NOVEL METHODS FOR ADDRESSING BIAS FROM MISCLASSIFIED EXPOSURE VARIABLES

A Dissertation

by

CHRISTOPHER MATTHEW MANUEL

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Samiran Sinha
Co-Chair of Committee,	Suojin Wang
Committee Members,	Raymond J. Carroll
	Joseph Orr
Head of Department,	Daren B.H. Cline

August 2020

Major Subject: Statistics

Copyright 2020 Christopher Matthew Manuel

ABSTRACT

Exposure variables are often misclassified in observational studies. Any analysis that does not make proper adjustments for misclassification may result in biased estimates of model parameters and that may lead to distorted inferences.

In this dissertation I study this bias in cohort and retrospective matched case-control study. For the matched case-control study, I consider a binary exposure variable whereas for the cohort study I consider a multicategory exposure variable that has more than two nominal categories. The novel aspect of this work is the use of instrumental variables to reduce the bias due to the misclassification of the exposure variable when no validation data are available. Each of the works, one involving matched case-control and the other involving the cohort data, consists of two major steps. First I study the parameter identifiability and obtain sufficient conditions for identifiability. Then I propose model estimation and inference methods after adopting the sufficient conditions of identifiability. In the first work, I use two methods of estimation including the efficient approach. In the second work, I use a variational Bayesian inference procedure aided by the automatic differentiation variational inference (ADVI) technique. Operating characteristics of the methods are assessed and compared with existing approaches through simulation studies. Simulation studies clearly indicate when and how the proposed methods are advantageous. Each of the methods are applied to analyze real datasets. For the matched case-control study scenario, the proposed methods are applied to the nested case-control data sampled from the 1989 United States birth registry where the reported smoking status of mothers during pregnancy is considered to be the misclassified exposure.

For the cohort data scenario, the proposed Bayesian method is applied to the US breast cancer mortality data sampled from the Surveillance Epidemiology and End Results (SEER) database, where reported treatment therapy is considered to be the misclassified exposure variable.

DEDICATION

To D.Wong

ACKNOWLEDGEMENTS

I first gives thanks to Jesus Christ.

In the Holy Catholic Bible, 1 Corinthians 13:13 mentions the important theological virtues: faith, hope, and charity (love). From my experiences during the PhD program, I've learned to acknowledge the presence of God in everything (like it says in Psalm 139:7-10). By extension, I've come to learn that these three virtues can be found everywhere if I only act greatful in everything bad and good.

I didn't believe that I would ever finish this program, but through the charity of people in the Statistics department, I learned to be faithful in those who guided and directed me in my bad and good times. First, I give thanks to my advisors, Dr. Samiran Sinha and Dr. Suojin Wang, for their patience and due diligence in working with me in the past four years, as a student and and as advisee and someone who asked for much direction. I would also like to thank Dr. Raymond Carroll and Dr. Joseph Orr for their time with me. Furthermore, I would like to give special thanks to Dr. Michael Longnecker who has always given the time to listen to my concerns, his never ending support of all my endeavors, and just for being someone to talk to about sports and life in general.

I also learned that the best way to endure any trial is through the support of family and friends who provide hope for the future. In that end I thank my, including my mom, dad, sister, and Ms. Cindy, as well as all my extended family especially Dan and Noel who came to visit me. I would also like to thank all my friends who have supported me along the way - Vennessa, Shelby, Lauren, Pat, and Eli, as well as Furong, Shubhadeep, Pallavi, Sandi, Kristyn, Minsuk, Raanju, Bob, Zijuan, and Marcin.

Last, I would like to thank the staff at St. Mary's Catholic Center for strengthening my faith in Jesus Christ for the last four years, and revealed the meaning of charity to me. The great services provided by the staff, especially those in the RCIA program are a blessing for all who partake in it. Special thanks to Kevin, Rohan and Barbara. I also particularly thank the priests, Fr. Brian, Fr. Ryan, and Fr. Augustine who showed me the meaning of charity through sacrifice and which is something for me to emulate in the future. I am especially thankful for the charity of prayers and encouragement from Fr. Charlie and the newly appointed vocations director Fr. Greg.

Thank you so much.

Luke 12:24

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Samiran Sinha, Dr. Suojin Wang and Dr. Raymond J. Carroll of the Department of Statistics, and Dr. Joseph Orr of the Department of Psychology. The data analyzed for Section 2 was obtained from NCHS' Vital Statistics Natality Birth Data 1989 dataset, which is available online through http://data.nber.org/data/vital-statistics-natality-data.html. The data analyzed for Section 3 was obtained from the National Cancer Institute Surveillance, Epidemiology and End Results (SEER) database, and access can be obtained through the website https://seer.cancer.gov/data/access.html. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by research assistant funding from the George P. Mitchell '40 Endowed Chair in Statistics.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

For this dissertation, I address the issue of using a misclassified exposure variable X in various logistic regression models, and novel ways to address to correct this bias. In this first chapter, I will review various logistic regression models commonly used for observational studies. I then discuss how misclassified exposure variables cause biased parameter estimates via a literature review, as well as providing an analytical proof of the effects of misclassification, and also by demonstrating this bias in a small simulation study. In chapter 2, I consider misclassification of X in the setting of matched case control data, where data are stratified and must be analyzed in the framework of the conditional logistic regression. I propose frequentist methods to adjust for misclassification in this setting. In chapter 3, I consider a general setting in which the exposure variable has more than two categories; then I study identification and develop an efficient Bayesian method to estimate the model parameters. Finally, in chapter 4, I discuss directions of future work.

1.1 Logistic regression estimates under the presence of misclassified covariates

1.1.1 Review of standard logistic regression

Logistic regression is a statistical method used to model the association between a set of covariates and a binary response variable [1]. Let Y be the response variable and X is a scalar predictor. Then the success probability under the logistic regression is

$$\text{pr}(Y = 1|X; \beta) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}, \tag{1.1}$$

where $\beta = (\beta_0, \beta_1)^T$. Suppose that X is also a binary variable. Then, the parameter β_1 represents the log odds ratio for $Y = 1$ from $X = 1$ to $X = 0$. That is,

$$\beta_1 = \log\left\{\frac{\text{odds}(X=1)}{\text{odds}(X=0)}\right\} = \log\left\{\frac{\text{pr}(Y=1|X=1)/\text{pr}(Y=0|X=1)}{\text{pr}(Y=1|X=0)/\text{pr}(Y=0|X=0)}\right\},$$

where the odds$(X = a)$ represents the ratio of probability of $Y = 1$ given $X = a$ to the probability of $Y = 0$ given $X = a$. The odds ratio, or OR, provides a measure of how likely a subject or observation would get a success $Y = 1$ given that they have some covariate $X = 1$ over how likely they would get a success given that their covariate $X = 0$. An odds ratio of 1 indicates that the odds of $Y = 1$ is the same regardless of whether $X = 0$ or 1, while an odds ratio greater than 1 indicates that those who have $X = 1$ are more likely to obtain $Y = 1$. Finally an odds ratio less than 1 indicates that those who have $X = 0$ are less likely to have $Y = 1$. We can estimate β_1 using the frequencies of different values of X and Y.

Suppose that we have a dataset that consists of n independently and identically distributed (iid) copies of (X, Y) sampled from an underlying population. Let n_{00} represent the total number of observations who exhibit $X = 0$ and $Y = 0$, n_{01} represent the total number of observations who exhibit $X = 0$ and $Y = 1$, n_{10} be the total number of $X = 1$ and $Y = 0$, and n_{11} be the total number of $X = 1$ and $Y = 1$. These numbers can be expressed in the following Table 1.1, referred the 2×2 contingency table:

Table 1.1: 2×2 Contingency Table

		Y	
		0	1
X	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

With these frequencies, the odds ratio is estimated by

$$\exp(\widehat{\beta}_1) = \frac{n_{11}n_{00}}{n_{10}n_{01}}.$$

And the estimator of the log-odds ratio β_1 is $\widehat{\beta}_1 = \log\{n_{00}n_{11}/n_{01}n_{10}\}$. These estimators are actually the maximum likelihood estimator, and can be obtained by maximizing the log-likelihood function $\mathcal{L}(\beta)$ with respect to β, where

$$\ell(\beta) = \prod_{i=1}^{n}\{\text{pr}(Y=1|X_i;\beta)\}^{Y_i}\{1-\text{pr}(Y=1|X_i;\beta)\}^{1-Y_i} = \prod_{i=1}^{n}\frac{\exp\{Y_i(\beta_0+\beta_1X_i)\}}{1+\exp(\beta_0+\beta_1X_i)}.$$

In general, when X consists of a set of predictor variables or X is a numeric predictor, the estimate of the set of regression parameter β_1 in a logistic regression model is obtained by maximizing the likelihood function. The maximization step is usually carried out using the iterative re-weighted least square (IWLS) method. It is well known that the MLE enjoys very nice properties. The details can be found in [25]. In this dissertation, we consider binary Y and use the logistic model in all chapters.

1.1.2 Polytomous logistic regression

When the response Y has more than two nominal categories, then one can use the polytomous (or multinomial) logistic model to write the probability. Typically the lowest category of Y is considered as the baseline (reference) category. For example, suppose the variable Y takes values in $(1,\ldots,J)$, and X is the vector of covariates. If 1 is selected as the baseline then the model for Y is

$$\text{pr}(Y=j|X;\beta) \;=\; \frac{\exp(\beta_{j0}+\beta_{j1}^TX)}{1+\sum_{r=2}^{J}\exp(\beta_{r0}+\beta_{r1}^TX)},$$

for $j = 2, \ldots, J$. Since $\sum_{j=1}^{J} \text{pr}(Y = j|X;\beta) = 1$, $\text{pr}(Y = 1|X)$ can be calculated using the formula $1 - \text{pr}(Y = 2|X) - \cdots - \text{pr}(Y = J|X)$. Here $\beta = (\beta_{20}, \beta_{21}^T, \ldots, \beta_{J0}, \beta_{J1}^T)^T$. Also, the logit is now defined as $\log\{\text{pr}(Y = j|X;\beta)/\text{pr}(Y = 1|X;\beta)\} = \beta_{j0} + \beta_{j1}^T X$ with $j \in (2, \ldots, J)$, where the probability on the denominator is that for the baseline category.

Assuming that one has a random sample of consisting of n iid copies of (X, Y) from the underlying population, the β-parameter can be estimated by maximizing the log-likelihood function

$$\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{n}\sum_{j=1}^{J} I(Y = j)\log\{\text{pr}(Y = j|X_i;\beta)\} \\
&= \sum_{i=1}^{n}\left[\sum_{j=2}^{J} I(Y = j)\log\{\text{pr}(Y = j|X_i;\beta)\} + I(Y = 1)\log\{\text{pr}(Y = 1|X_i;\beta)\}\right] \\
&= \sum_{i=1}^{n}\left[\sum_{j=2}^{J} I(Y = j)\log\{\text{pr}(Y = j|X_i;\beta)\}\right. \\
&\quad \left. + \sum_{j=2}^{J}\{1 - I(Y = j)\}\log\{\text{pr}(Y = 1|X_i;\beta)\}\right] \\
&= \sum_{i=1}^{n}\left[\sum_{j=2}^{J} I(Y = j)\log\left\{\frac{\text{pr}(Y = j|X_i;\beta)}{\text{pr}(Y = 1|X_i;\beta)}\right\} + \log\{\text{pr}(Y = 1|X_i;\beta)\}\right] \\
&= \sum_{i=1}^{n}\left[\sum_{j=2}^{J} I(Y = j)(\beta_{j0} + \beta_{j1}^T X_i) + \log\left\{1 - \sum_{j=2}^{J}\text{pr}(Y = j|X_i;\beta)\right\}\right].
\end{aligned}$$

For the work in this dissertation, the response Y will be considered binary. However, in Chapter 3, I will consider a scenario where X is a categorical variable and hence will need to use the multinomial logistic regression to model its probabilities.

1.1.3 Case-control studies

Besides through random samples from the targeted population, data may be collected through a case-control sample. In a case-control sample, usually, random

samples are collected from the case group ($Y = 1$) and control group ($Y = 0$) separately. Typically, case-control samples are used for studying a rare disease where chances of having $Y = 1$ is low in the population. Suppose that the exposure X is a binary variable, then the odds ratio of $Y = 1$ from $X = 1$ to $X = 0$ is

$$\begin{aligned}
\text{OR} &= \frac{\text{pr}(Y=1|X=1)/\text{pr}(Y=0|X=1)}{\text{pr}(Y=1|X=0)/\text{pr}(Y=0|X=0)} \\
&= \frac{\text{pr}(Y=1,X=1)/\text{pr}(Y=0,X=1)}{\text{pr}(Y=1,X=0)/\text{pr}(Y=0,X=0)} \\
&= \frac{\text{pr}(X=1|Y=1)\text{pr}(Y=1)/\text{pr}(X=1|Y=0)\text{pr}(Y=0)}{\text{pr}(X=0|Y=1)\text{pr}(Y=1)/\text{pr}(X=0|Y=0)\text{pr}(Y=0)} \\
&= \frac{\text{pr}(X=1|Y=1)/\text{pr}(X=1|Y=0)}{\text{pr}(X=0|Y=1)/\text{pr}(X=0|Y=0)}.
\end{aligned}$$

Observe that the OR is expressed in terms of conditional probabilities $\text{pr}(X = x|Y = 0)$ and $\text{pr}(X = x|Y = 1)$, and they can be consistently estimated from the control and case samples, respectively. Hence, the odds ratio can also be estimated from the case-control data. Thus, β_1 (or $\exp(\beta_1)$) of the logistic model (1.1) can be estimated from the case-control sample. The next question is if the intercept β_0 of model (1.1) can be estimated from the case-control sample. Clearly,

$$\exp(\beta_0) = \frac{\text{pr}(Y=1|X=0;\beta)}{\text{pr}(Y=0|X=0;\beta)} = \frac{\text{pr}(X=0|Y=1;\beta)\text{pr}(Y=1;\beta)}{\text{pr}(X=0|Y=1;\beta)\text{pr}(Y=1;\beta)}.$$

Although $\text{pr}(X = 0|Y = 0;\beta)$ and $\text{pr}(X = 0|Y = 1;\beta)$ can be estimated from the control and case samples, the absolute risk $\text{pr}(Y = 1;\beta)$ or $\text{pr}(Y = 0;\beta)$ cannot. Hence, β_0 cannot be estimated from the case-control data; but it is not a cause of concern when the interest lies in finding association between the response and exposure variables. This idea holds true for any type covariate X.

To avoid or reduce spurious association (the association that is mostly caused by

other variables) sometimes data are collected through a matched case-control study. In such a study, cases and controls are matched at various levels of the matching variables. Usually confounding variables are used for matching. So, at different values of the matching variables, cases and controls are identified and then exposure information is ascertained. For example, a person's age may affect the association between a therapy and a health outcome. To prevent the confounding effects of age to cause spurious association, cases and controls are identified at some age groups. Then the therapy information is collected from the cases and controls. If one case and M controls are collected at every level of the matching variables, referred to as an 1:M matched case-control study. A 1:1 matched case-control study occurs when M= 1 . Case(s) and controls at every value of the matching variables form a stratum. In a matched case-control data, the number of cases and controls are fixed at every stratum, but the number of strata is large, say n.

Suppose that the interest is in estimating the log-odds ratio parameter for the set of exposure variables X using a 1: M matched case-control data with n observations. The data from the jth subject in the ith stratum are $(Y_{i,j}, X_{i,j})$, for $j = 1, \ldots, M + 1$, and $i = 1, \ldots, n$. Since there are one case and M controls in every stratum, $\sum_{j=1}^{M+1} Y_{i,j} = 1$. The probability model for the response Y in the ith stratum is

$$\text{pr}(Y_{i,j} = 1 | X_{i,j}; \beta) = \frac{\exp(\beta_{0,i} + \beta_1^T X_{i,j})}{1 + \exp(\beta_{0,i} + \beta_1^T X_{i,j})}.$$

The intercept terms $\beta_{0,i}$s are allowed to vary with strata. Through this model, it is assumed that the log-odds ratio parameters do not change with strata. That means the association between X and Y is homogeneous across the strata. To estimate β_1 usually conditional logistic regression method is used. The conditional likelihood for the ith stratum is the probability of observing $Y_{i,j}$ given $X_{i,j}$ and the condition that

there is only one case in the ith matched set,

$$\begin{aligned}\mathcal{L}_i(\beta_1) &= \text{pr}(Y_{i,j} = y_{i,1}, \ldots, Y_{i,M+1} = y_{i,M+1} | X_{i,1}, \ldots, X_{i,M+1}, \sum_{j=1}^{M+1} Y_{i,j} = 1) \\ &= \frac{\exp(\sum_{j=1}^{M+1} \beta_1^T X_{i,j} y_{i,j})}{\sum_{j=1}^{M+1} \exp(\beta_1^T X_{i,j})}.\end{aligned}$$

The likelihood $\mathcal{L}_i(\beta_1)$ is completely free from the stratum specific intercept $\beta_{0,i}$. This is possible because for the logistic model $\sum_{j=1}^{M+1} Y_{i,j}$ is the sufficient statistic for $\beta_{0,i}$. Then the likelihood of all strata is $\mathcal{L}(\beta_1) = \prod_{i=1}^{n} \mathcal{L}_i(\beta_1)$, and then β_1 is estimated by solving the score equations

$$S_{\beta_1} = \sum_{i=1}^{n} \sum_{j=1}^{M+1} \left\{ Y_{i,j} - \frac{\exp(\beta_1^T X_{i,j})}{\sum_{j'=1}^{M+1} \exp(\beta_1^T X_{i,j'})} \right\} X_{i,j} = 0.$$

For a binary exposure variable X, the data from the kth stratum can be summarized as follows:

	Case($Y=1$)	Control($Y=0$)	Total
Exposed ($X=1$)	n_{k11}	n_{k10}	n_{k1+}
Unexposed ($X=0$)	n_{k01}	n_{k00}	n_{k0+}

Here n_{kjl} denotes the number of observations with $X = j$ and $Y = l$ in the kth stratum, $j, l = 0, 1$, and $k = 1, \ldots, n$, and $n_{kj+} = n_{kj0} + n_{kj1}$. Furthermore, n_{k01} and n_{k10} are called discordant pairs. Then the score function reduces to

$$S_{\beta_1} = \sum_{k=1}^{n} \left\{ n_{k11} - \frac{\exp(\beta_1) n_{k1+}}{\exp(\beta_1) n_{k1+} + n_{k0+}} \right\} = \sum_{k=1}^{n} \left\{ \frac{\exp(\beta_1) n_{k1+}(n_{k11} - 1) + n_{k0+}}{\exp(\beta_1) n_{k1+} + n_{k0+}} \right\}.$$

For a 1:1 matched case-control dataset, there are four types of strata: 1) $n_{k11} = 1, n_{k01} = 0, n_{k10} = 1, n_{k00} = 0$, 2) $n_{k11} = 0, n_{k01} = 1, n_{k10} = 1, n_{k00} = 0$, 3) $n_{k11} = 1, n_{k01} = 0, n_{k10} = 0, n_{k00} = 1$, 4) $n_{k11} = 0, n_{k01} = 1, n_{k10} = 0, n_{k00} = 1$. Suppose that

the number of strata of types 1, 2, 3, 4 are $s_{11}, s_{01}, s_{10}, s_{00}$, respectively. Then the score function can be written as

$$\begin{aligned} S_{\beta_1} &= s_{11} + s_{10} - s_{11} \sum_{j=1}^{2} \frac{\exp(\beta_1)}{\sum_{j'=1}^{2} \exp(\beta_1)} - s_{01} \frac{\exp(\beta_1)}{\exp(\beta_1)+1} - s_{10} \frac{\exp(\beta_1)}{\exp(\beta_1)+1} \\ &= s_{10} - (s_{01} + s_{10}) \frac{\exp(\beta_1)}{\exp(\beta_1)+1}. \end{aligned}$$

Hence, the conditional logistic regression (CLR) estimator of β_1 is the solution of

$$\begin{aligned} 0 &= s_{10} - (s_{01} + s_{10}) \frac{\exp(\beta_1)}{\exp(\beta_1)+1} = s_{10}\{\exp(\beta_1)+1\} - (s_{01} + s_{10}) \exp(\beta_1) \\ &= s_{10} - s_{01} \exp(\beta_1), \end{aligned}$$

and the estimator is $\widehat{\beta}_1 = \log(s_{10}/s_{01})$, where s_{01}, s_{10} are called discordant sets (strata). There is connection between the CLR estimator and test and the McNemar test for this 1:1 matched data.

