posterior computations:

$$(3.10) \qquad f(y, x_1, x_2, x \mid z) = \left\{ \prod_{i=1}^{n} f(y_i \mid x_i, z_i, \beta) f(x_i \mid z_i, \gamma) f(x_{1i}, x_{2i} \mid x_i, \alpha) \right\}$$

Based on the three submodels (the outcome model, the exposure model and the measurement model) specified above in equation 3.4, 3.5 and 3.9, the augmented data likelihood function can be expanded as following:

$$
\begin{aligned}
f(y, x_1, x_2, x | z) = &\prod_{i=1}^{n} \frac{exp\{y_i(\beta_0 + \beta_1 x_i + \beta_2' z_i)\}}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2' z_i)} \prod_{i=1}^{n} \frac{exp\{x_i(\gamma_0 + \gamma_1' z_i)\}}{1 + exp(\gamma_0 + \gamma_1' z_i)} \\
(3.11) \qquad &\prod_{i=1}^{n} \left\{ p_1^{x_{1i}}(1 - p_1)^{1-x_{1i}} p_2^{x_{2i}}(1 - p_2)^{1-x_{2i}} \right\}^{x_i} \\
&\prod_{i=1}^{n} \left\{ q_1^{1-x_{1i}}(1 - q_1)^{x_{1i}} q_2^{1-x_{2i}}(1 - q_2)^{x_{2i}} \right\}^{1-x_i}
\end{aligned}
$$

This likelihood is in the form of the product of a binomial likelihood for the outcome model, a binomial likelihood for the exposure model and an indicator function for the measurement model. Because the two binomial components depend only on the parameters β' and γ', and the indicator function depends only on the sensitivity and specificity of the two imperfect tests and true exposure status, we can separate these three components to calculate the conditional posterior distributions of the sensitivity and specificity of the two assessments, $\alpha' = (p_1, q_1, p_2, q_2)$, the outcome model parameters, β', and the exposure model parameters, γ'.

3.2   IDENTIFIABILITY OF PARAMETERS

Let X be a vector of observed random variables. Define f to be a probability distribution function for a model completely specified by parameters θ. If there exists some $\theta_1 \neq \theta_2$ satisfying

$$f(X \mid \theta_1) = f(X \mid \theta_2)$$

for all X, then the parameters of the model are not identifiable, i.e., all possible sets of observations have identical probabilities for two different sets of parameters. Intuitively, a nonidentifiable model can not yield consistent estimator of θ.

Identifiablity of parameters is an important issue for measurement error models, because some additional information must be known about the mismeasurement process in order to obtain a fully identified model. In our motivating example, suppose we do not have a covariate Z, so that we have 6 parameters to be estimated $(\beta_0, \beta_1, p_1, p_2, q_1, q_2)$. The observed data likelihood can be written in the following expression based on equation 3.4-3.6:

$$(3.12) \qquad f(y, x_1, x_2) = \int f(y \mid x, x_1, x_2) f(x_1, x_2 \mid x) dx$$

That is,

$$
(3.13) \qquad
\begin{aligned}
f(y, x_1, x_2) = \int & \frac{exp(\beta_0 + \beta_1 x)y}{1 + exp(\beta_0 + \beta_1 x)} \left\{ p_1^{x_1}(1-p_1)^{1-x_1} p_2^{x_2}(1-p_2)^{1-x_2} \right\}^{x} \\
& \left\{ q_1^{1-x_1}(1-q_1)^{x_1} q_2^{1-x_2}(1-q_2)^{x_2} \right\}^{1-x} dx
\end{aligned}
$$

In the motivating example, x is a Bernoulli variable taking value at 0 and 1, then

$$
(3.14) \qquad
\begin{aligned}
f(y, x_1, x_2) = & \frac{exp(\beta_0 y)}{1 + exp(\beta_0)} \left\{ q_1^{1-x_1}(1-q_1)^{x_1} q_2^{1-x_2}(1-q_2)^{x_2} \right\} \\
& + \frac{exp(\beta_0 + \beta_1)y}{1 + exp(\beta_0 + \beta_1)} \left\{ p_1^{x_1}(1-p_1)^{1-x_1} p_2^{x_2}(1-p_2)^{1-x_2} \right\}
\end{aligned}
$$

For the different combinations of x_1, x_2 and y based on possible value of 0 and 1 for each variable, there are 8 equations obtained as following:

$$(3.15) \qquad \frac{1}{1 + exp(\beta_0)} q_1 q_2 + \frac{1}{1 + exp(\beta_0 + \beta_1)}(1-p_1)(1-p_2)$$

$$\text{(3.16)} \qquad \frac{1}{1+exp(\beta_0)}q_1(1-q_2) + \frac{1}{1+exp(\beta_0+\beta_1)}(1-p_1)p_2$$

$$\text{(3.17)} \qquad \frac{1}{1+exp(\beta_0)}(1-q_1)q_2 + \frac{1}{1+exp(\beta_0+\beta_1)}p_1(1-p_2)$$

$$\text{(3.18)} \qquad \frac{1}{1+exp(\beta_0)}(1-q_1)(1-q_2) + \frac{1}{1+exp(\beta_0+\beta_1)}p_1p_2$$

$$\text{(3.19)} \qquad \frac{exp(\beta_0)}{1+exp(\beta_0)}q_1q_2 + \frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)}(1-p_1)(1-p_2)$$

$$\text{(3.20)} \qquad \frac{exp(\beta_0)}{1+exp(\beta_0)}q_1(1-q_2) + \frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)}(1-p_1)p_2$$

$$\text{(3.21)} \qquad \frac{exp(\beta_0)}{1+exp(\beta_0)}(1-q_1)q_2 + \frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)}p_1(1-p_2)$$

$$\text{(3.22)} \qquad \frac{exp(\beta_0)}{1+exp(\beta_0)}(1-q_1)(1-q_2) + \frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)}p_1p_2$$

It is obvious that the six parameters should be identified from the above eight equations with two redundant equations. Given additional information such as covariate Z (the three indicator variables for PSI class) in our motivating example, the data can be summarized as multiple sets of 2×2 tables and parameter identifiability will not be a problem.

3.3 SPECIFICATION OF PRIORS

We complete the Bayesian formulation of these models by assigning prior distributions for α', β' and γ'. The natural conjugate prior distribution for α is a Beta distribution, say $\alpha_i \sim$ beta $(\alpha_{0i}, \beta_{0i})$ for $i = 1, 2, 3, 4$. To obtain a noninformative prior distribution, we set $\alpha_{0i} = \beta_{0i} = 1$ for $i = 1, \cdots, 4$. For β' and γ', we use diffuse proper priors, the multivariate normal distribution with mean 0 and variance 10, (i.e, N(0, 10)) respectively. Placing proper priors on the parameters guarantees that the posterior distributions are proper. We assume that the prior distributions of all of these parameters are independent.

3.4 MCMC IMPLEMENTATION: FIXED EFFECT LOGISTIC REGRESSION

The MCMC algorithm is a powerful numerical tool for obtaining marginal posteriors of interest when analytical evaluations are difficult. This iterative algorithm builds approximations to the posterior distributions of interest based on Monte Carlo simulations. Compared to alternatives such as numerical integration, MCMC requires much less computing time and is easier to implement and customize. For the augmented data likelihood function and prior specification given above, the joint posterior of α, β, γ and x can be expressed as

$$(3.23) \quad f(x, \alpha, \beta, \gamma | x_1, x_2, y, z) \propto \{\prod_{i=1}^{n} f(x_{1i}, x_{2i} | x_i, \alpha) f(y_i | x_i, z_i, \beta) f(x_i | z_i, \gamma)\} f(\alpha, \beta, \gamma).$$

In our proposed method, the MCMC algorithm alternates between sampling new values for the latent data (e.g, true oxygenation status) and for the population parameters defining each component of the measurement error model based on the relevant full conditional distributions. This algorithm involves both Gibbs sampling (Gelfand and Smith, 1990) and Metropolis-Hastings (Hastings, 1970) steps. The full conditional distributions needed to implement the algorithm are summarized below.

After assigning an arbitrary initial value to each parameter, our MCMC algorithm proceeds as follows:

(1) For i=1, $\cdots$, n, sample x_i from the full conditional posterior distribution of x_i.

$$(3.24) \qquad\qquad f(x_i | \cdot) \sim Bin(1, \frac{w}{w + v})$$

where

$$w = \frac{exp(y_i \beta_1 + \gamma_0 + \gamma_1' z_i)}{1 + exp(\beta_0 + \beta_1 + \beta_2' z_i)} p_1^{x_{1i}} (1 - p_1)^{1 - x_{1i}} p_2^{x_{2i}} (1 - p_2)^{1 - x_{2i}}$$

$$v = \frac{1}{1 + exp(\beta_0 + \beta_2' z_i)} q_1^{1 - x_{1i}} (1 - q_1)^{x_{1i}} q_2^{1 - x_{2i}} (1 - q_2)^{x_{2i}}$$

(2) Sample β using the Metropolis algorithm where

$$(3.25) \qquad\qquad f(\beta | \cdot) \propto \prod_{i=1}^{n} \frac{exp\{y_i(\beta_0 + \beta_1 x_i + \beta_2' z_i)\}}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2' z_i)} exp(-\frac{1}{2} \beta' \Sigma_\beta^{-1} \beta)$$

In large samples, this can be approximated by a Gaussian distribution with mean $\hat{\beta}$, the maximum likelihood estimator, and a variance $\hat{V}_\beta$, the inverse of the Fisher information. That is, to sample from $(\beta|\cdot)$, we find $\hat{\beta}$ and $\hat{V}_\beta$ by performing logistic regression of y_i on x_i and z_i and then generate a random variate β from a multivariate Gaussian distribution, $N(\hat{\beta}, \hat{V}_\beta)$.

(3) Sample γ using the Metropolis algorithm where

(3.26)
$$f(\gamma|\cdot) \propto \prod_{i=1}^{n} \frac{exp\{x_i(\gamma_0 + \gamma_1' z_i)\}}{1 + exp(\gamma_0 + \gamma_1' z_i)} exp(-\frac{1}{2}\gamma' \Sigma_\gamma^{-1} \gamma)$$

We can use the same algorithm as in step (2) by performing logistic regression of x_i on z_i and sample from the multivariate normal approximation.

(4) Sample p_1 from its full conditional Beta distribution

(3.27)
$$f(p_1|\cdot) \propto Beta\{1 + \sum_{i=1}^{n}(x_{1i} = 1 \ \& \ x_i = 1), 1 + \sum_{i=1}^{n}(x_{1i} = 0 \ \& \ x_i = 1)\}$$

(5) Sample p_2 from its full conditional Beta distribution

(3.28)
$$f(p_2|\cdot) \propto Beta\{1 + \sum_{i=1}^{n}(x_{2i} = 1 \ \& \ x_i = 1), 1 + \sum_{i=1}^{n}(x_{2i} = 0 \ \& \ x_i = 1)\}$$

(6) Sample q_1 from its full conditional Beta distribution

(3.29)
$$f(q_1|\cdot) \propto Beta\{1 + \sum_{i=1}^{n}(x_{1i} = 0 \ \& \ x_i = 0), 1 + \sum_{i=1}^{n}(x_{1i} = 1 \ \& \ x_i = 0)\}$$

(7) Sample q_2 from its full conditional Beta distribution

(3.30)
$$f(q_2|\cdot) \propto Beta\{1 + \sum_{i=1}^{n}(x_{2i} = 0 \ \& \ x_i = 0), 1 + \sum_{i=1}^{n}(x_{2i} = 1 \ \& \ x_i = 0)\}$$

(8) Repeat steps 1-7 until apparent convergence and collect a large number of additional draws from which to calculate posterior summaries.

3.5 MCMC IMPLEMENTATION: RANDOM EFFECT LOGISTIC REGRESSION

It is straightforward to generalize the model and computational algorithm to incorporate a random effect. Due to the nature of the Gibb's sampler, we do not need to integrate out the random effects to make inferences about β. Thus, no intractable integral arises for Bayesian inference for GLMM. We just need to update the random effect and unobserved latent variable along with parameters from their posterior distributions.

