ABSTRACT

Bayesian Propensity Score Analysis for Clustered Observational Studies

Qi Zhou, Ph.D.

Chairperson: Joon Jin Song, Ph.D.

There is an increasing demand to investigate questions in observational studies. The propensity score is a popular confounding adjustment technique to ensure valid causal inference for observational studies. Observational data often have multilevel structure that would lead to one or more levels of confounding. Multilevel models are employed in Bayesian propensity score analysis to account for cluster and individual level confounding in the estimation of both the propensity score and in turn the exposure effect. In an extensive simulation study, several propensity score analaysis approches with varing degrees of complexity of multilevel modeling structures are examined in terms of average absolute bias and mean square error. The Bayesian propensity score analysis for multilevel data is further developed to accomodate misclassified binary responses. Errors in response can distort the exposure to response relationship. The true exposure-response surface can be recovered through two classification probabilities, the sensitivity and specificity. These link the observed misclassified response and the unobserved true response. Incorpating misclassification greatly reduces bias in exposure effect estimation and yields coverage rate of 95% credible sets close to the nomial level. Strong ignorability is the fundamental assumption

for propensity score. There is little literature that discusses this important but untestable assumption. Without the confidence that there are no unmeasured confounders, we assume the existence of unmeasured confounding and assess the sensitivity of exposure effect estimation to unmeasured confounding through two sensitivity parameters which characterize the associations of the unmeasured confounder with the exposure status and response variable. The influence of unmeasured confounding can be examined by possible change in exposure effect estimation with hypothetical values of sensitivity parameters.

Bayesian Propensity Score Analysis for Clustered Observational Studies

by

Qi Zhou, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

Joon Jin Song, Ph.D., Chairperson

_____

Jack D. Tubbs, Ph.D.

_____

James D. Stamey, Ph.D.

_____

David J. Kahle, Ph.D.

_____

Gerald B. Cleaver, Ph.D.

Accepted by the Graduate School
May 2018

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to thank my committee members, for their insightful suggestion and comments on my dissertation. My sincere thanks goes to my advisor, Dr. Song, for his support and guidance. He helps me develop my research and provides lots of great suggestions for my career and life.

Foremost, I would like to thank my parents, for bringing me to this world. Their love and unconditional support made me who I am now and will always inspire me to overcome any difficulties. I would also like thank my daughter, Karen, for so much joy and happiness she brings to me.

CHAPTER ONE

Introduction

In randomized controlled trials, the subjects are allocated to different groups by random chance. The distribution of confounders, measured and unmeasured, are theoretically balanced among groups. Therefore, the treatment (or exposure to distinguish 'treatment' effect in non-clinical setting) effect can be directly estimated by comparing the difference of treatment groups. However, a lot of data do not have the luxury of randomization, for instance, data collected from medical records and social science studies. Many questions of interest in observational studies face the difficulty that the direct comparison of treatment groups may not be reliable due to possible confounding among groups being compared. For observational studies, subjects have varying probabilities to go in any treatment group and therefore the confounding factors are not equally distributed among groups.

The propensity score is first proposed to adjust confounding when comparing groups. The propensity score is the probability of receiving a treatment status conditioned on the observed covariates. Conditional on the propensity score, the distribution of observed baseline covariates are similar between comparison groups. Subjects with the same propensity score have the same covariate distribution. Two quantities are usually inferred from the propensity score: average treatment effect (ATE) and average treatment effect on treated units (ATT). Traditionally, propensity score is estimated through logistic regression. There are four common methods to use the estimated propensity score for inference: matching, weighting, covariate adjustment using propensity score and stratification. In a frequentist framework, the estimated propensity score is assumed true to make inference thus the uncertainty in propensity score estimation is not integrated into the treatment effect estimation. Bayesian propensity score analysis combines propensity score estimation and treatment effect inference into a single framework. The marginal distribution of the treat-

ment effects is obtained by integrating out parameters in the propensity score model and incorporating uncertainty in propensity score estimation in treatment effect estimation.

Observational data are often collected in the setting with a cluster structure. For examples: patients are naturally clustered by health system, further by hospital within health system and by provider within hospital; Subjects are clustered by state and by household within state. The cluster structure confounds the inference when cluster level confounding is related to treatment assignment or potential outcome. The propensity score analysis is originally developed for unclustered structure. Multilevel modeling is a well developed tool to account for a cluster structure in the data. Employing multilevel modeling in propensity score analysis adjusts cluster level confounding and separates variation among clusters from random error and between-subject variance.

In epidemiology and survey data, outcome or response variables are often subject to misclassification. Misclassification is typically referred as error in categorical variables. Including variables contaminated with misclassification in the model can possibly bias estimation or reduce the efficiency of the estimates. The classical setting is non-differential misclassification: the misclassification error in one variable is independent of the other variables. Usually, the misclassified variable is observed and the true variables is not observable. For a dichotomous variable, the true variable is linked to the misclassified variable with false positive and false negative probabilities. Bayesian binomial regression can facilitate the modeling of misclassified binary response. Relatively informative priors are taken for misclassification parameters, employing the available information about misclassification parameters and incorporating uncertainty about those parameters in the covariate effect estimates.

The fundamental assumption for causal inference based on propensity score techniques is the strong ignorable assumption: overlap and unconfoundness. The overlap assumption states that the propensity score of all units must be between 0 and 1 and there is overlap of the propensity score for treated units and control units, ensuring every treated

unit has at lease one matching control unit. This assumption can be checked by histograms of propensity among the comparative groups. If there is no or little overlap of the propensity score of treated units in comparative groups, valid causal inference can not be made without strong extrapolation. The unconfoundess assumption assumes treatment assignment is independent of the outcomes condition on the observed covariates. Researchers are rarely confident of satisfaction of this assumption and usually conduct sensitivity analysis for unmeasured confounding. Sensitivity analysis assesses the possible change in parameter estimates assuming the existence of an unmeasured confounder. Two sensitivity parameters control how the unmeasured confounder enters the inferential models to characterize the associations of the unmeasured confounder with the treatment and the outcomes. The sensitivity parameters are unknown and there are two ways to deal with sensitivity parameters in the Bayesian framework. First, assume the sensitivity parameters are unknown and informative priors are assigned to avoid non-identifiability. Informative priors of sensitivity parameters incorporate the expert opinion about unmeasured confounding and provide a range of estimates in the presence of unmeasured confounding. This may bring in bias due to mis-specification of priors. Second, sensitivity parameters are taken as fixed. With a set of hypothetical values of sensitivity parameters, sensitivity analysis produces a table of estimates. The non-identifiability and mis-specification issue are avoided.

This dissertation consists of three projects. The first project is the study of Bayesian propensity score analysis for clustered observational data. The method is applied to investigate the effect of lipid screening on controlling LDL-C level in youth. Medical records of patients are extracted, including demographic information, vitals and test results, etc. Patients are clustered by health systems. Employing a multilevel model in propensity score analysis to adjust for confounding at the individual level and cluster level in both propensity score estimation and the treatment effect. Through extensive simulations, several propensity score analysis approaches are compared with varying complexity of multilevel modeling structures. Fixed effect propensity score model and random intercept outcome model

turn out to be the best combination with smallest average absolute bias and mean square error among all candidate combinations.

The second project is motivated by the study of the impact of female employment on the physical spousal violence towards women for India National Health Survey 2005-6. The females in the study are clustered by state and females are more similar in demographic characteristics and social environment within same state. Some women may not answer yes to the survey question of physical violence even when they truly suffer physical violence. The response variable is subject to misclassification. Bayesian multilevel propensity score analysis is extended to accommodate misclassification in the response. The covariate to misclassified response regression surface is adjusted with sensitivity and specificity to recover the true covariate to response surface. Relatively informative priors for sensitivity and specificity are extracted from external information. In the simulation, four scenarios of propensity score analysis are compared: ignoring misclassification and multilevel structure, incorporating misclassification only, incorporating multilevel structure only and incorporating both misclassification and multilevel structure. The scenario taking into account misclassification and multilevel structure yields smallest average absolute bias and closest coverage rate of credible intervals to the nominal level. It is found that correcting misclassification in response is more crucial to reduce bias and improve coverage rate than incorporating a multilevel structure. The contributors of this paper are: Dr. Joon Jin Song, Dr. James D. Stamey and Dr. Yoo-Mi Chin and Qi Zhou. Dr. Song and Qi Zhou did the literature review and worked on the methodology and simulation studies in this paper. Dr. Stamey provided prior information on the sensitivity parameters. Dr. Chin worked on data description and helped interpret empirical results.

The third project proposes Bayesian sensitivity analysis of unmeasured confounding for observational data with misclassified responses. The approach corrects bias from error in response and examines possible change in exposure effect estimation if a binary unmeasured confounder exists. We assess the influence of unmeasured confounding on

exposure effect estimation through two sensitivity parameters that enter inferential models as regression coefficients and characterize the associations of the unmeasured confounder with the exposure status and with the response variable. In the analysis, exposure effect estimators are produced from a range of hypothetical values of sensitivity parameters. The proposed approach is illustrated in the study of the effect of female employment status on the likelihood of domestic violence. An extensive simulation study is conducted to confirm the efficacy of the proposed approach. The simulation results indicate accounting for misclassification in response and unmeasured confounding significantly reduces the bias in exposure effect estimation.

CHAPTER TWO

Bayesian Propensity Score Analysis for Multilevel Observational Data

2.1 Introduction

Randomized controlled trials, when properly executed, support causal inference since random assignment of study subjects to comparison groups is free of confounding. Since some investigations necessarily require passive observation and nonrandomized group assignment, identified statistical associations may be confounded with distributions of confounders due to lack of randomization.

The method called propensity score, first proposed by Rosenbaum and Rubin (1983b), is a popular confounding adjustment technique for approaching causal inference in observational studies. The propensity score is the conditional probability of obtaining treatment assignment based on observed covariates, typically estimated through logistic regression. Then the estimated score, which should now account for the differences in measured confounder distributions among the compared groups, is used in the outcome model for treatment effect estimation. Since the estimated propensity score is treated as true in the outcome analysis, the uncertainty in propensity score estimation is ignored, and therefore variance in treatment effect estimates is underestimated. In the multilevel setting, ignoring uncertainty in propensity score estimation would confound with random effects and measurement error. The most common propensity score methods include using propensity score as a covariate, inverse probability weighting, matching and stratification. Propensity score stratification retains all the subjects in the study and sufficiently takes advantage of information in the data (Elze et al., 2017).

Several methods have been proposed to adjust the variance of treatment effect estimation. Abadie and Imbens (2008, 2009) criticized standard bootstrapping for not providing valid inference for standard errors, and instead proposed standard error estimators

of treatment effect for propensity score matching. The weakness of this method is that the variance estimate can be adjusted downward to be in a negative range. To avoid this issue of variance adjustment as proposed by Abadie and Imbens (2009), McCandless et al. (2009) introduced a Bayesian propensity score solution using stratification to handle uncertainty in the propensity score estimation. Bayesian propensity score analysis treats parameters in the propensity score as nuisance parameters then incorporated uncertainty in estimating propensity scores into treatment effect estimation in the outcome model by integrating out the nuisance parameters. An (2010) proposed Bayesian propensity score matching and covariate adjustment and an intermediate Bayesian propensity score estimator. An's work showed that a Bayesian approach provided improved variance estimates of the treatment effect. When the propensity score and the outcome model are estimated at same time in the Bayesian framework, a feedback occurs: updated parameters in the outcome model affect the update of parameters in the propensity score model. To avoid feedback of the outcome model onto the propensity score model, Kaplan and Chen (2012) proposed a two-step Bayesian propensity score analysis where the posterior samples of the estimated propensity scores are used for a Bayesian outcome counterpart by matching, weighting and stratification. The two-step Bayesian approach may provide overly high coverage rates if informative priors are used.

Data collected in many studies are clustered naturally, for example, by health systems or practice groups. Propensity scores have not been well studied in the setting of clustered data. If cluster-level confounding is associated with treatment assignment and goes unmeasured, then the assumption of strong ignorability is violated and treatment effect estimation might be biased. Multilevel models have been developed for clustered data to deal with hierarchical structures. Specifically for time-series cross-sectional data, employing multilevel modeling in both steps of propensity score matching results in less bias and yields more efficient estimates (Su and Cortina, 2009). Through intensive Monte Carlo simulations, Arpino and Mealli (2011) noted the necessity of addressing the cluster structure in

propensity score matching. Li et al. (2013) evaluated the performance of different propensity score weighting estimators and concluded that considering the cluster structure in one stage of propensity score weighting results in less bias, and the treatment effect estimate is more impacted by the choice of outcome model than by the propensity score model.

The objective of this paper is to propose a Bayesian propensity score analysis for multilevel, clustered observational data in order to improve treatment effect estimation by accounting for cluster and individual level confounding. Besides, the Bayesian approach naturally accounts for uncertainty in propensity score estimation in treatment effect estimates and distinguishes the uncertainty from random effects and measurement error. To highlight the performance of multilevel modeling applied to Bayesian propensity score stratification, we examine a scenario of lipid screening on managing LDL-C levels among all youth within three large integrated healthcare systems. We evaluated different multilevel models in both propensity score and outcome stages in the simulation study and applied these models to the case study. Section 2 introduces the motivating case study of lipid management in youth. In Section 3, a general modeling for propensity score and treatment effect estimation is proposed with related, simplified candidate models used in the case study and simulations. Section 4 presents simulation results. The analysis of the application is discussed in Section 5. The chapter concludes with discussion in Section 6.

2.2 Motivating Case Study

Premature adult atherosclerotic cardiovascular disease has been shown to begin as early as childhood, especially for youth with markedly elevated total and low density lipoprotein cholesterol (LDL-C) associated with an inherited condition called familial hypercholesterolemia. Lipid management requires blood cholesterol testing, provider recognition (diagnosis) of abnormal values, and proper treatment of the lipid disorders. Primary determinants of these decision points may occur at the child-parent level, provider level, geographic level, demographic level, or health system level and may be influenced by a number of known factors(child, parent, provider preferences and biases, insurance status

including copays, distance from a clinic and/or laboratory setting, etc.) and unknown variables. Our methodology is motivated by an observational study investigating lipid management in youth aged from 2 to 20 years from three health systems located in central Texas, north-central Pennsylvania and the Detroit metropolitan area in southeastern Michigan. From the total cohort of size 1,211,556, we obtained data from 2,349 youth who had an LDL-C test result or multiple results between 2001 and 2012. The data is naturally clustered by health systems and within each health system clustered at the clinic level and within the clinic, at the provider level although there can be and often is flux between clinics and especially between providers over time giving rise to very heterogeneous encounters. We consider the cluster at the health system level in the modeling. The study cohort is divided into two groups depending on whether they had a lipid screening any time between 2001 to 2011 – a tested (treatment) group and a control group if no lipid test was performed. For each patient, demographics information, history of diabetes, hypertension and dyslipidemia diagnostics, prescription medication claims and laboratory tests are retrieved. The number of outpatients in 2012 was obtained as a cluster level covariate.

Table 2.1 summarizes the characteristics of each group. Crudely, average LDL-C level in 2012 in the tested group is 8.1 mg/dL higher than the control group. The average age, average percentage of females and average percentage of whites are similar in the two groups. The average body mass index BMI of the tested group is higher, possibly because they are older (14 vs 13 years). Systolic and diastolic blood pressures in the test group are on average higher than those of the control group. The percentage of smoking exposure is similar in both groups. There are higher percentages of youth diagnosed with diabetes, hypertension or dyslipidemia (i.e., any lipid disorder not just familial hypercholesterolemia) in the tested group. In addition, relatively more youth are prescribed diabetes or lipid medication in the tested group. Testing occurred in about 50% of youth who came to a health care provider for a well-child visit as opposed to a visit for an acute illness and/or follow-up appointment. We employ Bayesian propensity score analysis to adjust individual level and

health system level observed pretreatment characteristics and to account for uncertainty in treatment effect estimation. The goal is to infer reliable effect of prior lipid screening in 2001 to 2011 on the LDL-C level in 2012 in the cohort.

Table 2.1: Descriptive statistics of lipid management data.

Variables	Lipid screening		No Lipid screening	
	Mean	SD	Mean	SD
a) Outcome				
LDL-C level in 2012	93.38	31.38	85.28	27.42
b) Confounding characteristics				
b1) Demographics				
Age	14	2.35	13	2.48
Gender (Female)	46.6 %	0.50	48.1 %	0.50
Race (White)	63.1 %	0.48	61.8%	0.49
b2) Vitals				
BMI	29.27	8.50	25.19	6.83
Systolic	116.40	13.63	111.46	13.22
Diastolic	68.77	9.12	66.41	8.65
b3) Diagnostics				
Diabetes	5.1%	0.22	1.2%	0.11
Hypertension	1.7%	0.13	1.1%	0.11
Dyslipidemia	5.3%	0.22	1.6%	0.12
b4) Medication prescription				
Lipid Medication	1.2%	0.11	0.4%	0.06
Diabetes Medication	4.9%	0.22	1.4%	0.12
b5) Other				
Well Child Visit	50.9%	0.5	52.5%	0.5
Smoke	0.7%	0.09	0.8%	0.09
b6) Site level variable				
Outpatient numbers	2.66	0.73	2.62	0.72
No. of observations	946		1404	

2.3 Methodology

2.3.1 General Modeling

Consider data consisting of ith subject nested within jth cluster ($i = 1, ...n_j, j = 1, ...J$). Let T_{ij} denote the dichotomous treatment assignment taking value 1 if subject is in the treatment group or 0 otherwise, and let Y_{ij} denote the corresponding outcome vari-

able. Let $\mathbf{X}_{ij}$ and $\mathbf{C}_j$ be subject-level and cluster-level measured covariates. To estimate treatment effects from observational data, we assume strong ignorable assumptions are met: first, unconfoundedness assumes no unmeasured confounders exist and, conditional on measured covariates, treatment assignment is independent of potential outcome; second, overlap assumes there is common support of propensity scores between treatment group and control group in the study population. Thus, each treatment unit would have at least one control counterpart. Last, let $z_{ij} = P(T_{ij} = 1|\mathbf{X}_{ij}, \mathbf{C}_j)$ denote the propensity score, which is usually estimated by a logistic model. The estimated propensity score can be used for matching, stratification and covariate adjustment in the outcome model. We focus on the case with binary treatment and a continuous outcome, such as LDL level. To facilitate modeling in a Bayesian framework, regression modeling is used in propensity score and outcome estimation, and the estimated propensity score is used to create subclasses, which allow the flexibility to estimate stratum-specific and overall treatment effects. According to Cochran (1968) and Rosenbaum and Rubin (1984), creating five subclasses based on estimated propensity scores can remove 90% of the bias due to measured confounding.

In this study, we propose a general framework for Bayesian propensity score analysis for multilevel data using random effects:

$$\text{logit}(z_{ij}) = \gamma_{0j} + (\gamma_1 + \mathbf{u}_j)^T \mathbf{X}_{ij} + {\gamma_2}^T \mathbf{C}_j,$$

$$y_{ij} = \beta_{0j} + (\beta_1 + v_j) T_{ij} + \xi^T g(z_{ij}),$$

where $g(z_{ij})$ is a function of estimated propensity scores z_{ij}. Treatment effect β_1 is of primary interest and assumed to be fixed. Random intercept $\gamma_{0j} \sim N(\mu_\gamma, \sigma_\gamma^2)$ absorbs heterogeneity of propensity scores among clusters. If a specific subject-level covariate is known to have varying influence on propensity scores across clusters, the random component $\mathbf{u}_j \sim N(0, \sigma_u^2)$ allows cluster-specific effects of subject-level covariates on propensity scores. In case cluster-specific treatment effect is of interest in the study, we can introduce the random effect $v_j \sim N(0, \sigma_v^2)$ in our outcome model. To implement the approach using stratification on estimated propensity scores, $g(z_{ij})$ is specified as a vector of stratum

membership indicators classified as:

$$g(z_{ij})^T = \begin{cases} (1,0,0,0) & \text{if } 0 < z_{ij} < q_1 \\ (0,1,0,0) & \text{if } q_1 < z_{ij} < q_2 \\ (0,0,1,0) & \text{if } q_2 < z_{ij} < q_3 \\ (0,0,0,1) & \text{if } q_3 < z_{ij} < q_4 \end{cases}$$

where (q_1, q_2, q_3, q_4) are the predetermined knots, which are typically based on quintiles of propensity scores by maximum likelihood estimation. The proposed framework enables us to propose a class of Bayesian propensity score analyses for multilevel observational data by varying multilevel modelling complexity.