Before discussing the connection I provide a brief overview of the McNemar test. The purpose of the McNemar test is to assess whether there is a difference in the case and control in terms of the exposure status for 1:1 matched paired data. To do this, the discordant pairs n_{k01} and n_{k10} are used in the McNemar test statistic $z_0^2 = ((n_{21} - n_{12})/\sqrt{n_{21} + n_{12}})^2$, which is distributed as χ^2 with 1 degree of freedom. However, the McNemar test can be directly applied in the conditional logistic regression setting by testing the hypothesis that $\beta_1 = 0$. This follows since $\widehat{\beta}_1$ is a function of the discordant sets s_{01}, s_{10}, from which s_{01} is the number of sets in which the discordant pairs n_{k01} and n_{k10} occur.

When there are M controls, this implies the use of 1:M matched paired data. Hence instead of using the McNemar test one would instead apply the Cochran Mantel-Haenszel (CMH) test. In this setting the hypothesis is $H_0 : OR = 1$, where

$OR = \exp(\beta_1)$ is the odds ratio parameter. Again, this can be tested directly in the conditional logistic setting by using a score test statistic for the hypothesis $H_0 : \beta_1 = 0$ [16, 39].

1.1.4 Presence of misclassified variables in logistic regression and its effect

Having discussed the types of logistic regression model, I now consider the scenario where instead of observing X a misclassified form of it, W is reported. First, consider how the misclassification causes the frequencies of W to differ from X as denoted below

		Y	
		0	1
W	0	n^*_{00}	n^*_{01}
	1	n^*_{10}	n^*_{11}

and so the estimate of the odds ratio using W instead of X is calculated as

$$\exp(\widehat{\beta^*_1}) = \frac{n^*_{11}n^*_{00}}{n^*_{10}n^*_{01}}.$$

The underlying parameter β^*_1 can be considered as coming from a mispecified model [65] (White, 1982), where the incorrect covariate is used in the logistic model. Hence, the resulting estimate does not converge to the true β_1. The expected difference between the estimator $\widehat{\beta}^*_1$ and the true β_1, i.e., $E(\widehat{\beta}^*_1) - \beta_1$, will be referred to as the misclassification bias. I will consider its form in the next section.

1.1.4.1 Analytical forms of the bias due to misclassified exposure in logistic regression

Before analyzing the form of $E(\widehat{\beta}_1^* - \beta_1)$, I consider work of [15] and [42]. [15] provided the analytical background for finding the tractable form of the analytical bias for a binary misclassified variable in a logistic regression setting. [42] further expanded on [15] by considering the analytical bias for a categorical misclassified variable. First, I assume that W and X are both discrete with categories (0,1). Define the joint probabilities as $\pi_{i,j} = \text{pr}(Y = i, X = j)$ and $p_{i,j} = P(Y = i, W = j)$ for the response Y and X, and Y and W, respectively. The probability models for Y given X and Y given W are as follows:

$$\begin{aligned} \text{pr}(Y = 1|X = j) \equiv \pi_{1|j} &= \frac{\exp(\beta_0 + \beta_1 I(j = 1))}{1 + \exp(\beta_0 + \beta_1 I(j = 1))}, \\ \text{pr}(Y = 1|W = j) \equiv p_{1|j} &= \frac{\exp(\beta_0^* + \beta_1^* I(j = 1))}{1 + \exp(\beta_0^* + \beta_1^* I(j = 1))}, \end{aligned}$$

where the parameters $\beta = (\beta_0, \beta_1)^T$ corresponding to the model using the true X and $\beta^* = (\beta_0^*, \beta_1^*)^T$ correspond to the model using the misclassified variable W. Next I define $\boldsymbol{\alpha} = (\alpha_{i00}, \alpha_{i01}, \alpha_{i11}, \alpha_{i10})$ as the set of misclassification and nonmisclassification probabilities. In particular, the nonmisclassification probabilities are $\alpha_{i11} = \text{pr}(W = 1|X = 1, Y = i)$ and $\alpha_{i00} = \text{pr}(W = 0|X = 0, Y = i)$, and the corresponding misclassification probabilities are $\alpha_{i01} = \text{pr}(W = 0|X = 1, Y = i)$ and $\alpha_{i10} = \text{pr}(W = 1|X = 0, Y = i)$. For the rest of this section I will consider the scenario with non differential misclassification, which means that $\alpha_{ijk} = \alpha_{jk}$, hence the misclassification is independent of the response Y. I will define the asymptotic bias in terms of these misclassification probabilities. Finally, I refer to the misclassification bias as $\Delta_j(\alpha) = \beta_j - \beta_j^*$ where $j = 0, 1$ corresponding to the model parameters

of β.

To obtain the form of the misclassification bias I first rewrite the parameters in terms of the misclassification probabilities. I first start with the parameters β. I will use the relation $\pi_{i|j}/\pi_{i'|j} = \pi_{i,j}/\pi_j/\pi_{i',j}\pi_j = \pi_{i,j}/\pi_{i',j}$, where $i' \neq i$ and $\pi_j = \text{pr}(X = j)$.

Then

$$\begin{aligned}
\frac{\pi_{1,0}}{\pi_{0,0}} &= \frac{\frac{\exp(\beta_0)}{1+\exp(\beta_0)}}{1-\frac{\exp(\beta_0)}{1+\exp(\beta_0)}} \\
&= \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \times \frac{1+\exp(\beta_0)}{1} \\
&= \exp(\beta_0)
\end{aligned}$$

which implies that $\beta_0 = \log(\frac{\pi_{1,0}}{\pi_{0,0}})$. Similarly,

$$\begin{aligned}
\frac{\pi_{1,1}}{\pi_{0,1}} &= \frac{\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}}{1-\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}} \\
&= \frac{\exp(\beta_0)}{1+\exp(\beta_0+\beta_1)} \times \frac{1+\exp(\beta_0+\beta_1)}{1} \\
&= \exp(\beta_0+\beta_1),
\end{aligned}$$

and together with the result for β_0 above it indicates that $\beta_1 = \log(\pi_{1,1}\pi_{0,0}/\pi_{0,1}\pi_{1,0})$. Using this logic, I define the parameters β^* using the relations $p_{i,j}/p_{i',j}$, giving the forms $\beta_0^* = \log(\frac{p_{1,0}}{p_{0,0}})$ and $\beta_1^* = \log(p_{1,1}p_{0,0}/p_{0,1}p_{1,0})$. Note that the joint probabilities of Y and W, $p_{i,j}$, can be written in terms of the misclassification probabilities α and the joint probabilities Y and X, $\pi_{i,j}$. To see this I use simple conditional probability properties on $p_{i,j}$:

$$p_{i,j} = \text{pr}(Y = i, W = j)$$

$$
\begin{aligned}
&= \sum_{x \in X} \text{pr}(Y = i, W = j | X = x)\text{pr}(X = x) \\
&= \text{pr}(Y = i, W = j | X = 0)\text{pr}(X = 0) + \text{pr}(Y = i, W = j | X = 1)\text{pr}(X = 1) \\
&= \text{pr}(Y = i, W = j, X = 0) + \text{pr}(Y = i, W = j, X = 1) \\
&= \text{pr}(W = j | Y = i, X = 0)\text{pr}(Y = i, X = 0) \\
&\quad + \text{pr}(W = j | Y = i, X = 1)\text{pr}(Y = i, X = 1) \\
&= \text{pr}(W = j | X = 0)\text{pr}(Y = i, X = 0) + \text{pr}(W = j | X = 1)\text{pr}(Y = i, X = 1) \\
&= \alpha_{j0}\pi_{i,0} + \alpha_{j1}\pi_{i,1},
\end{aligned}
$$

where I use the non-differential misclassification property on the second to last line. Using these properties I derive the form of the bias for each of the parameters. For Δ_0,

$$
\begin{aligned}
\Delta_0(\alpha) \equiv \beta_0 - \beta_0^* &= \log\left(\frac{\pi_{1,0}}{\pi_{0,0}}\right) - \log\left(\frac{p_{1,0}}{p_{0,0}}\right) = \log\left(\frac{\frac{\pi_{1,0}}{\pi_{0,0}}}{\frac{p_{1,0}}{p_{0,0}}}\right) = \log\left(\frac{\pi_{1,0}p_{0,0}}{\pi_{0,0}p_{1,0}}\right) \\
&= \log\left(\frac{\pi_{1,0} \times \{\alpha_{0,0}\pi_{0,0} + (1 - \alpha_{11})\pi_{0,1}\}}{\pi_{0,0} \times \{\alpha_{00}\pi_{10} + (1 - \alpha_{11})\pi_{1,1}\}}\right) \\
&= \log\left(\frac{\pi_{1,0}\alpha_{00}\pi_{0,0} \times \{1 + \frac{(1-\alpha_{11})}{\alpha_{00}\pi_{0,0}}\pi_{0,1}\}}{\pi_{0,0}\alpha_{00}\pi_{1,0} \times \{1 + \frac{(1-\alpha_{11})}{\alpha_{00}\pi_{1,0}}\pi_{1,1}\}}\right) \\
&= \log\left(\frac{1 + \xi_0\frac{\alpha_{01}}{\alpha_{00}}}{1 + \xi_1\frac{\alpha_{01}}{\alpha_{00}}}\right),
\end{aligned}
$$

where $\xi_i = \pi_{i,1}/\pi_{i,0} = \text{pr}(Y = i, X = 1)/\text{pr}(Y = i, X = 0) = \text{pr}(X = 1|Y = i)/\text{pr}(X = 0|Y = i,)$, which represents the retrospective odds of the true exposure X given Y. Hence there is no bias between the intercept terms β_0 and β_0^* when the inner term of the log function equals 1. This would occur if $\xi_0 = \xi_1$, or if $\alpha_{01} = 0$.

A similar derivation is done for the bias of Δ_1:

$$\begin{aligned}
\Delta_1(\alpha) \equiv \beta_1 - \beta_1^* &= \log\left(\frac{\pi_{1,1}\pi_{0,0}}{\pi_{0,1}\pi_{1,0}}\right) - \log\left(\frac{p_{1,1}p_{0,0}}{p_{0,1}p_{1,0}}\right) \\
&= \log\left(\frac{\pi_{1,1}\pi_{0,0}p_{0,1}p_{1,0}}{\pi_{0,1}\pi_{1,0}p_{1,1}p_{0,0}}\right) \\
&= \log\left(\frac{\pi_{1,1}\pi_{0,0}}{\pi_{0,1}\pi_{1,0}}\frac{\{\pi_{0,1}\alpha_{11}+\pi_{0,0}(1-\alpha_{00})\}\{\pi_{1,0}\alpha_{00}+\pi_{1,1}(1-\alpha_{11})\}}{\{\pi_{1,1}\alpha_{11}+\pi_{1,0}(1-\alpha_{00})\}\{\pi_{0,0}\alpha_{00}+\pi_{0,1}(1-\alpha_{11})\}}\right) \\
&= \log\left(\frac{\pi_{1,1}\pi_{0,0}}{\pi_{0,1}\pi_{1,0}}\frac{(\pi_{0,1}\alpha_{11})\{1+\frac{\pi_{0,0}(1-\alpha_{00})}{\pi_{0,1}\alpha_{11}}\}(\pi_{1,0}\alpha_{00})\{1+\frac{\pi_{1,1}(1-\alpha_{11})}{\pi_{1,0}\alpha_{00}}\}}{(\pi_{1,1}\alpha_{11})\{1+\frac{\pi_{1,0}(1-\alpha_{00})}{\pi_{1,1}\alpha_{11}}\}(\pi_{0,0}\alpha_{00})\{1+\frac{\pi_{0,1}(1-\alpha_{11})}{\pi_{0,0}\alpha_{00}}\}}\right) \\
&= \log\left(\frac{\{1+\frac{\pi_{0,0}(1-\alpha_{00})}{\pi_{0,1}\alpha_{11}}\}\{1+\frac{\pi_{1,1}(1-\alpha_{11})}{\pi_{1,0}\alpha_{00}}\}}{\{1+\frac{\pi_{1,0}(1-\alpha_{00})}{\pi_{1,1}\alpha_{11}}\}\{1+\frac{\pi_{0,1}(1-\alpha_{11})}{\pi_{0,0}\alpha_{00}}\}}\right) \\
&= \log\left(\frac{\{1+\xi_0^{-1}\frac{\alpha_{10}}{\alpha_{11}}\}\{1+\xi_1\frac{\alpha_{01}}{\alpha_{00}}\}}{\{1+\xi_1^{-1}\frac{\alpha_{10}}{\alpha_{11}}\}\{1+\xi_0\frac{\alpha_{01}}{\alpha_{00}}\}}\right).
\end{aligned}$$

Here there would be no bias between β_1 and β_1^* when either $\alpha_{10} = \alpha_{01} = 0$ or if $\xi_0 = \xi_1$.

Having shown the form of bias in a logistic model setting I consider the form of the bias $E(\widehat{\beta_1^*} - \beta_1)$ when I only considered count data. Because

$$\widehat{\beta_1^*} = \log\left(\frac{\widehat{p_{0,0}}\widehat{p_{1,1}}}{\widehat{p_{1,0}}\widehat{p_{0,1}}}\right) = \log\left(\frac{n_{00}^* n_{11}^*}{n_{10}^* n_{01}^*}\right) \quad \text{and} \quad \widehat{\beta_1} = \log\left(\frac{\widehat{\pi_{0,0}}\widehat{\pi_{1,1}}}{\widehat{\pi_{1,0}}\widehat{\pi_{0,1}}}\right) = \log\left(\frac{n_{00}n_{11}}{n_{10}n_{01}}\right),$$

then $E(\widehat{\beta_1^*} - \beta_1)$ can be expressed as

$$\begin{aligned}
E(\widehat{\beta_1^*} - \beta_1) &= E(\widehat{\beta_1^*} - \beta_1 + \widehat{\beta_1} - \widehat{\beta_1}) \\
&= E(\widehat{\beta_1^*} - \widehat{\beta_1}) + E(\widehat{\beta_1} - \beta_1) \\
&= E\left\{\log\left(\frac{\widehat{p_{0,0}}\widehat{p_{1,1}}}{\widehat{p_{1,0}}\widehat{p_{0,1}}}\right) - \log\left(\frac{\widehat{\pi_{0,0}}\widehat{\pi_{1,1}}}{\widehat{\pi_{1,0}}\widehat{\pi_{0,1}}}\right)\right\},
\end{aligned}$$

where $E(\widehat{\beta_1} - \beta_1) = 0$ follows because $\widehat{\beta_1}$ is the unbiased estimator based on the

correctly measured data. Next, applying the formulations from [15] yields

$$\begin{aligned} E(\widehat{\beta_1^*} - \beta_1) &= E\left\{\log\left(\frac{(\alpha_{10}\widehat{\pi_{1,0}} + \alpha_{11}\widehat{\pi_{1,1}})(\alpha_{00}\widehat{\pi_{0,0}} + \alpha_{01}\widehat{\pi_{0,1}})}{(\alpha_{10}\widehat{\pi_{0,0}} + \alpha_{11}\widehat{\pi_{0,1}})(\alpha_{00}\widehat{\pi_{1,0}} + \alpha_{01}\widehat{\pi_{1,1}})}\right)\right\} \\ &\quad -E\left\{\log\left(\frac{\widehat{\pi_{0,0}}\widehat{\pi_{1,1}}}{\widehat{\pi_{1,0}}\widehat{\pi_{0,1}}}\right)\right\} \\ &= E\left\{\log\left(\alpha_{10}\widehat{\pi_{1,0}} + \alpha_{11}\widehat{\pi_{1,1}}\right)\right\} + E\left\{\log\left(\alpha_{00}\widehat{\pi_{0,0}} + \alpha_{01}\widehat{\pi_{0,1}}\right)\right\} \\ &\quad -E\left\{\log\left(\alpha_{10}\widehat{\pi_{0,0}} + \alpha_{11}\widehat{\pi_{0,1}}\right)\right\} - E\left\{\log\left(\alpha_{00}\widehat{\pi_{1,0}} + \alpha_{01}\widehat{\pi_{1,1}}\right)\right\} \\ &\quad -E\left\{\log\left(\widehat{\pi_{0,0}}\right)\right\} - E\left\{\log\left(\widehat{\pi_{1,1}}\right)\right\} + E\left\{\log\left(\widehat{\pi_{1,0}}\right)\right\} + E\left\{\log\left(\widehat{\pi_{0,1}}\right)\right\}, \end{aligned}$$

where the log operator is applied to the insides. While calculation of these terms is nontrival; we can see from the last line that term $E(\widehat{\beta_1^*} - \beta_1)$ may equal 0 when the misclassification probabilities $\alpha_{01} = \alpha_{01} = 0$, implying that $\alpha_{00} = \alpha_{11} = 1$. The conclusion of this proof is that the difference between $\widehat{\beta_1^*}$ and β_1 is nonzero when misclassification exists. Finally [6] pointed out the effects of misclassified data not only causes bias in the estimates, but can reduce the power of a significance test.

1.1.5 Simulation

Having shown the misclassification issue theoretically, I demonstrate this issue concretely by conducting a few simple simulations. Define X as a binary variable with $\text{pr}(X = 1) = 0.55$. Let Y be a binary response with the model $\text{pr}(Y = 1|X) = [1 + \exp\{-(\beta_0 + \beta_1 X)\}]^{-1}$, where $\beta_0 = -1$ and $\beta_1 = 0.5$. We model W_a, the misclassified version of X, as follows: set the misclassification probabilities $\text{pr}(W_a = 1|X = 0)$ and $\text{pr}(W_a = 0|X = 1)$ to 0.3. So, the misclassified version W_a is generated by setting $W_a = X \times \tilde{V} + (1 - X) \times (1 - \hat{V})$, where $\tilde{V}$ and $\hat{V}$ are the Bernoulli random variables with success probabilities, $1 - \text{pr}(W_a = 1|X = 0) = 0.7$ and $1 - \text{pr}(W_a = 0|X = 1) = 0.7$. I also consider W_b, defined similarly as W_a but with the following misclassification probabilities $\text{pr}(W_b = 0|X = 1) = 0.3$, $\text{pr}(W_b = 1|X = 0) = 0.05$.

I generated a cohort of size $n = 1000$, which means each cohort consisted of n iid copies of (Y, X, W_a, W_b). To analyze each cohort data, I fit a logistic model for Y on X, Y on W_a and Y on W_b, and reported the average bias of the parameter estimates based on 500 replications. I also report the average standard deviation of the estimates in parenthesis.

Table 1.2: Average bias of logistic regression parameters using true exposure X and misclassified exposure W_a and W_b.

	$\widehat{\beta}_0$	$\hat{\beta}_1$
X	$-0.006(0.151)$	$0.011\ (0.196)$
W_a	$0.181\ (0.141)$	$-0.298(0.192)$
W_b	$0.145\ (0.127)$	$-0.164(0.193)$

The results are summarized in Table 1.2. Here I note two things. First, the bias of the estimates using the misclassified W_a and W_b is much larger than if X was used in the regression model. More importantly, it is evident that the bias occurs in both β_0 and β_1, hence the misclassification can impact all model parameters. These results show the importance of addressing this misclassified covariate issue. Additionally, I consider the power of the significance tests when using W_a and W_b. Under the scenario where the true variable X, the power is 100% for β_1. However, if the misclassified variables are used, the power for β_1 are 75.1% and 89.0%, corresponding to W_a and W_b, respectively.