Let y_{ij} denote the jth observation on the ith cluster for $i = 1, \cdots, I$ and $j = 1, \cdots, n_i$. Conditional on the random effects b_i, the observations in each cluster are independent. The random effect logistic model and the likelihood function for all the observations is given by

$$logit(Pr(y_i = 1|b_i)) = \beta_0 + \beta_1 x_{ij} + \beta_2' z_{ij} + b_i$$

and

$$P(y|\beta, b) = \prod_{i=1}^{I} \prod_{j=1}^{n_i} P(y_{ij}|\beta, b_i)$$

If we assume b_i follows a multivariate Gaussian distribution with mean 0 and variance D, i.e, $b_i \sim N(0, D)$, (where $D = I_{8X8} \otimes \sigma_b^2$), the joint likelihood function for the outcome model has the form

(3.31)

$$f(y_{ij}, b_i|x_{ij}, z_{ij}) \propto \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{exp\{y_i(\beta_0 + \beta_1 x_{ij} + \beta_2' z_{ij} + b_i)\}}{1 + exp(\beta_0 + \beta_1 x_{ij} + \beta_2' z_{ij} + b_i)} \prod_{i=1}^{I} |D|^{-1/2} exp(-\frac{1}{2} b_i' D^{-1} b_i)$$

The measurement model and exposure model are the same as in the fixed effect logistic model. Hence the full conditional sampling for parameters p_1, q_1, p_2, q_2, γ' is also the same. We need to focus on only the conditional distribution of β, b_i, σ_b^2, and $x_i j$ as follows:

(1) Sample β using the Metropolis algorithm,

Given the b_i, the random effect logistic model reduces to a logistic model with offset b_i for each observation. Still assuming a proper diffuse prior distribution N(0,10) for the likelihood function $\prod_{ij} f(y_{ij}|b_i)$, the full conditional distribution for β is

$$(3.32) \qquad f(\beta|\cdot) \propto \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{exp\{y_i(\beta_0 + \beta_1 x_i + \beta_2' z_i + b_i)\}}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2' z_i + b_i)} exp(-\frac{1}{2}\beta' \Sigma_\beta^{-1} \beta)$$

This still can be approximated by a Gaussian distribution with mean $\hat{\beta}_b$, the maximum likelihood estimator, and a variance $\hat{V}_{\beta b}$. That is, to sample from$(\beta|\cdot)$, we find $\hat{\beta}_b$ and $\hat{V}_{\beta b}$ by performing GLM regression of y_{ij} on x_{ij} using the simulated values b_i as offsets and then generate a random variate β from a multivariate Gaussian distribution, N($\hat{\beta}_b$, $\hat{V}_{\beta b}$).

(2) Sample b_i using the Metropolis algorithm,

Generating b_i from its full conditional distribution is the most time-consuming step. As described by Zeger and Karim (1991), we use the Metropolis algorithm to generate the posterior samples. Suppose the conditional moment $\mu_{ij} = E(y_{ij}|b_i)$ and $v_{ij} = var(y_{ij}|b_i)$ satisfy

$$h(\mu_{ij}) = \eta_{ij} = x_{ij}'\beta + z_{ij}'b_i$$

where h is the logit function.

Let $y_i = (y_{i1}, \cdots, y_{in_i})'$, $X_i = (x_i 1', \cdots, x_{in_i}')$, $Z_i = (z_{i1}', \cdots, z_{in_i}')$, $\mu_i = (\mu_{i1}, \cdots, \mu_{in_i})$, $\eta_i = (\eta_{i1}, \cdots, \eta_{in_i})$, and $V_i = diag(v_{i1}, \cdots, v_{in_i})$

If we let $y_i^\star$ be the linear approximation to $h(y_i)$ given by

$$y_i^\star = \eta_i + (\frac{\partial \mu_i}{\partial \eta_i})^{-1}(y_i - \mu_i)$$

then the maximum value of $p(b_i)$ occurs at $\hat{b}_i = (Z_i'V_iZ_i + D^{-1})^{-1}Z_iV_i(y_i^\star - X_i\beta)$, and its curvature is $\hat{v}_i = (Z_i'V_iZ_i + D^{-1})^{-1}$. Because $y_i^\star$ and v_i depend on b_i, the actual mode and curvature must be obtained by iterating the equations for $\hat{b}_i$ and $\hat{v}_i$.

(3) Sample σ_b^2 from Inverse Gamma distribution,

For the prior distribution of random effect variance, we assign inverse gamma IG $(1, 1)$, the most commonly used distribution for the variance of normally distributed random effects in Bayesian hierarchical models. This choice leads to a simple Gibbs sampling update for the variance component. Therefore, we can sample σ_b^2 from the posterior inverse gamma distribution.

$$(3.33) \qquad f(\sigma_b^2|\cdot) \propto IG((\frac{I+1}{2}, \frac{\sum_{i=1}^{I} b_i^2 + 1}{2}))$$

(4) Sample x_{ij} from the Bernoulli distribution,

The true unobserved quantity x_{ij} can be updated from the Bernoulli distribution as in the fixed effect logistic model, except that we include the offset b_i for each observation.

$$(3.34) \qquad f(x_i|\cdot) \sim Bin(1, \frac{w'}{w' + v'})$$

where

$$w' = \frac{exp(y_i\beta_1 + \gamma_0 + \gamma_1' z_i)}{1 + exp(\beta_0 + \beta_1 + \beta_2' z_i + b_i)} p_1^{x_{1i}}(1 - p_1)^{1-x_{1i}} p_2^{x_{2i}}(1 - p_2)^{1-x_{2i}}$$

$$v' = \frac{1}{1 + exp(\beta_0 + \beta_2' z_i + b_i)} q_1^{1-x_{1i}}(1 - q_1)^{x_{1i}} q_2^{1-x_{2i}}(1 - q_2)^{x_{2i}}$$

(5) Sample $p1$, $q1$, $p2$, $q2$ and γ' from the same posterior distributions as in the fixed effect logistic model.

4.0   RESULTS

We present an application to the motivating example and simulation studies to illustrate the proposed methodology. In the motivating example we analyze data from the control arm of the EDCAP trial. The MCMC algorithm described in chapter 3 was used to generate 25,000 MCMC iterates, with the first 500 discarded as a burn-in. The posteriors for the motivating example were computed using naive estimates as starting values in the MCMC algorithm. Stationarity in all cases were reached in about 500 iterations and was determined by inspection of trace plots. The computations were carried out by R, whose interactive graphics provided a convenient environment for monitoring the convergence of the MCMC algorithm. The R programs were run on a Pentium PC with 1.4G HZ with running time about 4 hours for fixed effect logistic model and 10 hours for random effect logistic model for our motivating example. Convergence was also assessed by a variety of diagnostic techniques (Gelman and Rubin, 1992; Geweke, 1992; Heidelberger and Welch, 1983; Raftery and Lewis, 1992), summarized by Cowles and Carlin (1996) and implemented using BOA (Smith, 2000) in the R software. The simulation studies use expected data (e.g, expected frequencies for each pattern of covariates) based on a known model. The purpose of this example is to check our algorithm and computational implementation.

4.1   RANDOM EFFECT LOGISTIC MODEL FOR MOTIVATING EXAMPLE

We applied the proposed approach to the pneumonia data described in the begining of section 2, accounting for clinical sites as a random effect. The parameter vector β' includes the effects of the true oxygenation status and indicator variables denoting the PSI risk classes

28

II, III and IV/V on inpatient status. The parameter vector γ' include the effects of PSI risk classes II, III and IV/V on the true oxygenation status. Figure 4.1 give the trace plots for all the parameters in our motivating example and we found no evidence of lack of convergence or of slow mixing based on examination of trace plots and standard diagnostic tests of the sampled parameters.

We compare the parameters by three Bayesian models: Treating x_{1i} (observed retrospective oxygenation status) as the true value; treating x_{2i} (observed prospective oxygenation status) as the true value; and the proposed method, treating both the retrospective and prospective assesments as imperfect measures of the latent true value. Posterior summaries of the parameters characterizing the relationship between inpatient status, oxygenation status and PSI risk class are shown in Table 4.1 (The estimated parameter vector γ' describing the exposure model are not shown). Posterior kernel densities of β_1 (the coefficient of mismeasured oxygenation) estimated by the three methods are plotted in Figure 4.2.

Patients with abnormal oxygenation status are more likely to be hospitalized (Table 4.1). The posterior mean and median of β_1, the log odds ratio of hospitalization for hypoxemic patient relative to otherwise similar non-hypoxemic patient in our proposed method (3.98, 3.71) are substantially higher than those assuming that retrospective (2.02, 2.00) and prospective (2.22, 2.17) assessments were accurate. The point estimate of log odds in our proposed method (3.98) corresponds to an odds ratio of 54. The 95 per cent credible interval for the random effect variance (σ_b^2) does not include 0 , indicating significant variance across clinical sites. The standard deviation and 95% credible interval for β_1 in proposed method (1.43 and (2.02, 7.55)) are much larger than those in the other two approaches (0.44, (1.22, 2.94)) and (0.63, (1.11, 3.59)), as shown in Figure 4.2. The proposed method correctly reflects the uncertainty about the latent variable x when only two noisy surrogates x_1 and x_2 are observed. In addition, the log odds of being hospitalized increases with increasing PSI risk class although the estimates are similar across all three approaches because this variable was measured without error. Both the retrospective and prospective assessments have high specificity (0.94 and 0.99) and relatively low sensitivity (0.77 and 0.68). Figure 4.2 shows the plots of posterior densities of β_1 under the three methods, which is consistent with the results in the table 4.1.

4.2 SIMULATION STUDIES FOR THE RANDOM EFFECT LOGISTIC MODEL

In simulation study 1, we consider a situation in which the unobserved binary predictor x is related to outcome y (1=disease, 0=no disease), one dichotomous covariates z (1=presence, 0=absence of certain characteristic) and a random intercept $b_i \sim N(0, 0.64)$ in the outcome model, with log odds 3.5, 1.5 which is similar to our motivating example. The covariate z is related to the true exposure x in the exposure model with a log odds of 0.5. The prevalence of the true exposure is chosen to be 0.3 for patients with covariate $z = 0$, and the prevalence of disease is chosen to be 0.3 with both covariate z and true exposure x set to zero. We assume that the sensitivity and specificity of two imperfect assessments x_1 and x_2 do not differ across levels of the y or z, i.e. nondifferential misclassification of the true variable x with respect to both disease and covariates. We assume that x_1 and x_2 are independent given the true exposure x. We chose sensitivity of 90 per cent and specificity 70 per cent for x_1 and sensitivity of 75 per cent and specificity of 95 per cent for x_2. The above assumptions and parameter values can be represented by the following components:

$$p(x_{1i}, x_{2i}|x_i = 1) = p_1^{x_{1i}}(1 - p_1)^{1-x_{1i}} p_2^{x_{2i}}(1 - p_2)^{1-x_{2i}}$$

and

(4.1) $$p(x_{1i}, x_{2i}|x_i = 0) = q_1^{1-x_{1i}}(1 - q_1)^{x_{1i}} q_2^{1-x_{2i}}(1 - q_2)^{x_{2i}}$$

(4.2) $$logit\{Pr(y = 1|x, z, b)\} = \beta_0 + \beta_1 x + \beta_2 z + b_i$$

(4.3) $$logit\{Pr(x = 1|z)\} = \gamma_0 + \gamma_1 z$$

Where $(\beta_0, \beta_1, \beta_2)$=(-0.70, 3.50, 1.50), (γ_0, γ_1)=(-0.85, 0.5) are parameters translated from the previously specified values of log odds and prevalences of exposure and disease. To simulate data based on the above model specification, we generated 8 random intercepts according $N(0, 0.64)$ and computed conditional probabilities $p(x|x_1, x_2)$ and $p(y = 1|x, x_1, x_2)$

given the covariate pattern. For each combination of b_i and z, we simulated 100 observations and computed the expected frequencies for different patterns of x_1 and x_2 based on conditional distribution equation 4.1. A total of 800 observations was generated.

With the same prior distributions described in the previous section, we used a burn-in period of 500 iterations and calculated posterior mean, standard deviation and 95 per cent credible intervals based on next 25,000 iterations. We took every 10th sample from 25,000 posterior samples to avoid correlation between the adjacent samples. Fitted parameters of the models are shown in Table 4.2 for three methods: (1) assuming that x_1 is measured without error, (2) assuming that x_2 is measured without error, and (3) the proposed method accounting for misclassification of x_1 and x_2. The posterior densities of β_1 under the three methods are plotted in Figure 4.3. If we assume the two imperfect measures as the true latent variable x, the point estimates for β_1 are around 1.44 and 2.25 respectively, which are far away from true value 3.50. The corresponding 95 per cent credible intervals do not cover the true value. But in our proposed method, the point estimates (mean 3.56 and median 3.45) are much closer to the true value and the 95 per cent credible interval does cover the true value. We also notice that the much bigger standard deviation for β_1 (0.66) in our proposed method compared to the two naive methods (0.18 and 0.28). It is reasonable since the true latent variable x is unobserved and we have only two noisy variables. The larger standard deviation correctly reflects uncertainty in the true latent variable x. The estimated sensitivities and specificities of two imperfect measures are almost identical to the true values. The posterior kernel density plots for β_1 under the three models in figure 4.3 are consistent with the results in the table. Therefore it appears that our MCMC algorithm is correctly specified and produces reasonable estimates of the parameters.

In order to check whether our proposed Bayesian method works well under different misclassification structure, we conducted sensitivity analysis for four different misclassification structures in additional simulation studies: In simulation study 2, both imperfect measurements have high sensitivity and high specificity, $(p_1, q_1, p_2, q_2) = (0.90, 0.95, 0.95, 0.90)$. In simulation study 3, both imperfect measurements have low sensitivity and low specificity, $(p_1, q_1, p_2, q_2) = (0.70, 0.80, 0.80, 0.75)$. In simulation study 4, both imperfect measurements have high sensitivity but low specificity, $(p_1, q_1, p_2, q_2) = (0.95, 0.75, 0.90, 0.80)$ and

in simulation study 5, both imperfect assessments have low sensitivity but high specificity, $(p_1, q_1, p_2, q_2) = (0.75, 0.90, 0.80, 0.95)$. Other parameters in these simulation studies are the same as those in simulation study 1.