2.3.2 Models for Propensity Score

In this paper, we consider four models for propensity score estimation. In the presence of cluster level confounding, propensity score modeling needs to suitably control it using cluster structure information. The first model is the simplest one including only subject-level information by fixed effects,

$$\text{PS1:logit}(z_{ij}) = \gamma_0 + {\gamma_1}^T \mathbf{X}_{ij}.$$

This model is inappropriate if a cluster-level covariate is associated with treatment assignment. Second, cluster-level information is incorporated into propensity score modeling by adding cluster-level covariates:

$$\text{PS2:logit}(z_{ij}) = \gamma_0 + {\gamma_1}^T \mathbf{X}_{ij} + {\gamma^T}_2 \mathbf{C}_j.$$

Two alternative models employ a multilevel structure by introducing random effects. A random intercept is used to control for heterogeneity of propensity scores among clusters and assumes a normal distribution, $N(0, \sigma^2_\gamma)$,

$$\text{PS3:logit}(z_{ij}) = \gamma_{0j} + {\gamma_1}^T \mathbf{X}_{ij}. \tag{2.1}$$

The last model contains both random intercept and fixed effects for cluster-level covariates in order to remove cluster-level confounding,

$$\text{PS4:logit}(z_{ij}) = \gamma_{0j} + {\gamma_1}^T \mathbf{X}_{ij} + \gamma_2 C_j.$$

2.3.3 Models for Outcome Estimation

The basic outcome model regresses the outcome variable on the treatment variable and indicators of subclass membership based on the estimated propensity scores ($\hat{z}_{ij}$),

$$\text{Out 1:}y_{ij} = \beta_0 + \delta T_{ij} + {\beta_1}^{\mathrm{T}} g(\hat{z}_{ij}).$$

A random effect outcome model employs the random intercepts to take into account the association of outcomes among clusters, which assume a normal distribution, $\beta_{0j} \sim N(0, \sigma_\beta^2)$,

$$\text{Out 2:}y_{ij} = \beta_{0j} + \delta T_{ij} + {\beta_\mathbf{1}}^{\mathrm{T}} g(\hat{z}_{ij}).$$

2.3.4 Posterior Inference

To complete the Bayesian model, we need to elicit prior distributions for the parameters in the propensity score and outcome models. Fixed parameters $\beta_1, \delta, \gamma_1$ and γ_2 are given independent normal priors. For hyperparameters, we consider inverse gamma priors: $\sigma_\gamma^2 \sim IG(a_\gamma, b_\gamma)$ and $\sigma_\beta^2 \sim IG(a_\beta, b_\beta)$, where IG refers to an inverse gamma distribution. We obtain the joint posterior density $p(\gamma, \beta, \delta|\text{data})$ by multiplying likelihood and priors,

$$p(\gamma, \beta, \delta|\text{data}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(Y_{ij}|\mathbf{X}_{ij}, C_{ij}, T_{ij}, \beta, \delta, \gamma) p(T_{ij}|X_{ij}, C_{ij}, \gamma) p(\beta) p(\delta) p(\gamma).$$

Since the posterior distributions of interest do not possess closed form, posterior samples are obtained via Markov chain Monte Carlo (MCMC) simulation. Full conditional distributions are updated successively using the Metropolis-Hastings algorithm from conditional densities $p(\gamma|\mathbf{Y}, \mathbf{T}, \mathbf{X}, \mathbf{C}, \beta, \delta)$ and $p(\beta, \delta|\mathbf{Y}, \mathbf{T}, \mathbf{X}, \mathbf{C}, \gamma)$. The conditional density of pa-

rameters in the propensity score model is given by

$$
\begin{aligned}
p(\gamma|\mathbf{Y},\mathbf{T},\mathbf{X},\mathbf{C},\beta,\delta) &= \prod_{j=1}^{J}\prod_{i=1}^{n_j} p(Y_{ij}|\mathbf{X}_{ij},C_{ij},T_{ij},\beta,\delta,\gamma)p(T_{ij}|X_{ij},C_{ij},\gamma)p(\gamma) \\
&= \prod_{j=1}^{J}\prod_{i=1}^{n_j}\left[\frac{\exp(Y_{ij}(\beta_{0j}+\delta T_{ij}+{\beta_1}^T g(z({\gamma_{0j}}+{\gamma_1}^T\mathbf{X}_{ij}+\gamma_2 C_{ij}))))}{1+\exp(Y_{ij}(\beta_{0j}+\delta T_{ij}+{\beta_1}^T g(z({\gamma_{0j}}+{\gamma_1}^T\mathbf{X}_{ij}+\gamma_2 C_{ij}))))}\right. \\
&\quad \left.\times\frac{\exp({\gamma_{0j}}+{\gamma_1}^T\mathbf{X}_{ij}+\gamma_2 C_{ij})}{1+\exp({\gamma_{0j}}+{\gamma_1}^T\mathbf{X}_{ij}+\gamma_2 C_{ij})}\right] \times p(\gamma_{0j})p(\gamma_1)p(\gamma_2),
\end{aligned}
$$

where $z({\gamma_{0j}}+{\gamma_1}^T\mathbf{X}_{ij}+\gamma_2 C_{ij})$ is the estimated propensity score. Updating parameters in the propensity score model is related to the updated parameters in the outcome model. The estimation of parameters in the outcome model assists estimation of parameters in the propensity score model. To update parameters in the outcome model, the full conditional is given by

$$
\begin{aligned}
p(\beta,\delta|\gamma,\mathbf{Y},\mathbf{T},\mathbf{X},\mathbf{C}) &= \prod_{j=1}^{J}\prod_{i=1}^{n_j} p(Y_{ij}|\mathbf{X}_{ij},C_{ij},T_{ij},\beta,\delta,\gamma)p(\beta)p(\delta) \\
&= \prod_{j=1}^{J}\frac{\exp(Y_{ij}(\beta_{0j}+\delta T_{ij}+{\beta_1}^T g(z({\gamma_{0j}}+{\gamma_1}^T\mathbf{X}_{ij}+\gamma_2 C_{ij}))))}{1+\exp(Y_{ij}(\beta_{0j}+\delta T_{ij}+{\beta_1}^T g(z({\gamma_{0j}}+{\gamma_1}^T\mathbf{X}_{ij}+\gamma_2 C_{ij}))))} \\
&\quad \times p(\beta_{0j})p(\beta_1)p(\delta).
\end{aligned}
$$

The marginal distribution of δ incorporates uncertainty in propensity score estimation by averaging over the uncertainty in posterior samples of γ.

2.4 Simulation

In the simulation, we evaluated the performance of the proposed models, with simulated data for different sample sizes and strengths of association of the cluster-level covariate with treatment assignment and outcome.

2.4.1 Simulation Design

In the simulation, we consider the case where there are two subject-level covariates (X_1 and X_2) and one cluster-level covariate (C). 100 data sets are generated for six simulation designs shown in Table 2.2. Data are generated using the following algorithm:

(1) Generate $X_1, X_2,$ and C independently from $N(0,1)$.

(2) For fixed values γ_1, γ_2 and γ_3 , treatment T_{ij} is simulated from a Bernoulli distribution with probability z_{ij} generated from the following propensity score model:

$$\text{logit}(z_{ij}) = \gamma_{0j} + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + \gamma_3 C_j,$$

where $\gamma_{0j} \sim N(0, 0.5)$.

(3) Generate outcome from a random intercept model with simulated T and random intercept $\beta_{0j} \sim N(0,2)$,

$$y_{ij} = \beta_{0j} + \delta T_{ij} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 C_j + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0,1).$$

Table 2.2 summarizes the fixed parameters used to generate the synthetic data. Three scenarios are considered: moderate number of clusters and moderate cluster size, which is commonly seen in studies with clustered structures (Scenario A); small number of clusters and large cluster size as in the case study (Scenario B); large number of clusters but small sample size within each cluster (Scenario C). For each scenario, we consider strong and weak associations of cluster-level covariate C with treatment assignment T and outcome Y.

Table 2.2: True parameters used for generating synthetic data for simulation designs.

Design	Number of Cluster	Cluster Size	δ	β	γ
A1	20	25	1	(1,1,5)	(0.5,0.5,1)
A2	20	25	1	(1,1,1)	(0.5,0.5,0.2)
B1	5	100	1	(1,1,5)	(0.5,0.5,1)
B2	5	100	1	(1,1,1)	(0.5,0.5,0.2)
C1	100	10	1	(1,1,5)	(0.5,0.5,1)
C2	100	10	1	(1,1,1)	(0.5,0.5,0.2)

2.4.2 Simulation Result

For each synthetic data set, eight models combining four propensity score models with two outcome models are fitted. Relatively non-informative priors are placed on parameters: N(0,10) for coefficients and IG(0.001,0.001) for variance component of random

intercept. In posterior inference using MCMC, two chains are run for 10,000 iterations with a burn-in of 5,000 samples discarded to ensure the data used are minimally influenced by the initial values chosen, whenever initial values are over-dispersed. Convergence is checked if the Gelman and Rubin R-hat statistics reach 1 and effective sample size exceeds 3,000 (Gelman et al., 2014). To compare the models, we assess the point estimate of treatment effect δ with average absolute bias (bias) and mean square error (mse). Tables 2.3-2.5 summarize simulation results of all models for each scenario. The rows and columns correspond to results of four propensity score models and two outcome models, respectively.

Table 2.3: Average absolute bias (bias) and mean square error (mse) of treatment effect estimator δ in simulations with strong and weak association between cluster-level covariate and treatment effect and outcome, under 20 clusters and cluster size 25.

Design	Propensity score model	Out1		Out2	
		Bias	mse	Bias	mse
A1	PS1	3.7495	16.1842	0.1086	0.0153
	PS2	0.6663	0.5981	0.0921	0.0102
	PS3	0.8543	0.8654	0.1264	0.0189
	PS4	0.8335	0.8021	0.1259	0.0178
A2	PS1	0.2855	0.1383	0.0955	0.0111
	PS2	0.2214	0.0821	0.0926	0.0099
	PS3	0.2717	0.1079	0.1199	0.0171
	PS4	0.2719	0.1034	0.1172	0.0162

Ignoring cluster structure in both propensity score model and outcome model (combination of PS1 and Out1) leads to greater average absolute bias and mean square error of treatment effect estimation than seen in other models which take cluster structure into account in at least one stage of propensity score analysis. Bias is especially problematic when a cluster-level covariate is strongly associated with treatment assignment and outcome (scenario A1, B1, and C1). In the scenarios where the cluster-level covariate is weakly associated with treatment assignment and outcome (A2, B2, and C2), the reduction of bias by cluster-level adjustment is much less evident. Generally, considering clustering

Table 2.4: Average absolute bias (bias) and mean square error (mse) of treatment effect estimator δ in simulations with strong and weak association between cluster-level covariate and treatment effect and outcome, under 5 clusters and cluster size 100.

Design	Propensity score model	Out1		Out2	
		Bias	mse	Bias	mse
B1	PS1	3.6219	18.2079	0.1092	0.0167
	PS2	0.5284	0.3654	0.0945	0.0106
	PS3	0.4932	0.2923	0.1068	0.0139
	PS4	0.4878	0.2846	0.1167	0.0171
B2	PS1	0.4481	0.2971	0.1086	0.0151
	PS2	0.2794	0.1062	0.0996	0.0112
	PS3	0.1942	0.0520	0.1051	0.0127
	PS4	0.2126	0.0604	0.1065	0.0133

Table 2.5: Average absolute bias (bias) and mean square error (mse) of treatment effect estimator δ in simulations with strong and weak association between cluster-level covariate and treatment effect and outcome, under 100 clusters and cluster size 10.

Design	Propensity score model	Out1		Out2	
		Bias	mse	Bias	mse
C1	PS1	5.0049	25.4614	0.1022	0.0123
	PS2	1.5305	2.4079	0.0934	0.0119
	PS3	1.0219	1.1619	0.1561	0.0290
	PS4	1.0092	1.1075	0.1499	0.0276
C2	PS1	1.3337	1.8151	0.0959	0.0115
	PS2	1.1392	1.3185	0.0888	0.0089
	PS3	0.3525	0.1467	0.1329	0.0212
	PS4	0.3607	0.1507	0.1290	0.0192

in either propensity score model or outcome model reduces bias and mean square error of treatment effect estimation.

In the results, we found that the outcome model influences treatment effect estimation more than the propensity score model does. Using the same propensity score model, comparison between fixed outcome model (Out1) and random intercept outcome model (Out2) showed the latter greatly reduced the average absolute bias and mean square error. For example, using random intercept propensity score model (PS3) in Scenario A2 where the cluster-level covariate is weakly associated with treatment assignment and outcome, the average absolute bias and mean square error of the random intercept outcome model is reduced from 0.2717 to 0.1199 and from 0.1079 to 0.0171 in the fixed outcome model. With the random intercept outcome model (Out2), the average absolute bias and mean square error of the random intercept propensity score model including cluster-level covariates (PS4) decreases from 0.1199 to 0.1172 and from 0.0171 to 0.0162 in the random intercept propensity score model (PS3). Incorporating multilevel structure in the outcome model improves estimation more significantly if only one stage could take into account cluster-level confounding. The improvement is more significant when cluster-level confounding is strongly related to treatment assignment and outcome (Scenario A1, B1 and C1).

Over all scenarios, the fixed-effect propensity score model (PS2) and random intercept outcome model (Out2) lead to the smallest average absolute bias and mean square error compared with all other models. For Scenario A with moderate number of clusters and cluster size, the fixed effect propensity score model (PS2) provides the smallest average absolute bias and mean square error compared to other propensity score models with the same outcome model, regardless of the degree of association of cluster-level covariate with treatment assignment and outcome. Fixed effect propensity score models do not always perform best with the same outcome model in scenario B and C. The impact of multilevel

modeling on bias reduction tends to weaken when the number of clusters is large relative to cluster size (Scenario C) compared to other scenarios (Scenario A and B).

2.5 Case Study

We applied the proposed methods to the lipid management data described in Section 2. The objective of this case study is to investigate the effect of prior lipid testing on the LDL-C level of children. The continuous outcome Y_{ij} is LDL level tested in 2012. The treatment is the binary indicator of lipid testing from 2001 to 2011 (prior testing), taking the value 1 (treatment) if the child had lipid testing in the pre-2012 time period, 0 (control) otherwise. During the period in question, children obtained lipid tests according to guidelines for lipid testing which include personal medical problems (i.e. diabetes, obesity, etc.) and family history (a family history of high cholesterol or cardiovascular diseases) or if the child is adopted and family history is unknown. Thus we elected to employ propensity scores in order to adjust for confounding characteristics before comparing LDL-C levels among the treatment (tested) and control (untested) groups. The individual level confounders $\mathbf{X}_{ij}$ we accounted for include demographic characteristics, BMI, systolic and diastolic blood pressures, lipid medication prescription claim, well-child visit, smoking status, and physician coded diagnosis of diabetes, hypertension, and dyslipidemia. The number of outpatient visits in each health system is used as a cluster-level covariate (C_j).

We consider eight models combining the propensity and outcome models proposed in Section 3 which cover four situations: first, the single-level propensity score model (PS1) combined with simple fixed outcome model (Out1) totally ignores cluster-level confounding; second, fixed effect propensity score model (PS2) and other two propensity score models employing multilevel structures (PS3 and PS4) combining with fixed outcome model (Out1) only consider data structure in the propensity score stage; third, single level propensity score model (PS1) united with random intercept outcome model (Out2) exploiting cluster structure only in outcome stage; last, random intercept outcome model (Out2) integrated with three propensity score models incorporating cluster-level information (PS2,

PS3, and PS4) takes cluster structure into account in both propensity score and outcome stage.

We ran three chains with 30,000 iterations and a burn-in of 5,000 for all models with over-dispersed initial values. Continuous cluster and individual level covariates are rescaled to have zero mean and unit variance. The quintiles of propensity scores to create subclasses are obtained by fitting (2.1) with maximum likelihood, and all subjects in the data are classified into subclasses based on the estimated propensity scores. Therefore, the probability of receiving lipid screening increases for children from subclass 1 to subclass 5. Table 2.6 and Table 2.7 display the posterior mean and 95% credible sets of parameters in the single-level outcome model and random intercept outcome model, respectively, with four propensity score models. Although the point estimates of the effect of lipid screening on LDL level are slightly different, the general trend is that the children who had lipid testing prior to 2012 have higher LDL-C levels than those who did not, and the increase in LDL-C levels is significant since the 95% credible sets exclude 0 for all analyses. The more likely a child was previously subject to lipid screening, the higher the LDL level in 2012. The amount of increase in LDL-C levels magnifies from subclass 1 to 5. Based on simulation results of scenarios with small number of clusters and large sample size, the fixed effect propensity score model and random intercept outcome model outperform other model combinations. From the result of superior models, the LDL-C level of youth in the treatment (i.e., previously tested) group is 4.65 mg/dL higher than that of children in the control group. This difference is significant with the 95% credible sets ranging from 2.15 mg/dL to 7.15 mg/dL. From subclass 1 to 5, children have increasing LDL level as the probability of receiving lipid screening increases.

2.6 Discussion

Propensity scores are widely applied in observational studies for confounder adjustment. A limited amount of literature investigates the use of propensity scores in multilevel data. At the same time, considering uncertainty in propensity score estimation has been

Table 2.6: Posterior mean and 95% credible sets (in parenthesis) of parameters in fixed single level outcome model (Out1) with different propensity score models (PS1-PS4).

Propensity score model	1	2	3	4
LDL screening effect δ	4.56 (2.08,7.04)	4.62 (2.12,7.11)	4.58 (2.08,7.07)	4.59 (2.09,7.07)
y-intercept β_{11}	80.15 (77.65,82.58)	80.05 (77.66,82.33)	80.04 (77.63,82.33)	80.02 (77.63,82.32)
Subclass 2 β_{12}	5.66 (1.26,9.70)	6.49 (2.41,10.23)	6.48 (2.43,10.24)	6.66 (2.77,10.33)
Subclass 3 β_{13}	8.37 (3.28,13.44)	8.02 (3.10,13.27)	7.9 (3.07,13.23)	7.91 (2.99,13.10)
Subclass 4 β_{14}	13.30 (8.35,18.06)	13.53 (8.19,18.40)	13.63 (8.35,18.52)	13.70 (8.35,18.50)
Subclass 4 β_{15}	18.38 (11.85,26.16)	17.25 (11.45,24.79)	17.50 (11.50,25.40)	17.28 (11.44,25.06)

Table 2.7: Posterior mean and 95% credible sets (in parenthesis) of parameters in fixed single level outcome model (Out2) with different propensity score models (PS1-PS4).

Propensity score model	1	2	3	4
LDL screening effect δ	4.59 (2.09,7.09)	4.65 (2.15,7.15)	4.61 (2.13,7.12)	4.63 (2.12,7.15)
y-intercept for site 1	79.38 (72.37,83.00)	77.81 (67.25,82.93)	78.57 (69.83,83.15)	78.20 (68.57,83.20)
y-intercept for site 2	80.39 (77.68,82.94)	80.70 (78.21,83.21)	80.56 (77.95,83.11)	80.67 (78.15,83.21)
y-intercept for site 3	79.76 (76.94,82.40)	78.95 (75.63,81.80)	79.00 (75.68,81.88)	78.83 (75.43,81.89)
Subclass 2 β_{12}	5.83 (1.55,9.79)	7.20 (3.14,11.15)	6.99 (2.87,10.93)	7.28 (3.19,11.20)
Subclass 3 β_{13}	8.23 (3.17,13.46)	7.80 (3.05,12.94)	8.17 (3.18,13.38)	7.98 (3.04,13.20)
Subclass 4 β_{14}	13.36 (8.41,18.14)	13.73 (8.46,18.41)	13.52 (8.30,18.34)	13.63 (8.39,18.37)
Subclass 4 β_{15}	18.25 (11.83,26.19)	17.20 (11.55,24.72)	17.95 (11.83,25.84)	17.54 (11.69,25.15)

studied only in unstructured data. In this paper, we proposed Bayesian propensity score stratification analysis for multilevel observational data. We expected multilevel models would incorporate clustering structure in the data and address cluster-level confounding. Also, we modeled propensity score and outcome at the same time in a Bayesian framework to integrate uncertainty in propensity score estimation into treatment effect estimation. Our application of these methods to youth lipid management provide the estimated effect of lipid testing with covariate confounding adjustment and clearer clinical picture of youth undergoing lipid testing.