Previously I considered only nondifferential misclassification. Next, I consider the effect of differential misclassification in a similar simulation as in Table 1. However, now the model for W_c, the misclassified version of X, is as follows: set the misclassification probabilities $\text{pr}(W_c = 1|X = 0, Y = 1)$ and $\text{pr}(W_c = 0|X = 1, Y = 1)$ to

0.3, and set $\text{pr}(W_c = 1|X = 0, Y = 0)$ and $\text{pr}(W_c = 0|X = 1, Y = 0)$ to 0.1. In this situation, the misclassification probabilities are dependent on the value of the Y. I regressed Y on X and Y on W_c in a logistic regression model and report average bias of parameter estimates based on 500 replications. I also report the average standard deviation of the estimates in Table 1.3. We see the inclusion of differential

Table 1.3: Average bias of logistic regression parameters using true exposure X and misclassified exposure W_c under differential misclassification.

	$\hat{\beta}_0$	$\hat{\beta}_1$
X	$-0.006(0.151)$	$0.011(0.196)$
W_c	$0.185(0.141)$	$-0.309(0.192)$

misclassification causes the average bias to increase for both β_0 and β_1 even more so than nondifferential misclassification. Additionally when using W_c, the power for β_1 is 75.8%.

1.1.6 Misclassification in real world data sets

I have demonstrated, theoretically and through some simulations, the implications of misclassified variables in a logistic regression setting; whether misclassified variables occur in reality is another question. A literature review shows the prevalence of misclassified variables in the realm of public health domain. This is due to the use of observational data recorded by incomplete or through inaccurate methods. In public health surveys, misclassified variables typically occur due the user response under duress or recall bias. For example, [18] studied the impact of misclassification due to subject fatigue from having to answer too many questions, and this fatigue negatively impacts statistical power in a logistic regression setting. [49] found higher rates of misclassification of self reported mammography use by black women than

white women in the Behavioral Risk Factor Surveillence System (BRFSS). [60] found that a retrospective questionnaire overestimated the intensity of children's headaches . However, misclassified data can also occur in patient records. One example is data obtained from Surveillance, Epidemiology and End Results (SEER), which contains cancer incidence data from specific regions in the United States. [35] studied the the rate of under-ascertainment of radiotherapy receipt in the SEER data for Los Angeles and Detroit subject diagnosed from June 2005 to February 2007. They found that higher rates of under reporting of radiotherapy receipt in the SEER data were mainly associated with age of the patient and patient's insurance type.

A similar example comes from [24], who found misclassification of race and ethnicity in the Greater Bay Area Cancer Registry, which may have been caused by faulty collection methods such as using the patient's surname in determining race/ethnicity.

1.2 Assessment of current methods to address misclassification

Much research has been conducted to address the issue of misclassification in covariates and responses for various models. [8] provided a classical review of measurement error in different regression settings and gave details on methods of adjustment for this error. [28] provided an overview of misclassified discrete variables as well as mismeasured continuous variables, and used Bayesian attempts to adjust for the misclassification in simple scenarios. [7] studied the effect of using a misclassified exposure on the bias of the parameters of a linear regression model. They provided both a mathematical and numerical example to illustrate the effect.

The most common approach in reducing the misclassification bias is the use of a subset of the data, referred to as validation data, which contains the true exposure along with the misclassified exposure variable. The justification is that estimated misclassification probabilities from the subsample can be used to make inferences on

the disease-exposure association for those not in the sample [9]. There are several approaches of making use of the validation data. Towards that goal, [47] proposed a matrix approach. In this method they estimated of the corrected log-odds ratio by solving a system of equations involving the sample proportions $\widehat{\text{pr}}(W = 1|\widetilde{X} = 1, Y = i)$ and $\widehat{\text{pr}}(W = 0|\widetilde{X} = 0, Y = i)$. Here the sample proportions are based on $\widetilde{X}$ which are measurements obtained from the validated data.

There are also approaches to modeling the bias inducing mechanisms parametrically, then estimating the error free covariates using these models. [57] considered a parametric approach that incorporated a likelihood for the participants from the full study and a likelihood for participates who also have validation data. Within each of the likelihoods are validation selection model, a logistic model for the binary response, and a reclassification model for estimating the values of the non-misclassified covariates. One assumption they make is that the model for being selected into the validation study is independent of the true non-misclassified covariates. Finally, another way of addressing the misclassification issue is through the MC-SIMEX approach [36].

In this approach, they proposed a two step process consisting of simulation and extrapolation. In the simulation step, the misclassified variable is simulated using a misclassification matrix, Π, whose diagonal elements are restricted to the range of $(0.5, 1)$ to create a pseudo data set. There are multiple pseudo data sets created for a fixed grid of values λ. In the extrapolation step, the data from each pseudo data set is used in a regression to generate parameter estimates. The parameter estimates obtained are then finally used in another least squares regression to obtain the MC-SIMEX estimator. The variance can be obtained through different methods including the use of a bootstrap technique.

An alternative approach to correcting the misclassification was suggested by [40].

Rather then trying to model the true exposure unknown variable, they suggested correcting the misclassification through a bias factor. This bias factor is dependent on the sensitivity of misclassification for different categories of the exposure. The sensitivity is modeled parametrically using a beta distribution.

The main difference between my approach and previous work is that I consider limited or no access to validation data to estimate the misclassification. In particular, I consider the amount of improvement possible if there is only access to instrumental variables that are also measured along with the misclassified exposure, binary response, and other prognostic factors. The motivating feature of instrumental variables is that they have an associated with the true exposure variable, but not with the outcome variable. Particularly, the conditional on the true exposure variable, the instrumental variable must be independent of the outcome. The use of instrumental variables comes primarily from the economic field - [55] suggested the use of instrumental variables to address measurement error in certain econometric models. Additionally he considered ways of reducing the size of the set of instruments required through dimension reduction tools such as canonical correlation. However, instrumental variables have also advocated for use in other areas of research. For example, [31] considered the use of instrumental variables in the field of causal inference as a technique to control bias due to unmeasured confounding variables. That said, my work will encompass the use of instrumental variables in both frequentist and Bayesian settings for bias adjustment due to misclassification.

2. MATCHED CASE-CONTROL DATA WITH A MISCLASSIFIED EXPOSURE: WHAT CAN BE DONE WITH INSTRUMENTAL VARIABLES?[1]

2.1 Introduction

The purpose of matched case-control designs are to determine the association between an outcome and an exposure from observational studies. To remove the influence of confounding variables, cases are matched with controls based on these potential confounding variables, that results in a stratum or a matched set. The matched case-control data set will then have many of these matched sets. A logistic model is typically used to model the disease incidence in terms of the exposure and other adjustment covariates (prognostic factors). Parameter estimation occurs by maximizing a conditional likelihood function. However, the estimators of the parameters are biased when the exposure variable is mismeasured or misclassified. This bias may occur due to many different reasons including but not limited to recall error and misreporting. For this work, I will focus on the bias due to a misclassified binary exposure variable in a matched case-control study.

Several Bayesian and frequentist methods have been developed to address the misclassification bias of a binary exposure variable in a matched study. For convention, I use the notations Y, X and W to refer to the binary disease indicator, the binary exposure variable, and the misclasssified version of X. Usually, X is not observed in the main study but Y and W are, while for the validation data Y, X, and W are observed. Also, the validation dataset has a much smaller sample size than

[1]Portions of this work reprinted with permission from Manuel, C.M. and Wang, S. and Sinha, S., "Matched Case-Control Data with a Misclassified Exposure: What can be done with Instrumental Variables?", Biostatistics, 2019, kxz012, by permission of Oxford University Press

the main study. However, this validation dataset can be used to estimate the misclassification probabilities. This is important because to remove the misclassification bias these estimated misclassification probabilities are used in the analysis.

I now discuss some previous research conducted in this field. [53] considered a full likelihood approach for handling a misclassified binary exposure in matched case-control data. He derived the likelihood in a multinomial model format, with the cell probabilities being functions of the common odds ratio and non-differential misclassification probabilities. Assuming the existence of validation data, he estimated the parameters using a Bayesian framework. [51] also developed a Bayesian method of estimating the parameters for a binary exposure variable that was misclassified in matched case-control data. However they assumed that the correct values of the exposure were available on a number of matched sets. In one of the three methods that the authors proposed, they estimated the parameter from the validation data only where the true values of the exposure were available. Afterwards, they utilized the first stage analysis result as a prior distribution for use in the second stage of the analysis. In the second stage the likelihood was based on the matched sets from which only the misclassified exposure was observed but not the true exposure. [11] developed a Bayesian approach where the main parameters of interest included the exposure prevalence, specificity, and sensitivities among cases and controls. They proposed a Bayesian inference with flexible priors for each of these parameters. [41] also considered a Bayesian method for misclassification bias in a 1:1 matched case-control study. Rather than assuming the existence of validation data, they supposed that expert prior knowledge exists and could be incorporated for the disease-exposure association and misclassification probabilities. [47] compared three methods of bias reduction: the matrix method, the inverse matrix method, and the maximum likelihood approach, under differential and non-differential misclassification scenarios.

In a regular case-control study with a misclassified binary exposure, [43] considered different robust weighted estimators for the odds ratio parameter based on two validation data sets. The types of validation data include an internal validation data set and an external but less precise validation data set.

Finally, in a different approach, [17] considered replacing some terms in the Mantel-Haenszel estimator with their corresponding conditional expected values given the observed data. To obtain these conditional expected values, the misclassification probabilities were required, and they were obtained from a validation data. Because of issues in computing the large sample variance, the authors suggest a Bayesian credible interval for the common odds ratio parameter with non-informative priors.

The main difference between the previous works and the problem I address is that I do not assume the existence of validation data. Hence, there is no need for the true binary exposure X to be available in any part of the data. Rather, I require that a set of instrumental variables are available for the study subjects. Using the definition given in [26], the instrumental variables are defined as variables that are correlated with X, yet are uncorrelated with the response as well as the confounding variables conditional on X. Therefore, my goal is to propose a statistical method of inference for this scenario. While instrumental variables are frequently used to solve issues of endogeneity, no one has used instrumental variables to reduce the misclassification bias from either a retrospective case-control or matched case-control study. In particular, instrumental variables have been used to estimate parameters in a linear regression or polynomial regression model with a mismeasured numeric covariate [5, 30]. However, the use of instrumental variables for addressing misclassification bias of a binary covariate in a retrospective study is limited [33, 56].

My work is inspired by the research of [33], who demonstrated that the misclassification probabilities and the latent model for X are nonparametrically identifiable

when there is a discrete instrumental variable. For parameter estimation, [33] provides a matrix diagonalization technique. In comparison to [33]'s work, I 1) assume the existence of many confounding variables as well as several instrumental variables, 2) I assume a parametric model for the distribution of X, and finally, 3) I assume that conditional on the true exposure, the misclassification probabilities do not depend on other variables. Based on these assumptions, I propose two methods of estimation in a non-trivial conditional likelihood setting. To make the method numerically work for a reasonable sample size, a parametric model assumption on the conditional distribution of X is necessary.

For this problem, I propose two estimation methods. The first method is a two step procedure. In the first step, I estimate the misclassification probabilities and the conditional distribution of the true exposure given the confounding variables and instrumental variables. In the second step, I obtain the induced model of the response given the observed variables. In doing so, I make use of the model estimated in the first step. Finally, a conditional likelihood is constructed and maximized to estimate the disease-exposure association parameters. The second method is an efficient method. I now outline the remainder of the article. Some further background information is provided in Section 2, while in Section 3 the methodology is discussed explicitly. In Section 4, I provide simulation studies and assessment. Finally, I apply the proposed methods to the analysis of a real data set in Section 5. The real data is a nested case-control data that is generated from the US birth cohort from the year of 1989. The data is analyzed to determine the effect of smoking during pregnancy, the main exposure of interest, on the incidence of low birth weight. I propose the use of instrumental variables to reduce the misclassification bias since the average number of cigarettes smoked daily cannot be measured accurately. Additionally, there are no validation data to access the misclassification probabilities for the exposure. Finally,

Section 6 contains a discussion of this work.

2.2 Models and Background

I assume that that there is 1:M matched case-control data set with n strata. Thus for each case subject, there are M control subjects within each stratum. The notation $a_{i,j}$ denotes the variable a for the jth subject in the ith stratum. I denote the matched data set by $Y_{i,j}, X_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{S}_i, j = 1, \ldots, (M+1)$, $i = 1, \ldots, n$. Y, X, and $\boldsymbol{Z}$ represent the response, the main binary exposure, and prognostic factors used for adjustment, and $\boldsymbol{S}$ represents the set of confounding variables used in matching cases and controls. The definition of confounding variables I use comes from [27], who states that the confounding variables $\boldsymbol{S}$ causally influence both X and Y [27]. Because the matching is done based on the values of the confounding variables, all subjects within a matched set have the same values of $\boldsymbol{S}$. Note that the value of $\boldsymbol{S}$ does not vary within a given stratum. I partition the prognostic factor $\boldsymbol{Z}$ as $\boldsymbol{Z} = (\boldsymbol{Z}_1^T, \boldsymbol{Z}_2^T)^T$, where $\boldsymbol{Z}_1$ is a prognostic factor that is causally independent of the exposure and $\boldsymbol{Z}_2$ is the prognostic factor that is a confounding variable. In our context, X is never observed but a misclassified version of X, W, is recorded instead. Additionally, I assume that W is independent of the other data $(\boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}, Y)$ conditioned on the true exposure X. I now define $\alpha_0 = \text{pr}(W = 1|X = 0) = \text{pr}(W = 1|X = 0, Y = 0) = \text{pr}(W = 1|X = 0, Y = 1)$ and $\alpha_1 = \text{pr}(W = 0|X = 1) = \text{pr}(W = 0|X = 1, Y = 0) = \text{pr}(W = 0|X = 1, Y = 1)$. Additionally, I assume the existence of a set of instrumental variables $\boldsymbol{X}^*$ for X is in the data. Following the definition [26], the instrumental variables satisfy the following conditions: a) $\boldsymbol{X}^*$ do have a direct influence on $\boldsymbol{X}$, b) $\boldsymbol{X}^*$ may have a direct influence on $\boldsymbol{Z}_1$ (a non-confounder), c) $\boldsymbol{X}^*$ is independent of all the confounding variables found in $\boldsymbol{S}$ and $\boldsymbol{Z}_2$, d) conditional independence of $\boldsymbol{X}^*$ and Y for a given X and $\boldsymbol{Z}_1$. However, I point out that that statistical validity of

the proposed methods does not rely on the dependence of $\boldsymbol{X}^*$ and any component of $\boldsymbol{Z}$. These assumptions are depicted in Figure 2.1 where the arrows indicate the variables of influence (the variable from which the arrow originates from) and the variables being influenced (the variable pointed at by the arrow). In this set-up,

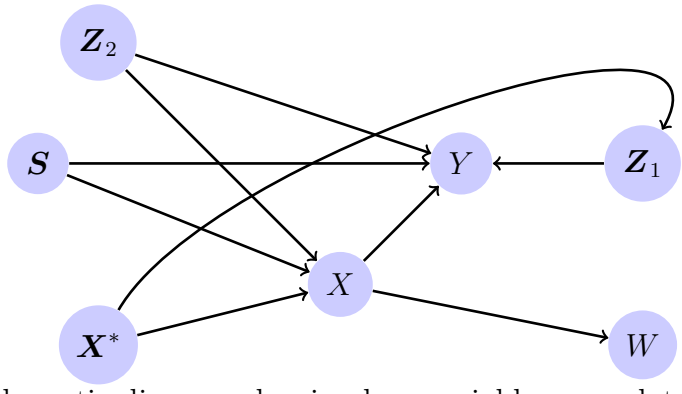

Figure 2.1: Schematic diagram showing how variables are related. (Manuel, Wang, Sinha 2019)

the observed data are $\{Y_{i,j}, \boldsymbol{X}^*_{i,j}, W_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{S}_i, j = 1, \ldots, (M+1), i = 1, \ldots, n\}$. In general, the assumed model of Y for the ith stratum is

$$\text{pr}(Y_{i,j} = 1 | \boldsymbol{S}_i, X_{i,j}, \boldsymbol{Z}_{i,j}) = H\{g_0(\boldsymbol{S}_i) + \beta_1 X_{i,j} + \boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j}\}, \tag{2.1}$$

where $H(u) = 1/\{1+\exp(-u)\}$. Here $g_0(\boldsymbol{S}_i)$ indicates how the stratification variable influences success probability of the response variable, whereas β_1 and $\boldsymbol{\beta}_2$ association parameters for X and $\boldsymbol{Z}$, respectively. The model can also be generalized through the inclusion of an interaction term between X and $\boldsymbol{Z}$.

When there are no $\boldsymbol{Z}_2$ in the data, then $\boldsymbol{S}$ will be the only set of confounding variables used for matching. This means that the regression parameter β_1 in Equation

(3.1) will have a causal interpretation conditional on $\boldsymbol{Z} = \boldsymbol{Z}_1$ (32). Also, regardless of the presence of the prognostic factors $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, the dependence between $\boldsymbol{X}^*$ and $\boldsymbol{Z}_1$, and the independence of $\boldsymbol{X}^*$ and $\boldsymbol{Z}_2$, the proposed estimation techniques are always valid. When the true X is observed in the data, the parameter $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T$ can be estimated through maximizing the conditional likelihood $\mathcal{L}_c(\boldsymbol{\beta})$. Conditioning on the number of cases for each stratum eliminates the nuisance parameter $g_0(\boldsymbol{S}_i)$. Thus,

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\beta}|X, \boldsymbol{Z}) &= \prod_{i=1}^{n} \text{pr}\{Y = Y_{i,j}, j = 1, \ldots, (M+1)|\boldsymbol{S}_i, X_{i,j}, \boldsymbol{Z}_{i,j}, \\
&\qquad j = 1, \ldots, (M+1), \sum_{j=1}^{M+1} Y_{i,j} = 1\} \\
&= \prod_{i=1}^{n} \frac{\prod_{j=1}^{M+1} \text{pr}^{Y_{i,j}}(Y = 1|\boldsymbol{S}_i, X_{i,j}, \boldsymbol{Z}_{i,j}) \text{pr}^{(1-Y_{i,j})}(Y = 0|\boldsymbol{S}_i, X_{i,j}, \boldsymbol{Z}_{i,j})}{\sum_{k=1}^{M+1} \text{pr}(Y = 1|\boldsymbol{S}_i, X_{i,k}, \boldsymbol{Z}_{i,k}) \prod_{r \neq k} \text{pr}(Y = 0|\boldsymbol{S}_i, X_{i,r}, \boldsymbol{Z}_{i,r})} \\
&= \prod_{i=1}^{n} \frac{\sum_{j=1}^{M+1} Y_{i,j} \exp(\beta_1 X_{i,j} + \boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j})}{\sum_{j=1}^{M+1} \exp(\beta_1 X_{i,j} + \boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j})}.
\end{aligned}$$

For the naive approach, X gets replaced by W in $\mathcal{L}_c$ and so the estimators are now defined as $\text{argmax}_{\boldsymbol{\beta}} \mathcal{L}_c(\boldsymbol{\beta}|W, \boldsymbol{Z})$. The degree of bias in naive estimators depends on the severity of the misclassification.

2.3 Proposed methodology

2.3.1 Intuitive estimator

I now discuss the first estimator. The goal is to estimate the regression parameters $\boldsymbol{\beta}$ in model (3.1), which requires the conditional distribution of Y given the observable random variables $\boldsymbol{S}, W, \boldsymbol{X}^*, \boldsymbol{Z}$. First I specify a model for X given $\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}$. This model along with the misclassification probabilities induces the model for X given $\boldsymbol{S}, W, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}$, and is denoted as $\text{pr}(X = 1|\boldsymbol{S}, W, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$. The

resulting induced model $\text{pr}(X = 1|\boldsymbol{S}, W, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$ along with model (3.1) yields the model for Y given $\boldsymbol{S}, W, \boldsymbol{X}^*, \boldsymbol{Z}$. Subsequently the estimation is carried out in two steps. To proceed, first I assume a logistic model for the probability of $X = 1$ given $\boldsymbol{S}$, $\boldsymbol{X}^*$, $\boldsymbol{Z}$ in the control population:

$$\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) = H(\gamma_0 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z}). \tag{2.2}$$

The term $H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})$ is used to denote $H(\gamma_0 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z})$, with $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T, \boldsymbol{\gamma}_3^T)^T$. Next I obtain the induced conditional probability model for the observed W:

$$\begin{aligned}
& \text{pr}(W = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
= \; & \text{pr}(W = 1|\boldsymbol{S}, X = 0, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})\text{pr}(X = 0|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
& + \text{pr}(W = 1|\boldsymbol{S}, X = 1, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
= \; & \text{pr}(W = 1|X = 0, Y = 0, \boldsymbol{Z})\text{pr}(X = 0|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
& + \text{pr}(W = 1|X = 1, Y = 0, \boldsymbol{Z})\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
= \; & \alpha_0\{1 - \text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})\} + (1 - \alpha_1)\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
= \; & \alpha_0 + (1 - \alpha_0 - \alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}),
\end{aligned} \tag{2.3}$$

where the second equality comes from the assumptions placed on the misclassification probability, and $\alpha_0 = \text{pr}(W = 1|X = 0)$ and $\alpha_1 = \text{pr}(W = 0|X = 1)$.