The parameters in the three models are compared under different misclassification structures: the first two naive methods treat the two imperfect assessments as the true values respectively, and our proposed method assumes that the true variable is unobserved. The results of the sensitivity analysis are listed in Table 4.3, 4.4, 4.5 and 4.6. The parameters listed in these tables are the same as in simulation study-1 (Table 4.2). Estimation of β_1 is summarized in Table 4.7 under all 4 scenarios. From the results of sensitivity analysis, the point parameter estimates from median of two naive methods in these simulation studies ((2.55, 2.25), (1.21, 1.19), (1.47, 1.61) and (1.85, 2.31)) are far away from the true value 3.50 and the 95 per cent credible intervals do not cover the true value. The median is more representative for the point estimate because the distribution of β_1 might not be symmetric. But in our proposed method, the point estimates of the median (3.53, 3.51, 3.59 and 3.45) are much closer to the true value and the 95 per cent credible intervals do cover the true value. Also the bigger standard deviation associated with our proposed method (0.42, 1.24, 0.94 and 0.55) compared to those in naive methods, is consistent with the results in simulation study 1. It is interesting to see the much bigger standard deviation in the simulation study 3 (1.24) in our proposed method. With the low sensitivities and low specificities specified for two imperfect assessments in simulation study 3, the data can be classified with lower accuracy and higher uncertainty. The posterior kernel density plots for the misclassified variable β_1 shown in Figure 4.4 are consistent with the results in the corresponding tables.
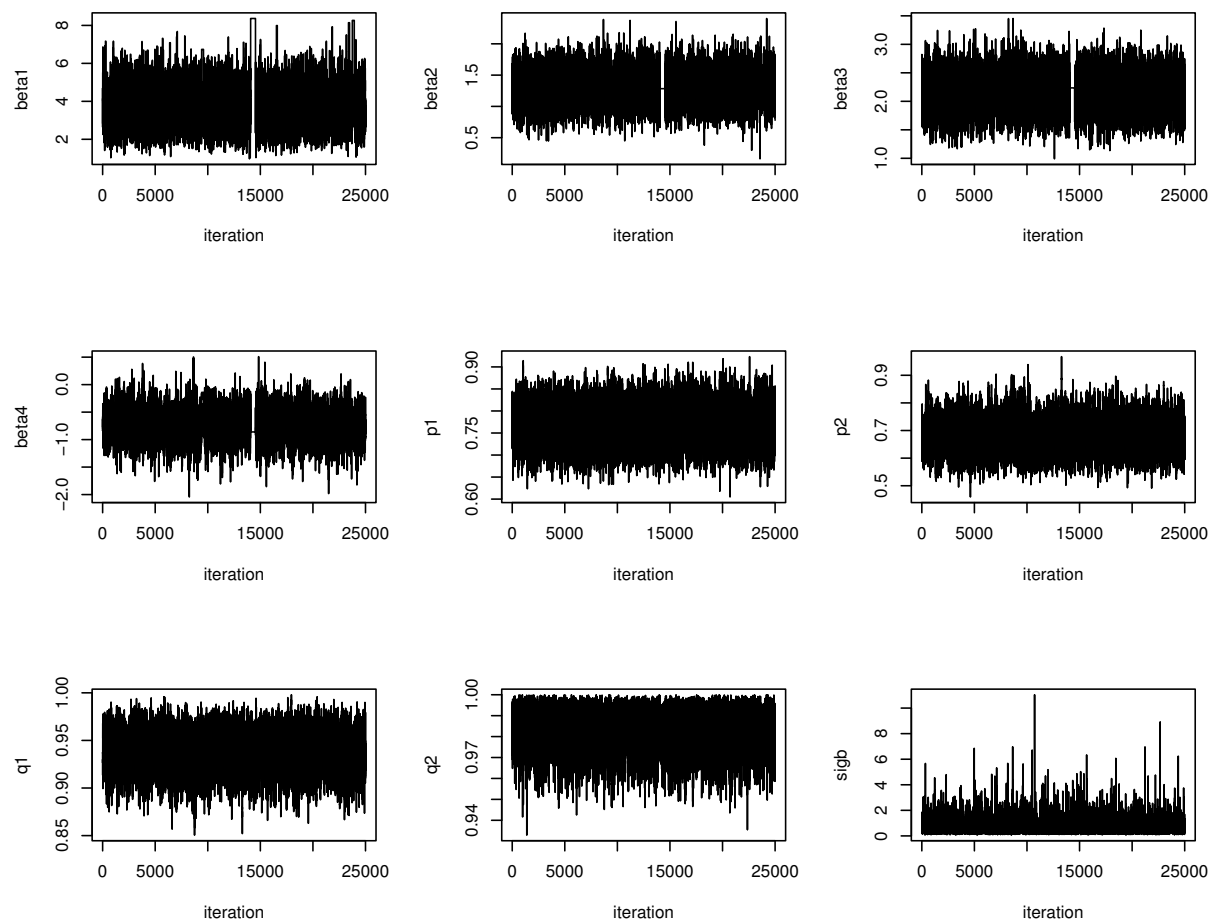
Figure 4.1: *Trace plots for the parameters in the motivating example*

Table 4.1: Posterior summaries of the parameters in random effect logistic model for the motivating example

Methods	Parameters	Posterior summaries			
		Mean	Median	SD	95% Credible interval
Retrospective assessment	β_0	-0.71	-0.71	0.26	(-1.20, -0.20)
	β_1	2.02	2.00	0.44	(1.22, 2.94)
	β_2	1.25	1.25	0.25	(0.77, 1.75)
	β_3	2.18	2.18	0.31	(1.57, 2.78)
	β_4	4.30	4.26	0.55	(3.35, 5.50)
	σ_b^2	0.59	0.48	0.43	(0.17, 1.71)
Prospective assessment	β_0	-0.59	-0.58	0.26	(-1.10, -0.06)
	β_1	2.22	2.17	0.63	(1.11, 3.59)
	β_2	1.26	1.25	0.24	(0.78, 1.76)
	β_3	2.21	2.20	0.30	(1.61, 2.79)
	β_4	4.26	4.23	0.54	(3.28, 5.38)
	σ_b^2	0.55	0.45	0.39	(0.16, 1.58)
Proposed method	β_0	-0.74	-0.74	0.27	(-1.28, -0.22)
	β_1	3.98	3.71	1.43	(2.02, 7.55)
	β_2	1.32	1.31	0.26	(0.82, 1.82)
	β_3	2.14	2.13	0.32	(1.54, 2.78)
	β_4	4.09	4.06	0.54	(3.08, 5.19)
	σ_b^2	0.61	0.49	0.43	(0.17, 1.75)
	p_1	0.77	0.77	0.04	(0.69, 0.85)
	p_2	0.68	0.68	0.06	(0.57, 0.79)
	q_1	0.94	0.94	0.02	(0.90, 0.98)
	q_2	0.99	0.99	0.01	(0.96, 1.00)

β_1: The log odds of inpatient status associated with mismeasured variable (oxygenation status)

β_2, β_3, β_4 : The log odds of inpatient status associated with risk class II, III and IV/V

σ_b^2: The variance of random effects

p_1, q_1 and p_2, q_2 are sensitivity and specificity of retrospective assessment and prospective assessment respectively.
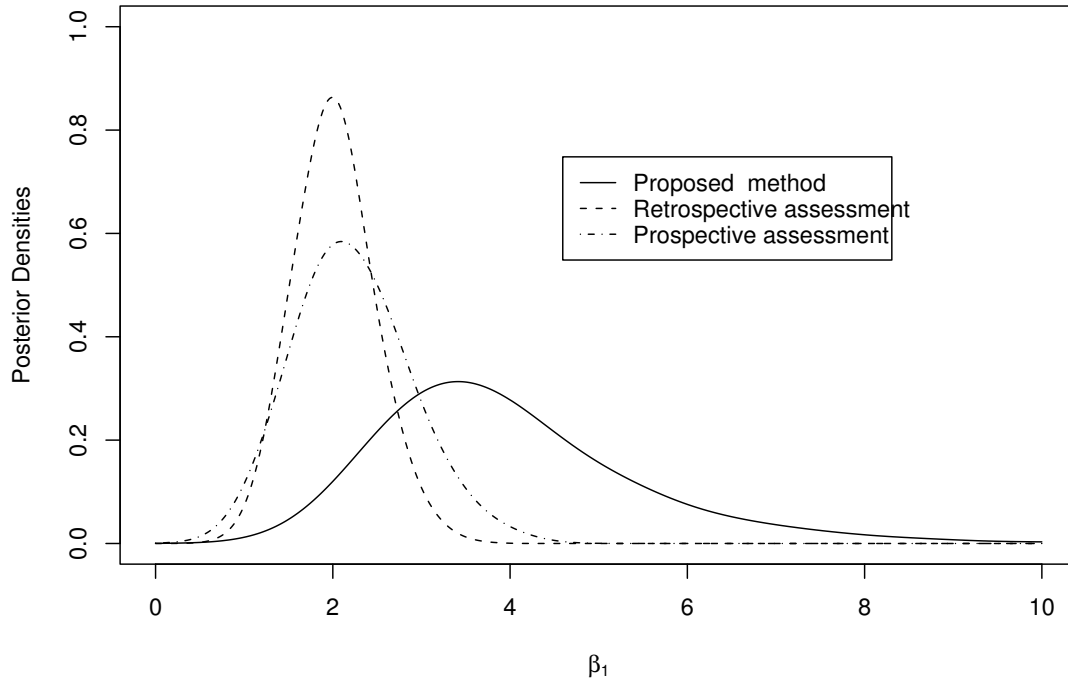
Figure 4.2: *Estimated posterior densities for the oxygen status parameter β_1 in a random effect logistic model for the motivating example under 3 different methods: Proposed method (solid curve), Retrospective assessment (dashed curve) and Prospective assessment (dotted curve).*

Table 4.2: Posterior summaries of the parameters in random effect logistic model for simulation study-1 with sensitivity 0.90, 0.70 and specificity 0.75, 0.95 for x_1 and x_2 respectively

Methods	Parameters	True value	Posterior summaries			
			Mean	Median	SD	95% Interval
"Imperfect	β_0	-0.70	-0.43	-0.43	0.23	(-0.88, 0.04)
-measure 1"	β_1	3.50	1.44	1.44	0.18	(1.09, 1.79)
	β_2	1.50	1.40	1.40	0.18	(1.05, 1.75)
	σ_b^2	0.64	0.60	0.49	0.42	(0.20, 1.63)
"Imperfect	β_0	-0.70	-0.29	-0.30	0.22	(-0.74, 0.13)
-measure 2"	β_1	3.50	2.25	2.24	0.28	(1.74, 2.83)
	β_2	1.50	1.44	1.44	0.18	(1.09, 1.78)
	σ_b^2	0.64	0.63	0.51	0.43	(0.20, 1.75)
Proposed	β_0	-0.70	-0.61	-0.60	0.27	(-1.14, -0.06)
method	β_1	3.50	3.56	3.45	0.66	(2.63, 5.20)
	β_2	1.50	1.46	1.46	0.21	(1.04, 1.88)
	σ_b^2	0.64	0.77	0.64	0.51	(0.24, 2.15)
	p_1	0.90	0.90	0.90	0.03	(0.84, 0.96)
	p_2	0.70	0.70	0.69	0.05	(0.61, 0.80)
	q_1	0.75	0.75	0.75	0.03	(0.69, 0.81)
	q_2	0.95	0.95	0.95	0.02	(0.91, 0.98)

β_1: The log odds of "disease" associated with the mismeasured variable.

β_2: The log odds of "disease" associated with covariate z.

σ_b^2: The variance of random effect.

p_1, q_1 and p_2, q_2 are sensitivity and specificity of two imperfect assessments respectively.

Figure 4.3: *Estimated posterior densities for the mismeasured variable parameter β_1 in a random effect logistic model for simulated data under 3 different methods: Proposed method (solid curve), Imperfect measure 1 and 2 ignoring misclassification (dashed curve and dotted curve). The vertical line denotes true value.*

Table 4.3: Posterior summaries of the parameters in random effect logistic model for simulation study-2 with sensitivity 0.90, 0.95 and specificity 0.95, 0.90 for x_1 and x_2 respectively

Methods	Parameters	True value	Posterior summaries			
			Mean	Median	SD	95% Interval
"Imperfect	β_0	-0.70	-0.48	-0.48	0.24	(-0.94, -0.01)
-measure 1"	β_1	3.50	2.55	2.55	0.26	(2.06, 3.08)
	β_2	1.50	1.43	1.43	0.19	(1.05, 1.79)
	σ_b^2	0.64	0.68	0.56	0.51	(0.22, 1.78)
"Imperfect	β_0	-0.70	-0.53	-0.54	0.24	(-1.00, -0.04)
-measure 2"	β_1	3.50	2.25	2.25	0.22	(1.82, 2.69)
	β_2	1.50	1.47	1.48	0.19	(1.09, 1.86)
	σ_b^2	0.64	0.70	0.59	0.49	(0.21, 1.98)
Proposed	β_0	-0.70	-0.61	-0.61	0.25	(-1.11, -0.14)
method	β_1	3.50	3.55	3.53	0.42	(2.83, 4.46)
	β_2	1.50	1.50	1.50	0.20	(1.11, 1.89)
	σ_b^2	0.64	0.80	0.65	0.53	(0.24, 2.31)
	p_1	0.90	0.90	0.90	0.03	(0.84, 0.97)
	p_2	0.95	0.94	0.94	0.02	(0.89, 0.99)
	q_1	0.95	0.95	0.95	0.01	(0.91, 0.97)
	q_2	0.90	0.90	0.90	0.02	(0.86, 0.94)

β_1: The log odds of "disease" associated with the mismeasured variable.