Totally ignoring cluster-level confounding in both stages of propensity score stratification results in severe bias in treatment effect estimation, especially when cluster-level confounding is strongly related to treatment assignment and outcome. Employing multilevel modeling in at least one stage greatly attenuated bias and mean square error. In circumstances where multilevel modeling can be applied in only one stage, use in the outcome stage is preferable to use in the propensity score stage.

There is controversy about feedback of the outcome model on the propensity score model when two stages of propensity score stratification are combined into one in the Bayesian framework. McCandless et al. (2009) point out that the outcome would assist propensity score estimation if the outcome depends heavily on the propensity score. In the view of Rubin (2007), propensity scores should be estimated without information from outcome data. Three possible ways are available to prevent feedback from the outcome model to the propensity score model. As described in Kaplan and Chen (2012), two stages of propensity score stratification are modeled separately in a Bayesian framework and outcome is modeled based on the posterior samples of the propensity score. Zigler et al. (2013) suggest cutting feedback by adding covariate adjustment in the outcome stage to recover true treatment and outcome association space. In McCandless et al. (2010), using a cut function in WinBUGS can disconnect the feedback from the outcome model to the propen-

sity score model. We envision some future study of feedback between the outcome stage and the propensity state in a multilevel setting.

In this paper, balanced cluster size scenarios are studied. It is of interest to look into the performance of multilevel modeling in Bayesian propensity score stratification when cluster sizes are unbalanced across clusters. It is worth exploring the effect of the imbalance on the estimation of treatment effect because the cluster sizes obtained from observational studies are commonly imbalanced.

CHAPTER THREE

Bayesian Misclassification and Propensity Score Methods for Clustered Observational Studies

This chapter published as: Zhou Qi, Yoo-Mi Chin, James D. Stamey, and Joon Jin Song. 2017 “Bayesian Misclassification and Propensity Score Methods for Clustered Observational Studies." *Journal of Applied Statistics* 1-14.

3.1 Introduction

Observational data are increasingly being used for causal inference in social science and public health studies. Reliable causal inference cannot be made directly in observational studies due to the lack of randomization. Also, errors in response distort the true association between response and intervention, leading to biased inferences.

Without adjustment, pre-existing differences in characteristics of exposure and non-exposure groups confound with treatment or intervention effects. The propensity score adjusting technique was first proposed by Rosenbaum and Rubin (1983b) and has become an increasingly popular technique for causal analysis in observational studies. The propensity score adjusts exposure effect in the presence of confounding among different groups and removes bias due to observed covariates if no unmeasured confounding exists. Traditionally, the propensity score is used for matching, stratification and covariate adjustment.

The propensity score is the probability of receiving treatment or exposure conditional on the observed covariates and this probability is used as a balancing score to remove confounding among exposures groups. Traditional propensity score analysis consists of two steps: estimating propensity score and applying estimated propensity score. The propensity score is generally estimated by logistic regression, and then the estimated propensity score is treated as true and used for matching, stratification or covariate adjustment. There are arguments that traditional two-step propensity score analysis ignores the uncertainty in the propensity score estimation. By deriving the multivariate normal distribution that ATE

and parameters in the propensity score model follows, Abadie and Imbens (2009) demonstrated that the variance of the estimated average treatment effect (ATE) based on matching on estimated propensity score needs to be adjusted downward except the case observed covariates and exposure is independent conditional on the propensity score. However, a drawback of this method is that the adjusted variance can be negative. The Bayesian approach is a natural way to appropriately accommodate uncertainty in propensity score analysis and to ensure positive variance of the estimators. McCandless et al. (2009) integrated two steps of propensity score stratification into one in Bayesian framework and integrated out propensity score in the treatment effect estimation as a nuisance parameter, incorporating the uncertainty of the propensity score estimation in treatment effect estimation. An (2010) proposed full Bayesian propensity score regression and matching and showed that Bayesian estimators provide correct standard errors of average treatment effect (ATE). An also came up with an intermediate propensity score approach, which only propensity score is estimated in Bayesian framework and estimated propensity score is applied in frequentist framework. Later, Kaplan and Chen (2012) proposed two-step Bayesian propensity score stratification, weighting, and matching. In their approach, the treatment effect is estimated with the posterior samples of the estimated propensity score. The performance of two-step Bayesian approach depends on the precision in the prior of the treatment effect. Compared with two Bayesian approaches above, higher precision leads to smaller variance and overly high coverage rates for two-step Bayesian stratification, weighting and matching. With non-informative prior, only two-step Bayesian propensity score stratification performs as good as An's Bayesian propensity score estimators.

Propensity score methods were originally developed in the unstructured data settings. However, data from a wide range of applications including medical and social sciences have multilevel structure. Analyses ignoring this structure would lead to inaccurate standard errors and biased estimation due to cluster-level confounding. There have been some studies on propensity score matching and weighting for the multilevel data from the frequentist

perspective. Arpino and Mealli (2011) evaluated the benefit of utilizing multilevel models in propensity score matching when unobserved cluster-level covariates were present. Li et al. (2013) compared different propensity score weighting estimators for multilevel data and concluded that employing multilevel structure can considerably reduce bias in causal estimation. Su and Cortina (2009) suggested that multilevel modeling in both stages of propensity score analysis is more effective in reducing bias in the estimation of treatment effect than multilevel modeling in only one stage, using time series cross sectional data.

In observational studies, response may not be reported correctly. For example, a woman may not answer yes when she truly experiences physical violence from her husband. It is shown in Neuhaus (1999) that ignoring errors in the responses would yield highly biased estimates of the covariate effects. Paulino et al. (2003) presented a Bayesian binomial regression approach to model the association between misclassified response and error-free covariates. Further, Paulino et al. (2005) extended the approach by allowing random effects to account for correlated misclassified binary responses.

In this paper, we propose a Bayesian propensity score regression approach for clustered observational studies, which utilizes multilevel modeling for hierarchical structures and corrects for misclassified responses. The propensity score is estimated through mixed effects logistic regression model with observed individual and cluster level covariates. The outcome model regresses on the exposure indicator and a function of estimated propensity score, which could be stratification membership indicator or covariate adjustment. Random effects are included in the outcome regression to account for possible correlations among responses. The true regression relationship between response and exposure is recovered by adjusting misclassified response with information of sensitivity and specificity.

India National Health Survey (NFHS-3) 2005-6 serves as an appropriate data set to be used to improve the shortcomings of existing studies. The data contain a rich set of demographic and socio-economic variables that can be used to control for confounders, geographic identifiers that can be used to differentiate state-level errors from individual-level

errors, and dichotomous spousal violence measures that are subject to a misclassification problem.

The paper is organized as follows. Section 2 presents the proposed approach in details. Section 3 provides an empirical application as an illustration for the proposed method. We close with a summary in Section 4.

3.2 Methodology

3.2.1 Assumptions

For the multilevel model we consider here, every individual in clusters receives aexposure, T and response Y is recorded. We assume both cluster level and individual level covariates and refer to these together as U. For the propensity score to be valid we require the ignorability assumption which consists of the following two assumptions:

$$\text{Unconfoundness: } Y \perp T|\mathbf{U},$$

$$\text{Overlap: } 0 < \text{P(T=1)} < 1.$$

The unconfoundness assumption states that conditioning on measured covariates there are no omitted covariates related to potential response, so that exposure assignment is independent of potential response. The overlap assumption ensures that there is common support for propensity scores of exposure units and non-exposure units for all observed covariates values in the study population, meaning an exposure unit always matches with at least one non-exposure counterpart. In presence of cluster level confounders and misclassified responses, model specification ignoring clustered structures or errors in response violates unconfoundness assumption, causing the estimated propensity score to yield biased exposure effects.

3.2.2 Model Specification

Suppose data consist of two levels: $i^{th}(i = 1, 2, ...,n_j)$ subject nested within $j^{th}(j = 1, 2, ..., N)$ cluster. Let T_{ij} denote a dichotomous exposure indicator taking a value of

1 when the subject receives exposure, and 0 when not. Let Y_{ij} denote a dichotomous response taking a value of 1 if the answer is yes and 0 if the answer is no. A $p \times 1$ vector $\mathbf{X_{ij}}$ and a $q \times 1$ vector $\mathbf{C_j}$ are individual and cluster level covariates, respectively. Let $z_{ij} = \mathrm{P}(T_{ij} = 1)$ be the probability of receiving the exposure and $\pi_{ij} = \mathrm{P}(Y_{ij} = 1)$ be the probability of truly answering yes to the question.

In observational data, misclassification is very common in response variables. To account for this, assume we do not observe the true response Y_{ij} but rather the misclassified response $\tilde{Y}_{ij}$ with the probability $\tilde{\pi}_{ij}$. We consider a setting in which the misclassified error in the response variable is independent of the other variables (non-differential) and define the sensitivity θ and specificity η as

$$\theta = \mathrm{P}(\tilde{Y}_{ij} = 1 | Y_{ij} = 1)$$

$$\eta = \mathrm{P}(\tilde{Y}_{ij} = 0 | Y_{ij} = 0)$$

By the law of total probability, the unconditional probability of the error prone response to be '1' is

$$\tilde{\pi}_{ij} = \theta\pi_{ij} + (1 - \eta)(1 - \pi_{ij})$$

The proposed models are the following:

$$\mathrm{logit}(z_{ij}) = \beta_{0j} + \beta_{\mathbf{1}}{}^{\mathbf{T}}\mathbf{X_{ij}} + \beta_{\mathbf{2}}^{\mathbf{T}}\mathbf{C_{ij}} \tag{3.1}$$

$$\mathrm{logit}(\pi_{ij}) = \gamma_{0j} + \gamma\, T_{ij} + \xi\hat{z}_{ij} \tag{3.2}$$

where $\beta_{0j} \sim \mathrm{N}(\mu_{\beta_0}, \sigma^2_{\beta_0})$ and $\gamma_{0j} \sim \mathrm{N}(\mu_{\gamma_0}, \sigma^2_{\gamma_0})$ are random intercepts. Estimated propensity score $\hat{z}_{ij}$ is used as a covariate in the outcome model.

In (3.1), the random intercept β_{0j} absorbs the effect of clustered structure and accounts for heterogeneity of exposure status regardless of clusters. With fixed slopes $\beta_{\mathbf{1}}$, same individual characteristics make the same contribution to propensity score regardless of cluster membership. The distributional assumption on random intercept greatly reduces the number of parameters to be estimated. The estimated propensity score is used as covari-

ate adjustment and removes confounding of pre-exposure covariates in outcome regression model (3.2). Cluster specific intercept γ_{0j} accommodates possible correlation of responses among clusters. Fixed parameter γ assumes that exposure effect is consistent for all clusters. The estimated propensity score $\hat{z}_{ij}$ can be generalized to a function of the estimated propensity score. An alternative to covariate adjustment would be propensity score stratification, in which the function is defined as subclass membership indicators based on the percentiles of estimated propensity score. According to Austin (2009), however, covariate adjustment removes more imbalance in covariates distribution than stratification. Therefore, we focus on covariate adjustment in this paper.

3.2.3 Prior Specification and Model Inference

For regression coefficients in models (3.1) and (3.2), normal priors are given to $\beta_{\mathbf{1}}$,$\beta_{\mathbf{2}}, \gamma,$ and ξ. Usually the variance of normal priors are set to be large so that priors are noninformative without available information about regression coefficients.

$$\beta_{\mathbf{1}} \sim \text{MVN}(\mathbf{0}, \sigma^2_{\beta_1}\mathbf{I})$$

$$\beta_{\mathbf{2}} \sim \text{MVN}(\mathbf{0}, \sigma^2_{\beta_2}\mathbf{I})$$

$$\gamma \sim \text{N}(0, \sigma^2_{\gamma})$$

$$\xi \sim \text{N}(0, \sigma^2_{\xi})$$

Beta priors are assigned on sensitivity and specificity:

$$\theta \sim \text{beta}(a_1, b_1),$$

$$\eta \sim \text{beta}(a_2, b_2).$$

The hyperparameters are typically determined by expert opinion and validation data.

The posterior distributions of individual parameters are obtained using Markov chain Monte Carlo (MCMC) algorithm as implemented in the freeware JAGS. Each update of γ is based on the updated β in propensity score model. The marginal distribution of γ

incorporates uncertainty in propensity score estimation by averaging over the uncertainty in posterior samples of β. The code is available from the authors upon request.

3.3 Application

3.3.1 Data Description

The method proposed in this paper is employed to analyze the effect of female employment on the odds of physical spousal violence. Numerous existing studies have examined the effect of women's employment on the likelihood of spousal violence towards them (Vyas and Watts, 2009), but many of these studies lack rigor in their analyses. Specifically, some studies do not account for confounding between working females and non-working females and compare their probability of suffering physical violence directly. Further, the analyses are often confined to the individual level and ignore a hierarchical structure of the data. More importantly, underreporting of spousal violence has not been systematically addressed, although underreporting is a chronic problem in domestic violence research (Palermo et al., 2013). Most domestic violence surveys measure intimate partner violence as an indicator variable that takes a value of one if an interviewee has an experience and zero otherwise. With considerable underreporting, these dichotomous domestic violence measures suffer from a typical misclassification problem. A recent study by Chin et al. (2017) addresses misclassification in violence reporting, but their study does not account for a multilevel structure of the data.

This paper uses demographic and socio-economic data of women from India National Health Survey 2005-6 (NFHS-3) and state-wise crime and unemployment data from National Crime Records Bureau (2005) and Singh and Kumar (2014). The study focuses on urban women defined as women who reside in mega city, large city, small city, and large town. Given that the outcome variable is physical spousal violence experience of married women, the sample is further restricted to currently married women who are included in the domestic violence module. After dropping the observations with missing variables, the

final sample is composed of 5,573 non-working women (non-exposure group) and 14,542 working women (exposure group).

Physical spousal violence is measured as an indicator of any of the following acts committed by a womanâs husband in the past 12 months: 1) slapping; 2) twisting her arm or pulling her hair; 3) pushing, shaking, or throwing something at her; 4) punching her with his fist or something harmful; 5) kicking or dragging her; 6) trying to choke or burn her; 7) threatening or attacking her with a knife/gun or other weapon. Exposure status is an indicator that takes the value of 1 if a woman worked in the past 12 months and 0 if not (control group: non-working women, treatment group: working women). Age and education variables are in years. Number of children/adult men/adult women are measured at the household level. Wealth index is reported by the DHS and represents a wealth status of a household in the population. It is an index based on household asset ownership and has a standard normal distribution. Low caste is an indicator for the household head belonging to scheduled castes and scheduled tribes. Hindu is an indicator for the household head being Hindu. Remarriage is an indicator for a woman having been married more than once. Polygamy is an indicator for a woman being in a polygamous marriage. Crime rate is state-level total cognizable crime rate per 100,000 population. Unemployment rate is state-wise unemployment rate per 1000 population.

Table 3.1 reports descriptive statistics of the main variables used in the analysis. Women and their husbands in the exposure group are older and less educated than their counterparts. They have more children but fewer adults in the household. Their households are more likely to be poor, belong to a low caste, and follow Hinduism. Women in the exposure group are more likely to be married more than once and have a polygamous marriage. Further, women in exposure group are more likely to reside in states with lower unemployment rate and crime rate. Given that there is heterogeneity in demographic and socio-economic characteristics between the exposure group and the non-exposure group, it

is necessary to obtain a balance in the distribution of pre-exposure characteristics between the two groups.

Table 3.1: Descriptive statistics. Source: India National Family Health Survey (NFHS-3) 2005-6.

Variables	Non-Exposure		Exposure	
	Mean	SD	Mean	SD
a) Response				
Physical violence	0.16	0.37	0.21	0.41
b) Pre-exposure characteristics				
b1) Demographic variables				
Female age	31.86	7.88	33.73	7.10
Female education	7.77	5.22	7.23	6.07
Male age	37.46	8.71	39.00	8.19
Male education	9.52	4.91	8.46	5.50
Number of children	2.16	1.39	2.28	1.43
Number of adult men	1.79	1.08	1.59	0.91
Number of adult women	1.61	0.89	1.53	0.77
b2) Socio-economic variables				
Wealth index	0.77	0.80	0.55	0.89
Indicator: low caste	0.49	0.50	0.60	0.49
b3) Cultural variables				
Indicator: Hindu	0.74	0.44	0.77	0.42
Indicator: remarriage	0.01	0.11	0.02	0.15
Indicator: polygamy	0.01	0.09	0.02	0.14
c) State level pre-exposure characteristics				
Unemployment rate/1000	6.41	4.90	6.13	4.76
Crime rate/100,000	466.99	304.91	501.77	311.39
No. of observations	5573		14542	

3.3.2 Empirical Results

This analysis aims at estimating the effect of the female working status on whether she suffers physical violence from her husband. The data consists of a sample of female individuals indexed by i from jth state in India. In the analysis, the exposure T_{ij} denotes employment status of a woman in the previous 12 month and z_{ij} is the probability of a woman working. The true response whether a woman experienced physical violence is the unobservable outcome denoted by Y_{ij} taking values of 1 or 0 (1, suffering physical vio-

lence in past 12 months, 0 not suffering physical violence) and the observed response $\tilde{Y}_{ij}$ is whether the woman reports experiencing physical violence from her husband in the same period. True physical experience status relates to reported status through sensitivity and specificity, which are obtained from external data source. For each woman, the individual level characteristics $\mathbf{X_{ij}}$ and state level covariates $\mathbf{C_j}$ are observed.

Four scenarios are compared in the data analysis: the first scenario is a Bayesian propensity score analysis without multilevel modeling and adjusting misclassification errors; the second scenario adjusts misclassified response with single level model in the analysis; third scenario employs multilevel modeling in propensity score and outcome model without misclassification error adjustment in response; fourth scenario utilizes both multilevel models and misclassification error adjustment in the analysis.

No previous information is available for coefficients in regression models so non-informative priors are placed on the parameters. Variance of 10 yields relatively non-informative normal priors on fixed coefficients: $\sigma^2_{\beta_1} = \sigma^2_{\beta_2} = \sigma^2_{\gamma} = \sigma^2_{\xi} = 10$. For random parameters, the hyper parameters are meant to result in non-informative priors. The mean of parameters are allowed to vary around 0 with variance 10 and the variance of parameters ranges from nearly 0 to 5, yielding relatively wide range for random parameters:

$$\mu_{\beta_{0j}} \sim \mathrm{N}(0, 10),$$
$$\sigma^2_{\beta_{0j}} \sim \mathrm{U}(0.001, 5),$$
$$\mu_{\gamma_{0j}} \sim \mathrm{N}(0, 10),$$
$$\sigma^2_{\gamma_{0j}} \sim \mathrm{U}(0.001, 5).$$

We require informative priors for the sensitivity and specificity. Rabin et al. (2009) summarizes the agreement of several approaches in assessing the occurrence of domestic violence. Using those results we assign priors for the sensitivity and specificity to be $\theta \sim$ beta (5.0,7.6) and $\eta \sim$ beta (165.7,9.7). The prior for the specificity is considerably more informative than the sensitivity. It is also centered considerably higher than the prior for

the sensitivity. This indicates in these previous studies violence was much more likely to be not-reported than falsely reported.