Using results from[29], I make the assumption that $0 < \alpha_0 + \alpha_1 < 1$, which guarantees model parameter identification. A detailed proof of the identifiability is given in Appendix A.1. [29] applied this restriction for parameter identification for the scenario of misclassified response variables. Next I write α_0 and α_1 using the formulas $\alpha_0 \equiv \alpha_0(\boldsymbol{\eta}) = \exp(\eta_0)/\{1 + \exp(\eta_0) + \exp(\eta_1)\}$ and $\alpha_1 \equiv \alpha_1(\boldsymbol{\eta}) =$

$\exp(\eta_1)/\{1+\exp(\eta_0)+\exp(\eta_1)\}$, where $\eta_0, \eta_1 \in \mathcal{R}$, and this formulation make (α_0, α_1) to automatically satisfy the restriction $0 < \alpha_0 + \alpha_1 < 1$. Next I denote $\mathrm{pr}(W = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$ by $p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})$. Then $\boldsymbol{\gamma}$ and $\boldsymbol{\eta} = (\eta_0, \eta_1)^T$ are estimated by maximizing the following likelihood:

$$\begin{aligned}\mathcal{L}_1(\boldsymbol{\gamma}, \boldsymbol{\eta}) &= \prod_{i=1}^{n}\sum_{j=1}^{M+1}\Bigg[\Big\{p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})\Big\}^{W_{i,j}} \\ &\quad \times\Big\{1 - p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})\Big\}^{1-W_{i,j}}\Bigg]^{(1-Y_{i,j})}.\end{aligned}$$

I define the estimators $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\eta}}$ for $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ as

$$(\boldsymbol{\gamma}, \boldsymbol{\eta}) = \arg\max_{\boldsymbol{\gamma}, \boldsymbol{\eta}} \mathcal{L}_1(\boldsymbol{\gamma}, \boldsymbol{\eta}).$$

In particular, I obtain $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\eta}}$ by solving $\boldsymbol{S}_\gamma(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \sum_{i=1}^n U_{i,\gamma}(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \boldsymbol{0}$ and $\boldsymbol{S}_\eta(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \sum_{i=1}^n U_{i,\eta}(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \boldsymbol{0}$, where

$$\begin{aligned}\boldsymbol{U}_{i,\gamma}(\boldsymbol{\gamma}, \boldsymbol{\eta}) &= \sum_{j=1}^{M+1}(1 - Y_{i,j})\Big\{W_{i,j} - p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})\Big\} \\ &\quad \times\frac{1}{p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})\{1 - p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, X^*_{i,j}, \boldsymbol{Z}_{i,j})\}} \\ &\quad \times\{1 - \alpha_0(\boldsymbol{\eta}) - \alpha_1(\boldsymbol{\eta})\}H(\boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}) \\ &\quad \times\{1 - H(\boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})\}\begin{pmatrix} 1 \\ \boldsymbol{S}_i \\ \boldsymbol{X}^*_{i,j} \\ \boldsymbol{Z}_{i,j}\end{pmatrix}, \\ \boldsymbol{U}_{i,\eta}(\boldsymbol{\gamma}, \boldsymbol{\eta}) &= \sum_{j=1}^{M+1}(1 - Y_{i,j})\Big\{W_{i,j} - p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})\Big\} \\ &\quad \times\frac{1}{p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})\{1 - p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})\}}\end{aligned}$$

$$\times \begin{bmatrix} \alpha_0(\boldsymbol{\eta})\{1-\alpha_0(\boldsymbol{\eta})\} - \alpha_0(\boldsymbol{\eta})\{1-\alpha(\boldsymbol{\eta})\}H(\boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}) \\ -\alpha_0(\boldsymbol{\eta})\alpha_1(\boldsymbol{\eta}) - \alpha_1(\boldsymbol{\eta})\{1-\alpha(\boldsymbol{\eta})\}H(\boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}) \end{bmatrix},$$

with $1-\alpha(\boldsymbol{\eta}) = 1-\alpha_0(\boldsymbol{\eta})-\alpha_1(\boldsymbol{\eta})$. I then obtain the induced model for X given $W=\omega$, $\boldsymbol{X}^*$, $\boldsymbol{S}$, and $\boldsymbol{Z}$ among the control subjects as follows:

$$\begin{aligned} & \text{pr}(X=1|\boldsymbol{S}, W=1, \boldsymbol{X}^*, Y=0, \boldsymbol{Z}) \\ = \; & \frac{\text{pr}(W=1|\boldsymbol{S}, X=1, \boldsymbol{X}^*, Y=0, \boldsymbol{Z})\text{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=0, \boldsymbol{Z})}{\text{pr}(W=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=0, \boldsymbol{Z})} \\ = \; & \frac{(1-\alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}{\alpha_0 + (1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}, \end{aligned} \tag{2.4}$$

$$\begin{aligned} & \text{pr}(X=1|\boldsymbol{S}, W=0, \boldsymbol{X}^*, Y=0, \boldsymbol{Z}) \\ = \; & \frac{\text{pr}(W=0|\boldsymbol{S}, X=1, \boldsymbol{X}^*, Y=0, \boldsymbol{Z})\text{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=0, \boldsymbol{Z})}{\text{pr}(W=0|\boldsymbol{S}, \boldsymbol{X}^*, Y=0, \boldsymbol{Z})} \\ = \; & \frac{\alpha_1 H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}{1-\alpha_0-(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}. \end{aligned} \tag{2.5}$$

Now, the induced model for the response Y given $\boldsymbol{S}$, W, $\boldsymbol{X}^*$, $\boldsymbol{Z}$ is provided in the following lemma.

Lemma 1. *Under the assumptions stated previously the induced model for Y given $\boldsymbol{S}$, W, $\boldsymbol{X}^*$, and $\boldsymbol{Z}$ is*

$$pr(Y=1|\boldsymbol{S}, W, \boldsymbol{X}^*, \boldsymbol{Z}) = H\{g_0(\boldsymbol{S}) + \boldsymbol{\beta}_2^T \boldsymbol{Z} + g_1(\beta_1, \boldsymbol{S}_i, W, \boldsymbol{X}^*, \boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}, \tag{2.6}$$

where

$$\begin{aligned} & \exp\{g_1(\beta_1, \boldsymbol{S}, W=1, \boldsymbol{X}^*, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\eta})\} \\ = \; & \frac{\exp(\beta_1)(1-\alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) + \alpha_0\{1-H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})\}}{\alpha_0 + (1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}, \end{aligned}$$

$$
\begin{aligned}
& \exp\{g_1(\beta_1, \boldsymbol{S}, W=0, \boldsymbol{X}^*, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\eta})\} \\
= & \frac{\exp(\beta_1)\alpha_1 H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) + (1-\alpha_0)\{1 - H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})\}}{1 - \alpha_0 - (1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}.
\end{aligned}
$$

Model (2.6) for the response Y is in terms of observed variables, $\boldsymbol{S}$, W, $\boldsymbol{X}^*$, and $\boldsymbol{Z}$, and it involves the main association parameters $\boldsymbol{\beta}$. To estimate $\boldsymbol{\beta}$, we form the conditional likelihood function based on this induced probability model and maximize with respect to $\boldsymbol{\beta}$. Now I define the estimator of $\boldsymbol{\beta}$ as

$$
\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}_2(\boldsymbol{\beta}|\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\gamma}}),
$$

where the conditional likelihood function is

$$
\begin{aligned}
\mathcal{L}_2(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\gamma}) &= \prod_{i=1}^{n} \mathrm{pr}\{Y_{i,j}, j=1,\ldots,(M+1)|\boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \\
&\qquad j=1,\ldots,(M+1), \sum_{j=1}^{M+1} Y_{i,j} = 1\} \\
&= \prod_{i=1}^{n} \frac{\prod_{j=1}^{M+1} \left\{ \mathrm{pr}^{Y_{i,j}}(Y=1|\boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}) \right.}{\sum_{k=1}^{M+1} \left\{ \mathrm{pr}(Y=1|\boldsymbol{S}_i, W_{i,k}, \boldsymbol{X}^*_{i,k}, \boldsymbol{Z}_{i,k}) \right.} \\
&\qquad \frac{\left. \times \mathrm{pr}^{(1-Y_{i,j})}(Y=0|\boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}) \right\}}{\left. \times \prod_{r\neq k} \mathrm{pr}(Y=0|\boldsymbol{S}_i, W_{i,r}, \boldsymbol{X}^*_{i,r}, \boldsymbol{Z}_{i,r}) \right\}} \\
&= \prod_{i=1}^{n} \frac{\sum_{j=1}^{M+1} Y_{i,j} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}}{\sum_{j=1}^{M+1} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\gamma}, \boldsymbol{\eta})\}} \\
&= \prod_{i=1}^{n} \frac{\exp[\sum_{j=1}^{M+1} Y_{i,j}\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})]}{\sum_{j=1}^{M+1} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}}.
\end{aligned}
$$

Therefore, $\widehat{\boldsymbol{\beta}}$ can be determined by solving $S_{\beta_1}(\boldsymbol{\beta}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} U_{i,\beta_1}(\boldsymbol{\beta}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\gamma}}) = 0$

and $\boldsymbol{S}_{\boldsymbol{\beta}_2}(\boldsymbol{\beta}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \boldsymbol{U}_{i,\boldsymbol{\beta}_2}(\boldsymbol{\beta}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\gamma}}) = \mathbf{0}$, where

$$
\begin{aligned}
U_{i,\beta_1}(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) = & \sum_{j=1}^{M+1} \left[Y_{i,j} - \frac{\exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}}{\sum_{k=1}^{M+1} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,k} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,k}, \boldsymbol{X}^*_{i,k}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}} \right] \\
& \times g_{\beta_1}(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma}), \\
\boldsymbol{U}_{i,\boldsymbol{\beta}_2}(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) = & \sum_{j=1}^{M+1} \left[Y_{i,j} - \frac{\exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}}{\sum_{k=1}^{M+1} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,k} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,k}, \boldsymbol{X}^*_{i,k}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}} \right] \\
& \times \boldsymbol{Z}_{i,j},
\end{aligned}
$$

with $g_{\beta_1}(\cdot) = \partial g_1(\cdot)/\partial \beta_1$. I now present the following main theorem whose proof can be found in the Appendix.

Theorem 1. *Under the standard regularity conditions and as $n \to \infty$, the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a mean-zero normal distribution. Additionally, the asymptotic variance of $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ can be consistently estimated by the last $(p+1)$ rows and the last $(p+1)$ columns of $\widehat{A}^{-1}(\sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{U}_i^T/n)\widehat{A}^{-T}$, where $\widehat{A} = -(1/n)\partial \boldsymbol{S}_\theta/\partial \boldsymbol{\theta}$, $\boldsymbol{S}_\theta = (\boldsymbol{S}_\gamma^T(\boldsymbol{\gamma}, \boldsymbol{\eta}), \boldsymbol{S}_\eta^T(\boldsymbol{\gamma}, \boldsymbol{\eta}), S_{\beta_1}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}), \boldsymbol{S}_{\boldsymbol{\beta}_2}^T(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}))^T$ and $\boldsymbol{U}_i = (\boldsymbol{U}_{i,\gamma}^T, \boldsymbol{U}_{i,\eta}^T, U_{i,\beta_1}, \boldsymbol{U}_{i,\boldsymbol{\beta}_2}^T)^T$.*

2.3.2 Efficient estimator

The estimator obtained from the second method requires the following.

Lemma 2. *For the proposed models (3.1) and (2.2) and the assumptions placed on the misclassification probabilities,*

$$
\begin{aligned}
&i) \quad pr(Y = 1 | \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) = H\{g_0(\boldsymbol{S}) + \boldsymbol{\beta}_2^T \boldsymbol{Z} + g_2(\boldsymbol{\gamma}, \beta_1, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})\}, \\
&ii) \quad pr(X = 1 | \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}, Y = 1) = H(\gamma_0 + \beta_1 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z}), \\
&iii) \quad pr(W = 1 | \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}, Y = 1) = \alpha_0 + (1 - \alpha) H(\gamma_0 + \beta_1 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z})
\end{aligned}
$$

where $(1-\alpha) = (1-\alpha_0-\alpha_1)$ *and*

$$g_2(\boldsymbol{\gamma}, \beta_1, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) = log\{1 - H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) + \exp(\beta_1) H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})\}.$$

The likelihood for the observed data $\{W_{i,j}, Y_{i,j}, i = 1, \ldots, (M+1)\}$ from the ith stratum conditional on $\boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, j = 1, \ldots, (M+1)$ is defined as

$$\begin{aligned}\mathcal{L}_i = \prod_{j=1}^{M+1} \text{pr}(W_{i,j}, Y_{i,j} | \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}) &= \prod_{j=1}^{M+1} \text{pr}(W_{i,j} | \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, Y_{i,j}, \boldsymbol{Z}_{i,j}) \\ &\quad \times \text{pr}(Y_{i,j} | \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}).\end{aligned}$$

Lemma 2 says that $\text{pr}(Y_{i,j} | \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})$ is in a logistic form that still contains the nuisance intercept parameter $g_0(\boldsymbol{S}_i)$. The dimension of this parameter increases with the number of matched sets n. However, note that $\sum_{j=1}^{M+1} Y_{i,j}$, the total number of successes in the ith stratum, is the complete sufficient statistic for the stratum specific intercept $g_0(\boldsymbol{S}_i)$. By conditioning on this complete sufficient statistic, the conditional likelihood becomes free of this stratum specific intercept. Now, from the arguments of [23] and [52], the maximum conditional likelihood estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \eta_0, \eta_1)^T$ (η_0, η_1 in lieu of α_0 and α_1) is semiparametric efficient. Subsequently, the conditional likelihood for the ith stratum is

$$\begin{aligned}\mathcal{L}_{i,c} &= \left\{ \prod_{j=1}^{M+1} \text{pr}(W_{i,j} | \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, Y_{i,j}, \boldsymbol{Z}_{i,j}) \right\} \\ &\quad \times \text{pr}(Y_{i,1}, \ldots, Y_{i,M+1} | \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, j = 1, \ldots, (M+1), \sum_{j=1}^{M+1} Y_{i,j} = 1) \\ &= \prod_{j=1}^{M+1} \Big\{ \text{pr}^{W_{i,j}}(W_{i,j} = 1 | \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, Y_{i,j} = 1, \boldsymbol{Z}_{i,j}) \\ &\quad \times \text{pr}^{(1-W_{i,j})}(W_{i,j} = 0 | \boldsymbol{S}_i, \boldsymbol{X}^*_{i,j}, Y_{i,j} = 1, \boldsymbol{Z}_{i,j}) \Big\}^{Y_{i,j}}\end{aligned}$$

$$\times\Bigg\{\mathrm{pr}^{W_{i,j}}(W_{i,j}=1|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},Y_{i,j}=0,\boldsymbol{Z}_{i,j})$$
$$\times\mathrm{pr}^{(1-W_{i,j})}(W_{i,j}=0|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},Y_{i,j}=0,\boldsymbol{Z}_{i,j})\Bigg\}^{1-Y_{i,j}}$$
$$\times\frac{\sum_{j=1}^{M+1}Y_{i,j}\exp\{\boldsymbol{\beta}_2^T\boldsymbol{Z}_{i,j}+g_2(\boldsymbol{\gamma},\beta_1,\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\}}{\sum_{j=1}^{M+1}\exp\{\boldsymbol{\beta}_2^T\boldsymbol{Z}_{i,j}+g_2(\boldsymbol{\gamma},\beta_1,\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\}}.$$

Following Equation (2.3) and Lemma 2, $\mathrm{pr}(W=1|\boldsymbol{S},\boldsymbol{X}^*,Y=0,\boldsymbol{Z})$ must be replaced by $\alpha_0+(1-\alpha_0-\alpha_1)H(\gamma_0+\boldsymbol{\gamma}_1^T\boldsymbol{S}+\boldsymbol{\gamma}_2^T\boldsymbol{X}^*+\boldsymbol{\gamma}_3^T\boldsymbol{Z})$ and $\mathrm{pr}(W=1|\boldsymbol{S},\boldsymbol{X}^*,Y=1,\boldsymbol{Z})$ by $\alpha_0+(1-\alpha_0-\alpha_1)H(\gamma_0+\beta_1+\boldsymbol{\gamma}_1^T\boldsymbol{S}+\boldsymbol{\gamma}_2^T\boldsymbol{X}^*+\boldsymbol{\gamma}_3^T\boldsymbol{Z})$.

Subsequently, the efficient estimator $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ for $\boldsymbol{\theta}$ can be obtained by solving $S_{\text{eff},\theta}=\sum_{i=1}^{n}\partial\log(\mathcal{L}_{i,c})/\partial\boldsymbol{\theta}=\mathbf{0}$. The asymptotic variance for $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ is estimated by inverting $-\sum_{i=1}^{n}\partial^2\log(\mathcal{L}_{i,c})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$, the negative of the Hessian matrix .

2.4 Simulation study

Simulation design: In order to simulate a matched case-control data set, I first generated a large population with a sample size $N=80000$ with 6 variables, (S,W,X,X^*,Z,Y), corresponding to the stratification variable, the misclassified binary exposure, the true binary exposure, the instrument, the prognostic factor, and the response of interest. From this population I sampled n cases ($Y=1$), and for each case I sampled $M=2$ controls randomly from the population by matching the value of the stratification variable S. Variables S, X^*, and Z were generated from a Uniform$(-1,1)$, Normal$(0,0.5^2)$ and Normal$(0,0.5^2)$ distributions, respectively. The binary exposure X was simulated from the Bernoulli distribution with the success probability $H(\gamma_0+\gamma_1S+\gamma_2X^*)$, $\gamma_0=-1$, $\gamma_1=1$. I set two values for γ_2, 1 and 2, which corresponds to a moderate and a strong association between the exposure and its instrument. Under this simulation setting, the marginal prevalence of X was approximately 30%. Then, Y was simulated from the Bernoulli distribution with

the success probability $H(-2-2S+X+0.5Z)$,with $\beta_1 = 1$ and $\beta_2 = 0.5$. The marginal prevalence of $Y = 1$ in the population was 20%. Finally, the surrogate (misclassified) variable was modeled as $W = B \times X + (1 - B^*) \times (1 - X)$, where $B \sim$ Bernoulli$(1 - \alpha_1)$ and $B^* \sim$Bernoulli$(1 - \alpha_0)$. Here I considered two scenarios, the first being $\alpha_0 = \alpha_1$ and the other scenario being $\alpha_0 \neq \alpha_1$. Under the first scenario, I looked at three different settings of the misclassification probabilities MC1: $\alpha_0 = \alpha_1 = 0.2$, MC2: $\alpha_0 = \alpha_1 = 0.1$, and MC3: $\alpha_0 = \alpha_1 = 0.05$. Under the second scenario, I considered three different settings considered MC1: $\alpha_0 = 0.2$, $\alpha_1 = 0.1$; MC2: $\alpha_0 = 0.2$, $\alpha_1 = 0.05$; MC3: $\alpha_0 = 0.1$, $\alpha_1 = 0.05$.