β_2: The log odds of "disease" associated with covariate z.

σ_b^2: The variance of random effect.

p_1, q_1 and p_2, q_2 are sensitivity and specificity of two imperfect assessments respectively.

Table 4.4: Posterior summaries of the parameters in random effect logistic model for simulation study-3 with sensitivity 0.70, 0.80 and specificity 0.80, 0.75 for x_1 and x_2 respectively

			Posterior summaries			
Methods	Parameters	True value	Mean	Median	SD	95% Interval
"Imperfect	β_0	-0.70	-0.24	-0.25	0.23	(-0.70, 0.21)
-measure 1"	β_1	3.50	1.22	1.21	0.18	(0.88, 1.58)
	β_2	1.50	1.41	1.41	0.17	(1.08, 1.76)
	σ_b^2	0.64	0.58	0.48	0.38	(0.18, 1.54)
"Imperfect	β_0	-0.70	-0.31	-0.31	0.23	(-0.75, 0.14)
-measure 2"	β_1	3.50	1.20	1.19	0.18	(0.86, 1.56)
	β_2	1.50	1.41	1.41	0.18	(1.08, 1.78)
	σ_b^2	0.64	0.58	0.48	0.42	(0.18, 1.56)
Proposed	β_0	-0.70	-0.61	-0.60	0.30	(-1.24, -0.05)
method	β_1	3.50	3.81	3.51	1.24	(2.25, 7.13)
	β_2	1.50	1.49	1.48	0.23	(1.06, 1.95)
	σ_b^2	0.64	0.78	0.65	0.54	(0.24, 2.17)
	p_1	0.70	0.71	0.71	0.06	(0.62, 0.85)
	p_2	0.80	0.80	0.80	0.06	(0.69, 0.92)
	q_1	0.80	0.80	0.80	0.03	(0.74, 0.86)
	q_2	0.75	0.75	0.75	0.03	(0.68, 0.82)

β_1: The log odds of "disease" associated with the mismeasured variable.

β_2: The log odds of "disease" associated with covariate z.

σ_b^2: The variance of random effect.

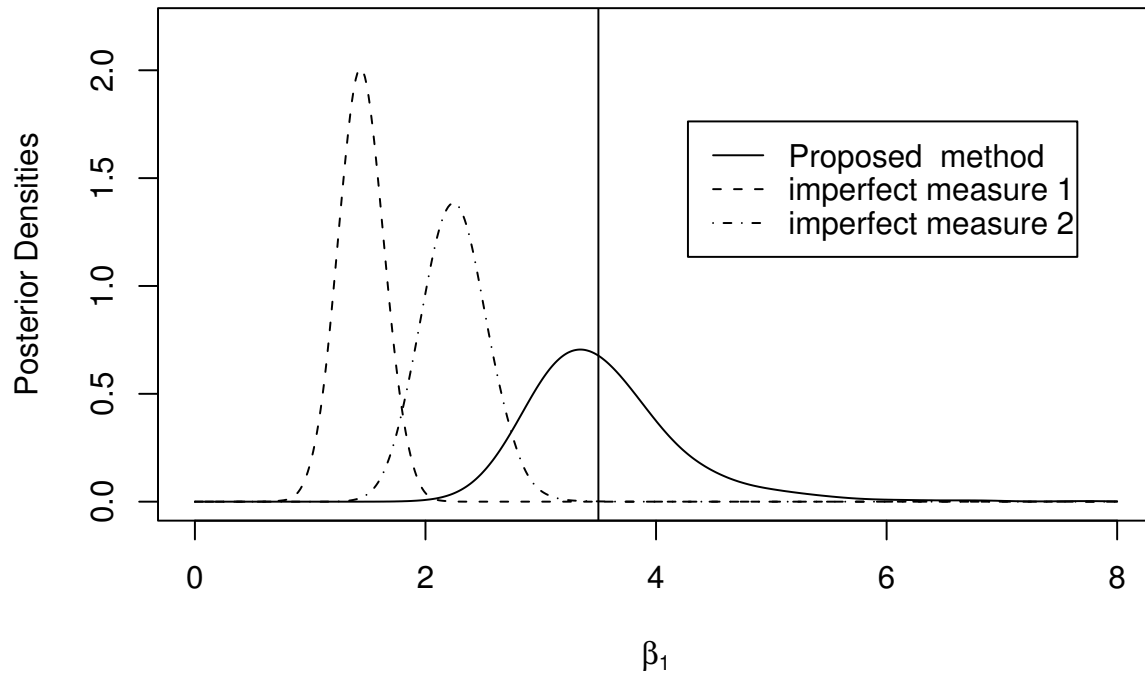p_1, q_1 and p_2, q_2 are sensitivity and specificity of two imperfect assessments respectively.

Table 4.5: Posterior summaries of the parameters in random effect logistic model for simulation study-4 with sensitivity 0.95, 0.90 and specificity 0.75, 0.80 for x_1 and x_2 respectively

			Posterior summaries			
Methods	Parameters	True value	Mean	Median	SD	95% Interval
"Imperfect	β_0	-0.70	-0.46	-0.46	0.23	(-0.90, 0.00)
-measure 1"	β_1	3.50	1.48	1.47	0.18	(1.15, 1.84)
	β_2	1.50	1.39	1.39	0.17	(1.06, 1.74)
	σ_b^2	0.64	0.60	0.50	0.39	(0.19, 1.62)
"Imperfect	β_0	-0.70	-0.45	-0.46	0.22	(-0.88, 0.00)
-measure 2"	β_1	3.50	1.61	1.61	0.18	(1.27, 1.97)
	β_2	1.50	1.40	1.40	0.19	(1.05, 1.75)
	σ_b^2	0.64	0.60	0.49	0.42	(0.19, 1.66)
Proposed	β_0	-0.70	-0.61	-0.60	0.26	(-1.13, -0.09)
method	β_1	3.50	3.79	3.59	0.94	(2.57, 6.34)
	β_2	1.50	1.45	1.45	0.21	(1.04, 1.87)
	σ_b^2	0.64	0.78	0.63	0.55	(0.23, 2.33)
	p_1	0.95	0.93	0.93	0.03	(0.86, 0.99)
	p_2	0.90	0.91	0.90	0.04	(0.82, 0.98)
	q_1	0.75	0.74	0.74	0.03	(0.69, 0.81)
	q_2	0.80	0.80	0.80	0.03	(0.74, 0.85)

β_1: The log odds of "disease" associated with the mismeasured variable.

β_2: The log odds of "disease" associated with covariate z.

σ_b^2: The variance of random effect.

p_1, q_1 and p_2, q_2 are sensitivity and specificity of two imperfect assessments respectively.

Table 4.6: Posterior summaries of the parameters in random effect logistic model for simulation study-5 with sensitivity 0.75, 0.80 and specificity 0.90, 0.95 for x_1 and x_2 respectively

Methods	Parameters	True value	Posterior summaries			
			Mean	Median	SD	95% Interval
"Imperfect	β_0	-0.70	-0.31	-0.31	0.23	(-0.73, 0.16)
-measure 1"	β_1	3.50	1.85	1.85	0.21	(1.45, 2.25)
	β_2	1.50	1.41	1.41	0.18	(1.08, 1.77)
	σ_b^2	0.64	0.62	0.52	0.40	(0.19, 1.70)
"Imperfect	β_0	-0.70	-0.37	-0.37	0.22	(-0.78, 0.09)
-measure 2"	β_1	3.50	2.32	2.31	0.26	(1.86, 2.87)
	β_2	1.50	1.41	1.41	0.18	(1.05, 1.74)
	σ_b^2	0.64	0.66	0.55	0.41	(0.21, 1.70)
Proposed	β_0	-0.70	-0.59	-0.60	0.27	(-1.20, -0.07)
method	β_1	3.50	3.48	3.45	0.55	(2.61, 4.75)
	β_2	1.50	1.45	1.45	0.20	(1.09, 1.86)
	σ_b^2	0.64	0.76	0.63	0.57	(0.22, 2.29)
	p_1	0.75	0.75	0.75	0.03	(0.69, 0.83)
	p_2	0.80	0.80	0.80	0.04	(0.73, 0.90)
	q_1	0.90	0.90	0.90	0.02	(0.86, 0.94)
	q_2	0.95	0.95	0.95	0.02	(0.91, 0.97)

β_1: The log odds of "disease" associated with the mismeasured variable.

β_2: The log odds of "disease" associated with covariate z.

σ_b^2: The variance of random effect.

p_1, q_1 and p_2, q_2 are sensitivity and specificity of two imperfect assessments respectively.

Table 4.7: Summary of parameter β_1 in the sensitivity analysis under different sensitivities and specificities.

Simulations			Posterior summaries			
(p_1, q_1, p_2, q_2)	Parameters	True value	Mean	Median	SD	95% Interval
"Imperfect 1"	β_1	3.50	2.55	2.55	0.26	(2.06, 3.08)
"Imperfect 2"	β_1	3.50	2.25	2.25	0.22	(1.82, 2.69)
(0.90,0.95,0.95,0.90)	β_1	3.50	3.55	3.53	0.42	(2.83, 4.46)
"Imperfect 1"	β_1	3.50	1.22	1.21	0.18	(0.88, 1.58)
"Imperfect 2"	β_1	3.50	1.20	1.19	0.18	(0.86, 1.56)
(0.70,0.80,0.80,0.75)	β_1	3.50	3.81	3.51	1.24	(2.25, 7.13)
"Imperfect 1"	β_1	3.50	1.48	1.47	0.18	(1.15, 1.84)
"Imperfect 2"	β_1	3.50	1.61	1.61	0.18	(1.27, 1.97)
(0.95,0.75,0.90,0.80)	β_1	3.50	3.79	3.59	0.94	(2.57, 6.34)
"Imperfect 1"	β_1	3.50	1.85	1.85	0.21	(1.45, 2.25)
"Imperfect 2"	β_1	3.50	2.32	2.31	0.26	(1.86, 2.87)
(0.75,0.90,0.80,0.95)	β_1	3.50	3.48	3.45	0.55	(2.61, 4.75)

β_1: The log odds of "disease" associated with the mismeasured variable.
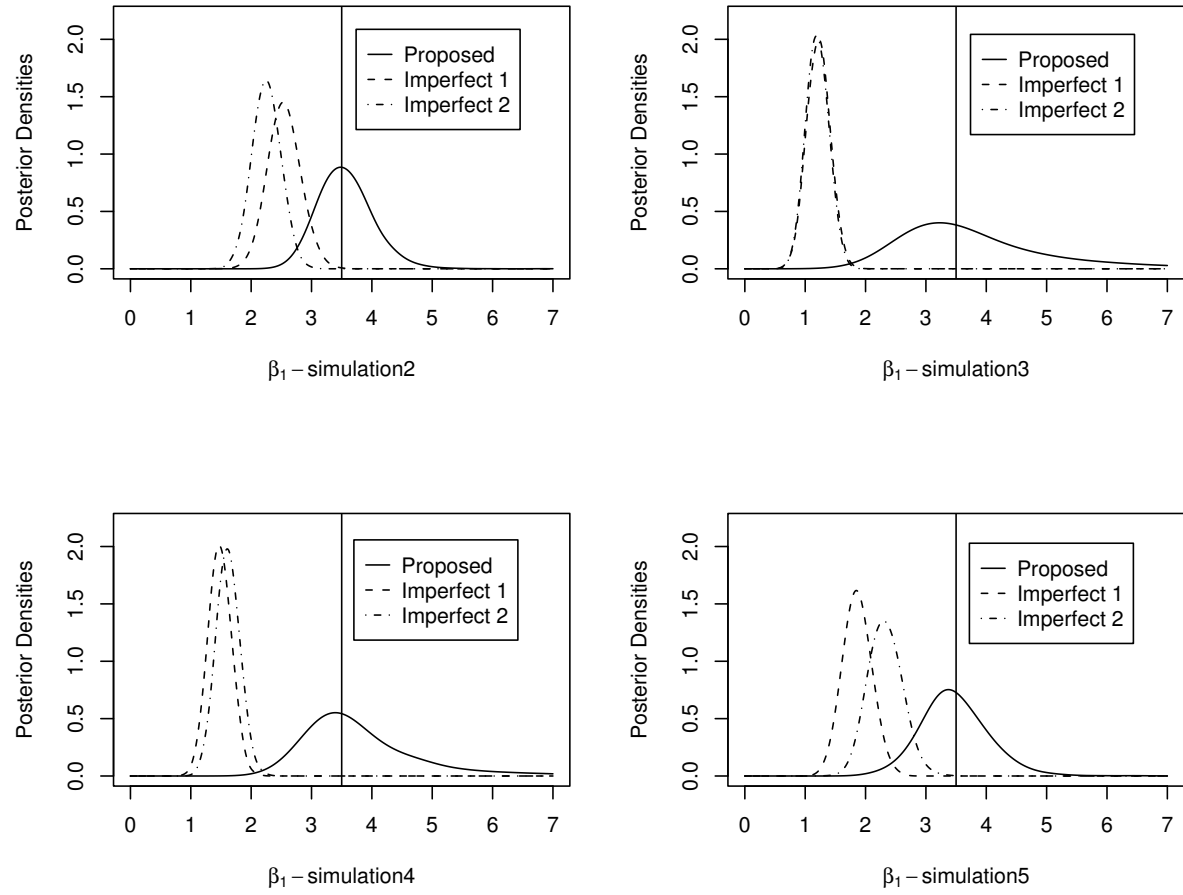
Figure 4.4: *Sensitivity analysis: Estimated posterior densities for mismeasured variable*