We generate two parallel independent MCMC runs of size 5,000 with widely spread initial values after a burn-in 1000. The convergence of MCMC chains is examined and monitoring results of exposure effect are reported in Figure 3.3.1 and Figure 3.3.2. The Gelman-Rubin statistics converge to 1 and the trace of chains mix well. Table 3.2 shows the posterior results of the exposure effect, odds ratio and DIC of the scenarios. Both scenario 1 with single level models and scenario 3 with multilevel models exclude response misclassification correction. The magnitudes of the estimated exposure effects in the two scenarios are very similar. For scenarios with misclassification correction, scenario 2 and 4 also have very similar magnitude of the estimated exposure effect. When the results from scenarios with and without misclassification component (1 vs 2, 3 vs 4) are compared, standard deviation of the exposure effect is found to be slightly larger in scenarios with misclassification component. This result is consistent with Neuhaus (1999)'s study, which found that errors in response lead to information loss of covariates effects. Comparing scenario 2 and 4 with misclassification component, accounting for multilevel structure reduces standard deviation of the estimate.

Table 3.2: Posterior results of exposure effect and DIC of the model.

Scenario	Mean of treatment (odds ratio)	SD (odds ratio)	95% Creditable Interval (odds ratio)	DIC
1	0.053 (1.055)	0.043 (0.045)	[-0.030,0.134] ([0.971,1.144])	40042.4
2	0.078 (1.085)	0.083 (0.091)	[-0.081,0.238] ([0.922,1.269])	39933.6
3	0.056 (1.058)	0.046 (0.049)	[-0.033,0.149] ([0.967,1.161])	39191.4
4	0.080 (1.086)	0.067 (0.073)	[-0.051,0.213] ([0.9501,1.238])	39162.5

In terms of model fitting, scenario 4, the propensity score analysis with multilevel modeling and misclassification correction, has the smallest DIC among four scenarios, in-

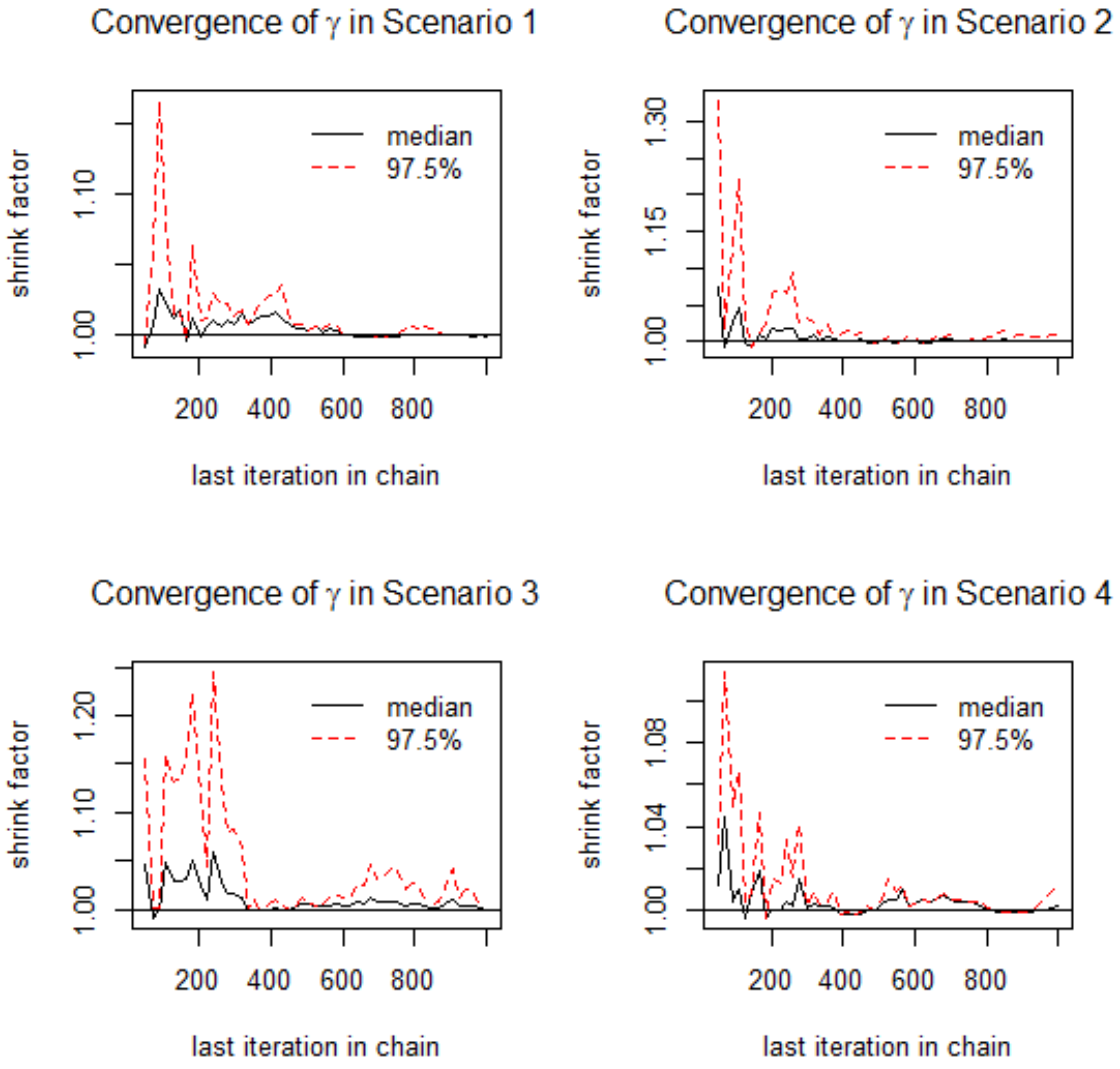

Figure 3.3.1: Gelman-Rubin convergence diagnostic statistic of exposure effect in four scenarios.

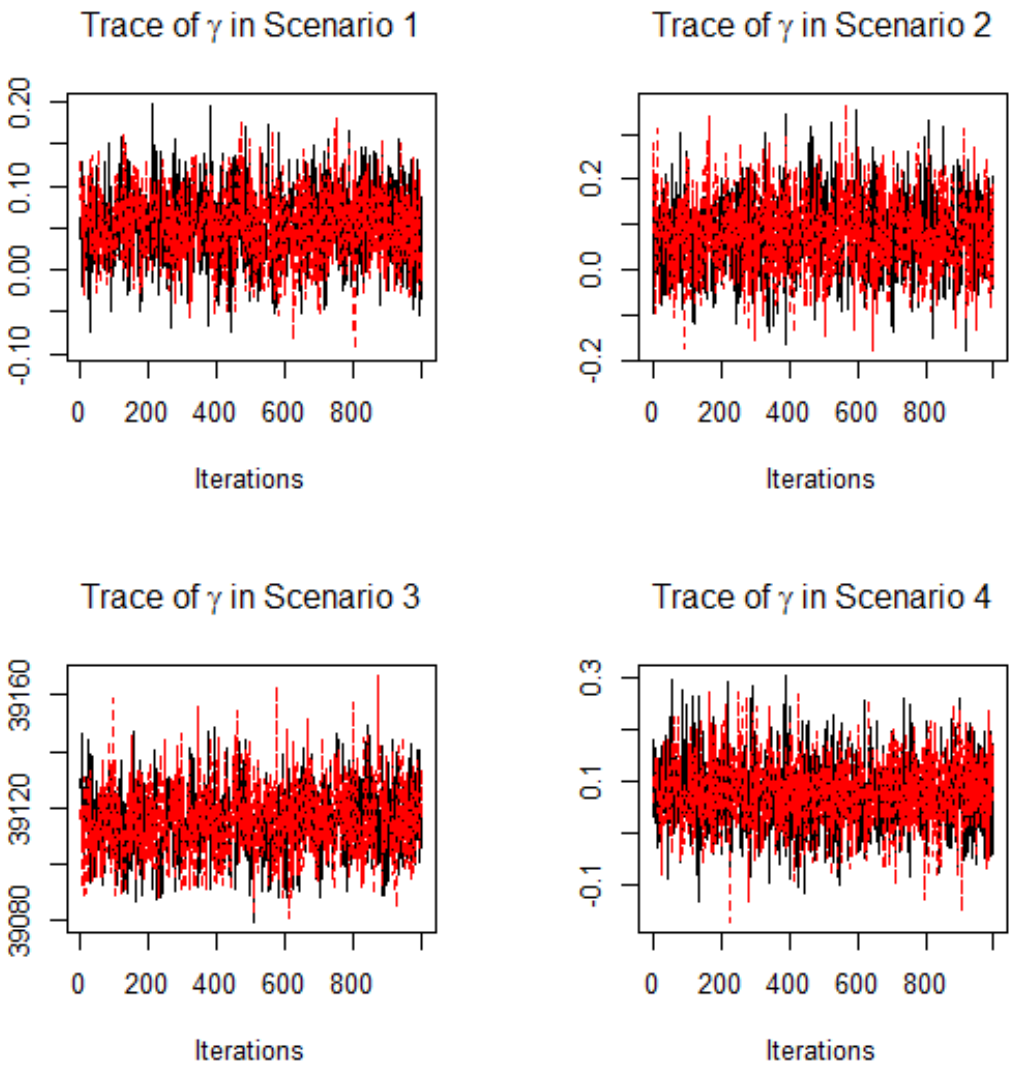

Figure 3.3.2: Trace of exposure effect in four scenarios.

dicating the best fit for the data. Scenario 2 and 3 have smaller DIC than scenario 1, suggesting that accounting for either multilevel structure of the data or misclassification of response improves the model fit. Based on the results of scenario 4, the odds of a working woman experiencing physical violence from her husband is 1.086 times higher than the odds of a woman without employment. The sign of the point estimate suggests that a wife's employment provokes more physical violence from the husband rather than to increase her bargaining power within the home and lower the likelihood of violence. In terms of confidence interval, however, this violence-provoking effect of female employment is statistically insignificant.

Table 3.3 displays the posterior results of sensitivity and specificity estimated in Scenario 2 and 4. The specificity estimates are very similar in two scenarios. The sensitivity estimate of scenario 4 with multilevel structure specification is larger than that of scenario 2. About 53% of abused women truly report their experience of physical violence, suggesting a fairly high level of underreporting.

Table 3.3: Estimated sensitivity and specificity in Scenario 2 and 4.

Scenario	Sensitivity	Specificity
2	0.389	0.983
4	0.531	0.985

3.4 Simulation Study

3.4.1 Simulation Design

We conducted a simulation study which mimics the data structure of empirical application to assess the performance of the proposed approach. The simulated data consists of 10 clusters index by j and 100 individuals index by i in each cluster. We assume two individual level covariates, C_1, C_2 and cluster level covariate Z distributed as independent standard normal distribution. Dichotomous exposure T and Y are generated with following

models:

$$\text{logit}(\text{P}(T_{ij}=1)) = \alpha_{0j} + \gamma_1 C_{1ij} + \gamma_2 C_{2ij} + \gamma_3 Z_j,$$
$$\text{logit}(\text{P}(Y_{ij}=1)) = \varphi_{0j} + \beta_1 C_{1ij} + \beta_2 C_{2ij} + \beta_3 Z_j + \delta T_{ij},$$

with $\alpha_0, \varphi_0 \sim N(0,2), \gamma = (0.3, 0.4, 0.5), \beta = (0.3, 0.3, 0.7)$ and $\delta = 1$. The true response Y relates to the observed response $\hat{Y}$ through sensitivity θ and specificity η and $\hat{Y}$ is generated from Bernoulli distribution with probability $P(\hat{Y}_{ij} = 1) = \theta P(Y_{ij} = 1) + (1-\eta)(1 - P(Y_{ij} = 1))$, where $\eta = 0.98$, the estimated specificity in the motivating example.

In the simulation, we consider two propensity score models:

(1) Single level propensity score model ignoring multilevel structure,

$$\text{logit}(\text{P}(T_{ij}=1)) = \gamma_0 + \gamma_1 C_{1ij} + \gamma_2 \text{C}_{2ij}.$$

(2) Random intercept propensity score model,

$$\text{logit}(\text{P}(T_{ij}=1)) = \alpha_{0j} + \gamma_1 C_{1ij} + \gamma_2 C_{2ij} + \gamma_3 Z_j,\ \alpha_{0_j} \sim \text{N}(\mu_\alpha, \sigma^2_\alpha).$$

The specification of outcome regression models are:

(1) Single level outcome regression model ignoring multilevel structure and misclassification in response:

$$\text{logit}(\text{P}(\hat{Y}_{ij}=1)) = \beta_0 + \beta_1 \hat{P}(T_{ij}=1) + \delta T_{ij}.$$

(2) Single level outcome regression model with misclassification correction:

$$\text{logit}(\text{P}(Y_{ij}=1)) = \beta_0 + \beta_1 \hat{P}(T_{ij}=1) + \delta T_{ij},$$
$$\text{P}(\hat{Y}_{ij}=1) = \theta \text{P}(Y_{ij}=1) + (1-\eta)(1 - \text{P}(Y_{ij}=1))$$

(3) Random intercept outcome model ignoring misclassification in response:

$$\text{logit}(\text{P}(\hat{Y}_{ij}=1)) = \varphi_{0j} + \beta_1 \hat{P}(T_{ij}=1) + \delta T_{ij},\ \varphi_{0j} \sim \text{N}(\mu_\varphi, \sigma^2_\varphi).$$

(4) Random intercept outcome model incorporating misclassification in response:

$$\text{logit}(\text{P}(\hat{Y}_{ij} = 1)) = \varphi_{0j} + \beta_1 \hat{P}(T_{ij} = 1) + \delta T_{ij},\ \varphi_{0j} \sim \text{N}(\mu_\varphi, \sigma^2_\varphi)$$

$$\text{P}(\hat{Y}_{ij} = 1) = \theta \text{P}(Y_{ij} = 1) + (1 - \eta)(1 - \text{P}(Y_{ij} = 1)).$$

We compare four scenarios in the simulation: first, single level propensity score model and outcome regression model ignoring multilevel structure and response misclassification; second, single level propensity score model and outcome regression model with misclassification correction; third, random intercept propensity score model and outcome regression model; fourth, random intercept propensity score model and outcome regression model including misclassification correcting component.

We consider three designs listed in Table 3.4 and generate 300 data sets for each scenario. In the design A and B, observed responses are generated with estimated sensitivity in the India data to validate conclusion in empirical application. The design A uses same prior for sensitivity in the empirical example and design B takes a slightly more informative prior for sensitivity. The design C further assesses proposed models with data generated with sensitivity as 0.7, a commonly seen value in applications.

Table 3.4: Sensitivity parameters used in data generating process and priors for sensitivity simulation study.

Design	Sensitivity to generate data	Prior for sensitivity
A	0.53	Beta(5,7.6)
B	0.53	Beta(15,21)
C	0.7	Beta(21,9)

3.4.2 Simulation Results

Table 3.5 includes simulation results of exposure effect estimation under three designs. Throughout three designs, we observed that scenario 4 which utilizes multilevel modeling and corrects misclassification in response outperforms other scenarios with smallest average absolute bias and closest coverage rate of 95% credible sets to the nominal level, for data with small number of clusters and large cluster size. Scenario 1 has smaller average

absolute bias than scenario 2 and 3 do and this indicates ignoring multilevel structure and misclassification leads to less biased point estimate of exposure effect than only incorporating either multilevel structure or misclassification. Comparing scenario 1 versus 2 and scenario 3 versus 4, incorporating misclassification in response would greatly improve the coverage rate of 95% credible sets.

Table 3.5: Average absolute bias (bias), mean square error (MSE), and coverage rate of 95% credible sets of exposure effect calculated from 300 simulated data sets under design A, B and C.

Exposure effect	Bias	MSE	Coverage Rate
Design A:			
Scenario 1	0.416	0.201	20%
Scenario 2	0.613	0.468	89.7%
Scenario 3	0.559	0.335	5.3%
Scenario 4	0.395	0.252	95.3%
Design B:			
Scenario 1	0.440	0.219	16%
Scenario 2	0.699	0.592	84.3%
Scenario 3	0.580	0.356	4%
Scenario 4	0.389	0.251	96.3%
Design C:			
Scenario 1	0.323	0.129	36.7%
Scenario 2	0.378	0.205	92%
Scenario 3	0.469	0.244	15.7%
Scenario 4	0.294	0.150	96%

Design A and B closely replicate the data structure of empirical example. Design B employs a more informative prior on sensitivity, approximately triple the equivalent sample size of prior of design A, which reduces the bias of and increases coverage rate in scenario 4. Simulation results of design A and B confirm our approach performs best among candidates for the empirical example.

Design C mimics a more common case in applications of which true sensitivity parameter is 0.7. With a moderate informative prior on sensitivity, scenario 4 still yields smallest average absolute bias and closest coverage rate of credible sets to the nominal level. Higher sensitivity reduces the absolute mean bias and mean square error reduces for

all 4 scenarios and increases the coverage rate in scenario 1, 2 and 3. Comparing scenario 2 and 3 throughout three designs, incorporating only misclassification reduces more bias in exposure effect estimation than incorporating only multilevel structure when sensitivity is higher in design C than that in design A and B.

3.5 Conclusion

In this paper, we address the problem of ignoring clustered structures and misclassified responses in the observational study. Ignoring clustered structures would confound cluster level characteristic with exposure effects. Furthermore, overlooking the misclassified responses distorts the true relationship between response and exposure effects. To deal with these issues, we proposed a Bayesian multilevel propensity score regression analysis with misclassification in response correction. The results of application in India National Health Survey 2005 strongly support the importance of misclassification correction component and multilevel structure specification in propensity score regression analysis in the clustered observational data with possible misreported responses. The simulation study indicates proposed approach yields exposure effect estimation with least average absolute bias and closest coverage rate to the nominal level.

When implementing the approach, we observe the exposure effect would be overestimated and its credible set tends to be overly wide if overly noninformative priors are placed on sensitivity, specificity, and regression coefficients of outcome model. Too informative priors would make coverage rate of creditable set much higher than the nominal level.

In the paper, we only consider random intercept multilevel model. For applications requiring more general models, $\beta_{\mathbf{1}}$ can be replaced by random slopes $\beta_{\mathbf{1j}} \sim$ MVN$(\mu_{\beta_{\mathbf{1}}}, \sigma^2_{\beta_1}\mathbf{I})$, which allow for varying effects of individual characteristics on propensity score for all clusters. Propensity score model with random slopes needs additional information about possible varying effects of individual covariates. Sometimes it is not easy to determine random slopes for too many covariates. Usually, the main interest is to estimate a universal exposure effect, not the variation of the exposure effect. In case a cluster-specific exposure

effect is of interest, the outcome model can accommodate the need by replacing γ with $\gamma_j \sim \text{MVN}(\mu_\gamma, \sigma^2_\gamma \mathbf{I})$.

The proposed approach is developed under the assumption that no unmeasured confounders exist. It would be interest to study the sensitivity of unmeasured confounders for multilevel observational study with misclassified response.

CHAPTER FOUR

Bayesian Sensitivity Analysis to Unmeasured Confounding for Misclassified Data

4.1 Introduction

Many questions of interest for observational study face the difficulty that causal inference cannot be made directly due to lack of randomization of exposure status. The ignorability assumption (Rosenbaum and Rubin, 1983b) ensures valid causal inference in observational studies but this untestable assumption is often violated. Rather than assume ignorability is met without testing, we can assume the presence of an unmeasured confounder and assess the sensitivity of violations to this assumption. In this approach, we specify how the unmeasured confounder enters inferential models through sensitivity parameters which characterize the relationships of the unmeasured confounder with both exposure and response. The study is deemed sensitive to violations of the ignorable assumption if the exposure estimates from models accounting for possible levels of unmeasured confounding are considerably different from original analysis.

Another common problem in observational data with binary outcomes is that of response misclassification. Errors in the response are often due to either an imperfect diagnostic test or when a sensitive question is asked in a survey. Adjusting for response misclassification has been addressed from both the frequentist (Magder and Hughes, 1997) and the Bayesian approaches (Paulino et al., 2003; McInturff et al., 2004).

Violation of the ignorability assumption for potential confounders and misclassification in response are possible sources of bias. Most previous work has dealt with one source of bias. Greenland (1996; 2005) discusses methods modeling several sources of bias, including unmeasured confounding, misclassification error and non-response. The departure from the true model is measured with the bias factor, an integrated correction from multi-

ple sources of bias. Besides reducing bias in parameter estimation, multiple bias modeling incorporates sources of uncertainty in addition to random error.