A control was considered to be matched with a given case when the absolute difference between the values of the confounding variable for the case and control subjects was less than 0.01. Two controls were randomly chosen from the set of all matched controls identified in the population for a given case. I also set two different sample sizes, $n = 200$ and 1000. Since, I considered 1:2 matched case-control studies, when $n = 200$, there were 200 cases and 400 controls for every matched data set, and when $n = 1000$, there were 1000 cases and 2000 controls in every matched data set.

Method of analysis: Every simulated data set was analyzed using four approaches. First I estimated $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ using the true X, referred to as M1. This method acted as the baseline for the other approaches. Note that in a real data analysis, M1 is unrealistic since X is not observed, . In the second method, referred to as M2, I replaced X by W in the conditional likelihood $\mathcal{L}_c(\boldsymbol{\beta}|X, \boldsymbol{Z})$, . In the third method, referred to as M3, I analyzed the simulated data sets using the proposed intuitive method (two step estimation). Finally, I analyzed the data sets by the efficient method, referred to as M4. For each one of the 4 scenarios (2 values for n, and 2 different associations between the instrument and the exposure variable) under

each set of misclassification probabilities, I conducted 5000 simulations. Since there is no validation data set or replicated data, I cannot use the regression calibration approach, commonly applied for reducing bias.

I compared the methods in terms of the operating characteristics of the estimator of $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. The statistics used as comparison include the relative median bias, a robust standard deviation calculated as $(Q3 - Q1)/1.349$ (SD* in Tables 2.1–2.3), where Q_1 and Q_3 are the first and third quartile of the 1000 estimates of the parameter, the standard deviation of the estimator using 1000 estimates, an average of the estimated standard errors, the 95% coverage probability using the Wald confidence interval, and the mean squared error. These summary measures were calculated based only on the converged data sets.

In the computation when either the absolute value of $\widehat{\beta}_1$ or the standard error of β_1 is greater than 5, I declared that data set as non-convergent. For a sample size of $n = 200$, roughly 7–8% data sets did not converge under M3; under M4 1–2% data sets had convergence issues. For a sample size of $n = 1000$, approximately 4–5% data sets did not converge in M3, while for M4 0–0.5% data sets faced convergence issues. There were no convergence problems for methods M1 and M2. The results presented in tables are based solely on the converged data sets.

Results: Results for the first scenario when the misclassification probabilities are the same are displayed in table 2.1. The performance of M1 is intuitively the best compared to M2, M3, and M4 in terms of all measures since the true values of X were used. Likewise, the estimator for β_2 performs equally under the different methods. However, the performance of $\widehat{\beta}_1$, the coefficient for X that is not observed in the data, varies greatly across the methods and scenarios. Regardless of the scenario and sample size used, the bias of $\widehat{\beta}_1$ under method M2 is large, leading to a coverage probability for β_1 far from the nominal level of 0.95. The bias and variance under M3

and M4 decreases as the strength of the association between X and X^* gets stronger. Moreover, as the association between X and X^* increases, the coverage probability gets closer to the nominal level. For a large sample size, M3 performs well, but for almost all scenarios M4 is the top performer in terms of bias reduction and low MSE. Intuitively , the estimator for M4 shows less variability than the estimator under M3. When $n = 1000$, bias decreases when the association between X and X^* changes from moderate (S1) to strong (S2) . However in all three scenarios: MC1, MC2, and MC3, it is evident that increasing the sample size leads to decrease in the standard errors of the proposed methods by nearly half. This naturally leads to a decrease in MSE, and shows the superiority of the M3 and M4 over the naive approach.

Table 2.1: Results of the simulation study under equal misclassification. (Manuel, Wang, Sinha 2019)
MT: method, B: relative median bias ×100, SD*: simulation standard deviation based on quantiles ×100, SD: simulation standard deviation ×100, SE: estimated standard error ×100,CP: 95% coverage probability based on the Wald confidence interval, MSE: mean squared error, M1: Conditional logistic analysis when true X is used, M2: Conditional logistic analysis when X is replaced by W, M3: proposed two-step method, M4: proposed efficient estimator, MC1: $\alpha_0 = \alpha_1 = 0.2$, MC2: $\alpha_0 = \alpha_1 = 0.1$, MC3: $\alpha_0 = \alpha_1 = 0.05$, $\text{pr}(X = 1|S, X^*) = H(-1 + S + \gamma_2 X^*)$, S1: $\gamma_2 = 1$, S2: $\gamma_2 = 2$, $\text{pr}(Y = 1|S, X, Z) = H(-2 - 2S + X + 0.5Z)$, $\alpha_0 = \text{pr}(W = 1|X = 0), \alpha_1 = \text{pr}(W = 0|X = 1)$

		MC1								MC2						MC3					
MT		M1		M2		M3		M4		M2		M3		M4		M2		M3		M4	
		β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
$n = 200$																					
S1	B	0.42	−0.25	−53.73	−2.42	−59.75	−0.47	7.12	2.11	−31.67	−1.28	−34.97	0.66	10.17	1.26	−17.72	−1.27	−20.35	1.48	12.11	1.39
	SD^*	22.26	18.23	18.89	18.24	47.53	19.51	64.97	21.90	20.01	18.18	57.68	19.70	47.50	19.74	20.26	18.27	62.18	19.69	39.17	19.21
	SD	22.25	18.72	18.70	18.43	56.18	20.18	82.86	25.48	19.92	18.53	57.82	20.52	59.68	21.31	21.02	18.62	59.35	20.50	48.64	20.04
	SE	21.90	18.48	18.66	18.06	78.78	24.55	63.68	23.14	19.79	18.22	71.11	25.27	49.20	21.60	20.66	18.33	67.07	23.93	42.38	22.41
	CP	94.95	94.61	18.30	94.41	67.19	95.95	85.45	93.20	63.81	94.50	77.65	95.67	91.67	93.71	85.72	94.82	81.26	95.68	93.74	93.72
	MSE	4.96	3.51	32.43	3.40	61.49	4.10	73.52	7.44	13.86	3.43	44.61	4.25	39.55	4.88	7.39	3.47	40.81	4.18	27.32	4.20
S2	B	0.53	0.99	−51.07	−1.93	−41.21	−1.17	1.34	0.60	−29.44	−0.97	−18.21	0.56	3.49	0.95	−15.92	−0.04	−7.54	1.12	4.11	1.21
	SD^*	21.09	18.32	18.48	17.89	39.79	19.36	40.17	19.48	19.15	18.05	37.62	18.79	30.77	18.64	20.33	17.88	32.92	19.15	27.33	18.55
	SD	21.43	18.75	18.56	18.19	42.51	19.94	46.85	21.07	19.29	18.41	38.34	19.51	32.55	19.22	20.27	18.55	36.23	19.38	28.30	18.94
	SE	21.15	18.53	18.50	18.08	56.23	21.88	39.81	21.23	19.44	18.25	45.23	21.23	32.25	21.76	20.16	18.37	38.96	20.23	28.90	20.44
	CP	95.22	94.68	20.47	94.64	72.54	95.01	90.62	93.83	66.64	94.60	83.99	95.17	93.63	94.25	86.83	94.80	89.13	95.02	94.39	94.12
	MSE	4.60	3.52	29.72	3.31	31.54	3.97	22.28	4.46	12.28	3.39	18.26	3.81	10.84	3.71	6.54	3.44	14.07	3.75	8.33	3.60
$n = 1000$																					
S1	B	0.17	0.04	−53.87	−2.59	−22.19	−0.84	1.21	-0.15	−31.75	−1.62	−2.51	0.52	1.69	0.10	−17.64	−0.94	4.27	1.15	2.31	0.19
	SD^*	9.69	8.40	8.41	8.20	38.90	9.13	26.14	8.97	8.62	8.27	27.63	9.00	20.49	8.54	9.09	8.35	24.30	8.85	18.16	8.39
	SD	9.83	8.32	8.33	8.17	36.14	9.42	28.20	8.81	8.83	8.22	28.59	9.05	21.14	8.44	9.25	8.25	27.07	8.93	18.09	8.36
	SE	9.71	8.21	8.30	8.02	41.00	10.57	24.89	9.66	8.79	8.09	32.33	9.76	20.52	9.55	9.17	8.14	26.73	10.09	18.79	9.40
	CP	94.83	94.89	0.04	94.37	73.76	95.50	89.82	94.57	4.98	94.68	81.63	95.56	87.85	94.24	51.33	94.88	85.05	95.49	89.30	94.04
	MSE	0.97	0.69	29.70	0.68	17.66	0.89	7.96	0.78	10.83	0.68	8.22	0.82	4.50	0.72	3.92	0.68	7.56	0.80	3.39	0.71
S2	B	0.09	0.02	−51.24	−2.66	−13.11	−0.88	0.19	−0.30	−29.54	−1.72	−2.74	0.00	0.28	−0.20	−16.03	−0.91	−0.18	0.18	0.55	−0.24
	SD^*	9.19	8.28	8.18	8.12	22.34	8.85	16.99	8.58	8.80	8.15	15.61	8.53	13.92	8.29	8.85	8.23	13.77	8.42	12.36	8.26
	SD	9.54	8.29	8.34	8.09	22.47	8.91	17.74	8.60	8.79	8.15	16.55	8.51	14.13	8.34	9.07	8.21	16.09	8.42	12.34	8.25
	SE	9.38	8.22	8.23	8.03	24.14	9.46	16.10	8.45	8.64	8.10	17.48	9.07	14.00	8.57	8.95	8.15	14.98	8.47	13.27	9.44
	CP	94.36	95.08	0.00	94.44	81.92	95.14	93.10	94.61	7.80	94.76	90.35	95.24	93.38	94.65	56.33	94.88	91.34	94.88	94.34	94.34
	MSE	0.91	0.69	27.04	0.67	7.19	0.79	3.15	0.74	9.47	0.67	2.83	0.73	2.00	0.70	3.39	0.68	2.59	0.71	1.53	0.69

Now I consider Table 2.2, containing the simulation results for the unequal misclassification probabilities scenarios. Under both sample sizes, M3 performs better than M2 in terms of bias for scenarios with a large sample size and moderate to strong association between X and X^*. Under M3, the bias, variability, and the distance between the coverage probability and its nominal level for the estimator of β_1 decrease as association between X and X^* increases. Similarly to the equal misclassification probability setting, increasing n from $n = 200$ to $n = 1000$ as well as increasing the association between X and X^* substantially decreases the standard errors and MSE of the proposed estimator. This gives M3 and M4 an advantage over M2. Note that M4 performs the best in terms of bias, MSE as well as variability.

Table 2.2: Results of the simulation study under unequal misclassification. (Manuel, Wang, Sinha 2019)
MT: method, B: relative median bias ×100, SD*: simulation standard deviation based on quantiles ×100, SD: simulation standard deviation ×100, SE: estimated standard error ×100,CP: 95% coverage probability based on the Wald confidence interval, MSE: mean squared error, M1: Conditional logistic analysis when true X is used, M2: Conditional logistic analysis when X is replaced by W, M3: two-step method, M4: proposed efficient estimator, MC1: $\alpha_0 = 0.2, \alpha_1 = 0.1$, MC2: $\alpha_0 = 0.2, \alpha_1 = 0.05$, MC3: $\alpha_0 = 0.1, \alpha_1 = 0.05$, $\text{pr}(X = 1|S, X^*) = H(-1 + S + \gamma_2 X^*)$, S1: $\gamma_2 = 1$, S2: $\gamma_2 = 2$, $\text{pr}(Y = 1|S, X, Z) = H(-2 - 2S + X + 0.5Z)$, $\alpha_0 = \text{pr}(W = 1|X = 0), \alpha_1 = \text{pr}(W = 0|X = 1)$

		MC1								MC2						MC3					
MT		M1		M2		M3		M4		M2		M3		M4		M2		M3		M4	
		β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
$n = 200$																					
S1	B	0.39	−0.26	−47.20	−1.93	−48.22	0.21	8.87	1.67	−43.91	−1.77	−42.71	0.28	10.33	1.83	−28.53	−1.37	−28.40	0.82	11.10	1.65
	SD^*	22.28	18.24	18.68	18.31	51.93	19.88	57.30	20.61	18.54	18.36	52.82	19.66	54.69	20.29	19.55	18.34	58.46	19.65	44.95	19.50
	SD	22.25	18.72	18.61	18.47	56.22	20.27	71.53	23.31	18.49	18.49	56.68	20.06	68.58	22.59	19.79	18.55	58.85	20.46	55.63	20.64
	SE	21.90	18.48	18.54	18.12	75.03	23.97	57.93	23.41	18.51	18.14	58.27	22.19	54.87	22.94	19.66	18.25	60.48	23.25	47.34	22.72
	CP	94.95	94.61	28.55	94.49	70.98	95.59	87.61	93.72	34.30	94.41	74.19	95.89	88.33	93.51	68.95	94.61	78.57	95.20	91.96	94.10
	MSE	4.96	3.51	25.60	3.41	50.83	4.09	55.76	5.97	22.51	3.42	47.61	4.08	51.93	5.28	11.99	3.44	42.69	4.19	34.68	4.55
S2	B	0.53	0.89	−44.02	−1.43	−29.97	−0.49	3.06	0.97	−40.83	−1.09	−23.31	0.36	3.55	0.81	−25.92	−0.57	−12.76	0.91	3.81	1.02
	SD^*	21.08	18.31	18.60	17.90	39.34	19.45	36.23	19.27	18.61	17.91	39.64	19.45	34.34	19.01	19.16	17.87	35.94	18.91	29.23	18.50
	SD	21.43	18.76	18.44	18.26	39.98	19.86	39.44	19.95	18.49	18.30	40.45	19.77	36.70	19.66	19.21	18.44	37.49	19.47	30.88	19.14
	SE	21.15	18.53	18.40	18.14	52.46	22.56	35.98	21.75	18.39	18.17	42.85	21.80	34.25	21.43	19.34	18.29	42.48	22.05	31.32	21.04
	CP	95.22	94.68	32.97	94.70	78.31	94.98	91.63	93.92	40.42	94.74	80.39	95.08	92.01	94.18	71.98	94.63	86.48	95.22	93.53	94.19
	MSE	4.60	3.52	22.92	3.34	23.57	3.91	15.83	3.99	19.90	3.35	21.35	3.89	13.78	3.87	10.45	3.40	16.14	3.79	9.85	3.66
$n = 1000$																					
S1	B	0.17	0.04	−47.34	−2.15	−11.89	−0.53	0.96	−0.15	−44.07	−2.04	−6.44	−0.17	1.96	−0.10	−28.78	−1.48	1.66	0.71	2.40	0.28
	SD^*	9.69	8.40	8.35	8.19	35.52	9.15	23.67	8.79	8.32	8.19	33.76	9.20	22.94	8.71	8.62	8.31	26.86	8.94	19.84	8.49
	SD	9.83	8.32	8.29	8.19	33.47	9.24	25.54	8.61	8.27	8.20	31.60	9.17	24.51	8.56	8.78	8.24	26.87	9.02	20.32	8.40
	SE	9.71	8.21	8.24	8.04	36.78	9.88	23.03	11.61	8.23	8.06	34.39	9.95	21.52	10.34	8.73	8.10	30.32	9.55	20.06	8.86
	CP	94.84	94.88	0.06	94.54	78.15	95.65	88.47	94.19	0.08	94.64	80.30	95.41	87.16	94.09	9.54	94.74	82.84	95.51	87.46	94.37
	MSE	0.97	0.69	23.07	0.68	12.73	0.86	6.53	0.74	20.02	0.68	10.55	0.84	6.03	0.73	8.99	0.68	7.25	0.81	4.19	0.72
S2	B	0.10	0.01	−44.44	−2.39	−7.25	−0.47	−0.04	−0.10	−40.83	−2.15	−4.05	−0.17	0.38	−0.13	−26.45	−1.47	−0.86	0.00	0.58	−0.24
	SD^*	9.19	8.28	8.15	8.14	18.86	8.73	15.65	8.4	8.15	8.14	17.53	8.66	14.77	8.35	8.58	8.17	14.96	8.43	13.05	8.35
	SD	9.54	8.29	8.27	8.11	19.71	8.74	16.56	8.44	8.23	8.13	18.74	8.68	16.03	8.39	8.72	8.17	15.53	8.47	13.62	8.30
	SE	9.38	8.22	8.18	8.05	20.41	19.77	15.27	9.25	8.18	8.07	18.52	10.33	15.52	11.18	8.59	8.12	16.46	8.97	13.70	9.08
	CP	94.36	95.06	0.06	94.48	86.47	95.37	92.53	94.38	0.16	94.56	87.46	95.27	91.98	94.19	14.28	94.80	92.08	95.33	92.91	94.26
	MSE	0.91	0.69	20.35	0.67	4.53	0.76	2.74	0.71	17.32	0.67	3.76	0.75	2.57	0.70	7.68	0.67	2.42	0.72	1.86	0.69

Finally, I considered the scenario in which there are multiple instruments and multiple confounding variables. To do this the simulated data sets were generated by closely mimicking the real data set with multiple confounding and instrumental variables; see Section 5 which discusses the data more thoroughly. I only considered newborns whose mother was Black, and after applying the necessary exclusions I had $N = 42933$ subjects. From this population I generated the a pseudo-population by sampling 42933 subjects with replacement . The values corresponding to confounding variables, covariates, and instrumental variables were automatically assigned for subjects selected in this pseudo-population, . To generate the true exposure variable X, the logistic model was applied

$$\begin{aligned} \text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*) &= H(-3.39 + 1.29S_1 + 0.36S_2 + 1.66S_3 + 0.67S_4 + 0.91S_5 \\ &\quad +1.09S_6 + 0.28S_7 + 0.14X_1^* - 0.32X_2^* - 0.21X_3^* \\ &\quad +0.99X_4^*), \end{aligned} \tag{2.7}$$

with $S_1 = 1$ when the number of years of education of the mother was less than 12 and 0 otherwise, S_2, S_3 corresponding to the indicator variables representing prenatal care beginning in either second or third trimester, respectively , S_4, S_5, S_6, S_7 representing indicator variables corresponding to the mother's age in [23, 26], [27, 30], [31, 39], ≥ 40, $X_1^* =$ is the instrumental variable for cigarette state tax rate for 1989, X_2^* being the instrumental variable, expressed in thousands, for logarithm of the median family income of the county where the child was born divided by 1000, minus the average of the logarithm of the median family income, $X_3^* = 1$ the instrumental variable indicating a black father and 0 for white father, and finally $X_4^* = 1$ the instrument if the number of years of education of the father was less than 12 and 0 otherwise. Note that the coefficients applied in model (2.7) are close to the estimates of $\boldsymbol{\gamma}$ parameters

from the real data analysis of the Black mother data. I then generated the surrogate variable W as $W = B \times X$, where $B \sim \text{Bernoulli}(1 - \alpha_1)$. Following the real data analysis using method M3, I set $\alpha_0 = 0$ and $\alpha_1 = 0.45$. Next the binary response Y was simulated with the following model

$$\begin{aligned} \text{pr}(Y = 1|\boldsymbol{S}, X, Z) &= H(-3.2 + 0.19S_1 - 0.03S_2 + 0.37S_3 + 0.09S_4 + 0.32S_5 \\ &\quad +0.57S_6 + 0.55S_7 + 0.69X + 0.33Z). \end{aligned}$$

I obtained the intercept and regression coefficients corresponding to $\boldsymbol{S}$ by regressing Y on $\boldsymbol{S}$ in the population of the Black mothers. The coefficients which correspond to X and Z are the estimates of β_1 and β_{22} for the Black mothers in the real data analysis when using the proposed approach with $\alpha_0 = 0$. Additionally, set $Z = 1$ if the father's age was greater or equal to 31 and 0 otherwise. This resulted in approximately 4.5% of the data with $Y = 1$. After creating this pseudo-population, nested case-control data were generated by sampling $n = 1800$ cases with 3600 matched controls. This process of generating a pseudo-population and subsequent sampling of a nested case-control data was replicated 5000 times.