5.0 APPLICATION

The specific aim of the EDCAP study was to compare the effectiveness and safety of three interventions of increasing intensity (low-intensity, moderate-intensity and high-intensity) to implement a PSI-based guideline incorporating arterial oxygenation information for the initial site of treatment. Outpatient care was recommended for low risk patients, defined as non-hypoxemic patients in PSI risk classes I–III. Inpatient care was recommended for higher risk patients (hypoxemic and/or PSI risk classes IV and V). To be consistent with the published paper for EDCAP study, we use outpatient status as outcome variable instead inpatient status in our motivating example. We estimate the odds of outpatient status (1=outpatient, 0=inpatient) as a function of intervention arm and risk status in the 32 sites, accounting for site as a random effect. The model is expressed as following:

$$(5.1) \quad logit\left\{Pr(y=1)\right\} = \beta_0 + \beta_1 * risk + \beta_2 * intervention2 + \beta_3 * intervention3 + b_i$$

This data analysis is complicated by the misclassification of binary variable of risk status (1=high risk, 0=low risk). Similar to our motivating example, the true risk status is unobserved and ascertained by two imperfect measurements: prospectively assessed in the Emergency Department and retrospectively assessed from medical chart review. We assume the two assessments are independent conditional on the true unobserved risk status. Table 5.1 displays the data for 3201 patients enrolled from 16 sites in western Pennsylvania and 16 sites in Connecticut on outpatient status, intervention arm and the two imperfect assessments of risk status.

Although the prospective assessment was used in the primary analysis of the EDCAP trial, subsequent analyses have demonstrated discrepancies between the prospective and retrospective assessments. Due to time lapses in data recording, risk status may be ascertained

retrospectively from ED medical chart review using data that were not in fact available in real time. Among these 3201 patients, 84% were classified concordently (52% as low risk and 34% as high risk) on the two assessments and 16% were classified discordently, with 7% classified as low risk by the prospective but not the retrospective assessment, and 7% classified as high risk by the retrospective but not the prospective assessment.

The proposed Bayesian methods were applied to the EDCAP trial data to evaluate the true intervention effects after adjusting for potential misclassification of risk status and accounting for the 32 clinical sites as a random effect. The parameter vector β' includes the effects of risk status, moderate intensity intervention and high intensity intervention on outpatient status. Figure 5.1 give the trace plots for all the parameters in this EDCAP study and we found no evidence of lack of convergence or of slow mixing based on examination of trace plots and standard diagnostic tests of the sampled parameters.

We compare the parameters by three Bayesian models: Treating retrospectively observed risk status as the true value; treating prospectively observed risk status as the true value; and the proposed method, treating both the retrospective and prospective assesments as imperfect measures of the latent true value. Posterior summaries of the parameters characterizing the relationship between outpatient status, risk status and intervention effects are shown in Table 5.2. Posterior kernel densities of β_1 (the coefficient of mismeasured risk status), β_2 (the coefficient of the moderate-intensity intervention), β_3 (the coefficient of the high-intensity intervention) estimated by the three methods are plotted in Figures 5.2, 5.3 and 5.4.

Low-risk patients are less likely to be hospitalized (Table 5.2). The posterior mean and median of β_1, the log odds of outpatient care for a high risk patient relative to an otherwise similar low-risk patient, in our proposed method (-3.81, -3.81) are lower than those assuming that retrospective (-3.02, -3.03) and prospective (-3.29, -3.28) assessments were accurate. The 95 per cent credible interval for random effect variance σ_b^2 (0.24, 0.81) does not include 0, indicating significant variance across sites. The standard deviation and 95% credible interval for β_1 in proposed method (0.17 and (-4.13, -3.49)) are larger than those in the other two approaches (0.12, (-3.27, -2.78)) and (0.13, (-3.56, -3.02)), as shown in Figure 5.2. The proposed method appropriately reflects the uncertainty about the latent variable

risk status when only two noisy surrogates are observed. The log odds of outpatient care in the moderate and high intensity arms are (0.89, 0.48) with corresponding odds ratios (2.44, 1.62) from the retrospective assessment, (0.97, 0.82) with corresponding odds ratios (2.64, 2.27) from the prospective assessment and (0.98, 0.74) with corresponding odds ratios (2.66, 2.10) from our proposed method. It is not too surprising that the results from prospective assessment are closer to the proposed method than estimates from the retrospective assessment because the prospective assessment has somewhat higher sensitivity and specificity (0.86, 0.99) than the retrospective assessment (0.85, 0.97). We also cross validated the first two naive Bayesian methods with maximum likelihood approaches by using SAS PROC NLMIXED procedure (Table 5.3) and obtained similar parameter estimates. This demonstrates that our computation algorithm and non-informative prior specification in our Bayesian approaches are reasonable.

We also found that log odds of outpatient care for patients in the moderate intensity intervention arm is a little higher relative to the high intensity intervention arm under three Bayesian methods. One important question is whether they are significantly different and we tested the hypothesis: $\beta_2 - \beta_3 = 0$. The 95 per cent credible intervals for $\beta_2 - \beta_3$ are (-0.01, 0.85) assuming that retrospective assessment is correct, (-0.24, 0.56) assuming that prospective assessment is correct and (-0.22, 0.69) in our proposed method. All these intervals including 0 show that there is no significant difference between the moderate and high intensity interventions based on this main effects model. But we noticed that in the retrospective assessment (the less accurate assessment), the 95 per cent credible interval (-0.01, 0.85) barely excludes 0, compared to intervals from the prospective assessment (-0.24, 0.56) and our proposed method (-0.22, 0.69).

One interesting issue in this study is the possible differential misclassification of risk status across the intervention arms. Table 5.4, 5.5 and 5.6 list the 2×2 tables of two imperfect assessments for the low-intensity intervention arm, moderate-intensity intervention arm and high-intensity intervention arm. Among these patients in different intervention arms, 16% were classified discordently in the low-intensity intervention arm, with 10% classified as low risk by the prospective but not the retrospective assessment, and 6% classified as high risk by the retrospective but not the prospective assessment. In the moderate-intensity

intervention arm, 13% were classified discordently , with 8% classified as low risk by the prospective but not the retrospective assessment, and 5% classified as high risk by the retrospective but not the prospective assessment. In the high-intensity intervention arm, 13% were classified discordently, with 4% classified as low risk by the prospective but not the retrospective assessment, and 9% classified as high risk by the retrospective but not the prospective assessment. In this EDCAP study we are also interested in assessing whether the effect of risk status varies by intervention arm. Because the guideline recommendations differ in low-risk and higher-risk patients, with outpatient care recommended for low-risk patients and inpatient care recommended for higher-risk patients, one would expect the intervention effect to vary by risk status. In our proposed Bayesian method such relationships can not be assessed because the true variable of risk status is unobserved and updating multiple latent variable is impossible. And the sensitivities and specificities of two imperfect assessments can not be estimated for three intervention arms separately due to lack of parameter identifiability. Here we used the prospective assessment (more accurate assessment) to evaluate whether risk status is differential to intervention arms by fitting the model:

$$(5.2) \quad \begin{aligned} logit\left\{Pr(y=1)\right\} = \beta_0 + \beta_1 risk + \beta_2 intervention2 + \beta_3 intervention3 \\ \beta_4 intervention2 * risk + \beta_5 intervention3 * risk + b_i \end{aligned}$$

The results for this model are listed in Table 5.7. The 95 per cent credible interval for β_4 and β_5 are (-0.72, 0.98) and (-1.46, 0.28), which include 0. The prospective assessment shows no evidence of a significant interaction.

Table 5.1: Frequencies of observed patterns of covariates and outpatient status for the full EDCAP data

Covariate pattern		Risk status assessment		
Outpatient status	Intervention	Prospective	Retrospective	Frequency
1	1	0	0	159
1	1	0	1	8
1	1	1	0	2
1	1	1	1	5
1	2	0	0	444
1	2	0	1	15
1	2	1	0	10
1	2	1	1	29
1	3	0	0	420
1	3	0	1	13
1	3	1	0	9
1	3	1	1	11
0	1	0	0	210
0	1	0	1	68
0	1	1	0	39
0	1	1	1	249
0	2	0	0	221
0	2	0	1	73
0	2	1	0	48
0	2	1	1	319
0	3	0	0	221
0	3	0	1	43
0	3	1	0	106
0	3	1	1	479
			Total	3201

Outpatient status (1=outpatient, 0=inpatient)

Intervention arm (1=low-intensity, 2=moderate-intensity, 3=high-intensity)

Prospective assessment (1=high risk, 0=low risk)

Retrospective assessment (1=high risk, 0=low risk)

Table 5.2: Posterior summaries of the parameters in random effect logistic model for the EDCAP data using a Bayesian approach

| Methods | Parameters | Posterior summaries | | | |
		Mean	Median	SD	95% Credible interval
Retrospective assessment	β_0	-0.34	-0.34	0.19	(-0.72, 0.02)
	β_1	-3.02	-3.02	0.12	(-3.27, -2.78)
	β_2	0.89	0.89	0.25	(0.38, 1.35)
	β_3	0.48	0.48	0.25	(0.02, 0.96)
	β_2-β_3	0.42	0.42	0.22	(-0.01, 0.85)
	σ_b^2	0.41	0.38	0.14	(0.22, 0.73)
Prospective assessment	β_0	-0.46	-0.46	0.19	(-0.83, -0.09)
	β_1	-3.29	-3.28	0.13	(-3.56, -3.02)
	β_2	0.97	0.98	0.23	(0.49, 1.42)
	β_3	0.82	0.81	0.23	(0.34, 1.29)
	β_2-β_3	0.15	0.15	0.20	(-0.24, 0.56)
	σ_b^2	0.37	0.36	0.11	(0.20, 0.64)
Proposed method	β_0	-0.15	-0.15	0.21	(-0.58, 0.25)
	β_1	-3.81	-3.81	0.17	(-4.13, -3.49)
	β_2	0.98	0.97	0.28	(0.40, 1.51)
	β_3	0.74	0.74	0.27	(0.22, 1.27)
	β_2-β_3	0.23	0.23	0.23	(-0.22, 0.69)
	σ_b^2	0.46	0.44	0.15	(0.24, 0.81)
	p_1	0.85	0.85	0.01	(0.83, 0.87)
	p_2	0.86	0.86	0.01	(0.84, 0.89)
	q_1	0.97	0.97	0.006	(0.95, 0.98)
	q_2	0.99	0.99	0.004	(0.98, 1.00)

β_1: The log odds of outpatient care associated with mismeasured variable (risk status)

β_2: The log odds of outpatient care associated with moderate intensity intervention

β_3: The log odds of outpatient care associated with high intensity intervention

σ_b^2: The variance of random effects

p_1, q_1 and p_2, q_2 are sensitivity and specificity of retrospective assessment and prospective assessment respectively.

Table 5.3: The naive parameter estimates from maximum likehood method in random effect logistic model for the EDCAP data

Methods	Parameters	Estimate	S.E.	95% Confidence Interval
Retrospective assessment	β_0	-0.32	0.23	(-0.79, 0.14)
	β_1	-3.02	0.13	(-3.28, -2.76)
	β_2	0.88	0.29	(0.29, 1.47)
	β_3	0.48	0.29	(-0.11, 1.07)
	σ_b^2	0.32	0.10	(0.11, 0.53)
Prospective assessment	β_0	-0.44	0.22	(-0.89, -0.001)
	β_1	-3.28	0.14	(-3.56, -2.99)
	β_2	0.94	0.28	(0.38, 1.51)
	β_3	0.81	0.28	(0.24, 1.38)
	σ_b^2	0.29	0.09	(0.10, 0.49)

β_1: The log odds of outpatient care associated with risk status.