Motivated by the study of the influence of female employment on the likelihood of domestic violence, we propose a Bayesian approach to conduct a sensitivity analysis accounting for multiple sources of bias. Specifically, the approach accounts for two sources of bias, misclassification in response and potential unmeasured confounding. The Bayesian framework allows for informative priors based on expert opinion and prior data to correct for the response misclassification. We characterize the unmeasured confounding via its associations with the exposure status and response with what are referred to as sensitivity parameters. We evaluate potential change in exposure effect estimation from a grid of hypothetical sensitivity parameters values, which are assumed to be known in the analysis. The simulation results show our approach yields least average absolute bias and coverage probability to the nominal level compared with approaches ignoring misclassification, unmeasured confounding, or both. The advantage of this approach is that researchers are able to study the influence of unmeasured confounding with interested values of sensitivity parameters, without need of prior information.

The paper is organized as follows. In Section 2 we provide the model and discuss the overall approach to the bias adjustment and sensitivity analysis. The approach is further illustrated with an empirical example in Section 3. In Section 4 we present simulation results to confirm the efficacy of the method. Section 5 concludes the paper.

4.2 Methodology

4.2.1 Review of Sensitivity Analysis Framework

Sensitivity analysis goes back to at least Cornfield et al. (1959) who discussed whether the effect of smoking on lung cancer could be zero after adjusting for an unmeasured covariate. Sensitivity analysis is often categorized into primal, dual and simultaneous analysis based on the number of sensitivity parameters, which controls how the unmeasured

confounder enters the inferential models. Usually one or two sensitivity parameters are specified. The primal and dual approaches work with a single parameter. The primal approach specifies the association between the unmeasured confounder and exposure status and assumes the unmeasured confounder has a nearly perfect correlation with response. Paralleling to the primal approach, the dual method is obtained with inverse sets of associations (Gastwirth, 1998). Most sensitivity analyses with one parameter are based on matched samples and rely on the nonparametric randomization test such as McNemar's test for binary response and Wilcoxon sign-rank test for continuous response. The primal and dual methods can be computationally intensive and sensitive to the choice of test statistic. The works by Carnegie et al. (2016); Dorie et al. (2016); Gustafson et al. (2010); Lin et al. (1998); McCandless et al. (2009); Rosenbaum and Rubin (1983a) discuss simultaneous sensitivity analysis which employs two sensitivity parameters to characterize the associations of an unmeasured confounder with exposure and response. The advantage of the simultaneous approach is that matched samples are not required and sensitivity parameters expressed as regression coefficients or partial correlation are easily interpreted. This approach also avoids assuming strict relationship between the unmeasured confounder and exposure or response. The trade-off is that the method relies strongly on parametric assumptions.

Sensitivity analysis in a Bayesian framework has been explored extensively recently. The sensitivity parameters enter the models for the exposure and covariate relationship and the response and exposure relationship and are specified as regression coefficients controlling the associations of the unmeasured confounder with the exposure and response. In the Bayesian sensitivity analysis proposed in McCandless et al. (2009) and Gustafson et al. (2010), sensitivity parameters are assumed to be unknown and the models become non-identifiable. Faries et al. (2013) use relatively informative priors obtained from external information for the sensitivity parameters. This provides a range for the exposure effect estimates based on the likely range of the sensitivity parameters and incorporates

uncertainty about the unmeasured confounding in the posterior distribution. However, the posterior results may be biased by misspecification of priors. Also, there might be convergence issues when a diffuse prior is used on sensitivity parameters. In the work of Dorie et al. (2016), the sensitivity parameters are fixed at some hypothetical values in Bayesian models, avoiding potential issues of assigning priors on parameters possibly very little is known about. The drawbacks are that credible intervals do not reflect uncertainty due to unmeasured confounding and it is hard to interpret results from a variety of sensitivity parameters combinations. A possible middle ground is Monte Carlo sensitivity analysis where sensitivity parameters are sampled from specified priors and use those samples in Bayesian models (Greenland, 2003; Steenland and Greenland, 2004). With modification, Monte Carlo sensitivity analysis can approximate posterior inference from Bayesian sensitivity analysis. The posterior distribution incorporates both uncertainty about unmeasured confounding and random error and provides a distribution of exposure effect estimates. The limitation is that uncertainty of unmeasured confounding may be underestimated if samples from priors are not large enough.

4.2.2 Model Specification

Suppose we are interested in a binary response, Y, with a binary exposure, Z. For observational studies of the type we consider, we are only able to identify a causal effect if the ignorability assumption is satisfied. The ignorable assumption states that all confounders are measured and conditioned on the measured confounders $\mathbf{X}$, the outcome Y and exposure are independent,

$$\mathrm{Y(1),Y(0)} \perp \mathrm{Z} | \mathbf{X},$$

where $Y(1)$ or $Y(0)$ are the response of a person in presence or absence of an exposure (Rubin, 1978). In reality, researchers rarely have confidence in satisfaction of the ignorability assumption. To assess the sensitivity of this assumption, we assume an unmeasured

confounder U exists and the ignorable assumption holds if U is included,

$$\text{Y(1),Y(0)}\perp Z \mid \mathbf{X}, \text{U}.$$

Following the complete factorization in Dorie et al. (2016), the joint distribution of the observed data and the unmeasured confounder is specified,

$$\text{P(Y,Z,U}|\mathbf{X})\text{=P(Y}|\text{Z,U,}\mathbf{X})\text{P(Z}|\text{ U,}\mathbf{X})\text{P(U}|\mathbf{X}).$$

We model the exposure-covariate and response-exposure surfaces incorporating the unmeasured confounder through logistic regression models:

$$\begin{aligned} \text{logit}(\text{P(Y=1}|\text{Z,U,}\mathbf{X})) &= \beta_0 + \delta\text{Z} + {\beta_1}^T\mathbf{X} + \zeta^{\text{y}}\text{U} \\ \text{logit}(\text{P(Z=1}|\text{U,}\mathbf{X})) &= \gamma_0 + {\gamma_1}^T\mathbf{X}\text{+}\zeta^{\text{z}}\text{U}, \end{aligned} \tag{4.1}$$

where partial correlations ζ^z and ζ^y serve as sensitivity parameters, characterizing the association of the unmeasured confounder with the exposure status and the response variable. In many cases, researchers do not have information about unmeasured confounders. The sensitivity parameters are set to be hypothetical values of interest so that researchers are able to explore how results vary with different degrees of unmeasured confounding. We assume U represents the combination of one or more unmeasured confounders beyond the observed covariates, so the unmeasured confounder is independent of the observed covariates. To simplify the computation, the unmeasured confounder is specified as binary to indicate the presence or absence of unmeasured confounding.

For observational data, especially survey data, the response is often reported in error and we only observe the misreported response $\tilde{Y}$ rather than true response Y. Using observed misreported response directly in (4.1) distorts the response and exposure relationship. The unobservable, true response Y is linked with misreported response $\tilde{Y}$ through the sensitivity θ and specificity η, which are defined as

$$\theta = \text{P}(\tilde{Y} = 1|Y = 1)$$

$$\eta = \text{P}(\tilde{Y} = 0|Y = 0).$$

In this paper, we consider non-differential misclassification, that is, we assume the misclassification error in the response is independent of other variables. By the law of total probability, the unconditional probability of the misclassified response to be 1 is

$$P(\tilde{Y} = 1|Z, U, \mathbf{X}) = \theta P(Y = 1|Z, U, \mathbf{X}) + (1 - \eta)(1 - P(Y = 1|Z, U, \mathbf{X})) \quad (4.2)$$

The true response and exposure relationship surface can be recovered by linking the misclassified response with the true response in (4.2).

4.2.3 Prior Specification and Model Inference

In general, normal priors are assigned for the logistic regression coefficients,

$$\delta \sim \mathrm{N}(0, \sigma_\delta^2),$$

$$\beta_0 \sim \mathrm{N}(0, \sigma_{\beta_0}^2),$$

$$\boldsymbol{\beta_1} \sim \mathrm{MVN}(\mathbf{0}, \sigma_{\beta_1}^2 \mathbf{I}),$$

$$\gamma_0 \sim \mathrm{N}(0, \sigma_{\gamma_0}^2),$$

$$\boldsymbol{\gamma_1} \sim \mathrm{MVN}(\mathbf{0}, \sigma_{\gamma_1}^2 \mathbf{I}),$$

where $\sigma_\delta^2, \sigma_{\beta_0}^2, \sigma_{\beta_1}^2, \sigma_{\gamma_0}^2,$ and $\sigma_{\gamma_1}^2$ are user-specified hyper parameters. With little information about regression coefficients, hyper parameters are set to be large to make priors relatively non-informative. The prior distribution for the unmeasured confounder is $U \sim \text{Bernoulli}(\pi_u)$ and the hyper parameter π_u represents information about the prevalence of unmeasured confounder. Beta priors are assigned on sensitivity and specificity with the hyper parameters determined with a combination of expert opinion and external data,

$$\theta \sim \text{ beta}(a_1, b_1),$$

$$\eta \sim \text{ beta}(a_2, b_2).$$

We use Markov chain Monte Carlo (MCMC) methods to obtain the posterior distribution of model coefficients. The sampling starts with generating initial values of unmeasured

confounder U then sample $(\delta, \beta_0, \beta_1, \gamma_0, \gamma_1, \theta, \eta)$ from distribution obtained by logistic regression of $\tilde{Y}$ on X, Z and generated U, and Z on X and generated U. The update of U is then based on updated sample of $(\delta, \beta_0, \beta_1, \gamma_0, \gamma_1, \theta, \eta)$. After the burn-in, samples of $(\delta, \beta_0, \beta_1, \gamma_0, \gamma_1, \theta, \eta)$ approximate the target distribution.

4.3 Simulation Study

We conducted an extensive simulation study to assess the performance of the proposed approach with misclassified data under two settings: first where the sensitivity parameters match the parameters used in the data generation and second, where the sensitivity parameters are misspecified. Throughout the simulation, we compared the proposed approach which simultaneously accounts for misclassification and unmeasured confounding with the naïve model which ignores both and models that account for one type of bias, but not both.

4.3.1 Simulation Setting

The two sensitivity parameters are set to range from 0 to 0.75 incremented by 0.25, which yields common parameter values in logistic regression, yielding a total of 16 combinations. For each combination, 600 synthetic data sets are generated and the sample size of each data set is 400. Three observed covariates are generated from the standard normal distribution and the unmeasured confounder is generated from a Bernoulli distribution with 0.5 as probability of success. Exposure status and true response are generated from the following models:

$$\begin{aligned} \text{logit}(P(Z = 1|U, \mathbf{X})) &= 0.25 + 0.25X_1 + 0.25X_2 + 0.25X_3 + \zeta^z U, \\ \text{logit}(P(Y = 1|Z, U, \mathbf{X})) &= 0.4 + Z + 0.4X_1 + 0.4X_2 + 0.4X_3 + \zeta^y U. \end{aligned}$$

The values of the sensitivity and specificity are set as 0.7 and 0.9. The misclassified response is generated from a Bernoulli distribution with probability $P(\tilde{Y} = 1|Z, U, X)$,

where

$$P(\tilde{Y} = 1|Z, U, \mathbf{X}) = 0.7P(Y = 1|Z, U, \mathbf{X}) + (1 - 0.9)(1 - P(Y = 1|Z, U, \mathbf{X})).$$

The models used to analyze the data are as follows: first, no misclassification and no unmeasured confounding (M1); second, no misclassification but considering unmeasured confounding (M2); third, no unmeasured confounding but considering misclassification (M3); last, considering both misclassification and unmeasured confounding (M4). We compare four scenarios above in two simulation settings. First, correct sensitivity parameters are plugged in the simulation to assess the performance of four scenarios. Second, several false sensitivity parameters are used to investigate the effect of misspecification of sensitivity parameters. The true sensitivity parameters are 0.5 whereas false sensitivity parameters take values 0, 0.25 and 0.75 to show the cases that underestimate and overestimate the effect of unmeasured confounding. In the simulation, we specify the prior distribution for the coefficients in the exposure model to be $N(0, 10)$ and coefficients in the response model are assigned $\mathrm{N}(0, 2)$ priors. The sensitivity and specificity are assigned relatively informative priors, $\mathrm{Beta}(70, 30)$ and $\mathrm{Beta}(90, 10)$.

4.3.2 Simulation Result

Figure 4.3.1 and Figure 4.3.2 display simulation results that compare the four scenarios in terms of average absolute bias. Each figure presents four grids of sensitivity parameters which describe the associations of unmeasured confounder with exposure and response, in total of 16 combinations. In Figure 4.3.1, the sensitivity parameters are correctly specified as shown. For Figure 4.3.2, the data are generated with sensitivity parameters as 0.5 and misspecified sensitivity parameters shown in the figure are used in modeling. For each combination of sensitivity parameters, four boxplots are displayed and each boxplot corresponds to one scenario. In Figure 4.3.1, the red line indicates the median average absolute bias of M4 for each combination of sensitivity parameters. In Figure 4.3.2, the red reference line is the median average absolute bias of M4 with true sensitivity parameters as

0.5. Tables 4.1 and 4.2 present coverage probabilities of the 95% credible intervals for the exposure effect.

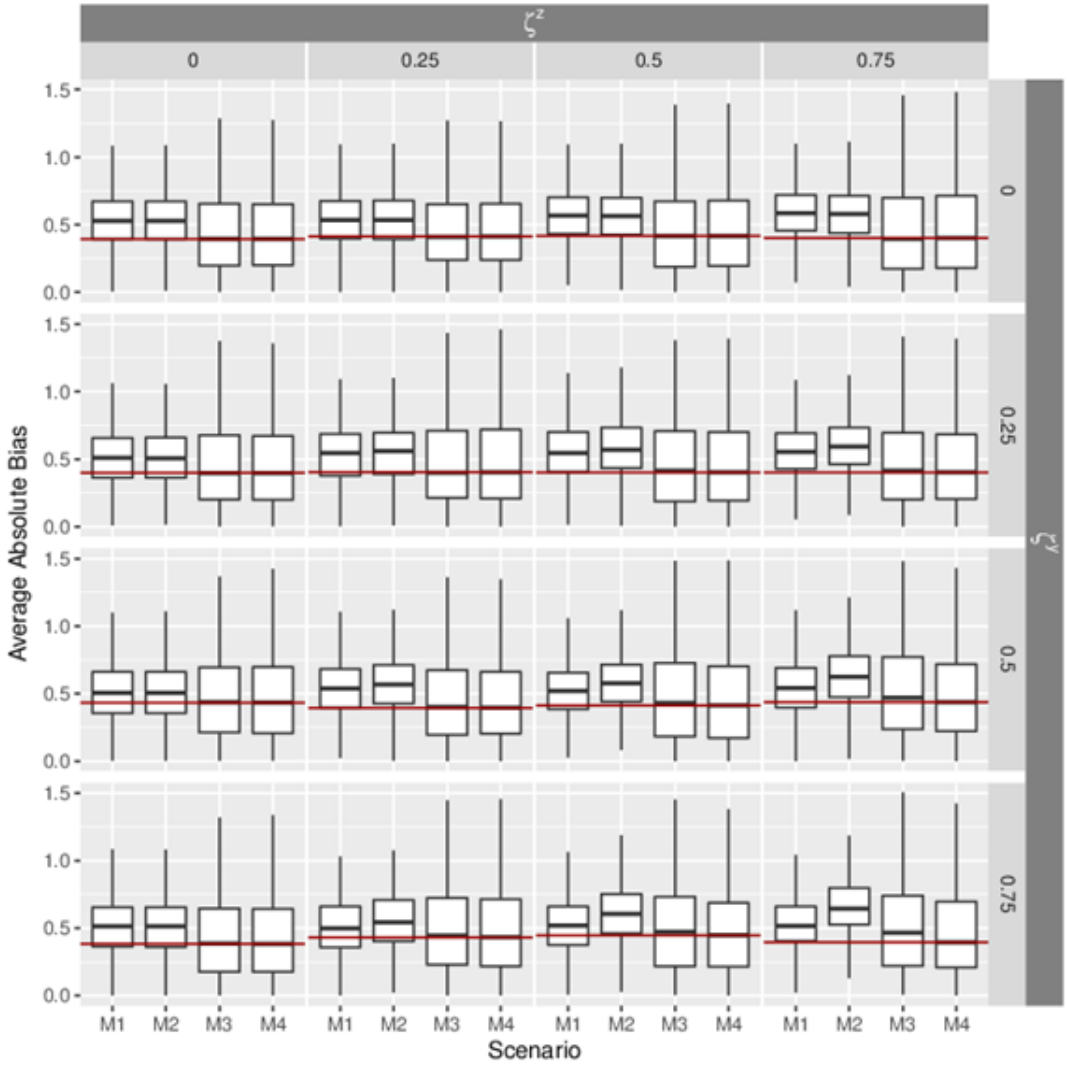

Figure 4.3.1: Average absolute bias for exposure effect estimation when sensitivity parameters are correctly specified in simulations.

In Figure 4.3.1 where the sensitivity parameters are correctly specified in the model, our approach incorporating misclassification and unmeasured confounding (M4) outperforms the other scenarios with the smallest median of average absolute bias when the sensitivity parameters are nonzero. As the magnitude of sensitivity parameters increase, incorporating unmeasured confounding reduces more bias than ignoring it. When one of

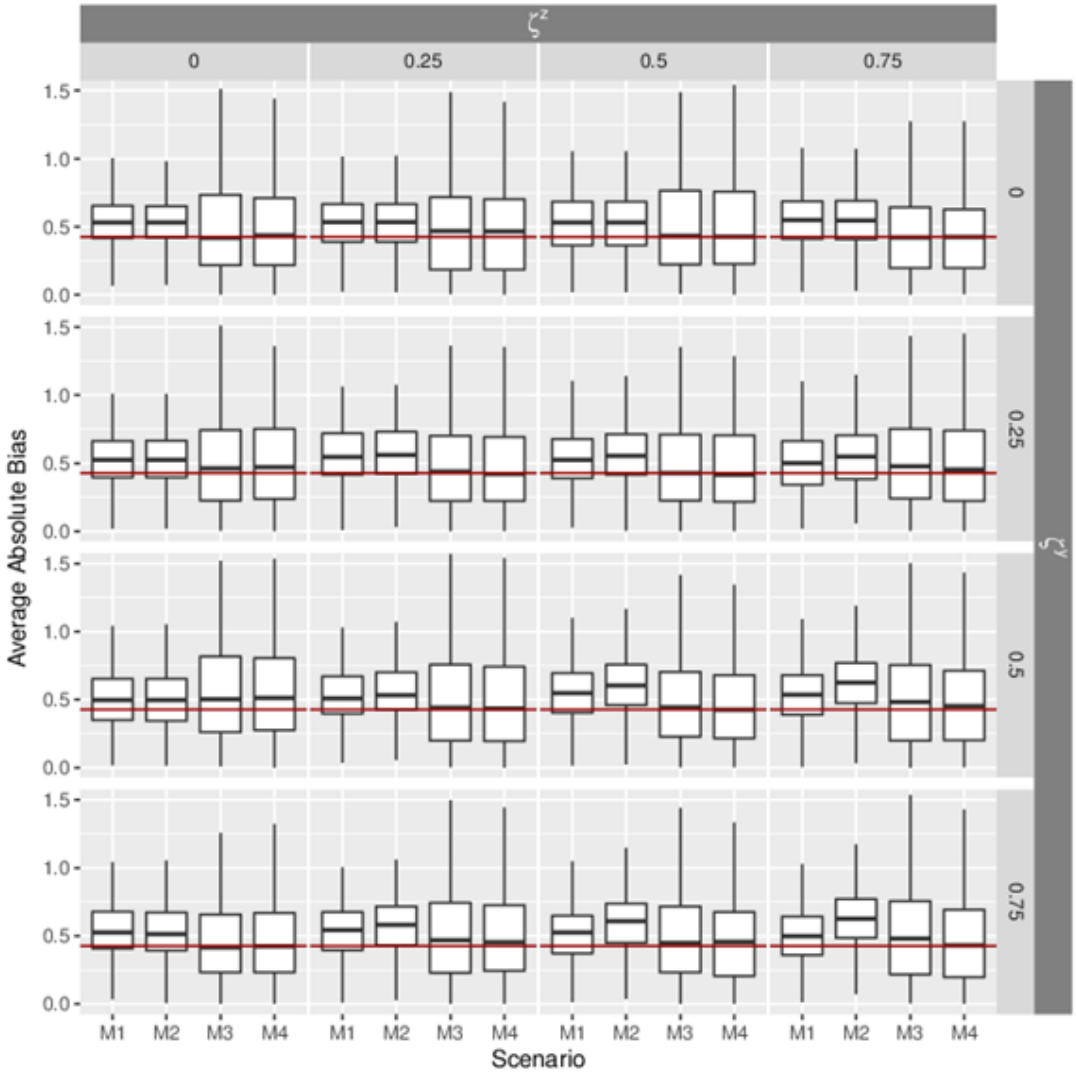

Figure 4.3.2: Average absolute bias of exposure effect estimation when true sensitivity parameters are 0.5. In simulation, sensitivity parameters used range from 0 to 0.75 with increment of 0.25.