I analyzed each simulated nested case control data set using the four methods mentioned earlier, M1, M2, M3, M4. For M3 and M4 I used $\alpha_0 = 0$. The first panel of Table 2.3 displays the results. The bias of estimating β_1 is greatly reduced in M3 and M4 when compared with M2, and they also have coverage probabilities much closer to the nominal level than that of M2. The reduced MSE also demonstrates the substantial bias reduction in M3 and M4. In particular, when compared to M2, the MSE for β_1 is 10% less in M3 and 14% less in M4. Once again, M4 turns out to be the best performing method.

Finally, within the last simulation design, I considered another case where S_1,

one of the confounding variables, was not used to form matched sets. Instead, S_1 was only used in the incidence model for Y as a prognostic factor (like Z_2 according to our notation). I estimated the corresponding regression parameter, denoted by β_3, as well as the regression parameters for X and that for the original Z variable. The second panel of Table 2.3 displays the results for this situation. As in the other simulations, both M3 and M4 perform very well and M4 is slightly better than M3.

In short, the results of the simulation studies demonstrate that these proposed methods perform better in reducing the bias significantly when compared to the naive approach. The proposed methods appear to work very well when there is a strong association between the true exposure and the instruments and the sample size is large. Incorporating any available prior information on the misclassification probabilities would help to improve the performance of these methods.

2.5 Real Data Analysis

Description of the data and variables: I consider the data from the 1989 U.S. Natality Birth Records [48], which was introduced in the Introduction. This data contains information on the birth records for infants who were born to residents and nonresidents within the United States during the year 1989, and provides information on the mother, the father and the child. For this analysis I define the binary response $Y = 1$ when a newborn's birth weight is less than 2500 grams, and 0 otherwise. Additionally, I defined $X = 1$ when a mother smoked more than 2 cigarettes daily during the pregnancy and 0 otherwise. Because there is no consensus on the definition of various levels of smoking, I took 2 as a cut-off which distinguished between 1) no, intermittent, and very light smokers, and 2) light, moderate and heavy smokers. The surrogate variable W is defined based on the reported average daily number of cigarettes smoked ($>$ 2 cigarettes as 1 and $<=$ 2 as 0) . Here the intuition is

Table 2.3: Results of the simulation study with $n = 1800$ in the case of multiple instruments and multiple confounding variables. (Manuel, Wang, Sinha 2019)
Panel 1: all confounding variables are used in matching, Panel 2: all but one confounding variables are used in matching and the other confounding variable is used as a covariate for adjustment, MT: method, Bias: relative median bias $\times 100$, SD*: simulation standard deviation based on quantiles $\times 100$, SD: simulation standard deviation $\times 100$, SE: estimated standard error $\times 100$, CP: 95% coverage probability based on the Wald confidence interval, MSE: mean squared error, M1: Conditional logistic analysis when true X is used, M2: Conditional logistic analysis when X is replaced by W, M3: proposed two-step method, M4: proposed efficient estimator

MT	M1		M2		M3		M4	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
Bias	−0.03	0.34	−13.98	−0.04	−0.99	0.38	0.80	0.34
SD*	7.88	8.12	9.84	8.13	12.90	8.44	12.48	8.27
SD	7.94	8.13	9.93	8.08	13.24	8.34	12.84	8.22
SE	7.80	7.88	9.67	7.83	13.45	8.09	12.63	7.98
CP	94.25	94.47	82.46	94.55	94.18	94.50	94.60	94.64
MSE	0.63	0.66	1.93	0.65	1.75	0.70	1.66	0.68

MT	M1			M2			M3			M4		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
Bias	−0.52	−0.32	1.02	−13.78	0.05	53.11	−1.17	0.06	1.94	0.20	−0.01	0.04
SD*	8.12	7.96	10.96	10.20	7.95	0.90	12.97	8.14	13.64	12.47	8.07	12.66
SD	8.03	8.07	11.13	10.16	8.03	10.93	13.10	8.22	13.71	12.61	8.16	12.85
SE	7.83	7.89	11.04	9.77	7.84	10.85	14.79	8.19	15.71	12.51	7.93	12.50
CP	97.00	94.48	94.96	82.56	94.46	83.30	94.66	94.94	94.66	94.42	94.22	93.96
MSE	0.65	0.65	1.24	1.93	0.64	2.29	1.72	0.68	1.88	1.59	0.67	1.65

that those who smoked more than 2 cigarettes or smoked regularly during pregnancy were more likely to suffer from recall bias in reporting their daily average number than those who were non smokers or smoked 2 or less. The following four variables were selected as instruments: cigarette state tax rate for 1989, the logarithm of the median family income of the county where the child was born divided by 1000 minus the average of the logarithm of the median family income expressed in thousands, father's race, and the father's number of years of education coded as a binary variable (< 12 and $>= 12$). In the literature cigarette tax, father's education and family income have been demonstrated as potential instrumental variables for the mother's smoking (59, 19, 46). I don't consider family income since it could directly impact birthweight, instead, we used the median family income of the county where the mother lived. Additionally, family income was not reported in the dataset, so there is no question of using it as an instrument or a confounding variable. Mother's age, mother's education, and when prenatal care began were selected as confounding variables since intuitively they could influence birthweight as well as the mother's smoking behavior. I defined the age of the mother into 5 categories, with 18–22 as the reference category, 23–26, 27–30, 31–40, and > 40. Mother's education was coded as a binary variable: 0 for < 12 years and 1 for ≥ 12. Finally, when prenatal care began was coded as 3 category variable: 1st trimester set as the reference category, 2nd trimester, and 3rd trimester. A few subjects (Black: 1.0%, White: 0.3%) had no prenatal care and so I combined that category with the prenatal care that began in the 3rd trimester category. To carry out this analysis, I assume that 1) conditional on the true smoking level, the reported smoking level, and all other variables are independent and 2) conditional on the true smoking level, the instrumental variables and the response are independent. Note that these are plausible assumptions. But truly there is no way to verify these assumptions without knowing the true smoking

level, and violation of these assumptions likely to cause bias. Also, I want to point out that in this data example the nondifferential misclassification assumption may be violated due to dichotomization of the numeric exposure variable [21], and developing a proper methodology for dealing with this scenario is a part of my future research.

As some exclusion criteria, I dropped subjects whose mother and father had a race other than Black and White, which represents about 3% newborns. This percentage includes the scenario where one of the parents is Black/White and the other parent is neither Black nor White, but does not include the scenario where both parents are neither Black nor White. Additionally, I removed the newborns whose mother or father was less than 18 years of age, and those who were not the first child of the parents. Newborns that were born in the states of Alaska, Delaware, Montana, New Hampshire, and Oregon were also excluded since these states did not have state tax on cigarettes in 1989. Moreover, newborns born in Colorado, Maryland, New Jersey, Rhode Island, and Wyoming were excluded because these state's tax rates did not apply to cigarettes. Finally, I divided the birth cohort into two groups: Black and White mothers, and conducted the analysis of these two groups separately. After applying these exclusions, my data contained 42933 newborns to Black mothers (group 1) and 347041 newborns to White mothers (group 2).

For forming matched data, I proceeded as follows. First, I identified all 2021 newborns with $Y = 1$ from group 1. I then randomly selected 2000 newborns out of 2021 newborns as cases to be included in my matched set for black group, and for each selected subject (newborn), $M = 2$ controls from 40912 were randomly sampled by matching the confounding variables. This then yielded 2000 cases and 4000 controls for the matched data set. In group 2, there were 6646 newborns with $Y = 1$, and subsequently 340395 newborns with $Y = 0$. Utilizing the same sampling mechanism a 1:2 matched case-control data was created for group 2, meaning there

were also 2000 cases and 4000 controls .

Results: Three methods of analysis were conducted: M2 (naive), M3 (intuitive two stage method) and M4 (efficient method). Note that method M1 cannot be used, like in the simulations, since the true X is never observed in this real data. The results for black and white mothers are presented in the first and second panels, respectively, of Table 2.4.

Table 2.4: Analysis of the low birthweight data sampled from 1989 US birth cohort. (Manuel, Wang, Sinha 2019) Here β_1 is the regression coefficient for mother's smoking, and β_{21} and β_{22} are the regression coefficients for father's age in $[31, 40]$ and > 40, respectively, while the father's age in $[18, 30]$ is treated as the reference category. MT: Method, M2: naive approach, M3: proposed two-step method, M4: proposed efficient estimator, EST: estimate, SE: standard error, PV: p-value

MT		M2			M3: $0 < \alpha_0 + \alpha_1 < 1$			M3: $\alpha_0 = 0, 0 < \alpha_1 < 1$			M4: $0 < \alpha_0 + \alpha_1 < 1$			M4: $\alpha_0 = 0, 0 < \alpha_1 < 1$		
Mother		EST	SE	PV	EST	SE	PV	EST	SE	PV	EST	SE	PV	EST	SE	PV
Black	β_1	0.532	0.097	0.000	0.824	0.170	0.000	0.764	0.130	0.000	0.613	0.117	0.000	0.685	0.141	0.000
	β_{21}	−0.096	0.067	0.149	−0.119	0.075	0.114	−0.139	0.079	0.078	−0.088	0.078	0.263	−0.127	0.077	0.099
	β_{22}	0.362	0.133	0.007	0.309	0.144	0.031	0.218	0.161	0.177	0.425	0.129	0.001	0.288	0.129	0.025
White	β_1	0.396	0.072	0.000	0.451	0.092	0.000	0.460	0.130	0.000	0.488	0.073	0.000	0.468	0.084	0.000
	β_{21}	0.420	0.065	0.000	0.424	0.072	0.000	0.427	0.072	0.000	0.414	0.073	0.000	0.429	0.065	0.000
	β_{22}	0.659	0.143	0.000	0.667	0.145	0.000	0.683	0.149	0.000	0.652	0.151	0.000	0.667	0.144	0.000

First, I discuss the results for the black mothers. Under the the regular identifiability condition $0 < \alpha_0 + \alpha_1 < 1$, the estimates for α_0 and α_1 in M3 were 0.014 (s.e. 0.013) and 0.521 (s.e. 0.148). It is to be noted that the estimates of α_0 and α_1 in M3 and M4 are quite close and therefore I exclude the estimates for M4. Additionally, I concluded that α_0 was not significantly different from zero based on the aforementioned results. Subsequently, I reanalyzed the data with the constraint that $\alpha_0 = 0$, and proceeded to estimate all the other parameters along with $\alpha_1 \in (0, 1)$. Under this situation estimated α_1 was 0.46 (s.e. 0.112).

Under all three methods of analysis, M2, M3, M4, I found that mother's smoking has a positive association with low birthweight. When setting $\alpha_0 = 0$, the estimated odds ratios for smoking are 1.7, 2.15 , and 1.99 for M2, M3 and M4 respectively. This indicates that there is approximately 26% and 17% increase in the odds ratio estimate when comparing M2 to M3 and M2 to M4. Additionally, the father's age has a statistically significant association with the risk of low birthweight under M2, M3, and M4 but not under M3 and when I set $\alpha_0 = 0$. Note that the standard error for β_1 is slightly larger under M3 and M4 than in M2. We expect this increase in M3 and M4 since the methods consider the uncertainty of not observing the true values of X.

For the white women, applying the regular identifiability condition $0 < \alpha_0 + \alpha_1 < 1$ in M3 yields estimates for α_0 and α_1 were 0.014 (s.e. 0.003) and 0.271 (s.e. 0.096) respectively. Again, because of these results I also conducted the second analysis setting $\alpha_0 = 0$. From this second setting I found $\widehat{\alpha}_1 = 0.251$ (s.e. 0.089) for M3. Note that all three methods M2, M3, M4 indicate that white mother's smoking is positively associated with low birthweight, just like the black mothers. However, our analysis indicates that both categories of father's age are significantly associated with the low birthweight in all methods of analysis as well as the cases where $\alpha_0 = 0$

and $\alpha_0 > 0$.

2.6 Discussion

In this chapter, I have proposed two consistent methods for reducing the bias when estimating disease-exposure association parameters in a matched case-control study. The novelty of the methods is to make use of instruments to recover the measurement uncertainty when the data are not accompanied by any validation data. The methodology contains with an uncertainty measure of the estimators, and contains a theoretical justification of the large sample properties. The realistic simulation studies clearly show the advantages and when the proposed methods are effective in reducing bias.

The basic idea of this work can be generalized to a multicategory exposure variable with added complexity of estimating comparatively a large number of misclassification probabilities. The proposed methodology can easily be extended to the case when the instrumental variables are observed for a subset of the main data set. I think, to handle potential convergence problems in the proposed methods one may consider a penalized estimator using [20]'s penalty function.

3. ADDRESSING MISCLASSIFICATION BIAS OF AN EXPOSURE VARIABLE WITH MULTIPLE CATEGORIES

3.1 Introduction

Misclassification in covariates is a known problem, and much research has been conducted to address this issue. In particular, regression models with misclassified binary regressors have been studied by researchers from epidemiology and econometrics. In this chapter I will focus on misclassification of a multicategory exposure variable, and how to reduce the bias in the model estimation with the help of instrumental variables. I begin by discussing some pertinent literature.

In the presence of a misclassified binary covariate, [44] considered how the model parameters from a nonparametric model can be identifiable using instrumental variables. His model was nonparametric in the sense that the form of the conditional mean of the response, given the true non-misclassified covariate and other correctly measured covariates, is unknown. For identification to occur in this nonparametric setting, [44] made the following set of assumptions: (I), the conditional expectation function is identified given knowledge of the population distribution of the response, the true non-misclassified covariate, and other correctly measured covariates; (II) there are restrictions on the total amount of misclassification that occur in the misclassified covariate; (III) there is a conditional independence of the misclassified covariate and the instrument variables given the true exposure and other covariates; (IV) the instrument is informative about the true non-misclassified covariate; and (V) the true non-misclassified covariate is associated with the mean of the response variable. Model estimation of the parameters was done nonparametrically using a kernel density estimator. [44] also considered a parametric model for the regression

function, and suggested to use a sieve maximum likelihood method for estimation purposes.

[10] considered identification and estimation of the nonparametric conditional mean model of a continuous response when a binary exposure is misclassified. Thus the observed data contain the measurements on the response, the misclassified binary exposure and a set of error-free covariates. The beauty of this approach is that the authors did not require any validation data, replicated measurements or instrumental variables for identification of the model parameters. They obtained a novel identification results under some restrictive moment conditions. In the discussion of the paper, [10] proposed model estimation through sieve estimators or generalized method of moments.

[28] considered the Bayesian estimation of the association between a disease status and a binary exposure using a case-control data set where the observed exposure was misclassified. He then considered two scenarios. In scenario one, he assumed that sensitivity - the conditional probability that the misclassified covariate equals one given the true non-misclassified covariate equal one, and the specificity - the conditional probability that the misclassified covariate equals zero given the true covariate are equal zero, are known. In scenario two, he assumed the sensitivity and specificity are unknown. [28] then modeled the number of controls who are apparently exposed, and the number of cases who are apparently exposed as binomial random variables. The parameter used for these binomial random variables is the probability of apparent exposure given the disease status; and this probability is a function of the sensitivity and specificity. Both of these probabilities are modeled using beta priors. In the second scenario, [28] suggested "adjusting with uncertainty" (AWU) method, that is assigning priors to these parameters but centering them at the guessed value.

Assessing the effects of misclassification in variables with more than two categories

has been studied in epidemiological research. For example, [3] studied the effect of non-differential misclassified categorical covariates, and derived analytical forms of the odds ratios in this misclassification setting and it was expressed as a weighted averages of the set of the correctly classified odds ratios for all the pairs of categories.

[64] studied the effects of non-differential misclassification in a categorical exposure on the direction of the bias. They noted that when a non-differential misclassified binary exposure is used, the bias will tend towards the null hypothesis. However, they mentioned that this does not occur when the exposure is categorical, and that the direction of the bias can reverse. Therefore, [64] was interested in determining when the trend of association between the misclassified categorical exposure and the response matches the trend between the true non-misclassified exposure and the response. They found that the trends match only when the mean of the response, conditional on the true exposure, is monotonically increasing and when the mean of the misclassified exposure, conditional on the true exposure, is monotonically increasing .

In spite of these studies, work to correct the bias due to misclassification in variables with more than two categories is not as prevalent. In a Poisson regression model, [63] derived analytical forms of the bias of categorical exposure variables subject to misclassification.

My work is closely related to the approach of [33], who obtained nonparametric identification of the conditional probability or the density of the response given a categorical exposure and a set of error-free covariates based on the observable data on a misclassified version of the categorical exposure with two or more categories, error-free covariates, the response and a set of instrumental variables. He obtained identification results under a set of constraints. Besides this elegant result, I was curious if there is any identification issue in a parametric setting. Therefore, I assumed

a parametric model for the response given the categorical exposure and the error-free covariates and another parametric model for the true categorical exposure given the error-free covariates and the instrumental variables. It turned out the identification issue is still present in my parametric setting. Then I obtained parametric identification under a condition and this condition was one of the conditions that [33] used. Due to my parametric setting, I was able to prove the result using the connection between model identification and information matrix [54]. Beyond identification, I considered a novel estimation procedure that seamlessly allows the integration of the prior information in making inference. In order to apply the proposed method to large scale epidemiological datasets, I have applied a variational Bayesian inference procedure using the automatic differentiation approach of [38].

As mentioned by [28], parameter identification is an important issue in modeling misclassified data, as nonidentifiability can lead to poor estimation results. My work addresses this issue by considering sufficient conditions for parameter identification in the widely used logistic model and integrating these conditions into model estimation. The identification result and the scalable estimation procedure will be useful to practitioners who often use parametric models for their analysis.

The structure for the rest of this chapter is as follows. In Section 2, I discuss the issue and sufficient conditions for identification. Parameter estimation using the variational inference approach is discussed in Section 3. Section 4 contains some simulations to assess the performance of the proposed approach, while Section 5 contains an application of the novel method to a real world data set. Concluding remarks are given in Section 6.

3.2 Parameter Identification

3.2.1 Background

Define Y, X, Z as the binary outcome, categorial exposure, and a set of other covariates. Let W be the misclassified version of X, and X^* be a set of instrumental variables for X. Following the definition of instrumental variables, X^* only affects X, but it has no direct link with W, Y or Z [26]. Here I list a set of assumptions that are used in this chapter.