β_2: The log odds of outpatient care associated with moderate intensity intervention.

β_3: The log odds of outpatient care associated with high intensity intervention.

σ_b^2: The variance of random effect.

Table 5.4: Classification by two imperfect measurements in the low-intensity intervention arm

	Retrospective 0	Retrospective 1
Prospective 0	369 (50%)	76 (10%)
Prospective 1	41 (6%)	254 (34%)

Discordance:16%

Table 5.5: Classification by two imperfect measurements in the moderate-intensity intervention arm

	Retrospective 0	Retrospective 1
Prospective 0	665 (57%)	88 (8%)
Prospective 1	58 (5%)	348 (30%)

Discordance:13%

Table 5.6: Classification by two imperfect measurements in the high-intensity intervention arm

	Retrospective 0	Retrospective 1
Prospective 0	641 (49%)	56 (4%)
Prospective 1	115 (9%)	490 (38%)

Discordance:13%

Table 5.7: Posterior summaries of the parameters in random effect logistic interaction model for EDCAP data from prospective assessment

Methods	Parameters	Mean	Median	SD	95% Credible interval
			Posterior summaries		
Prospective assessment	β_0	-0.54	-0.54	0.10	(-0.72, -0.34)
	β_1	-3.48	-3.46	0.39	(-4.31, -2.74)
	β_2	1.11	1.12	0.13	(0.86, 1.37)
	β_3	1.23	1.23	0.13	(0.97, 1.46)
	β_4	0.15	0.14	0.44	(-0.72, 0.98)
	β_5	-0.59	-0.59	0.46	(-1.46, 0.28)
	σ_b^2	0.52	0.51	0.13	(0.31, 0.84)

β_1: The log odds of outpatient care associated with mismeasured variable (risk status)

β_2: The log odds of outpatient care associated with moderate intensity intervention

β_3: The log odds of outpatient care associated with high intensity intervention

β_4: The interaction between moderate intensity intervention and risk status

β_5: The interaction between high intensity intervention and risk status

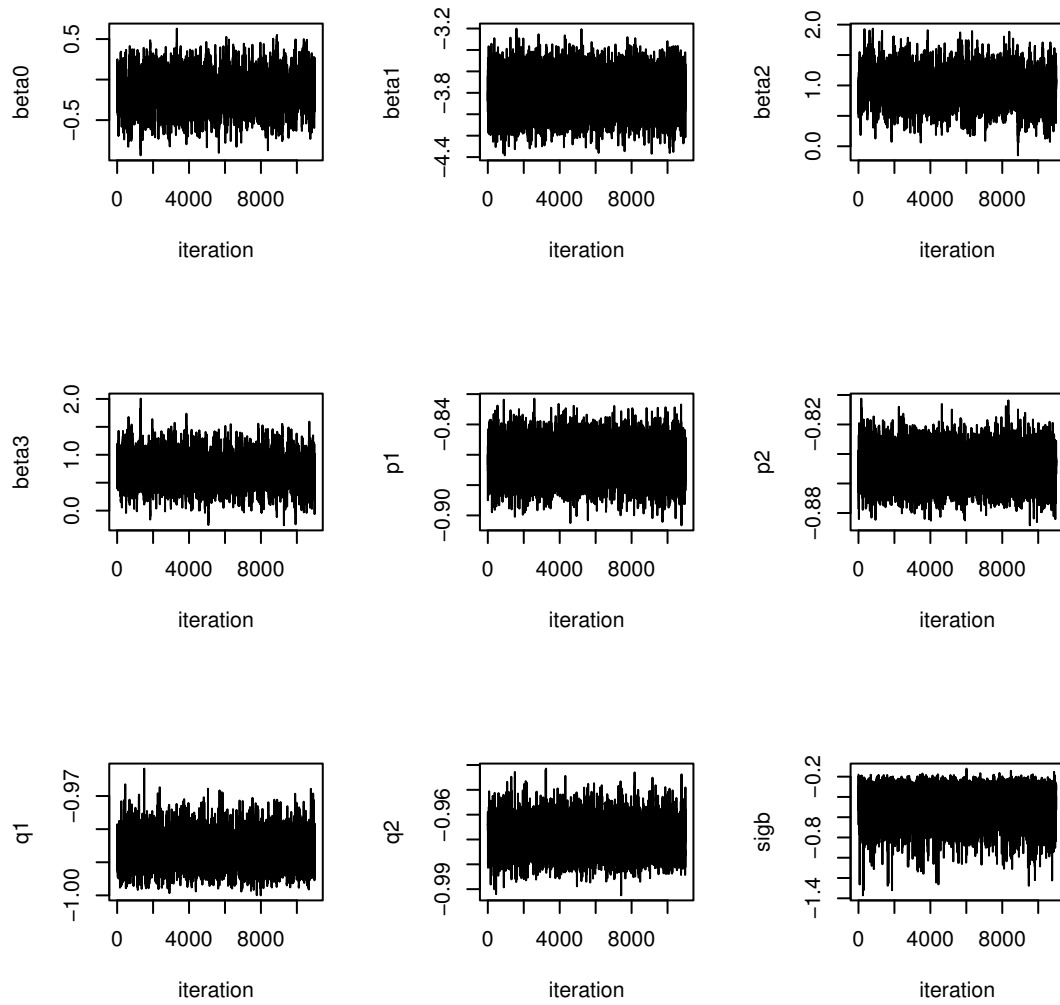σ_b^2: The variance of random effects

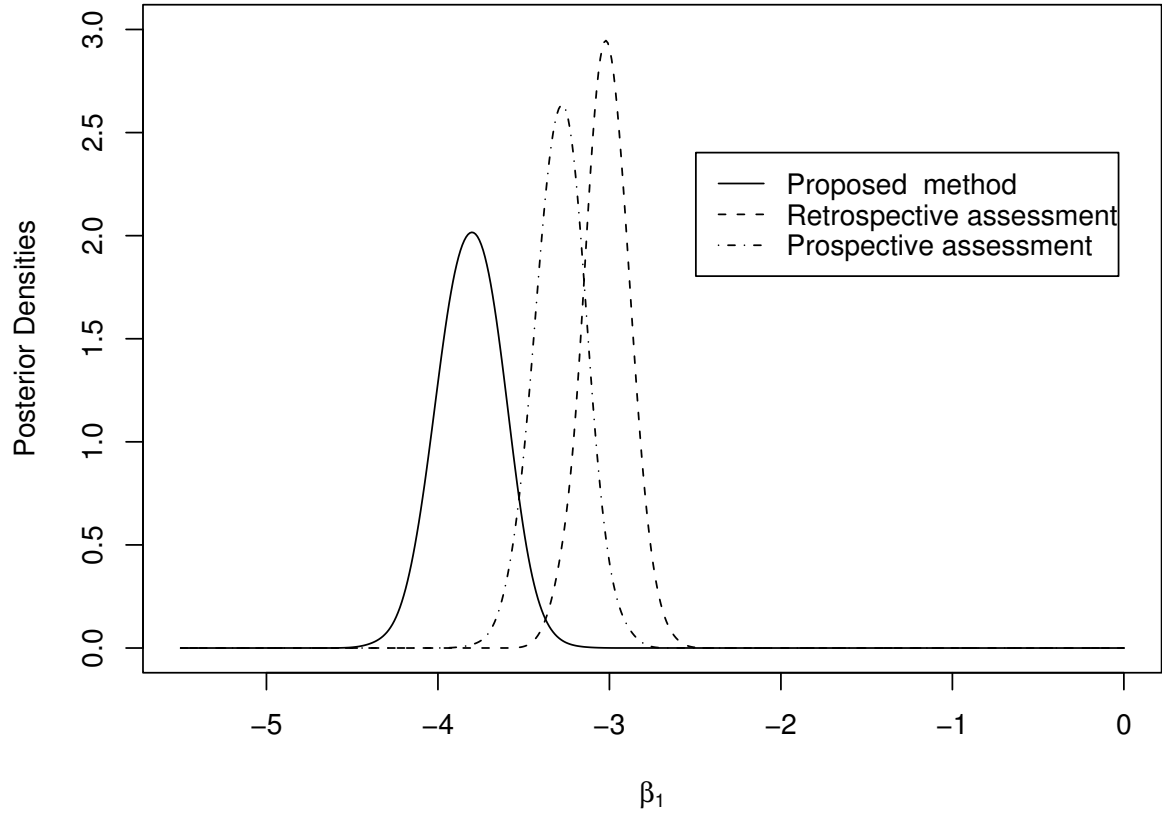Figure 5.1: *Trace plots of parameters in EDCAP data*

Figure 5.2: *Estimated posterior densities for the risk class status parameter β_1 in a random effect logistic model for the full EDCAP data under 3 different methods: Proposed method (solid curve), Retrospective assessment (dashed curve) and Prospective assessment (dotted curve).*
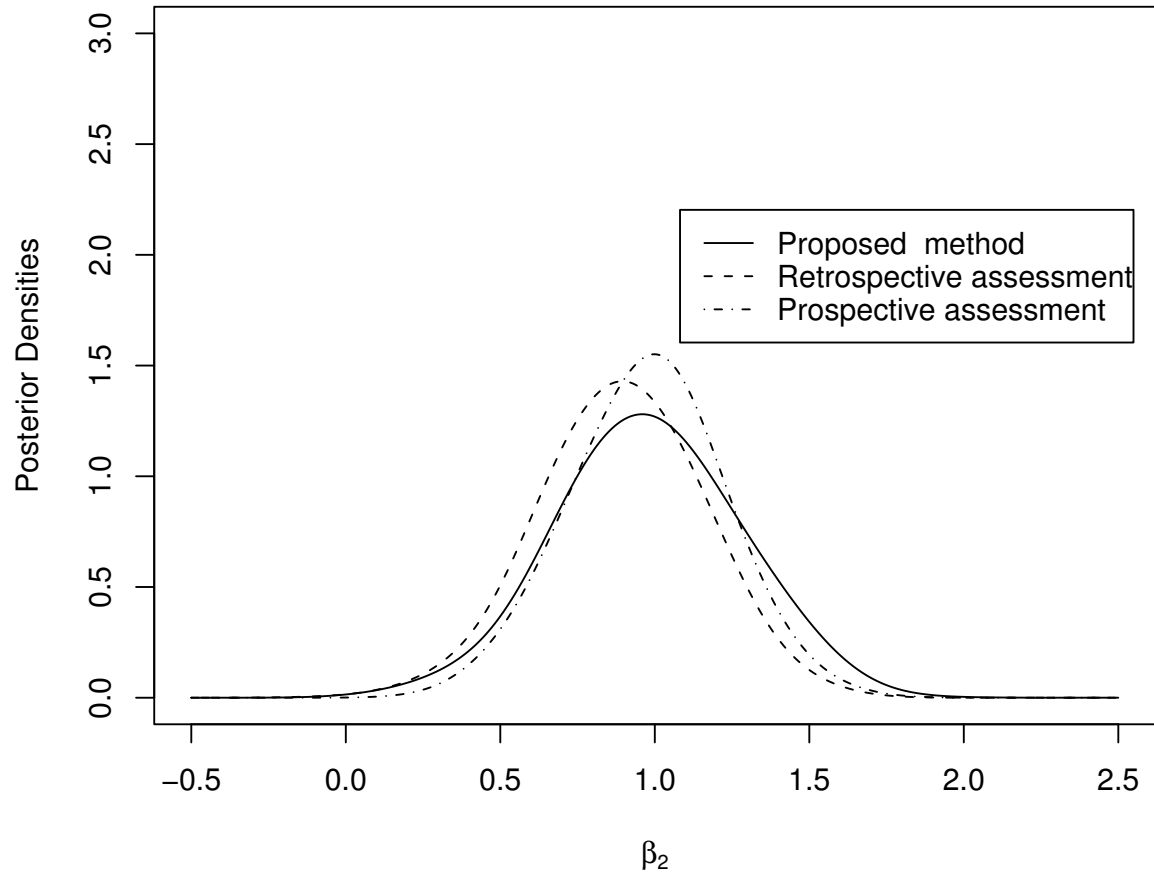
Figure 5.3: *Estimated posterior densities for the moderate intensity intervention effect parameter β_2 in a random effect logistic model for the full EDCAP data under 3 different methods: Proposed method (solid curve), Retrospective assessment (dashed curve) and Prospective assessment (dotted curve).*
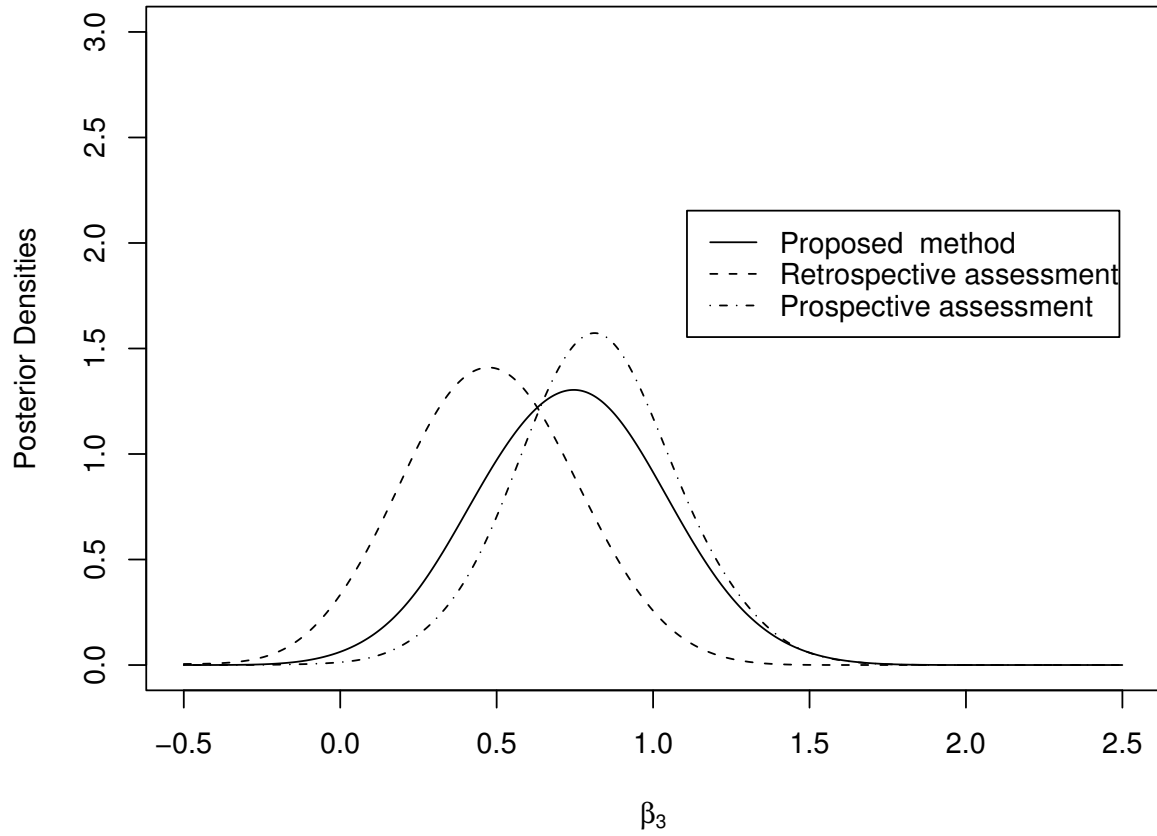
Figure 5.4: *Estimated posterior densities for the high intensity intervention effect parameter β_3 in a random effect logistic model for the full EDCAP data under 3 different methods: Proposed method (solid curve), Retrospective assessment (dashed curve) and Prospective assessment (dotted curve).*