Table 4.1: Coverage probability (%) of 95% posterior credible sets for exposure effect estimation when sensitivity parameters are correctly specified in simulations.

ζ^y	ζ^z	0	0.25	0.5	0.75
0	M1	29.3	33.5	33.8	35.2
	M2	28.3	33.0	33.2	34.3
	M3	97.7	96.5	96.5	96.0
	M4	97.8	96.7	96.7	96.2
0.25	M1	28.2	32.7	29.2	35.0
	M2	28.5	30.8	23.8	29.3
	M3	97.5	95.3	96.3	96.3
	M4	97.2	95.7	97.2	96.7
0.5	M1	21.8	26.8	32.0	32.0
	M2	23.5	24.0	23.0	19.3
	M3	97.7	97.8	97.0	95.7
	M4	97.7	98.5	97.2	96.2
0.75	M1	18.0	23.2	28.3	29.3
	M2	22.3	21.3	20.0	15.2
	M3	97.2	96.8	96.8	95.2
	M4	96.7	97.2	97.5	97.3

Table 4.2: Coverage probability (%) of 95% posterior credible sets for exposure effect estimation with mis-specified sensitivity parameters in simulations when true sensitivity parameters are 0.5.

ζ^y	ζ^z	0	0.25	0.5	0.75
0	M1	25.5	30.7	30.7	26.3
	M2	25.0	30.7	30.0	27.7
	M3	95.7	96.0	95.7	96.0
	M4	95.0	96.0	96.0	95.7
0.25	M1	30.3	26.3	31.3	35.7
	M2	29.0	24.7	26.3	30.7
	M3	95.7	97.7	95.3	96.3
	M4	95.7	97.7	96.0	97.0
0.5	M1	36.7	28.0	28.0	32.0
	M2	39.0	26.0	21.3	18.3
	M3	92.3	97.0	95.0	94.7
	M4	93.3	97.3	96.3	96.0
0.75	M1	27.0	28.0	31.0	33.3
	M2	32.0	25.7	22.0	20.7
	M3	98.0	95.3	96.0	93.0
	M4	98.0	96.0	96.3	96.3

sensitivity parameters is zero, ignoring the unmeasured confounder in the models does not inflate the bias for the two scenarios either ignoring or incorporating misclassification at the same time. Ignoring the misclassification yields considerable bias. Incorporating misclassification into the response model greatly reduces the average absolute bias and improves coverage probabilities to the nominal level. The reduction of bias by correcting for the misclassification error is greater than that due to the unmeasured confounding. Comparing scenarios M1 versus M3 and M2 versus M4, most bias in exposure effect estimation could be reduced by incorporating misclassification. The variability of the average absolute bias increases for scenarios correcting for misclassification errors because the estimate integrates uncertainty of the classification probabilities. In scenarios M1 and M2, ignoring misclassification error results in very low coverage probability. Both scenarios M3 and M4 which accommodate misclassification have coverage probability close to the nominal level. Scenario M3 which only considers misclassification is a very strong competitor to Scenario M4 which accounts for both misclassification and unmeasured confounding. Scenario M3 has slightly higher bias but closer coverage probability to the nominal level than scenario M4.

In Figure 4.3.2, we see that scenario leads to the least biased results compared with the other three scenarios even if one or two sensitivity parameters are misspecified. Interestingly, for the scenarios M3 and M4 which correct for misclassification, overestimating or underestimating the unmeasured confounding has little influence on bias and coverage probability.

4.4 Case Study

4.4.1 Data Description

The case study aims at investigating the effect of female employment on the likelihood of physical violence. The data come from the India National Health Survey (NFHS-3) 2005-6 which covers a range of health-related issues. We focus on currently married

women who are involved in the domestic violence module in urban areas. After dropping observations with missing variables, the sample is refined to 5573 non-working females and 14542 working females. The data contain a variety of demographic and socio-economic variables including the exposure, female employment status and the dichotomous response, spousal violence, which is subject to misclassification. The spousal violence indicator takes 1 if a woman's husband has harmed or threatened her physically and 0 otherwise. Exposure status takes value 1 or 0 to indicate a woman worked or not in the past 12 months. Education variables are number of years of education received. Number of children, adult men and adult women are measured for each household. Wealth index measures asset ownership at the household level. Low caste indicates whether the household head belongs to scheduled castes and scheduled tribes. Hindu is an indicator on whether the household head is Hindu. Remarriage indicates whether a woman has been married more than once. Polygamy is an indicator on whether a woman is in polygamous marriage. Table 4.3 reports the descriptive statistics of variables used in the analysis. A higher percentage of working women suffer physical violence from their spouse than non-working women. Working women tend to be older and have fewer years of education than non-working women. The age and education years of husbands for working women and non-working women show the same trend. On average, working women have more children and less adults in household. The household of working women are more likely to be poor, belong to low caste, and believe in Hindu. Higher percentage of working women married more than once and have a polygamous marriage.

In a recent study, Zhou et al. (2017) addresses misclassification in violence experience reporting and multilevel structure in the data but ignores the potential unmeasured confounding. As addressed earlier, it is unfeasible to test the ignorability assumption in observational studies, which is not often met. In the analysis, we intend to study the sensitivity

of the effect of female working status on spousal violence to the presence of misclassification and an unmeasured confounder.

Table 4.3: Descriptive statistics. Source: India National Family Health Survey (NFHS-3) 2005-6.

	Non-Exposure		Exposure	
	Mean	SD	Mean	SD
a) Response				
Physical violence	0.16	0.37	0.21	0.41
b) Pre-exposure characteristics				
b1) Demographic variables				
Female age	31.86	7.88	33.73	7.10
Female education	7.77	5.22	7.23	6.07
Male age	37.46	8.71	39.00	8.19
Male education	9.52	4.91	8.46	5.50
Number of children	2.16	1.39	2.28	1.43
Number of adult men	1.79	1.08	1.59	0.91
Number of adult women	1.61	0.89	1.53	0.77
b2) Socio-economic variables				
Wealth index	0.77	0.80	0.55	0.89
Indicator: low caste	0.49	0.50	0.60	0.49
b3) Cultural variables				
Indicator: Hindu	0.74	0.44	0.77	0.42
Indicator: remarriage	0.01	0.11	0.02	0.15
Indicator: polygamy	0.01	0.09	0.02	0.14
No. of observations	5573		14542	

4.4.2 Result

Figure 4.4.1 shows the sensitivity analysis for the effect of female employment on the likelihood of suffering physical violence over different values of the sensitivity parameters ranging from -2 to 2 by increments of 0.5. The left panel displays the results obtained without adjusting for misclassification in the response, while the right panel shows results of the same combination of sensitivity parameters when correcting for misclassification. The contour lines show the estimated odds ratio of suffering spousal violence for the working female to non-working female, for the levels of unmeasured confounding on the axes.

Red dots on the plots are points where levels of confounding cause the female employment effect estimates to be insignificant.

The left panel indicates the odds of a working female suffering physical violence is 1.2 times that of a non-working woman without adjusting for the binary unmeasured confounder. Across the hypothetical values of the sensitivity parameter, the effect of female working status changes substantially. For instance, when the sensitivity parameter in the exposure model is 0.5 and the sensitivity parameter in the response model is greater than 0, the unmeasured confounder can drive the exposure effect to be insignificant. Correcting for misclassification in the right panel, the odds ratio of experiencing physical violence for a working woman increases to 1.45, ignoring the unmeasured confounding. The effect would be insignificant if the unmeasured confounder has positive or negative associations with both the exposure and response. On the other hand, when the unmeasured confounder is positively associated with either exposure or response and negatively associated with the other, the odds ratio of suffering physical violence is elevated. Comparing the two figures, the odds ratio of suffering spousal violence increases after taking into account misclassification for the same combination of sensitivity parameter values. When one of sensitivity parameters is 0, unmeasured confounding does not change the results. Whether or not misclassification is accounted for, the exposure effect estimation is not robust to the violation of ignorability assumption.

These two plots demonstrate that the analysis of the influence of female employment on the likelihood of physical violence could substantially change depending on the levels of unmeasured confounding. Including misclassification correction in the analysis could also influence the female employment effect and its sensitivity to unmeasured confounding.

4.5 Discussion

In this paper, we propose a Bayesian sensitivity analysis for two sources of bias, misclassification and unmeasured confounding, in observational data. Our approach utilizes two sensitivity parameters to model the association of the unmeasured confounder

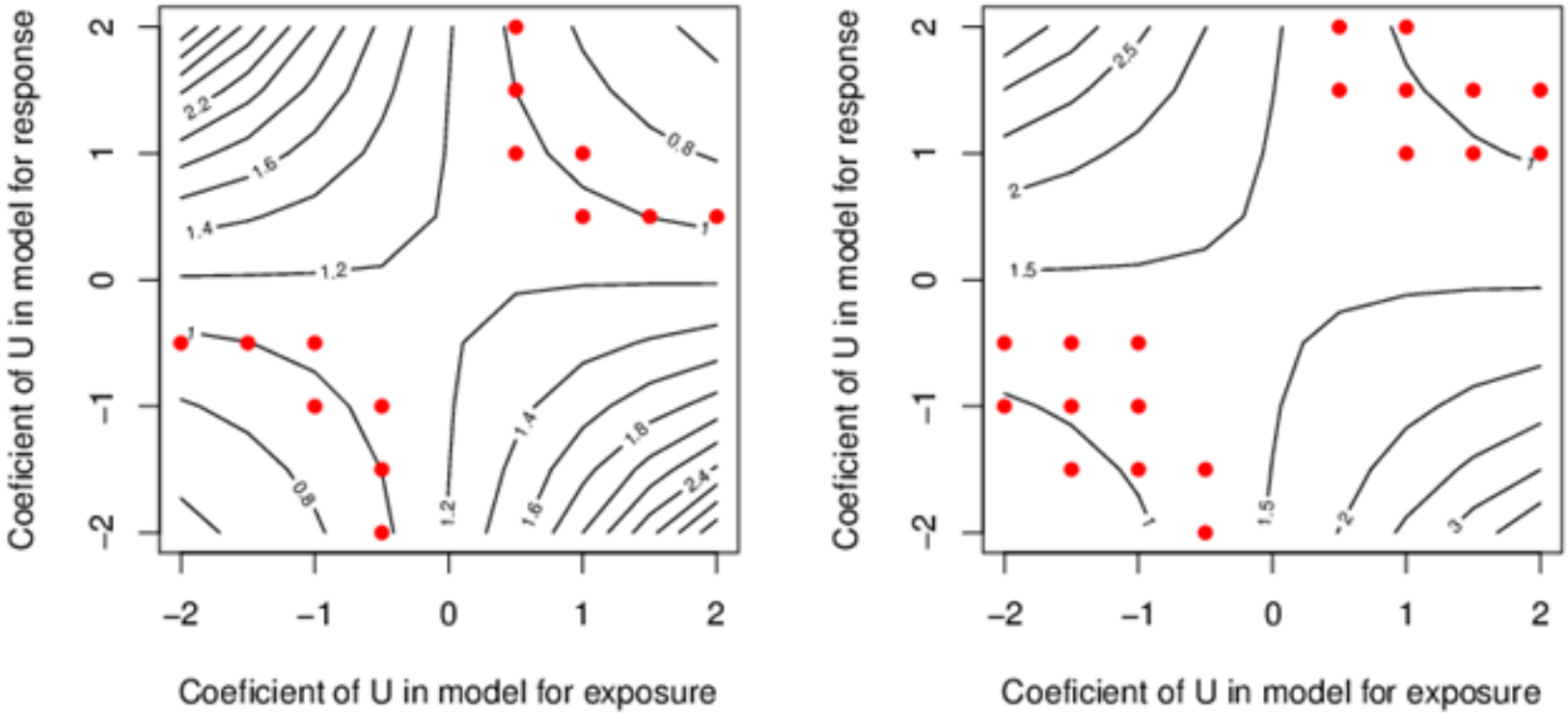

Figure 4.4.1: Left panel is the contour plot of effect of a woman working status on whether she had physical violence without correcting misclassification in response for sensitivity parameter ranging from -2 to 2 by grid of 0.5. The right panel is the contour plot of the exposure effect with misclassification correction component. Right dots are sensitivity parameter combinations which make exposure effect insignificant.

with the exposure status and with the response variable subject to misclassification. We evaluate the change in the exposure effect estimation with hypothetical levels of unmeasured confounding.

The proposed approach is illustrated with the study of the influence of female employment on the likelihood of suffering domestic violence. The results change substantially after correcting for errors in the response and are sensitive to a binary unmeasured confounder. The simulation study confirms the efficacy of the proposed method. Correcting for misclassification in the response and accounting for unmeasured confounding reduces bias in exposure effect estimation and the reduction of bias improves as the sensitivity parameters strengthen. Incorporating misclassification error appears plays to play a more important role in bias reduction and improving coverage probabilities to the nominal level.

In this study, the sensitivity parameters are set to be fixed. This avoids the issue of non-identifiability and lack of convergence in MCMC methods, but leads to difficulty in interpreting results from different combinations of the sensitivity parameters. In the future, we plan to specify priors on sensitivity parameters to get a possible range of change in estimation due to unmeasured confounding and incorporate the uncertainty in the sensitivity parameters.

The data in the motivating example has a clustered structure. Incorporating this structure removes cluster level confounding and ensures the validity of causal inference. It is of interest to extend our approach for cluster structured data. We can consider the cluster level unmeasured confounding and investigate the influence of multilevel modeling on exposure effect estimation.

CHAPTER FIVE

Conclusion

The dissertation proposes Bayesian propensity score analysis for clustered observational data investigates multiple source of bias: multilevel confounding and misclassification. Multilevel models are employed to account for cluster level confounding in propensity score analysis. Also, random coefficients in multilevel models borrow strength from the study population to improve the precision of estimates. The simulation study shows employing multilevel models in both propensity score and exposure effect estimation results in least absolute average bias. The reduction in bias increases as the association of cluster level confounding with response strengthens. Misclassification in response distort the exposure-response regression surface. Adjusting misclassified response greatly reduces the bias in estimates. If only one source of bias can be taken into account, it is better to incorporate misclassification in modeling. Correcting misclassification reduces more bias than incorporating multilevel structure and increases coverage rate of credible intervals to the nominal level.

A big challenge in inference for observational study that the unconfoundness assumption cannot be ensured. Sensitivity analysis envisions how analysis results would change due to unmeasured confounding. To perform the sensitivity analysis, assume the existence of an unmeasured confounder and specify the sensitivity parameters as regression coefficients to enter the inferential models. The sensitivity parameters controls the strength of associations of the unmeasured confounder with exposure status and response. Incorporating both misclassification and unmeasured confounding outputs the estimates with smallest median of average absolute bias. As the degree of unmeasured confounding gets stronger, the reduction in bias increases. If unmeasured confounding is only associated with exposure or response, unmeasured confounding do not interfere the inference.

For observational data, multilevel structure and misclassification are important source of bias to be accounted for in order to deliver reliable inference. The case study and simulations confirm the efficacy of the proposed Bayesian misclassification and propensity score methods for clustered observational data. The inference from propensity score analysis should be further investigated with sensitivity analysis to unmeasured confounding. If the estimates change substantially in existence of unmeasured confounder, the results are sensible to possible violation of unconfoundess assumption.