1. The variables X and W have the equal number of categories.

2. Conditional on X, W does not depend on other variables.

The observed data are n iid copies of (X^*, W, Y, Z). Suppose that both X and W have r categories and they are denoted by $1, \ldots, r$. Assume the following model for the outcome

$$\begin{aligned} \text{pr}(Y = 1|X, Z; \beta) &= 1 - \text{pr}(Y = 0|X, Z) \\ &= \frac{\exp\{\beta_0 + \sum_{k=2}^{r} I(X = k)\beta_{x,k} + \beta_z^T Z\}}{1 + \exp\{\beta_0 + \sum_{k=2}^{r} I(X = k)\beta_{x,k} + \beta_z^T Z\}}. \end{aligned} \tag{3.1}$$

Here $\beta_{x,2}, \ldots, \beta_{x,r}$ are the regression parameters corresponding to X and β_z is the regression parameter corresponding to Z containing numeric or binary 0-1 variables. Here $\beta = (\beta_0, \beta_{x,2}, \ldots, \beta_{x,r}, \beta_z^T)^T$. Conditional on the covariate Z, the odds ratio of the outcome for changing X from k' to k is $\exp(\beta_{x,k} - \beta_{x,k'})$. I also assume the following probability model for X given X^* and Z,

$$\begin{aligned} \text{pr}(X = r|X^*, Z; \gamma) &= \frac{\exp(\gamma_{r0} + \gamma_{r1}^T X^* + \gamma_{r2}^T Z)}{1 + \sum_{k=2}^{r} \exp(\gamma_{k0} + \gamma_{k1}^T X^* + \gamma_{k2}^T Z)}, \text{ for } r = 2, 3, \ldots, r, \\ \text{pr}(X = 1|X^*, Z; \gamma) &= \frac{1}{1 + \sum_{k=2}^{r} \exp(\gamma_{k0} + \gamma_{k1}^T X^* + \gamma_{k2}^T Z)}, \end{aligned} \tag{3.2}$$

where X^* is a p-vector consisting of a set of binary 0-1 or numeric instrumental variables, and $\gamma = (\gamma_{20}, \gamma_{21}^T, \gamma_{22}^T, \gamma_{30}, \gamma_{31}^T, \gamma_{32}^T, \ldots, \gamma_{r0}, \gamma_{r1}^T, \gamma_{r2}^T)^T$. The misclassification probability matrix is

	X				
W	1	2	3	...	r
1	α_{11}	α_{12}	α_{13}	...	α_{1r}
2	α_{21}	α_{22}	α_{23}	...	α_{2r}
$\vdots$	$\vdots$				
r	α_{r1}	α_{r2}	α_{r3}	...	α_{rr}

with $\alpha_{1j} = 1 - \alpha_{2j} - \alpha_{3j} - \cdots - \alpha_{rj}$, $j = 1, \ldots, r$. To be specific here

$$\alpha_{i,j} = \text{pr}(W = i | X = j), i, j = 1, \ldots, r. \tag{3.3}$$

The goal is estimation of β-parameters of model (3.1) along with the other parameters for the model for X and the misclassification matrix. Before proceeding for estimation, I investigate if the model parameters are identifiable. The reason is that if the parameters are not identifiable, then there are many more model parameters than what are defined through our models, and consequently the concept of estimation is meaningless. For identification I use the seminal paper,[54] that formalized the definitions of identifiability and provided conditions of identifiability. First I state two definitions from [54] :

Definition 1. *Two structures are said to be observationally equivalent if they imply the same probability distribution for the observable random variable V.*

Definition 2. *A parameter point ω^* is locally identifiable if there exists an open neighborhood of ω^*, $\Omega_{\omega^*} \subset \Omega$ containing no other $\omega \in \Omega$, that are not observationally equivalent.*

Define $\theta = (\alpha_{21}, \alpha_{22}, \alpha_{23}, \ldots, \alpha_{2r}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \ldots, \alpha_{3r}, \ldots, \alpha_{r1}, \alpha_{r2}, \alpha_{r3}, \ldots, \alpha_{rr}, \gamma^T)^T$. If the parameters (β, θ) are not identifiable, then for every (β, θ) I can find another (β^*, θ^*) that are observationally equivalent. More particularly, if (β, θ) and (β^*, θ^*) are observationally equivalent, then

$$\begin{bmatrix} \text{pr}(Y=1, W=1|X^*, Z; \beta, \theta) \\ \vdots \\ \text{pr}(Y=1, W=r|X^*, Z; \beta, \theta) \\ \text{pr}(Y=0, W=1|X^*, Z; \beta, \theta) \\ \vdots \\ \text{pr}(Y=0, W=r|X^*, Z; \beta, \theta) \\ \text{pr}(W=1|X^*, Z; \beta, \theta) \\ \vdots \\ \text{pr}(W=r|X^*, Z; \beta, \theta) \end{bmatrix} = \begin{bmatrix} \text{pr}(Y=1, W=1|X^*, Z; \beta^*, \theta^*) \\ \vdots \\ \text{pr}(Y=1, W=r|X^*, Z; \beta^*, \theta^*) \\ \text{pr}(Y=0, W=1|X^*, Z; \beta^*, \theta^*) \\ \vdots \\ \text{pr}(Y=0, W=r|X^*, Z; \beta^*, \theta^*) \\ \text{pr}(W=1|X^*, Z; \beta^*, \theta^*) \\ \vdots \\ \text{pr}(W=r|X^*, Z; \beta^*, \theta^*) \end{bmatrix} \tag{3.4}$$

for every X^* and Z. Here

$$\begin{aligned} \text{pr}(Y=1, W=r'|X^*, Z; \beta, \theta) &= \sum_{j=1}^{r} \text{pr}(Y=1|X=j, Z; \beta)\text{pr}(W=r'|X=j, X^*, Z; \theta) \\ &\quad \times \text{pr}(X=j|X^*, Z; \gamma), \end{aligned}$$

$$\text{pr}(W=r'|X^*, Z; \beta, \theta) = \sum_{j=1}^{r} \text{pr}(W=r'|X=j, X^*, Z; \theta)\text{pr}(X=j|X^*, Z; \gamma).$$

for $r' = 1, \ldots, r$. So, the question is if (β, θ) are really non-identifiable, and which is now investigated on a case-by-case basis.

3.2.2 Non-identifiability of model parameters

Case 1: Both W and X have 2 categories

Suppose $\text{pr}(X = 2|X^*, Z; \gamma) = H(\gamma_0 + \gamma_1^T X^* + \gamma_2^T Z)$, and the misclassification probability matrix is

	X	
W	1	2
1	α_{11}	α_{12}
2	α_{21}	α_{22}

with $\alpha_{11} = 1 - \alpha_{21}$ and $\alpha_{12} = 1 - \alpha_{22}$. Define $\theta = (\alpha_{21}, \alpha_{22}, \gamma^T)^T$, where $\gamma = (\gamma_0, \gamma_1^T, \gamma_2^T)^T$. Then

$$\begin{aligned}
\text{pr}(W = 2|X^*, Z; \theta) &= \alpha_{21}\text{pr}(X = 1|X^*, Z; \gamma) + \alpha_{22}\text{pr}(X = 2|X^*, Z; \gamma) \\
&= \alpha_{21}\{1 - \text{pr}(X = 2|X^*, Z; \gamma)\} + \alpha_{22}\text{pr}(X = 2|X^*, Z; \gamma) \\
&= (\alpha_{22} - \alpha_{21})\text{pr}(X = 2|X^*, Z; \gamma) + \alpha_{21}.
\end{aligned}$$

Observe that a different set of parameters $\theta^\dagger = (\alpha_{21}^\dagger, \alpha_{22}^\dagger, \gamma^{\dagger,T})^T$, where $\gamma^\dagger = -\gamma$, $\alpha_{22}^\dagger = \alpha_{21}$ and $\alpha_{21}^\dagger = (\alpha_{22} - \alpha_{21})$, gives back the same probability as before, i.e., $\text{pr}(W = 2|X^* Z; \theta^\dagger) = \text{pr}(W = 2|X^*, Z; \theta)$. That means θ and $\theta^\dagger$ are observationally equivalent. Hence, the misclassification probabilities and the regression parameter γ are not identifiable. This non-identifiability holds irrespective of X^* being a categorical or continuous variable.

For brevity of notations define $p_{1*,i} = \text{pr}(X = 1|X_i^*, Z_i; \gamma)$, $p_{2*,i} = P(X = 2|X_i^*, Z_i; \gamma)$, $p_{1*,i}^w = \text{pr}(W = 1|X_i^*, Z_i; \theta)$, $p_{2*,i}^w = \text{pr}(W = 2|X_i^*, Z_i; \theta)$, $\alpha_{22}^d = \alpha_{22} - \alpha_{21}$,

and $a^{\otimes 2} = aa^T$ for any generic vector a. The likelihood function is

$$\mathcal{L} = \prod_{i=1}^{n} \{\text{pr}(W = 1|X_i^*, Z_i; \theta)\}^{I(W_i=1)} \{\text{pr}(W = 2|X_i^*, Z_i; \theta)\}^{I(W_i=2)}.$$

Then the information matrix is

$$I(\theta) = \sum_{i=1}^{n} \frac{1}{p_{1*,i}^w p_{2*,i}^w} \begin{bmatrix} A_{1,i} & A_{2,i} X_i^{*,T} & A_{2,i} Z_i^T \\ A_{2,i}^T X_i^* & (\alpha_{2d} p_{1*,i} p_{2*,i})^2 (X_i^*)^{\otimes 2} & (\alpha_{2d} p_{1*,i} p_{2*,i})^2 X_i^* Z_i^T \\ A_{2,i}^T Z_i & (\alpha_{2d} p_{1*,i} p_{2*,i})^2 Z_i X_i^{*,T} & (\alpha_{2d} p_{1*,i} p_{2*,i})^2 Z_i^{\otimes 2} \end{bmatrix}, \quad (3.5)$$

where

$$A_{1,i} = \begin{pmatrix} p_{1*,i}^2 & p_{1*,i} p_{2*,i} & \alpha_{2d} p_{1*,i}^2 p_{2*,i} \\ p_{1*,i} p_{2*,i} & p_{2*,i}^2 & \alpha_{2d} p_{1*,i} p_{2*,i}^2 \\ \alpha_{2d} p_{1*,i}^2 p_{2*,i} & \alpha_{2d} p_{1*,i} p_{2*,i}^2 & (\alpha_{2d} p_{1*,i} p_{2*,i})^2 \end{pmatrix}, \; A_{2,i} = \begin{pmatrix} \alpha_{2d} p_{1*,i}^2 p_{2*,i} \\ \alpha_{2d} p_{1*,i} p_{2*,i}^2 \\ (\alpha_{2d} p_{1*,i} p_{2*,i})^2 \end{pmatrix}.$$

If X^* or Z contains at least one single numeric (continuous) component, $I(\theta)$ is non-singular. I want to point out that this non-singularity of $I(\theta)$ does not contradict with the non-identification of parameters. The reason is explained in Theorem 1 of [54]. However before explaining the theorem, we provide the following definition.

Definition 3. *Let $M(\omega)$ be a matrix whose elements are continuous functions of ω everywhere in Ω. Then the point $\omega^* \in \Omega$ is said to be a regular point of the matrix if there exists an open neighborhood of ω^* in which $M(\omega)$ has constant rank.*

The result of Theorem 1 of [54] states that if θ is a *regular point* of $I(\theta)$, then non-singularity and parameter identification are equivalent. If p and q are the dimensions of X^* and Z, then $I(\theta)$ has rank $3 + p + q$ when $\alpha_{21} \neq \alpha_{22}$ and all components of X^* and Z are numeric, otherwise its rank is 2, hence it is not a matrix of constant rank. So, according to the definition of a regular point, θ is not a regular point of

$I(\theta)$ [54].

Case 2: Both W and X have 3 categories

Now suppose the following model:

$$\begin{aligned}
\text{pr}(X = r|X^*, Z; \boldsymbol{\gamma}) &= \frac{\exp(\gamma_{r0} + \gamma_{r1}^T X^* + \gamma_{r2}^T Z)}{1 + \sum_{k=2}^{3} \exp(\gamma_{k0} + \gamma_{k1}^T X^* + \gamma_{k2}^T Z)}, \text{ for } r = 2, 3, \\
\text{pr}(X = 1|X^*, Z; \boldsymbol{\gamma}) &= \frac{1}{1 + \sum_{k=2}^{3} \exp(\gamma_{k0} + \gamma_{k1}^T X^* \gamma_{k2}^T Z)},
\end{aligned}$$

where X^* is a p-vector consists of a set of binary or numeric instrumental variables, and the misclassification probability matrix is

	X		
W	1	2	3
1	α_{11}	α_{12}	α_{13}
2	α_{21}	α_{22}	α_{23}
3	α_{31}	α_{32}	α_{33}

with $\alpha_{11} = 1 - \alpha_{21} - \alpha_{31}$, $\alpha_{12} = 1 - \alpha_{22} - \alpha_{32}$, $\alpha_{13} = 1 - \alpha_{23} - \alpha_{33}$. Define $\theta = (\alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \gamma^T)^T$, where $\gamma = (\gamma_{20}, \gamma_{21}^T, \gamma_{22}^T, \gamma_{30}, \gamma_{31}^T, \gamma_{32}^T)^T$. Then

$$\begin{aligned}
\text{pr}(W = 2|X^*, Z; \theta) &= \alpha_{21}\text{pr}(X = 1|X^*, Z; \gamma) \\
&\quad + \alpha_{22}\text{pr}(X = 2|X^*, Z; \gamma) + \alpha_{23}\text{pr}(X = 3|X^*, Z; \gamma) \\
&= \alpha_{21}\{1 - \text{pr}(X = 2|X^*, Z; \gamma) - \text{pr}(X = 3|X^*, Z)\} \\
&\quad + \alpha_{22}\text{pr}(X = 2|X^*, Z; \gamma) + \alpha_{23}\text{pr}(X = 3|X^*, Z; \gamma) \\
&= \alpha_{21} + (\alpha_{22} - \alpha_{21})\text{pr}(X = 2|X^*, Z; \gamma) \\
&\quad + (\alpha_{23} - \alpha_{21})\text{pr}(X = 3|X^*, Z; \gamma),
\end{aligned}$$

and similarly,

$$\begin{aligned}\text{pr}(W=3|X^*,Z;\theta) &= \alpha_{31}+(\alpha_{32}-\alpha_{31})\text{pr}(X=2|X^*,Z;\gamma)\\ &\quad +(\alpha_{33}-\alpha_{31})\text{pr}(X=3|X^*,Z;\gamma).\end{aligned}$$

Observe that a different set of parameter $\theta^\dagger=(\alpha_{21}^\dagger,\alpha_{22}^\dagger,\alpha_{23}^\dagger,\alpha_{31}^\dagger,\alpha_{32}^\dagger,\alpha_{33}^\dagger,\gamma^{\dagger,T})^T$, $\gamma^\dagger=(\gamma_{20}^\dagger,\gamma_{21}^{\dagger,T},\gamma_{30}^\dagger,\gamma_{31}^{\dagger,T})^T$, $\gamma_{20}^\dagger=\gamma_{30}$, $\gamma_{30}^\dagger=\gamma_{20}$, $\gamma_{21}^\dagger=\gamma_{31}$, $\gamma_{31}^\dagger=\gamma_{21}$, $\gamma_{22}^\dagger=\gamma_{32}$, $\gamma_{32}^\dagger=\gamma_{22}$, $\alpha_{21}^\dagger=\alpha_{21}$, $\alpha_{22}^\dagger=\alpha_{23}$, $\alpha_{23}^\dagger=\alpha_{22}$, $\alpha_{31}^\dagger=\alpha_{31}$, $\alpha_{32}^\dagger=\alpha_{33}$, $\alpha_{33}^\dagger=\alpha_{32}$, gives back the same probability as before, i.e., $\text{pr}(W=2|X^*,Z;\theta^\dagger)=\text{pr}(W=2|X^*,Z;\theta)$ and $\text{pr}(W=3|X^*,Z;\theta^\dagger)=\text{pr}(W=3|X^*,Z;\theta)$. For example, we examine $\text{pr}(W=3|X^*,Z;\theta^\dagger)$:

$$\begin{aligned}\text{pr}(W=3|X^*,Z;\theta^\dagger) &= \alpha_{31}^\dagger+(\alpha_{32}^\dagger-\alpha_{31}^\dagger)\text{pr}(X=2|X^*,Z;\gamma^\dagger)\\ &\quad +(\alpha_{33}^\dagger-\alpha_{31}^\dagger)\text{pr}(X=3|X^*,Z;\gamma^\dagger)\\ &= \alpha_{31}+(\alpha_{33}-\alpha_{31})\text{pr}(X=2|X^*,Z;\gamma^\dagger)\\ &\quad +(\alpha_{32}-\alpha_{31})\text{pr}(X=3|X^*,Z;\gamma^\dagger)\\ &= \alpha_{31}+(\alpha_{33}-\alpha_{31})\left\{\frac{\exp(\gamma_{20}^\dagger+\gamma_{21}^{\dagger T}X^*+\gamma_{22}^{\dagger T}Z)}{1+\sum_{k=2}^3\exp(\gamma_{k0}+\gamma_{k1}^TX^*+\gamma_{k2}^TZ)}\right\}\\ &\quad +(\alpha_{32}-\alpha_{31})\left\{\frac{\exp(\gamma_{30}^\dagger+\gamma_{31}^{\dagger T}X^*+\gamma_{32}^{\dagger T}Z)}{1+\sum_{k=2}^3\exp(\gamma_{k0}+\gamma_{k1}^TX^*+\gamma_{k2}^TZ)}\right\}\\ &= \alpha_{31}+(\alpha_{33}-\alpha_{31})\left\{\frac{\exp(\gamma_{30}+\gamma_{31}^TX^*+\gamma_{32}^TZ)}{1+\sum_{k=2}^3\exp(\gamma_{k0}+\gamma_{k1}^TX^*+\gamma_{k2}^TZ)}\right\}\\ &\quad +(\alpha_{32}-\alpha_{31})\left\{\frac{\exp(\gamma_{20}+\gamma_{21}^TX^*+\gamma_{22}^TZ)}{1+\sum_{k=2}^3\exp(\gamma_{k0}+\gamma_{k1}^TX^*+\gamma_{k2}^TZ)}\right\}\\ &= \alpha_{31}+(\alpha_{33}-\alpha_{31})\text{pr}(X=3|X^*,Z;\gamma^\dagger)\\ &\quad +(\alpha_{32}-\alpha_{31})\text{pr}(X=2|X^*,Z;\gamma^\dagger)\\ &= \text{pr}(W=3|X^*,Z;\theta).\end{aligned}$$

The same derivation can be used to show $\text{pr}(W=2|X^*,Z;\theta^\dagger)=\text{pr}(W=$

$2|X^*, Z; \theta)$. Subsequently θ and $\theta^\dagger$ are observationally equivalent. Hence, the misclassification probabilities and the regression parameter are not identifiable.

Case 3: Both W and X have $r > 3$ categories

I now consider the following model. Suppose

$$\begin{aligned}
\text{pr}(X = r|X^*, Z; \boldsymbol{\gamma}) &= \frac{\exp(\gamma_{r0} + \gamma_{r1}^T X^* + \gamma_{r2}^T Z)}{1 + \sum_{k=2}^r \exp(\gamma_{k0} + \gamma_{k1}^T X^* + \gamma_{k2}^T Z)}, \text{ for } r = 2, 3, \ldots, r, \\
\text{pr}(X = 1|X^*, Z; \boldsymbol{\gamma}) &= \frac{1}{1 + \sum_{k=2}^r \exp(\gamma_{k0} + \gamma_{k1}^T X^* + \gamma_{k2}^T Z)},
\end{aligned}$$

where X^* is a p-vector consists of a set of binary or numeric instrumental variables, and the misclassification probability matrix is

	X				
W	1	2	3	...	r
1	α_{11}	α_{12}	α_{13}	...	α_{1r}
2	α_{21}	α_{22}	α_{23}	...	α_{2r}
$\vdots$	$\vdots$				
r	α_{r1}	α_{r2}	α_{r3}	...	α_{rr}

with $\alpha_{1j} = 1 - \alpha_{2j} - \alpha_{3j} - \cdots - \alpha_{rj}$, $j = 1, \ldots, r$. Define $\gamma = (\gamma_{20}, \gamma_{21}^T, \gamma_{30}, \gamma_{31}^T, \gamma_{40}, \gamma_{41}^T, \ldots, \gamma_{r0}, \gamma_{r1}^T)^T$ and then $\theta = (\alpha_{21}, \alpha_{22}, \alpha_{23}, \ldots, \alpha_{2r}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \ldots, \alpha_{3r}, \ldots, \alpha_{r1}, \alpha_{r2}, \alpha_{r3}, \ldots, \alpha_{rr}, \gamma^T)^T$. Then for any $k = 2, \ldots, r$,

$$\begin{aligned}
\text{pr}(W = k|X^*, Z; \theta) &= \alpha_{k1}\text{pr}(X = 1|X^*, Z; \gamma) + \cdots + \alpha_{kr}\text{pr}(X = r|X^*, Z; \gamma) \\
&= \alpha_{k1} + (\alpha_{k2} - \alpha_{k1})\text{pr}(X = 2|X^*, Z; \gamma) + (\alpha_{k3} - \alpha_{k1}) \\
&\quad \times \text{pr}(X = 3|X^*, Z; \gamma) + \cdots + (\alpha_{kr} - \alpha_{k1})\text{pr}(X = r|X^*, Z; \gamma).
\end{aligned}$$

Observe that a different set of parameter $\theta^\dagger = (\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2r}, \alpha_{31}, \alpha_{32}, \ldots, \alpha_{3r}, \alpha_{41}, \alpha_{42}, \ldots, \alpha_{4r} \ldots, \alpha_{r1}, \alpha_{r2}, \ldots, \alpha_{rr}, \gamma^{\dagger,T})^T$, where $\gamma^\dagger = (\gamma_{30}^\dagger, \gamma_{31}^{\dagger,T}, \gamma_{20}^\dagger, \gamma_{21}^{\dagger,T}, \gamma_{40}, \gamma_{41}^T, \ldots,$

$\gamma_{r0}, \gamma_{r1}^T)^T$ gives back the same probability as before, i.e., $\text{pr}(W = k|X^*, Z; \theta^\dagger) = \text{pr}(W = k|X^*, Z; \theta)$ for $k = 1, \ldots, r$. Thus θ and $\theta^\dagger$ are observationally equivalent. Hence, the misclassification probabilities and the regression parameter are not identifiable.