6.0 CONCLUSIONS

The main advantage of a Bayesian approach for measurement error problems, as discussed by Richardson and Gilks (1993), is that it allows the problem to be modelled in a conceptually straightforward way without approximations. All the available information is utilized and the uncertainty from different sources is properly reflected in the parameter estimates. Moreover, it works under more complicated model frameworks (e.g, GLMM) and accounts for measurement error in a relatively straightforward way. In our example, attenuated estimates of the mismeasured covariate are not surprising because nondifferential misclassification tends to attenuate the parameter estimate to the null. Similarly, the wider credible intervals for mismeasured variable β_1 under the proposed method are expected because a strength of the Bayesian approach is the ability to appropriately propagate uncertainty.

This dissertation has described a flexible Bayesian approach to adjust for misclassification of a dichotomous explanatory variable in the absence of a gold standard. This approach allows adjustment for other discrete or continuous covariates when evaluating the relationship between outcome and covariates. The sensitivities and specificities of two imperfect assessments can also be estimated from our proposed method. After data augmentation, the unobserved variables can be sampled along with the model parameters from their full posterior distribution and the posterior computation can proceed via a straightforward Gibbs sampling algorithm with a Metropolis step. This algorithm is shown to produce reasonable results for real and simulated data.

Parameter identifiablity is an important issue in the models that account for the mismeasurement, because some additional information must be known about the mismeasurement process in order to obtain a fully identified model. If there is no additional covariate Z in the model, the likelihood function of the random effect logistic model lacks parameter identifi-

ablity, and the parameters can not be consistently estimated. Given additional information covariate Z in our motivating example (the three indicator variables for PSI risk class) and EDCAP trial (the two indicator variables for intervention arm), the data can be summarized multiple sets of 2×2 tables and parameter identifiability will not be a problem.

In the EDCAP trial, one interesting issue is to evaluate the possible differential misclassification of risk status across different intervention arms. In our proposed Bayesian method such relationships can not be assessed because the true variable of risk status is unobserved and updating this latent variable is impossible in our algorithm. And the sensitivities and specificities of two imperfect assessments can not be estimated for three intervention arms separately due to lack of parameter identifiability. But if we have additional covariates in the model, it is worth to be further investigated in the future.

In this dissertation, it has been assumed that the two imperfect tests are conditionally independent given the true latent measurement. This assumption is not empirically verifiable without any gold-standard measurements and it may be dubious in some applications (Fryback 1978, Brenner 1996, Vacek 1985). We will consider extending the model to allow for conditional dependence between the two tests given the true variable in future work.

APPENDIX A

POSTERIOR CONDITIONAL DISTRIBUTION

The key step in our MCMC algorithm is to impute the latent variable x from their full conditional distribution. The posterior distribution of x can be simplified and updated from the Bernoulli distribution.

$$
\begin{aligned}
p(x_i|\cdot) &\propto \prod_{i=1}^{n} \frac{exp\{y_i(\beta_0 + \beta_1 x_i + \beta_2' z_i)\}}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2' z_i)} \prod_{i=1}^{n} \frac{exp\{x_i(\gamma_0 + \gamma_1' z_i)\}}{1 + exp(\gamma_0 + \gamma_1' z_i)} \\
&\quad \prod_{i=1}^{n} \left\{ p_1^{x_{1i}}(1-p_1)^{1-x_{1i}} p_2^{x_{2i}}(1-p_2)^{1-x_{2i}} \right\}^{x_i} \left\{ q_1^{1-x_{1i}}(1-q_1)^{x_{1i}} q_2^{1-x_{2i}}(1-q_2)^{x_{2i}} \right\}^{1-x_i} \\
&\propto \prod_{i=1}^{n} \left\{ \frac{exp(y_i \beta_1 + \gamma_0 + \gamma_1' z_i)}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2' z_i)} \right\}^{x_i} \left\{ \frac{1}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2' z_i)} \right\}^{1-x_i} \\
&\quad \left\{ p_1^{x_{1i}}(1-p_1)^{1-x_{1i}} p_2^{x_{2i}}(1-p_2)^{1-x_{2i}} \right\}^{x_i} \left\{ q_1^{1-x_{1i}}(1-q_1)^{x_{1i}} q_2^{1-x_{2i}}(1-q_2)^{x_{2i}} \right\}^{1-x_i} \\
&\propto \prod_{i=1}^{n} \left\{ \frac{exp(y_i \beta_1 + \gamma_0 + \gamma_1' z_i)}{1 + exp(\beta_0 + \beta_1 + \beta_2' z_i)} p_1^{x_{1i}}(1-p_1)^{1-x_{1i}} p_2^{x_{2i}}(1-p_2)^{1-x_{2i}} \right\}^{x_i} \\
&\quad \left\{ \frac{1}{1 + exp(\beta_0 + \beta_2' z_i)} q_1^{1-x_{1i}}(1-q_1)^{x_{1i}} q_2^{1-x_{2i}}(1-q_2)^{x_{2i}} \right\}^{1-x_i}
\end{aligned}
$$

APPENDIX B

MCMC ALGORITHM

In Bayesian inference, the parameter is assumed to be random with some known distribution and its estimate ends up being an expected value under the posterior distribution. For example, given a sample y from a distribution with likelihood $L(y \mid \theta)$ and a prior density for θ, the posterior distribution of parameter θ is proportional to the product of data likelihood function and prior distribution by Baye's theorem, i.e,

$$\pi(\theta \mid y) \propto L(y \mid \theta)\pi(\theta)$$

Given the posterior, and in the case where $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$ is multivariate, for example, we may be interested in the marginal posterior distributions of θ_1, such as

$$p(\theta_1 \mid y) = p(\theta \mid y)d\theta_2 \cdots d\theta_p$$

Computing this marginal posterior distribution leads to a very high dimensional integral that is totally intractable. Traditionally we would be forced to use numerical integration or analytic approximation techniques. However, the Markov Chain Monte Carlo (MCMC) method provides an alternative by performing the integration implicitly. We can sample from the posterior distribution directly and obtain parameter estimates of the quantities of interest. Gibb's sampler (Hastings, 1970) and Metropolis-Hasting (MH) algorithm (Tierney, 1994) are most often used MCMC methods. I will give a little more details in the following:

B.1   GIBB'S SAMPLER

The Gibb's sampler is an algorithm for simulating a Markov chain $x_1, \cdots, x_k$ which is converging to $f(x)$, by successively sampling from the full conditional component distributions $f(x_i \mid x_{-i}), i = 1, \cdots, p$ where x_{-i} denotes the components of x other than x_i. For example, suppose that $\theta = (\theta_1, \theta_2, \theta_3)$, and the target distribution is $p(\theta \mid y)$, the Gibb's sampler algorithm proceeds as following:

1. Start with initials $(\theta_1^0, \theta_2^0, \theta_3^0)$.
2. Draw θ_1^1 from $p(\theta_1 \mid \theta_2 = \theta_2^0, \theta_3 = \theta_3^0, y)$
3. Draw θ_2^1 from $p(\theta_2 \mid \theta_1 = \theta_1^1, \theta_3 = \theta_3^0, y)$
4. Draw θ_3^1 from $p(\theta_3 \mid \theta_1 = \theta_1^1, \theta_2 = \theta_2^1, y)$

At the beginning all the parameters were assigned an arbitrary initial values. A cycle of Gibb's sampler is completed when all the parameters were sampled once. Repeat this procedure a large times after convergence. The samples generated form posterior distributions can be considered from the the target distribution $p(\theta \mid y)$.

Conceptually, the Gibb's sampler is fairly straightforward. Ideally, the conditional distribution $f(x_i \mid x_{-i})$ will be of the form of a standard distribution and a suitable prior specification often ensures that this is the case. However, in the cases where it is non-standard, there are many ways to sample from the appropriate conditionals and Metropolis-Hastings (M-H) algorithm is often applied.

B.2   METROPOLIS-HASTING (M-H) ALGORITHM

An alternative, and more general, updating scheme is as a form of generalized rejection sampling, where values are drawn from arbitrary distributions and rejected or accepted according to the importance ratio. Asymptotically the samples generated behave as random observations from the target distribution. The M-H algorithm proceeds as follwing:

1. Given a draw θ_t in iteration t, sample a candidate draw θ^* from a proposal distribution $J(\theta^* \mid \theta)$

2. Accept the draw with probability

$$r = \frac{p(\theta^* \mid y)/J(\theta^* \mid \theta)}{p(\theta \mid y)/J(\theta \mid \theta^*)}$$

3. Stay in place (do not accept the draw) with probability $1 - r$, i.e., $\theta^{(t+1)} = \theta^{(t)}$.

By construction, the algorithm does not depend on the normalization constant. The variation of the algorithm in which the proposal distribution is not symmetric is due to Hasting (1970) and for this reason the algorithm is often also called Metropolis-Hasting.

Using MCMC for a complex problem is not an automatic procedure and some tuning is needed. One of the important things to choose with care is the proposal function. If too small jumps are proposed, the chain moves too slowly, and can even remain trapped in a subregion and never sample the rest of the parameter space if the probability distribution is defined over disconnected regions. If too large steps are proposed, the proposed points could often fall in very low probability regions and not be accepted, in which case the chain remains stuck in a point for many cycles.

APPENDIX C

COMPUTATION ALGORITHM FOR RANDOM EFFECT LOGISTIC MODEL

After assigning an initial value to each parameter, our MCMC algorithm for random effect logisitic model proceeds as follows:

(1) Sample β using the Metropolis algorithm from the full conditional distribution

(C.1)
$$f(\beta|\cdot) \propto \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{exp\{y_i(\beta_0 + \beta_1 x_i + \beta_2' z_i + b_i)\}}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2' z_i + b_i)} exp(-\frac{1}{2}\beta'\Sigma_\beta^{-1}\beta)$$

(2) Sample b_i using the Metropolis algorithm Suppose the conditional moment $\mu_{ij} = E(y_{ij}|b_i)$ and $v_{ij} = var(y_{ij}|b_i)$ satisfy

$$h(\mu_{ij}) = \eta_{ij} = x_{ij}'\beta + z_{ij}'b_i$$

where h is the logit function.

If we let $y_i^\star$ be the linear approximation to $h(y_i)$ given by

(C.2)
$$y_i^\star = \eta_i + (\frac{\partial \mu_i}{\partial \eta_i})^{-1}(y_i - \mu_i)$$

then the maximum value of $p(b_i)$ occurs at $\hat{b}_i = (Z_i'V_iZ_i + D^{-1})^{-1}Z_iV_i(y_i^\star - X_i\beta)$, and its curvature is $\hat{v}_i = (Z_i'V_iZ_i + D^{-1})^{-1}$. Because $y_i^\star$ and v_i depend on b_i, the actual mode and curvature must be obtained by iterating the equations for $\hat{b}_i$ and $\hat{v}_i$.