APPENDIX

APPENDIX A

R and JAGS code for Bayesian Multilevel Propensity Score Analysis

The code below are used for simulation in Bayesian multilevel propensity score analysis in chapter two.

```
###Chapter Two:Simulation for Bayesianl Multilevel  ###
###                Propensity Score                ###
###                  R Code                        ###
### # of replications: nr
### # of cluster: nc = 20
### cluster size: ns = 25
### individual level: c1,c2 ~ N(0,1)
### cluster level: z ~ N(0,1)
### treatment assignment: x
### outcome: y

library(rjags)
library(R2jags)
load.module("dic")
nc = 20
ns = 25
n = nc*ns
set.seed(111)
r0 = rep(rnorm(nc,0,0.5), times = ns)
r1 = 0.5
r2 = 0.5
```

```
r3 = 1
b0 = rep(rnorm(nc,0,2), times = ns)
b1 = 1
b2 = 1
b3 = 1
b4 = 3
b5 = 0
###nr = 50
###re1 <- vector("list", nr)
sim <- function(){
c_ind <- rep(1:nc, times = ns)
c1 <- rnorm(n,0,1)
c2 <- rnorm(n,0,1)
###  z uncorrelated with c1,c2
z_c <- rnorm(nc,0,1)
z <- rep(z_c, times = ns)
ps <- exp(r0+r1*c1+r2*c2+r3*z)/
      (1+exp(r0+r1*c1+r2*c2+r3*z))
quat <- exp(sqrt(0.5^2+r1^2+r2^2+r3^2)*
qnorm(c(0.2,0.4,0.6,0.8)))/(1+exp(sqrt
(0.5^2+r1^2+r2^2+r3^2)*qnorm(c(0.2,0.4,0.6,0.8))))
x <- rbinom(n,1,ps)
y <- b0+b1*x+b2*c1+b3*c2+b4*z+b5*x*z
fit1 <- jags(data = data1, inits = inits1, params1,
n.chain = 2, n.burnin = 5000,  n.iter = 10000,n.thin=1,
model.file = m1)
fit2 <- jags(data = data2, inits = inits2, params2,
```

```
n.chain = 2, n.burnin = 5000,  n.iter = 10000,n.thin=1,
model.file = m2)
fit3 <- jags(data = data3, inits = inits3, params3,
n.chain = 2, n.burnin = 5000, n.iter = 10000,n.thin=1,
model.file = m3)
fit4 <- jags(data = data4, inits = inits4, params4,
n.chain = 2, n.burnin = 5000, n.iter = 10000,n.thin=1,
model.file = m4)
fit5 <- jags(data = data5, inits = inits5, params5,
n.chain = 2, n.burnin = 5000, n.iter = 10000,n.thin=1,
model.file = m5)
fit6 <- jags(data = data6, inits = inits6, params6,
n.chain = 2, n.burnin = 5000, n.iter = 10000,n.thin=1,
model.file = m6)
fit7 <- jags(data = data7, inits = inits7, params7,
n.chain = 2, n.burnin = 5000, n.iter = 10000,n.thin=1,
model.file = m7)
fit8 <- jags(data = data8, inits = inits8, params8,
n.chain = 2, n.burnin = 5000, n.iter = 10000,n.thin=1,
model.file = m8)
return(list(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8))
}
re <- replicate(50, sim())

###                  JAGS Code                        ###
###  PS0 + Out1
m1 <- function(){
```

```
for (i in 1:n)
    {
     x[i] ~ dbin(p[i],1)
     logit(p[i]) <- r0 + r[1]*c1[i] + r[2]*c2[i]

     y[i] ~ dnorm(mu[i], tau.l)
     g2[i] <- step(p[i]-quat[1])-step(p[i]-quat[2])
     g3[i] <- step(p[i]-quat[2])-step(p[i]-quat[3])
     g4[i] <- step(p[i]-quat[3])-step(p[i]-quat[4])
     g5[i] <- step(p[i]-quat[4])
     mu[i] <- delta*x[i]+eta[1]+eta[2]*g2[i]+
        eta[3]*g3[i]+eta[4]*g4[i]+eta[5]*g5[i]
    }

r0 ~ dnorm(0, 0.1)
for (j in 1:2)
    {
     r[j] ~ dnorm(0, 0.1)
    }
tau.l ~  dunif(0.01, 100)
sig.l <- 1/tau.l
delta ~ dnorm(0, 0.01)
or.delta <- exp(delta)
for (k in 1:5)
    {
     eta[k] ~ dnorm(0, 0.01)
    }
```

```
}
data1 <- list("x", "c1", "c2", "quat", "y", "n")
inits1 <- function(){
list(r0=rnorm(1),r=rnorm(2),delta=rnorm(1),
eta=rnorm(5),tau.l=dunif(1))}
params1 <- c("r0", "r", "delta", "or.delta", "eta",
 "sig.l")
###  PS1 + Out1
m2 <- function(){
  for (i in 1:n)
      {
       x[i] ~ dbin(p[i],1)
       logit(p[i]) <- r0+r[1]*c1[i]+r[2]*c2[i]+r[3]*z[i]

       y[i] ~ dnorm(mu[i], tau.l)
       g2[i] <- step(p[i]-quat[1])-step(p[i]-quat[2])
       g3[i] <- step(p[i]-quat[2])-step(p[i]-quat[3])
       g4[i] <- step(p[i]-quat[3])-step(p[i]-quat[4])
       g5[i] <- step(p[i]-quat[4])
       mu[i] <- delta*x[i]+eta[1]+eta[2]*g2[i]+
                eta[3]*g3[i]+eta[4]*g4[i]+eta[5]*g5[i]
      }
  r0 ~ dnorm(0, 0.1)
  for (j in 1:3)
      {
       r[j] ~ dnorm(0, 0.1)
      }
```

```
  tau.l ~  dunif(0.01, 100)
  sig.l <- 1/tau.l
  delta ~ dnorm(0, 0.01)
  or.delta <- exp(delta)
  for (k in 1:5)
      {
       eta[k] ~ dnorm(0, 0.01)
      }
}
data2 <- list("x", "c1", "c2", "z", "quat", "y",
 "n")
inits2 <- function(){
 list(r0=rnorm(1),r=rnorm(3),delta=rnorm(1),
 eta=rnorm(5),tau.l=dunif(1))}
params2 <- c("r0", "r", "delta", "or.delta", "eta",
 "sig.l")

###  PS2 + Out1
m3 <- function(){
  for (i in 1:n)
      {
       x[i] ~ dbin(p[i],1)
       logit(p[i])<-r0[c_ind[i]]+r[1]*c1[i]+r[2]*c2[i]
       y[i] ~ dnorm(mu[i], tau.l)
       g2[i] <- step(p[i]-quat[1])-step(p[i]-quat[2])
       g3[i] <- step(p[i]-quat[2])-step(p[i]-quat[3])
       g4[i] <- step(p[i]-quat[3])-step(p[i]-quat[4])
```

```
        g5[i] <- step(p[i]-quat[4])
        mu[i] <- delta*x[i]+eta[1]+eta[2]*g2[i]+
                 eta[3]*g3[i]+eta[4]*g4[i]+eta[5]*g5[i]
       }
  for (h in 1:nc)
     {
      r0[h] ~ dnorm(m1.r, tau.r)
      r0.adj[h] <- r0[h] - mean(r0[])
     }
  m1.r ~ dnorm(0, 0.1)
  tau.r ~  dunif(0.1, 10)
  sig.r <- 1/tau.r
  for (j in 1:2)
      {
       r[j] ~ dnorm(0, 0.1)
      }
  tau.l ~  dunif(0.01, 100)
  sig.l <- 1/tau.l
  delta ~ dnorm(0, 0.01)
  or.delta <- exp(delta)
  for (k in 1:5)
      {
       eta[k] ~ dnorm(0, 0.01)
      }
}
data3 <- list("x", "c1", "c2", "quat", "y", "n", "nc",
 "c_ind")
```

```
inits3 <- function(){
 list(r0 = rnorm(nc), r = rnorm(2), delta = rnorm(1),
 eta= rnorm(5), tau.l = dunif(1), m1.r = rnorm(1),
 tau.r = dunif(1))}
params3 <- c("r0", "r", "delta", "or.delta", "eta",
 "m1.r", "sig.r", "sig.l")
###  PS3 + Out1
m4 <- function(){
  for (i in 1:n)
      {
       x[i] ~ dbin(p[i],1)
       logit(p[i])<-r0[c_ind[i]]+r[1]*c1[i]+r[2]*c2[i]
       y[i] ~ dnorm(mu[i], tau.l)
       g2[i] <- step(p[i]-quat[1])-step(p[i]-quat[2])
       g3[i] <- step(p[i]-quat[2])-step(p[i]-quat[3])
       g4[i] <- step(p[i]-quat[3])-step(p[i]-quat[4])
       g5[i] <- step(p[i]-quat[4])
       mu[i] <- delta*x[i]+eta[1]+eta[2]*g2[i]+
         eta[3]*g3[i]+eta[4]*g4[i]+eta[5]*g5[i]
      }
  for (h in 1:nc)
     {
      r0[h] ~ dnorm(m[h], tau.r)
      m[h] <- b0 + r[3]*z_c[h]
     }
  b0 ~ dnorm(0, 0.1)
  tau.r ~ dunif(0.1, 10)
```

```
  sig.r <- 1/tau.r
  for (j in 1:3)
      {
       r[j] ~ dnorm(0, 0.1)
      }
  tau.l ~ dunif(0.01, 100)
  sig.l <- 1/tau.l
  delta ~ dnorm(0, 0.01)
  for (k in 1:5)
      {
       eta[k] ~ dnorm(0, 0.01)
      }
  or.delta <- exp(delta)
}
data4 <- list("x", "c1", "c2", "z_c", "quat", "y", "n",
 "nc","c_ind")
inits4 <- function(){
list(r0 = rnorm(nc), b0 = rnorm(1), r = rnorm(3),
delta = rnorm(1), eta= rnorm(5), tau.l = dunif(1),
tau.r = dunif(1))}

params4 <- c("r0", "b0", "r", "delta", "or.delta",
 "eta", "sig.r", "sig.l")

###  PS0 + Out2
m5 <- function(){
  for (i in 1:n)
```

```
    {
     x[i] ~ dbin(p[i],1)
     logit(p[i])<-r0+r[1]*c1[i]+r[2]*c2[i]
     y[i] ~ dnorm(mu[i], tau.l)
     g2[i] <- step(p[i]-quat[1])-step(p[i]-quat[2])
     g3[i] <- step(p[i]-quat[2])-step(p[i]-quat[3])
     g4[i] <- step(p[i]-quat[3])-step(p[i]-quat[4])
     g5[i] <- step(p[i]-quat[4])
     mu[i] <- delta*x[i]+eta0[c_ind[i]]+eta[1]*g2[i]
     +eta[2]*g3[i]+eta[3]*g4[i]+eta[4]*g5[i]
    }
r0 ~ dnorm(0, 0.1)
for (j in 1:2)
    {
     r[j] ~ dnorm(0, 0.1)
    }
tau.l ~  dunif(0.01, 100)
sig.l <- 1/tau.l
delta ~ dnorm(0, 0.01)
for (h in 1:nc)
    {
     eta0[h] ~ dnorm(m1.e, tau.e)
     eta0.adj[h] <- eta0[h] - mean(eta0[])
    }
m1.e ~ dnorm(0, 0.1)
tau.e ~ dunif(0.01, 10)
sig.e <- 1/tau.l
```

```
  for (k in 1:4)
      {
       eta[k] ~ dnorm(0, 0.01)
      }
  or.delta <- exp(delta)
}
data5 <- list("x", "c1", "c2", "quat", "y", "n",
 "nc", "c_ind")
inits5 <- function(){
 list(r0 = rnorm(1), r = rnorm(2), delta = rnorm(1),
 eta0 = rnorm(nc), eta= rnorm(4), tau.l = dunif(1),
 m1.e = rnorm(1), tau.e = dunif(1))}
params5 <- c("r0", "r", "delta", "or.delta", "eta0",
"eta", "m1.e", "sig.e", "sig.l")
###  PS1 + Out2
m6 <- function(){
  for (i in 1:n)
      {
       x[i] ~ dbin(p[i],1)
       logit(p[i])<-r0+r[1]*c1[i]+r[2]*c2[i]+r[3]*z[i]
       y[i] ~ dnorm(mu[i], tau.l)
       g2[i]<-step(p[i]-quat[1])-step(p[i]-quat[2])
       g3[i]<-step(p[i]-quat[2])-step(p[i]-quat[3])
       g4[i]<-step(p[i]-quat[3])-step(p[i]-quat[4])
       g5[i]<-step(p[i]-quat[4])
       mu[i]<-delta*x[i]+eta0[c_ind[i]]+eta[1]*g2[i]+
              eta[2]*g3[i]+eta[3]*g4[i]+eta[4]*g5[i]
```

```
      }
  r0 ~ dnorm(0, 0.1)
  for (j in 1:3)
      {
       r[j] ~ dnorm(0, 0.1)
      }
  tau.l ~ dunif(0.01, 100)
  sig.l <- 1/tau.l
  delta ~ dnorm(0, 0.01)
  for (h in 1:nc)
      {
       eta0[h] ~ dnorm(m1.e, tau.e)
       eta0.adj[h] <- eta0[h] - mean(eta0[])
      }
  m1.e ~ dnorm(0, 0.1)
  tau.e ~ dunif(0.01, 10)
  sig.e <- 1/tau.l
  for (k in 1:4)
      {
       eta[k] ~ dnorm(0, 0.01)
      }
  or.delta <- exp(delta)
}
data6 <- list("x", "c1", "c2", "z", "quat", "y", "n",
              "nc", "c_ind")
inits6 <- function(){
         list(r0 = rnorm(1), r = rnorm(3), delta = rnorm(1),
```

```
        eta0 = rnorm(nc), eta= rnorm(4), tau.l = dunif(1),
        m1.e = rnorm(1), tau.e = dunif(1))}
params6 <- c("r0", "r", "delta", "or.delta", "eta", "eta0",
            "m1.e", "sig.e", "sig.l")
###  PS2 + Out2
m7 <- function(){
  for (i in 1:n)
      {
       x[i] ~ dbin(p[i],1)
       logit(p[i]) <- r0[c_ind[i]]+r[1]*c1[i]+r[2]*c2[i]

       y[i] ~ dnorm(mu[i], tau.l)
       g2[i] <- step(p[i]-quat[1])-step(p[i]-quat[2])
       g3[i] <- step(p[i]-quat[2])-step(p[i]-quat[3])
       g4[i] <- step(p[i]-quat[3])-step(p[i]-quat[4])
       g5[i] <- step(p[i]-quat[4])
       mu[i] <- delta*x[i]+eta0[c_ind[i]]+eta[1]*g2[i]+
                eta[2]*g3[i]+eta[3]*g4[i]+eta[4]*g5[i]
      }
  for (h in 1:nc)
     {
      r0[h] ~ dnorm(m1.r, tau.r)
      r0.adj[h] <- r0[h] - mean(r0[])
      eta0[h] ~ dnorm(m1.e, tau.e)
      eta0.adj[h] <- eta0[h] - mean(eta0[])
     }
  m1.r ~ dnorm(0, 0.1)
```

```
  tau.r ~ dunif(0.1, 10)
  sig.r <- 1/tau.r
  m1.e ~ dnorm(0, 0.1)
  tau.e ~ dunif(0.01, 100)
  sig.e <- 1/tau.e
  for (j in 1:2)
      {
       r[j] ~ dnorm(0, 0.1)
      }
  tau.l ~ dunif(0.01, 100)
  sig.l <- 1/tau.l
  delta ~ dnorm(0, 0.01)
  for (k in 1:4)
      {
       eta[k] ~ dnorm(0, 0.01)
      }
  or.delta <- exp(delta)
}
data7 <- list("x", "c1", "c2", "quat", "y", "n", "nc",
             "c_ind")
inits7 <- function(){
 list(r0 = rnorm(nc), r = rnorm(2), delta = rnorm(1),
 eta0 = rnorm(nc), eta= rnorm(4), m1.r = rnorm(1),
 tau.r=dunif(1), tau.l = dunif(1), m1.e = rnorm(1),
 tau.e = dunif(1))}
params7 <- c("r0", "r", "delta", "or.delta", "eta0",
"eta", "m1.r", "sig.r", "m1.e", "sig.e", "sig.l")
```

```
###  PS3 + Out2
m8 <- function(){
  for (i in 1:n)
      {
       x[i] ~ dbin(p[i],1)
       logit(p[i]) <- r0[c_ind[i]]+r[1]*c1[i]+r[2]*c2[i]
       y[i] ~ dnorm(mu[i], tau.l)
       g2[i] <- step(p[i]-quat[1])-step(p[i]-quat[2])
       g3[i] <- step(p[i]-quat[2])-step(p[i]-quat[3])
       g4[i] <- step(p[i]-quat[3])-step(p[i]-quat[4])
       g5[i] <- step(p[i]-quat[4])
       mu[i] <- delta*x[i]+eta0[c_ind[i]]+eta[1]*g2[i]
                +eta[2]*g3[i]+eta[3]*g4[i]+eta[4]*g5[i]
      }
  for (h in 1:nc)
     {
      r0[h] ~ dnorm(m[h], tau.r)
      m[h] <- b0 + r[3]*z_c[h]
      eta0[h] ~ dnorm(m1.e, tau.e)
      eta0.adj[h] <- eta0[h] - mean(eta0[])
     }
  b0 ~ dnorm(0, 0.1)
  tau.r ~ dunif(0.1, 10)
  sig.r <- 1/tau.r
  m1.e ~ dnorm(0, 0.1)
  tau.e ~ dunif(0.01, 10)
  sig.e <- 1/tau.e
```

```
  for (j in 1:3)
      {
       r[j] ~ dnorm(0, 0.1)
      }
  tau.l ~ dunif(0.01, 100)
  sig.l <- 1/tau.l
  delta ~ dnorm(0, 0.01)
  for (k in 1:4)
      {
       eta[k] ~ dnorm(0, 0.01)
      }
  or.delta <- exp(delta)
}
data8 <- list("x", "c1", "c2", "z_c", "quat", "y",
              "n", "nc", "c_ind")
inits8 <- function(){
 list(r0 = rnorm(nc), b0 = rnorm(1), r = rnorm(3),
 delta = rnorm(1),eta0 = rnorm(nc), eta= rnorm(4),
 tau.r = dunif(1), tau.l = dunif(1),
 m1.e = rnorm(1), tau.e = dunif(1))}
params8 <- c("r0", "b0", "r", "delta", "or.delta",
"eta0", "eta", "sig.r", "m1.e", "sig.e", "sig.l")
```

APPENDIX B

R and JAGS code for Bayesian Multilevel Propensity Score Analysis with misclassified response

The code below is for Bayesian multilevel propensity score analysis with misclassified response in Chapter Three. Data generation is performed in R and models are fitted in JAGS.

```
###Chapter Three:Simulation for Bayesianl Multilevel   ###
###Propensity Score Analysis with Miclassified Response###
###                        R Code                      ###
### multilevel, misclassification simulation
### # of replications: nr
### # of cluster: nc = 10
### cluster size: ns = 100
### individual level: c1,c2 ~ N(0,1)
### cluster level: z ~ N(0,1)
### treatment assignment: x
### outcome: y
### se: sensitivity
### sp: specificity
se = 0.53
sp = 0.98
nr = 100
nc = 10
ns = 100
n = nc*ns
```

```
set.seed(1234)
r0 = rep(rnorm(nc), times = ns)
r1 = 0.3
r2 = 0.4
r3 = 0.5
set.seed(4321)
b0 = rep(rnorm(nc), times = ns)
b1 = 1
b2 = 0.3
b3 = 0.3
b4 = 0.7
b5 = 0
library(rjags)
library(R2jags)
load.module("dic")
sim <- function()
{
c_ind <- rep(1:nc, times = ns)
c1 <- rnorm(n,0,1)
c2 <- rnorm(n,0,1)
###  z uncorrelated with c1,c2
z_c <- rnorm(nc,0,1)
z <- rep(z_c, times = ns)
ps <- exp(r0+r1*c1+r2*c2+r3*z)/(1+exp(r0+r1*c1+r2*c2+r3*z))
x <- rbinom(n,1,ps)
tos <- exp(b0+b1*x+b2*c1+b3*c2+b4*z+b5*x*z)/
(1+exp(b0+b1*x+b2*c1+b3*c2+b4*z+b5*x*z))
```

```
os <- tos*se + (1-sp)*(1-tos)
### observed y
y <- rbinom(n,1,os)
fit1 <- jags(data=bugs.data1, inits=bugs.inits1,
bugs.params1, n.chain=2, n.burnin=1000, n.iter=6000,
n.thin=1, model.file=m1)
fit2 <- jags(data=bugs.data2, inits=bugs.inits2,
bugs.params2, n.chain=2, n.burnin=1000, n.iter=6000,
n.thin=1, model.file=m2)
fit3 <- jags(data=bugs.data3, inits=bugs.inits3,
bugs.params3, n.chain=2, n.burnin=1000, n.iter=6000,
n.thin=1, model.file=m3)
fit4 <- jags(data=bugs.data4, inits=bugs.inits4,
bugs.params4, n.chain=2, n.burnin=1000, n.iter=6000,
n.thin=1, model.file=m4)
return(list(fit1$BUGSoutput$summary,fit2$BUGSoutput$summary,
fit3$BUGSoutput$summary,fit4$BUGSoutput$summary))
}
system.time(re1 <- replicate(50,sim()))

###                  JAGS Code                          ###
### no misclassification, no multilevel
m1 <- function(){
  for (i in 1:n)
      {
        x[i] ~ dbin(p[i],1)
        logit(p[i]) <- r0 + r[1]*c1[i] + r[2]*c2[i]
```

```
        y[i] ~ dbin(op[i], 1)
        logit(op[i]) <- b0 + b[1] * x[i] + b[2]*p[i]
        }
  r0 ~ dnorm(0, 0.1)
  for (j in 1:2)
      {
       r[j] ~ dnorm(0, 0.1)
      }
  b0 ~ dnorm(0, 0.1)
  b[1] ~ dnorm(0,0.5)
  b[2] ~ dnorm(0,0.1)
or.trt<-exp(b[1])
}
bugs.data1 <- list("x", "c1", "c2", "y", "n")
bugs.params1 <- c("r0", "r", "b0", "b", "or.trt")
bugs.inits1 <- function(){
 list(r0 = rnorm(1), r = rnorm(2), b0 = rnorm(1),
 b= rnorm(2))
}
### misclassification, no multilevel                ###
m2 <- function(){
  for (i in 1:n)
      {
        x[i] ~ dbin(p[i],1)
        logit(p[i]) <- r0 + r[1]*c1[i] + r[2]*c2[i]
        y[i] ~ dbin(op[i], 1)
       logit(tp[i]) <- b0 + b[1] * x[i] + b[2]*p[i]
```

```
      op[i]<- se1*tp[i] +(1-sp1)*(1-tp[i])
      }
  ##se1 ~ dbeta(u1+0.5,50.5-u1)
  ##sp1 ~ dbeta(50.5-u2,u2+0.5)
  se1 ~ dbeta(5,7.6)
  sp1 ~ dbeta(165.7,9.7)
  r0 ~ dnorm(0, 0.1)
  for (j in 1:2)
      {
       r[j] ~ dnorm(0, 0.1)
      }
  b0 ~ dnorm(0, 0.1)
  b[1] ~ dnorm(0,0.5)
  b[2] ~ dnorm(0,0.1)
or.trt<-exp(b[1])
}
bugs.data2 <- list("x", "c1", "c2", "y", "n")
bugs.params2 <- c("r0", "r", "b0", "b", "or.trt",
                  "se1", "sp1")
bugs.inits2_1 <- list(r0 = rnorm(1), r = rnorm(2),
 b0 = rnorm(1), b=c(rnorm(1),1), se1 = 0.53, sp1 = 0.98)
bugs.inits2_2 <- list(r0 = rnorm(1), r = rnorm(2),
b0 = rnorm(1),b=c(rnorm(1),0.98), se1 = 0.51, sp1 = 0.99)
bugs.inits2 <- list(bugs.inits2_1,bugs.inits2_2)
### multilevel, no misclassification                ###
m3 <- function(){
  for (i in 1:n)
```

```
      {
        x[i] ~ dbin(p[i],1)
        logit(p[i]) <- r0[c_ind[i]] + r[1]*c1[i] + r[2]*c2[i]
        y[i] ~ dbin(op[i], 1)
        logit(op[i]) <- b0[c_ind[i]] + b[1]*x[i] + b[2]*p[i]
      }
  for (l in 1:nc)
     {
     r0[l] ~ dnorm(mu1[l], tau1)
     mu1[l] <- a[1] + a[2]*z_c[l]
     b0[l] ~ dnorm(mu2[l], tau2)
     mu2[l] <- d
     }
  tau1 ~ dgamma(0.001,0.001)
  tau2 ~ dgamma(0.001,0.001)
  d ~ dnorm(0, 0.1)
  for (j in 1:2)
      {
       r[j] ~ dnorm(0, 0.1)
       a[j] ~ dnorm(0, 0.1)
      }
  b[1] ~ dnorm(0,0.5)
  b[2] ~ dnorm(0,0.1)
or.trt<-exp(b[1])
}
bugs.data3 <- list("x", "c1", "c2", "y", "n", "nc",
                   "z_c", "c_ind")
```

```
bugs.params3 <- c("a", "r", "d", "b","or.trt", "tau1",
                  "tau2")

bugs.inits3 <- function(){
 list(a = rnorm(2), r = rnorm(2), d = rnorm(1),
 b= rnorm(2), tau1 = runif(1), tau2 = runif(1))
}
### multilevel, misclassification                    ###
m4 <- function(){
  for (i in 1:n)
      {
       x[i] ~ dbin(p[i],1)
       logit(p[i]) <- r0[c_ind[i]] + r[1]*c1[i] + r[2]*c2[i]
       y[i] ~ dbin(op[i], 1)
       logit(tp[i]) <- b0[c_ind[i]] + b[1]*x[i] + b[2]*p[i]
       op[i] <- se1*tp[i] + (1-sp1)*(1-tp[i])
       }
  ##se1 ~ dbeta(u1+0.5,50.5-u1)
  ##sp1 ~ dbeta(50.5-u2,u2+0.5)
  se1 ~ dbeta(5,7.6)
  sp1 ~ dbeta(165.7,9.7)
  for (l in 1:nc)
     {
     r0[l] ~ dnorm(mu1[l], tau1)
     mu1[l] <- a[1] + a[2]*z_c[l]
     ##r0.adj[l] <- r0[l] - mean(r0[])
     b0[l] ~ dnorm(mu2[l], tau2)
```

```
      mu2[l] <- d
      ##b0.adj[l] <- b0[l] - mean(b0[])
      }
  tau1 ~ dgamma(0.001,0.001)
  tau2 ~ dgamma(0.001,0.001)
  for (j in 1:2)
      {
       r[j] ~ dnorm(0, 0.1)
       a[j] ~ dnorm(0, 0.1)
      }
  d ~ dnorm(0, 0.1)
  b[1] ~ dnorm(0,0.5)
  b[2] ~ dnorm(0,0.1)
  or.trt<-exp(b[1])
}
bugs.data4 <- list("x", "c1", "c2", "y","n", "nc", "c_ind",
                   "z_c")
bugs.params4 <- c("a", "r", "b", "d", "or.trt", "se1", "sp1",
                  "tau1", "tau2")
bugs.inits4_1 <- list(a = rnorm(2), d=rnorm(1), r = rnorm(2),
                      b=c(rnorm(1),1),
se1 = 0.53, sp1 = 0.98,tau1=0.5, tau2=0.8)
bugs.inits4_2 <- list(a = rnorm(2), d=dnorm(1), r = rnorm(2),
                      b=c(rnorm(1),0.99),
se1 = 0.52, sp1 = 0.99,tau1=1,tau2=1)
bugs.inits4 <- list(bugs.inits4_1,bugs.inits4_2)
```