(3) Sample σ_b^2 from Inverse Gamma distribution,

(C.3)
$$f(\sigma_b^2|\cdot) \propto IG((\frac{I+1}{2}, \frac{\sum_{i=1}^{I} b_i^2 + 1}{2}))$$

(4) Sample x_{ij} from the Bernoulli distribution

(C.4)
$$f(x_i|\cdot) \sim Bin(1, \frac{w'}{w' + v'})$$

where

$$w' = \frac{exp(y_i\beta_1 + \gamma_0 + \gamma_1'z_i)}{1 + exp(\beta_0 + \beta_1 + \beta_2'z_i + b_i)} p_1^{x_{1i}}(1 - p_1)^{1-x_{1i}} p_2^{x_{2i}}(1 - p_2)^{1-x_{2i}}$$

$$v' = \frac{1}{1 + exp(\beta_0 + \beta_2'z_i + b_i)} q_1^{1-x_{1i}}(1 - q_1)^{x_{1i}} q_2^{1-x_{2i}}(1 - q_2)^{x_{2i}}$$

(5) Sample γ using the Metropolis algorithm

(C.5)
$$f(\gamma|\cdot) \propto \prod_{i=1}^{n} \frac{exp\{x_i(\gamma_0 + \gamma_1'z_i)\}}{1 + exp(\gamma_0 + \gamma_1'z_i)} exp(-\frac{1}{2}\gamma'\Sigma_\gamma^{-1}\gamma)$$

(6) Sample p_1 from its full conditional Beta distribution

(C.6)
$$f(p_1|\cdot) \propto Beta\{1 + \sum_{i=1}^{n}(x_{1i} = 1 \,\&\, x_i = 1), 1 + \sum_{i=1}^{n}(x_{1i} = 0 \,\&\, x_i = 1)\}$$

(7) Sample p_2 from its full conditional Beta distribution

(C.7)
$$f(p_2|\cdot) \propto Beta\{1 + \sum_{i=1}^{n}(x_{2i} = 1 \,\&\, x_i = 1), 1 + \sum_{i=1}^{n}(x_{2i} = 0 \,\&\, x_i = 1)\}$$

(8) Sample q_1 from its full conditional Beta distribution

(C.8)
$$f(q_1|\cdot) \propto Beta\{1 + \sum_{i=1}^{n}(x_{1i} = 0 \,\&\, x_i = 0), 1 + \sum_{i=1}^{n}(x_{1i} = 1 \,\&\, x_i = 0)\}$$

(9) Sample q_2 from its full conditional Beta distribution

(C.9)
$$f(q_2|\cdot) \propto Beta\{1 + \sum_{i=1}^{n}(x_{2i} = 0 \,\&\, x_i = 0), 1 + \sum_{i=1}^{n}(x_{2i} = 1 \,\&\, x_i = 0)\}$$

(10) Repeat steps 1-9 until apparent convergence and collect a large number of additional draws from which to calculate posterior summaries.

BIBLIOGRAPHY

[1] B. Armstrong. Measurement error in generalized linear models. *Communications in Statistics, part B-Simulation and Computation*, 14:529–544, 1985.

[2] D. Aujesky, R. A. Stone, D. S. Obrosky, D. M. Yealy, T. E. Auble, T. P. Meehan, L. G. Graff, J. M. Fine, and M. J. Fine. Using randomized controlled trial data, the agreement between retrospectively and prospectively collected data comprising the pneumonia severity index was substantial. *Journal of Clinical Epidemiology*, 58:357–363, 2005.

[3] B. A. Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, 33:414–418, 1977.

[4] J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society B*, 61:265–285, 1999.

[5] H. Brenner. How independent are mutiple 'independent' diagnostic classification? *Statistics in Medicine*, 15:1377–1386, 1996.

[6] N. E. Breslow and R. Holubkov. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine*, 16:103–116, 1997.

[7] B. P. Carlin and A. E. Gelfand. An iterative monte carlo method for nonconjugate bayesian analysis. *Statistics and Computing*, 1:119–128, 1991.

[8] R. J. Carrol, M. H. Gail, and J. H. Lubin. Case-control studies with error in covariates. *Journal of the American Statistical Association*, 88:185–199, 1993.

[9] R. J. Carrol, H. Kuchenhoff, F. Lombard, and L. A. Stefansi. Asymptotics for the simex estimator in structural measurement error models. *Journal of the American Statistical Association*, 91:242–250, 1996.

[10] R. J. Carroll, D. Ruppert, and L. A. Stefanski. In *Measurement errors in nonlinear models*. Chapman and Hall, New York, 1995.

[11] R. J. Carroll and L. A. Stefanski. Approximate quasilikehood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85:652–663, 1990.

[12] T. T. Chen. A review of methods for misclassified categorical data and epidemiology. *Statistics in Medicine*, 8:1095–1106, 1989.

[13] J. Cook and L. A. Stefanski. A simulation estrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328, 1995.

[14] K. T. Copeland, H. Checkoway, A. J. McMichael, and R. H. Holbrook. Bias due to misclassification in the estimation of relative risks. *American Journal of Epidemiology*, 105:488–495, 1977.

[15] M. K. Cowles and B.P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904, 1996.

[16] P. Dellaportas and D. A. Stephens. Bayesian analysis of errors-in-variables regression models. *Biometrics*, 51:1085–1095, 1995.

[17] Laird N. Dempster, A. and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[18] C. D. Drews, W. D. Flanders, and A. S. Kosinski. Use of two data sources to estimate odds ratios in case-control studies. *Epidemiology*, 4:327–335, 1993.

[19] B. Efron and R. Tibshirani. In *Introduction to the Bootstraps*. Chapman and Hall, New York, 1993.

[20] M. J. Fine, Auble T. E., and D. M. et al Yearly. A prediction rule to identify low-risk patients with community-acquired pneumonia. *The New England Journal of Medicine*, 336:243–250, 1997.

[21] K. M. Flegal, C. Brownie, and J. D. Haas. The effects of exposure misclassification on estimate of relative risk. *American Journal of Epidemiology*, 123:736–751, 1986.

[22] D.G. Fryback. Bayes theorem and conditional nonindependence of data in medical diagnosis. *Biomedical Research*, 11:423–434, 1978.

[23] A. E. Gelfand, S. E. Hills, A. Racing-Poon, and A. F. M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85:972–985, 1990.

[24] A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[25] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

[26] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Institute of Electrical and Electronic Engineers Transactions in Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[27] J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, 4:169–193, 1992.

[28] W. R. Gilks, D. G. Clayton, D. J. Spiegehalter, N. G. Best, A. J. McNeil, L. D. Sharples, and J. Kirby. Modelling complexity: applications of gibbs sampling in medicine. *Journal of Royal Statistical Society, Series B*, 55:39–52, 1993.

[29] W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 41:337–348, 1992.

[30] L. J. Gleser. In P. J. Brown and W.A. Fuller, editors, *Statistical Analysis of Measurement Error Models and Application*. American Mathmatical Society, Providence, 1990.

[31] S. Greenland. The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, 112:564–569, 1980.

[32] S. Greenland. The effects of misclassification in matched-pair case-control studies. *American Journal of Epidemiology*, 116:402–406, 1982.

[33] S. Greenland and D.G. Kleinbaum. Correcting for misclassification in two way tables and matched-pair studies. *International Journal of Epidemiology*, 12:93–97, 1983.

[34] W. H. Gullen, J. E. Berman, and E. A. Johnson. Effects of misclassification in epidemiologic studies. *Public Health Reports*, 53:1956–1965, 1968.

[35] W. K. Hastings. Monte-carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[36] J. K. Haukka. Correction for covariate measurement error in generalized linear model-a bootstrap approach. *Biometrics*, 51:1127–1132, 1995.

[37] P. Heidelberger and P. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31:1109–1114, 1983.

[38] S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171, 1980.

[39] H. Ko and M. Davidian. Correcting for measurement error in individual-level covariates in nonlinear mixed effects models. *Biometrics*, 56:368–375, 2000.

[40] A. S. Kosinski and W. D. Flanders. Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: a regression approach. *Statistics in Medicine*, 18:2795–2808, 1999.

[41] P. Kristensen. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology*, 3:210–215, 1992.

[42] J. Kuha. Corrections for esposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine*, 13:1135–1148, 1994.

[43] J. Kuha. Estimation by data augmentation in regression models with continuous and discrete covariates measured with error. *Statistics in Medicine*, 16:189–201, 1997.

[44] R. J. A. Little and D. B. Rubin. In *Statistical analysis with missing data*. John Wiley and Sons, New York, 2002.

[45] T. A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B*, 44:226–233, 1982.

[46] B. K. Mallick and A. E. Gelfand. Semiparametric errors-in-varaibles models:a bayesian approach. *Journal of Statistical planning and Inference*, 52:307–321, 1996.

[47] P. Muller and K Roeder. A bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, 84:523–537, 1997.

[48] D. J. Newell. Errors in the interpretation of errors in epidemiology. *American Journal of Public Health*, 52:1925–1928, 1962.

[49] M. S. Pepe and T. R. Fleming. A nonparametric method for dealing with mismeasured data. *Journal of the American Statistical Association*, 86:108–113, 1991.

[50] A. L. Raftery and S. Lewis. In J.M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*. Oxford University Press, Oxford, 1992.

[51] M. Reilly and M. S. Pepe. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82:299–314, 1995.

[52] S. Richardson and W. R. Gilks. A bayesian approach to measurement error probalems in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138:430–442, 1993.

[53] S. Richardson and W. R. Gilks. Conditional independence models for epidemilogical studies with covariate measurement error. *Statistics in Medicine*, 12:1703–1722, 1993.

[54] S. Richardson and L. Leblond. Some comments on misspecification of priors in bayesian modelling of measured error problems. *Statistics in Medicine*, 16:203–213, 1997.

[55] S. Richardson, L. Leblond, I. Jaussent, and P. J. Green. Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society A*, 165:549–566, 2002.

[56] B. Rosner, W. C. Willett, and D. Spiegelman. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8:1051–1069, 1989.

[57] D. B. Rubin. In J.M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Multiple imputation for nonresponse in surveys*. wiley, New York, 1987.

[58] G. A. Satten and L. L. Kupper. Inferences about exposure-disease association using probability of exposure information. *Journal of the American Statistical Association*, 88:200–208, 1993.

[59] J. H. Sepanski, R. Knickerbocker, and R. J. Carroll. A semiparametric correction for attenuation. *Journal of the American Statistical Association*, 89:1366–1373, 1994.

[60] S. B. Sheps and M. T. Schechter. The assessment of diagnostic tests. *Journal of the American Medical Association*, 17:2418–2422, 1984.

[61] A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: a sampling-resampling perspective. *American Statistician*, 46:84–88, 1992.

[62] A. F. M. Smith and G. O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society, series B*, 55:3–23, 1993.

[63] B. J. Smith. In *Bayesian Output Analysis Program (BOA) Version 0.5.0 User Manual*. Department of Biosttistics, College of Public Health, University of Iowa, Ames, 2000.

[64] D. J. Spiegelhalter, A. Thomas, and N. G. Best. In *Win-BUGS, Version 1.2*. University of Cambridge MRC Biostatistics Unit, Cambridge, U.K, 1999.

[65] Rosner B. Spiegelman, D. and R. Logan. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95:51–61, 2000.

[66] L. A. Stefanski and J. Cook. Simulation extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90:1247–1256, 1995.

[67] M. A.. Tanner. In *Tools for Statistical Inference: Methods for Exploration of Posterior Distribtution and Likelihood Functions*. Springer-verlag, New York, 1996.

[68] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22:1701–1702, 1994.

[69] P. M. Vacek. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41:959–968, 1985.

[70] M. A. G. Viana. Bayesian small-sample estimation of misclassified multinomial data. *Biometrics*, 50:237–243, 1994.

[71] S. Wacholder, B. Armstrong, and P. Hartge. Validation studies using an alloyed gold standard. *American Journal of Epidemiology*, 137:1251–1258, 1993.

[72] A. M. Walker and M. Blettner. Comparing imperfect measures of exposure. *American Journal of Epidemiology*, 121:783–790, 1985.

[73] A. S. Whittmore and J. B. Keller. Approximations for regression with covariate measurement error. *Journal of the American Statistical Association*, 83:1057–1066, 1988.

[74] D. M. Yealy, T. E. Auble, R. A. Stone, J. R. Lave, T. P. Meehan, L. G. Graff, J. M. Fine, D. S. Obrosky, S. M. Edick, L. J. Hough, K. Tuozzo, and M. J Fine. The emergency department community-acquired pneumonia trial: Methodgy of a quality improvement intervention. *Annals of Emergency Medicine*, 43:770–781, 2004.

[75] S. Zeger and M. R. Karim. Generalized linear models with random effects: A gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86, 1991.