APPENDIX C

R and JAGS code for Bayesian Sensitivity Analysis with misclassified response

The code below is for Bayesian sensitivity Analysis to unmeasured confounding in chapter four. In the simulation, R code is for data generation and JAGS code is for model fitting.

```
### Chapter Four:Simulation for Bayesianl Sensitivity ###
### Analysis to Unmeasured Confounding                ###
###                        R Code                     ###
###sensitivity of unmeasured confounding+misclassification
###propensity score model:
###logit(P(z=1))= r0+r1*x1+r2*x2+r3*x3+r*u
###outcome model:logit(P(y=1))= b0+b1*x1+b2*x2+b3*x3+b4*z+b*u
###misclassification:P(y_hat = 1)=se*P(y=1)+(1-sp)*(1-P(y=1))
###pretend U is unknown
### sample size
n = 400
### paramters in outcome model
b0 = 0.4
b1 = 0.4
b2 = 0.4
b3 = 0.4
### trt ###
b4 = 1
### association between y and u
###b
```

```
### paramters in propensity score model
r0 =0.25
r1 = 0.25
r2 = 0.25
r3 = 0.25
### association between z and u
###r
### hyperparamter for u
pu = 0.5
### misclassification sensitivity and specificity
se = 0.7
sp = 0.9
### jags model###
##se.model1:no misclassification,
##no unmeasured confounder U
##se.model:no misclassification,
##with unmeasured confounder U
##with prior U~dbin(pu,1)
##se.mis.model1:with misclassification,
##no unmeasured confounder U
##se.mis.model:with misclassification,
##with unmeasured confounder U
##with prior U~dbin(pu,1)
### no mis no U ###
se.model1 <- function(){
  for (i in 1:n)
      {
```

```
        z[i] ~ dbin(p[i],1)
        logit(p[i])<-r[1]+r[2]*x1[i]+r[3]*x2[i]+r[4]*x3[i]
        y[i] ~ dbin(tp[i], 1)
        logit(tp[i])<-delta*z[i]+b[1]+b[2]*x1[i]+b[3]*x2[i]
                      +b[4]*x3[i]
        }
  for (j in 1:4)
      {
        r[j] ~ dnorm(0, 0.1)
        b[j] ~ dnorm(0, 0.5)
      }
  delta ~ dnorm(0, 0.5)
  or.delta<-exp(delta)
}
params.se1 <- c("r", "b", "delta", "or.delta")
inits.se1 <- function(){
list(r=rnorm(4), b=rnorm(4), delta=rnorm(1))
}
### no mis ###
se.model <- function(){
  for (i in 1:n)
      {
        u[i] ~ dbin(pu,1)
        z[i] ~ dbin(p[i],1)
        logit(p[i])<-r[1]+r[2]*x1[i]+r[3]*x2[i]+r[4]*x3[i]
                     +rz*u[i]
        y[i] ~ dbin(tp[i], 1)
```

```
        logit(tp[i])<-delta*z[i]+bu*u[i]+b[1]+b[2]*x1[i]+
                      b[3]*x2[i]+b[4]*x3[i]
        }
  for (j in 1:4)
      {
       r[j] ~ dnorm(0, 0.1)
       b[j] ~ dnorm(0, 0.5)
      }
  delta ~ dnorm(0, 0.5)
  or.delta<-exp(delta)
}
params.se <- c("r", "b", "delta", "or.delta")
inits.se <- function(){
list(r=rnorm(4), b=rnorm(4), delta=rnorm(1),
     u=rbinom(n,1,0.5))
}


### mis no U ###
se.mis.model1 <- function(){
  for (i in 1:n)
      {
       z[i] ~ dbin(p[i],1)
       logit(p[i])<-r[1]+r[2]*x1[i]+r[3]*x2[i]+r[4]*x3[i]
       y[i] ~ dbin(op[i], 1)
       op[i] <- se1*tp[i] +(1-sp1)*(1-tp[i])
       logit(tp[i])<-delta*z[i]+b[1]+b[2]*x1[i]+b[3]*x2[i]
```

```
                    +b[4]*x3[i]
      }

  se1 ~ dbeta(70,30)
  sp1 ~ dbeta(90,10)

  for (j in 1:4)
      {
       r[j] ~ dnorm(0, 0.1)
       b[j] ~ dnorm(0, 0.5)
      }

  delta ~ dnorm(0, 0.5)
or.delta<-exp(delta)
}
params.se.mis1 <- c("r", "b", "delta", "or.delta","se1","sp1")
inits.se.mis1 <- function(){
list(r=rnorm(4), b=rnorm(4), delta=rnorm(1), se1=runif(1),
     sp1=runif(1))
}
### mis ###
se.mis.model <- function(){
  for (i in 1:n)
      {
       u[i] ~ dbin(pu,1)
       z[i] ~ dbin(p[i],1)
       logit(p[i])<-r[1]+r[2]*x1[i]+r[3]*x2[i]+r[4]*x3[i]
```

```
                    +rz*u[i]
        y[i] ~ dbin(op[i], 1)
        op[i] <- se1*tp[i] +(1-sp1)*(1-tp[i])
        logit(tp[i])<-delta*z[i]+bu*u[i]+b[1]+b[2]*x1[i]
                     +b[3]*x2[i]+b[4]*x3[i]
        }
  se1 ~ dbeta(70,30)
  sp1 ~ dbeta(90,10)
  for (j in 1:4)
      {
       r[j] ~ dnorm(0, 0.1)
       b[j] ~ dnorm(0, 0.5)
      }
  delta ~ dnorm(0, 0.5)
  or.delta<-exp(delta)
}
params.se.mis <- c("r", "b", "delta", "or.delta","se1","sp1")
inits.se.mis <- function(){
list(r=rnorm(4), b=rnorm(4), delta=rnorm(1), u=rbinom(n,1,0.5),
se1=runif(1),sp1=runif(1))
}

library(R2jags)
library(rjags)
load.module("dic")
### simulation ###
sim <- function(b=1,r=1){
```

```
x1 <- rnorm(n,0,1)
x2 <- rnorm(n,0,1)
x3 <- rnorm(n,0,1)
u <- rbinom(n,1,0.5)
pz <- exp(r0 + r1*x1 + r2*x2 + r3*x3 + r*u)/
      (1+exp(r0 + r1*x1 + r2*x2 + r3*x3 + r*u))
z <- rbinom(n,1,pz)
py <- exp(b0 + b1*x1 + b2*x2 + b3*x3 + b*u + b4*z)/
      (1+exp(b0 + b1*x1 + b2*x2 + b3*x3 + b*u + b4*z))
op <- py*se + (1-py)*(1-sp)
y <- rbinom(n,1,op)
se.data1 <- list("z", "y", "x1", "x2", "x3", "n")
se.fit1 <- jags(data=se.data1, inits=inits.se1, params.se1,
n.chain=2, n.burnin=1000, n.iter=6000, n.thin=1,
model.file=se.model1)
mis.se.data1 <- list("z", "y", "x1", "x2", "x3", "n")
mis.se.fit1 <- jags(data=mis.se.data1, inits=inits.se.mis1,
params.se.mis1,n.chain=2, n.burnin=1000, n.iter=6000,
n.thin=1, model.file=se.mis.model1)
rz = r
bu = b
se.data2 <- list("z", "y", "x1", "x2", "x3", "n", "rz",
                 "bu", "pu")
se.fit2 <- jags(data=se.data2, inits=inits.se, params.se,
n.chain=2, n.burnin=1000,n.iter=6000, n.thin=1,
model.file=se.model)
mis.se.data2 <- list("z", "y", "x1", "x2", "x3", "n", "pu",
```

```
                    "rz", "bu")
mis.se.fit2 <- jags(data=mis.se.data2, inits=inits.se.mis,
params.se.mis, n.chain=2, n.burnin=1000, n.iter=6000,
n.thin=1, model.file=se.mis.model)
return(list(se.fit1$BUGSoutput$summary,
se.fit2$BUGSoutput$summary,mis.se.fit1$BUGSoutput$summary,
mis.se.fit2$BUGSoutput$summary))
}
```

BIBLIOGRAPHY

A. Abadie and G. W. Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008. ISSN 0012-9682. doi: 10.3982/Ecta6474.

Alberto Abadie and Guido W. Imbens. Matching on estimated propensity score. *NBER working paper series working paper (15301)*, 2009.

Weihua An. Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40:151–189, 2010.

B. Arpino and M. Cannas. Propensity score matching with clustered data. an application to the estimation of the impact of caesarean section on the apgar score. *Stat Med*, 35(12):2074–91, 2016. ISSN 1097-0258 (Electronic) 0277-6715 (Linking). doi: 10.1002/sim.6880.

Bruno Arpino and Fabrizia Mealli. The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis*, 55(4):1770–1780, 2011.

Peter C. Austin. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, 29(6):661–77, 2009.

N. B. Carnegie, M. Harada, and J. L. Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016. ISSN 1934-5747. doi: 10.1080/19345747.2015.1078862.

Yoo-Mi Chin, Joon Jin Song, and James Stamey. A bayesian approach to misclassified binary response: Female employment and intimate partner violence in urban india. *Applied Economics Letters*, forthcoming, 2017.

W. G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24(2):295–313, 1968. ISSN 0006-341X (Print) 0006-341X (Linking).

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst*, 22(1):173–203, 1959. ISSN 0027-8874 (Print) 0027-8874 (Linking).

V. Dorie, M. Harada, N. B. Carnegie, and J. Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Stat Med*, 35(20):3453–70, 2016. ISSN 1097-0258 (Electronic) 0277-6715 (Linking). doi: 10.1002/sim.6973.

M. C. Elze, J. Gregson, U. Baber, E. Williamson, S. Sartori, R. Mehran, M. Nichols, G. W. Stone, and S. J. Pocock. Comparison of propensity score methods and covariate adjustment evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology*, 69(3):345–357, 2017. ISSN 0735-1097. doi: 10.1016/j.jacc.2016.10.060.

D. Faries, X. Peng, M. Pawaskar, K. Price, J. D. Stamey, and Jr. Seaman, J. W. Evaluating the impact of unmeasured confounding with internal validation data: an example cost evaluation in type 2 diabetes. *Value Health*, 16(2):259–66, 2013. ISSN 1524-4733 (Electronic) 1098-3015 (Linking). doi: 10.1016/j.jval.2012.10.012.

J. Gastwirth. Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85 (4):907–920, 1998. ISSN 0006-3444 1464-3510. doi: 10.1093/biomet/85.4.907.

Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis, Third Edition (Chapman Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2014. ISBN 1439840954. doi: citeulike-article-id:12855856.

D. Gouet, J. Rouffineau, C. Chauvin, J. C. Abadie, M. Ribet, and B. Becq Giradon. [repeated recurrences of hepatic amoebiasis with failure of metronidazole treatment]. *Nouv Presse Med*, 11(45):3349, 1982. ISSN 0301-1518 (Print) 0301-1518 (Linking).

S. Greenland. Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, 25(6):1107–1116, 1996. ISSN 0300-5771. doi: DOI 10.1093/ije/25.6.1107-a.

S. Greenland. The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *Journal of the American Statistical Association*, 98(461):47–54, 2003. ISSN 0162-1459. doi: 10.1198/01621450338861905.

S. Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 168:267–291, 2005. ISSN 0964-1998. doi: DOI 10.1111/j.1467-985X.2004.00349.x.

L. Griffon, A. Amaddeo, G. Mortamet, C. Barnerias, V. Abadie, J. Olmo Arroyo, L. de Sanctis, S. Renolleau, and B. Fauroux. Sleep study as a diagnostic tool for unexplained respiratory failure in infants hospitalized in the picu. *J Crit Care*, 42:317–323, 2017. ISSN 1557-8615 (Electronic) 0883-9441 (Linking). doi: 10.1016/j.jcrc.2016.04.003.

P. Gustafson, L. C. McCandless, A. R. Levy, and S. Richardson. Simplified bayesian sensitivity analysis for mismeasured and unobserved confounders. *Biometrics*, 66(4): 1129–1137, 2010. ISSN 0006-341x. doi: 10.1111/j.1541-0420.2009.01377.x.

T. Hoshino. A bayesian propensity score adjustment for latent variable modeling and mcmc algorithm. *Computational Statistics Data Analysis*, 52(3):1413–1429, 2008. ISSN 0167-9473. doi: 10.1016/j.csda.2007.03.024.

David Kaplan and Jianshen Chen. A two-step bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, 77(3):581–609, 2012.

Fan Li, Alan M. Zaslavsky, and Mary Beth Landrum. Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387, 2013.

D. Y. Lin, B. M. Psaty, and R. A. Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54(3):948–963, 1998. ISSN 0006-341x. doi: Doi 10.2307/2533848.

L. S. Magder and J. P. Hughes. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146(2):195–203, 1997. ISSN 0002-9262.

L. C. McCandless, P. Gustafson, and A. Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, 26(11):2331–2347, 2007. ISSN 0277-6715. doi: 10.1002/sim.2711.

L. C. McCandless, I. J. Douglas, S. J. Evans, and L. Smeeth. Cutting feedback in bayesian regression adjustment for the propensity score. *Int J Biostat*, 6(2):Article 16, 2010. ISSN 1557-4679 (Electronic) 1557-4679 (Linking).

Lawrence C. McCandless, Paul Gustafson, and Peter C. Austin. Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28:94–112, 2009.

P. McInturff, W. O. Johnson, D. Cowling, and I. A. Gardner. Modelling risk when binary outcomes are subject to error. *Statistics in Medicine*, 23(7):1095–1109, 2004. ISSN 0277-6715. doi: 10.1002/sim.1656.

John M. Neuhaus. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4):843–855, 1999.

Tia Palermo, Jennifer Bleck, and Amber Peterman. Tip of the iceberg: Reporting and gender-based violence in developing countries. *AMERICAN JOURNAL OF EPIDEMIOLOGY*, 179:602–612, 2013.

Carlos Daniel Paulino, Paulo Soares, and John Neuhaus. Binomial regression with misclassification. *Biometrics*, 59:670–675, 2003.

Carlos Daniel Paulino, Giovani Silva, and Jorge Alberto Achcar. Bayesian analysis of correlated misclassified binary data. *Computational Statistics Data Analysis*, 49:1120–1131, 2005.

R. F. Rabin, J. M. Jennings, J. C. Campbell, and M. H. Bair-Merritt. Intimate partner violence screening tools: a systematic review. *Am J Prev Med*, 36(5): 439–445 e4, 2009. ISSN 1873-2607 (Electronic) 0749-3797 (Linking). doi: 10.1016/j.amepre.2009.01.024.

P. R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987. ISSN 0006-3444. doi: DOI 10.1093/biomet/74.1.13.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–304, 2002. ISSN 0883-4237.

P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B-Methodological*, 45(2):212–218, 1983a. ISSN 0035-9246.

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79 (387):516–524, 1984. ISSN 0162-1459. doi: Doi 10.2307/2288398.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983b.

D. B. Rubin. Bayesian-inference for causal effects - role of randomization. *Annals of Statistics*, 6(1):34–58, 1978. ISSN 0090-5364. doi: DOI 10.1214/aos/1176344064.

D. B. Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*, 26(1):20–36, 2007. ISSN 0277-6715 (Print) 0277-6715 (Linking). doi: 10.1002/sim.2739.

K. Steenland and S. Greenland. Monte carlo sensitivity analysis and bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology*, 160(4):384–392, 2004. ISSN 0002-9262. doi: 10.1093/aje/kwh211.

Yu-Sung Su and Jeronimo Cortina. What do we gain? combining propensity score methods and multilevel modeling. *APSA 2009 Toronto Meeting Paper*, 2009.

S. Vyas and C. Watts. How does economic empowerment affect womenâs risk of intimate partner violence in low and middle income countries? a systematic review of published evidence. *Journal of International Development*, 21(5):577–602, 2009.

Qi Zhou, Yoo-Mi Chin, James D. Stamey, and Joon Jin Song. Bayesian misclassification and propensity score methods for clustered observational studies. *Journal of Applied Statistics*, pages 1–14, 2017. ISSN 0266-4763. doi: 10.1080/02664763.2017.1380786.

C. M. Zigler, K. Watts, R. W. Yeh, Y. Wang, B. A. Coull, and F. Dominici. Model feedback in bayesian propensity score estimation. *Biometrics*, 69(1):263–73, 2013. ISSN 1541-0420 (Electronic) 0006-341X (Linking). doi: 10.1111/j.1541-0420.2012.01830.x.