3.2.3 Parameter identifiability under constraints

The parameters can be locally identifiable under constraints. Suppose that there are c constraints on the parameters, $G_k(\theta) = 0$, $k = 1, \ldots, c$, and define

$$\boldsymbol{\psi}(\theta) = \begin{bmatrix} \partial G_1(\theta)/\partial\theta^T \\ \vdots \\ \partial G_c(\theta)/\partial\theta^T \end{bmatrix}.$$

Theorem 2 of [54] says that if θ in the constrained parameter space is a *regular point* for $I(\theta)$ and $\boldsymbol{\psi}(\theta)$, then θ is identifiable if and only if the rank of $V(\theta)$ is d_θ, the dimension of θ, where

$$V(\theta) = \begin{pmatrix} I(\theta) \\ \boldsymbol{\psi}(\theta) \end{pmatrix}_{(d_\theta + c) \times d_\theta}.$$

Now, I introduce the following constraint.

(C1) The misclassification probability matrix $((\alpha_{ij}))$ is strictly diagonally dominant. That is $\alpha_{ii} > \sum_{j \neq i} \alpha_{ji}$ or equivalently $\alpha_{ii} > 0.5$.

Assumption (C1) is one of the assumptions that [33] used for nonparametric identification of the probability models $\text{pr}(Y|X, Z)$, $\text{pr}(W|X, Z)$, and $\text{pr}(X|X^*, Z)$ when W, X, X^* are all categorial variables. Now, I state the result in the following theorem.

Theorem 1. *If pr$(Y|X, Z)$ and pr$(X|X^*, Z)$ follow the models specified in (3.1) and*

(3.2), respectively, and assumption (C1) holds, then all the parameters are identifiable.

Proof: The proof of this result will partly be based on the method of induction and partly on rigorous mathematics. First, I consider the case where both X and W have two categories.

Both W and X have two categories: Let us adopt constraint (C1). Then $\alpha_{11} > 0.5$ and $\alpha_{22} > 0.5$, that means $\alpha_{21} < 0.5$ and $\alpha_{22} > 0.5$. So,

$$\boldsymbol{\psi}(\theta) = \begin{bmatrix} -1 & 0 & \mathbf{0}^T & \mathbf{0}^T \\ 0 & 1 & \mathbf{0}^T & \mathbf{0} \end{bmatrix}.$$

Observe that in the constrained space, θ is a regular point for $I(\theta)$ given in (3.5) and $\boldsymbol{\psi}(\theta)$. Now, the rank of $V(\theta)$ is d_θ. Hence, the parameters are identifiable. I want to point out that [45] obtained parameter identifiability under another condition that is, in our current notations, $\alpha_{12} + \alpha_{21} < 1$. However, this condition and constraint (C1) are not equivalent. Particularly, (C1) implies $\alpha_{12} + \alpha_{21} < 1$ but not vice versa.

Both W and X have three categories: Model for X given instrumental variable X^* is

$$\begin{aligned} p_{k*,i} \equiv \text{pr}(X = k|X_i^*, Z_i; \gamma) &= \frac{\exp(\gamma_{k0} + \gamma_{k1}^T X_i^* + \gamma_{k2}^T Z_i)}{1 + \sum_{r=2}^3 \exp(\gamma_{r0} + \gamma_{r1}^T X_i^* + \gamma_{r2}^T Z_i)}, \; k = 2, 3, \\ p_{1*,i} \equiv \text{pr}(X = 1|X_i^*, Z_i; \gamma) &= \frac{1}{1 + \sum_{r=2}^3 \exp(\gamma_{r0} + \gamma_{r1}^T X_i^* + \gamma_{r2}^T Z_i)}, \end{aligned}$$

and the induced model for W given X^* is

$$\begin{aligned} p_{k*,i}^w \equiv \text{pr}(W = k|X_i^*, Z_i; \theta) &= \alpha_{k1} + (\alpha_{k2} - \alpha_{k1})\text{pr}(X = 2|X_i^*, Z_i; \gamma) \\ &\quad + (\alpha_{k3} - \alpha_{k1})\text{pr}(X = 3|X_i^*, Z_i; \gamma), \end{aligned}$$

for $k = 2$ and 3 and $p^w_{1*,i} = 1 - p^w_{2*,i} - p^w_{3*,i}$. Define $\alpha^d_{22} = \alpha_{22} - \alpha_{21}$, $\alpha^d_{23} = \alpha_{23} - \alpha_{21}$, $\alpha^d_{32} = \alpha_{32} - \alpha_{31}$ and $\alpha^d_{33} = \alpha_{33} - \alpha_{31}$. The information matrix $I(\theta)$ can be partitioned as follows:

$$I(\theta) = \begin{pmatrix} A_1 & A_2 & B_2 & B_3 \\ A_2^T & A_3 & C_2 & C_3 \\ B_2^T & C_2^T & A_4 & A_5 \\ B_3^T & C_3^T & A_5^T & A_6 \end{pmatrix},$$

where

$$A_1 = \sum_{i=1}^{n} \left(\frac{1}{p^w_{1*,i}} + \frac{1}{p^w_{2*,i}} \right) \begin{bmatrix} p^2_{1*,i} & p_{1*,i}p_{2*,i} & p_{1*,i}p_{3*,i} \\ p_{1*,i}p_{2*,i} & p^2_{2*,i} & p_{2*,i}p_{3*,i} \\ p_{1*,i}p_{3*,i} & p_{2*,i}p_{3*,i} & p^2_{3*,i} \end{bmatrix},$$

$$A_2 = \sum_{i=1}^{n} \frac{1}{p^w_{1*,i}} \begin{bmatrix} p^2_{1*,i} & p_{1*,i}p_{2*,i} & p_{1*,i}p_{3*,i} \\ p_{1*,i}p_{2*,i} & p^2_{2*,i} & p_{2*,i}p_{3*,i} \\ p_{1*,i}p_{3*,i} & p_{2*,i}p_{3*,i} & p^2_{3*,i} \end{bmatrix},$$

$$A_3 = \sum_{i=1}^{n} \left(\frac{1}{p^w_{1*,i}} + \frac{1}{p^w_{3*,i}} \right) \begin{bmatrix} p^2_{1*,i} & p_{1*,i}p_{2*,i} & p_{1*,i}p_{3*,i} \\ p_{1*,i}p_{2*,i} & p^2_{2*,i} & p_{2*,i}p_{3*,i} \\ p_{1*,i}p_{3*,i} & p_{2*,i}p_{3*,i} & p^2_{3*,i} \end{bmatrix},$$

$$A_4 = \sum_{i=1}^{n} \left[\frac{\{-(\alpha^d_{22} + \alpha^d_{32})(1 - p_{2*,i}) + (\alpha^d_{23} + \alpha^d_{33})p_{3*,i}\}^2}{p^w_{1*,i}} \right.$$

$$+\frac{\{\alpha_{22}^d(1-p_{2*,i})-\alpha_{23}^d p_{3*,i}\}^2}{p_{2*,i}^w}+\frac{\{\alpha_{32}^d(1-p_{2*,i})-\alpha_{33}^d(p_{3*,i})\}^2}{p_{3*,i}^w}\Bigg]p_{2*,i}^2\begin{pmatrix}1\\X_i^*\end{pmatrix}^{\otimes 2},$$

$$\begin{aligned}A_5 &= \sum_{i=1}^n\Bigg[\frac{1}{p_{1*,i}^w}\{-(\alpha_{22}^d+\alpha_{32}^d)(1-p_{2*,i})+(\alpha_{23}^d+\alpha_{33}^d)p_{3*,i}\}\\
&\quad\times\{-(\alpha_{23}^d+\alpha_{33}^d)(1-p_{3*,i})+(\alpha_{22}^d+\alpha_{32}^d)p_{2*,i}\}\\
&\quad+\frac{\{\alpha_{22}^d(1-p_{2*,i})-\alpha_{23}^d p_{3*,i}\}\{\alpha_{23}^d(1-p_{3*,i})-\alpha_{22}^d p_{2*,i}\}}{p_{2*,i}^w}\\
&\quad+\frac{\{\alpha_{32}^d(1-p_{2*,i})-\alpha_{33}^d p_{3*,i}\}\{\alpha_{33}^d(1-p_{3*,i})-\alpha_{32}^d p_{2*,i}\}}{p_{3*,i}^w}\Bigg]p_{2*,i}p_{3*,i}\begin{pmatrix}1\\X_i^*\end{pmatrix}^{\otimes 2},\end{aligned}$$

$$\begin{aligned}A_6 &= \sum_{i=1}^n\Bigg[\frac{\{(\alpha_{22}^d+\alpha_{32}^d)p_{2*,i}-(\alpha_{23}^d+\alpha_{33}^d)(1-p_{3*,i})\}^2}{p_{1*,i}^w}\\
&\quad+\frac{\{\alpha_{23}^d(1-p_{3*,i})-\alpha_{22}^d p_{2*,i}\}^2}{p_{2*,i}^w}+\frac{\{\alpha_{33}^d(1-p_{3*,i})-\alpha_{32}^d p_{2*,i}\}^2}{p_{3*,i}^w}\Bigg]p_{3*,i}^2\begin{pmatrix}1\\X_i^*\end{pmatrix}^{\otimes 2},\end{aligned}$$

$$\begin{aligned}B_2 &= \sum_{i=1}^n\left\{\frac{(\alpha_{22}^d+\alpha_{32}^d)(1-p_{2*,i})-(\alpha_{23}^d+\alpha_{33}^d)p_{3*,i}}{p_{1*,i}^w}+\frac{\alpha_{22}^d(1-p_{2*,i})-\alpha_{23}^d p_{3*,i}}{p_{2*,i}^w}\right\}\\
&\quad\times\begin{pmatrix}p_{1*,i}p_{2*,i}\\p_{2*,i}^2\\p_{3*,i}p_{2*,i}\end{pmatrix}(1\ X_i^{*,T}),\end{aligned}$$

$$B_3 = \sum_{i=1}^n\left\{\frac{(\alpha_{23}^d+\alpha_{33}^d)(1-p_{3*,i})-(\alpha_{22}^d+\alpha_{32}^d)p_{2*,i}}{p_{1*,i}^w}+\frac{\alpha_{23}^d(1-p_{3*,i})-\alpha_{22}^d p_{2*,i}}{p_{2*,i}^w}\right\}$$

$$\times \begin{pmatrix} p_{1*,i}p_{3*,i} \\ p_{2*,i}p_{3*,i} \\ p_{3*,i}^2 \end{pmatrix} (1 \; X_i^{*,T}),$$

$$C_2 = \sum_{i=1}^{n} \left\{ \frac{(\alpha_{22}^d + \alpha_{32}^d)(1 - p_{2*,i}) - (\alpha_{23}^d + \alpha_{33}^d)p_{3*,i}}{p_{1*,i}^w} + \frac{\alpha_{32}^d(1 - p_{2*,i}) - \alpha_{33}^d p_{3*,i}}{p_{3*,i}^w} \right\}$$
$$\times \begin{pmatrix} p_{1*,i}p_{2*,i} \\ p_{2*,i}^2 \\ p_{3*,i}p_{2*,i} \end{pmatrix} (1 \; X_i^{*,T}),$$

$$C_3 = \sum_{i=1}^{n} \left\{ \frac{(\alpha_{23}^d + \alpha_{33}^d)(1 - p_{3*,i}) - (\alpha_{22}^d + \alpha_{32}^d)p_{2*,i}}{p_{1*,i}^w} + \frac{\alpha_{33}^d(1 - p_{3*,i}) - \alpha_{32}^d p_{2*,i}}{p_{3*,i}^w} \right\}$$
$$\times \begin{pmatrix} p_{1*,i}p_{3*,i} \\ p_{2*,i}p_{3*,i} \\ p_{3*,i}^2 \end{pmatrix} (1 \; X_i^{*,T}).$$

Note that when $\alpha_{31} = \alpha_{32} = \alpha_{33}$ and $\alpha_{21} = \alpha_{22} = \alpha_{23}$, the information matrix is singular. Thus, θ is not a regular point for the information matrix. Now, assume that the misclassification matrix is diagonally dominant. That means $\alpha_{11} > 0.5$, $\alpha_{22} > 0.5$ and $\alpha_{33} > 0.5$. Under this condition,

$$\boldsymbol{\psi}(\theta) = \begin{bmatrix} \text{-1} & 0 & 0 & \text{-1} & 0 & 0 & \mathbf{0}^T & \mathbf{0}^T \\ 0 & 1 & 0 & 0 & 0 & 0 & \mathbf{0}^T & \mathbf{0}^T \\ 0 & 0 & 0 & 0 & 0 & 1 & \mathbf{0}^T & \mathbf{0}^T \end{bmatrix}.$$

Observe that in the constrained space, θ is a regular point for $I(\theta)$ and $\boldsymbol{\psi}(\theta)$. Now,

the rank of $V(\theta)$ is d_θ. Hence, the parameters are identifiable.

Both W and X have four categories: Define

$$\begin{aligned} p_{k*,i} \equiv \text{pr}(X = k|X_i^*, Z_i; \gamma) &= \frac{\exp(\gamma_{k0} + \gamma_{k1}^T X_i^* + \gamma_{k2}^T Z_i)}{1 + \sum_{r=2}^4 \exp(\gamma_{r0} + \gamma_{r1}^T X_i^* + \gamma_{r2}^T Z_i)}, \; k = 2, 3, 4 \\ p_{1*,i} \equiv \text{pr}(X = 1|X_i^*, Z_i; \gamma) &= \frac{1}{1 + \sum_{r=2}^4 \exp(\gamma_{r0} + \gamma_{r1}^T X_i^* + \gamma_{r2}^T Z_i)}, \end{aligned}$$

and the induced model for W given X^* is

$$\begin{aligned} p_{k*,i}^w \equiv \text{pr}(W = k|X_i^*, Z_i; \theta) &= \alpha_{k1} + (\alpha_{k2} - \alpha_{k1})\text{pr}(X = 2|X_i^*, Z_i; \gamma) \\ &\quad + (\alpha_{k3} - \alpha_{k1})\text{pr}(X = 3|X_i^*, Z_i; \gamma) \\ &\quad + (\alpha_{k4} - \alpha_{k1})\text{pr}(X = 4|X_i^*, Z_i; \gamma), \end{aligned}$$

for $k = 2, \ldots, 4$, and $p_{1*,i}^w = 1 - p_{2*,i}^w - p_{3*,i}^w - p_{4*,i}^w$. Define $\alpha_{22}^d = \alpha_{22} - \alpha_{21}$, $\alpha_{23}^d = \alpha_{23} - \alpha_{21}$, $\alpha_{24}^d = \alpha_{24} - \alpha_{21}$, $\alpha_{32}^d = \alpha_{32} - \alpha_{31}$, $\alpha_{33}^d = \alpha_{33} - \alpha_{31}$, $\alpha_{34}^d = \alpha_{34} - \alpha_{31}$, $\alpha_{42}^d = \alpha_{42} - \alpha_{41}$, $\alpha_{43}^d = \alpha_{43} - \alpha_{41}$, and $\alpha_{44}^d = \alpha_{44} - \alpha_{41}$. The information matrix $I(\theta)$ can be partitioned as follows:

$$I(\theta) = \begin{pmatrix} A_1 & A_2 & A_2 & B_1 & B_2 & B_3 \\ A_2^T & A_3 & A_2 & C_1 & C_2 & C_3 \\ A_2^T & A_2^T & A_4 & D_1 & D_2 & D_3 \\ B_1^T & C_1^T & D_1^T & A_5 & A_6 & A_7 \\ B_2^T & C_2^T & D_2^T & A_6^T & A_8 & A_9 \\ B_3^T & C_3^T & D_3^T & A_7^T & A_9^T & A_{10} \end{pmatrix},$$

where

$$A_1 = \sum_{i=1}^{n} \left(\frac{1}{p^w_{1*,i}} + \frac{1}{p^w_{2*,i}} \right) \begin{bmatrix} p^2_{1*,i} & p_{1*,i}p_{2*,i} & p_{1*,i}p_{3*,i} & p_{1*,i}p_{4*,i} \\ p_{1*,i}p_{2*,i} & p^2_{2*,i} & p_{2*,i}p_{3*,i} & p_{2*,i}p_{4*,i} \\ p_{1*,i}p_{3*,i} & p_{2*,i}p_{3*,i} & p^2_{3*,i} & p_{3*,i}p_{4*,i} \\ p_{1*,i}p_{4*,i} & p_{2*,i}p_{4*,i} & p_{3*,i}p_{4*,i} & p^2_{4*,i} \end{bmatrix},$$

$$A_2 = \sum_{i=1}^{n} \left(\frac{1}{p^w_{1*,i}} \right) \begin{bmatrix} p^2_{1*,i} & p_{1*,i}p_{2*,i} & p_{1*,i}p_{3*,i} & p_{1*,i}p_{4*,i} \\ p_{1*,i}p_{2*,i} & p^2_{2*,i} & p_{2*,i}p_{3*,i} & p_{2*,i}p_{4*,i} \\ p_{1*,i}p_{3*,i} & p_{2*,i}p_{3*,i} & p^2_{3*,i} & p_{3*,i}p_{4*,i} \\ p_{1*,i}p_{4*,i} & p_{2*,i}p_{4*,i} & p_{3*,i}p_{4*,i} & p^2_{4*,i} \end{bmatrix},$$

$$A_3 = \sum_{i=1}^{n} \left(\frac{1}{p^w_{1*,i}} + \frac{1}{p^w_{3*,i}} \right) \begin{bmatrix} p^2_{1*,i} & p_{1*,i}p_{2*,i} & p_{1*,i}p_{3*,i} & p_{1*,i}p_{4*,i} \\ p_{1*,i}p_{2*,i} & p^2_{2*,i} & p_{2*,i}p_{3*,i} & p_{2*,i}p_{4*,i} \\ p_{1*,i}p_{3*,i} & p_{2*,i}p_{3*,i} & p^2_{3*,i} & p_{3*,i}p_{4*,i} \\ p_{1*,i}p_{4*,i} & p_{2*,i}p_{4*,i} & p_{3*,i}p_{4*,i} & p^2_{4*,i} \end{bmatrix},$$

$$A_4 = \sum_{i=1}^{n} \left(\frac{1}{p^w_{1*,i}} + \frac{1}{p^w_{4*,i}} \right) \begin{bmatrix} p^2_{1*,i} & p_{1*,i}p_{2*,i} & p_{1*,i}p_{3*,i} & p_{1*,i}p_{4*,i} \\ p_{1*,i}p_{2*,i} & p^2_{2*,i} & p_{2*,i}p_{3*,i} & p_{2*,i}p_{4*,i} \\ p_{1*,i}p_{3*,i} & p_{2*,i}p_{3*,i} & p^2_{3*,i} & p_{3*,i}p_{4*,i} \\ p_{1*,i}p_{4*,i} & p_{2*,i}p_{4*,i} & p_{3*,i}p_{4*,i} & p^2_{4*,i} \end{bmatrix},$$

$$B_1 = \sum_{i=1}^{n} \left[\frac{1}{p^w_{1*,i}} \left\{ (\alpha^d_{22} + \alpha^d_{32} + \alpha^d_{42})(1 - p_{2*,i}) - (\alpha^d_{23} + \alpha^d_{33} + \alpha^d_{43})p_{3*,i} \right.\right.$$

$$-(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)p_{4*,i}\Big\}+\frac{\alpha_{22}^d(1-p_{2*,i})-\alpha_{23}^d p_{3*,i}-\alpha_{24}^d p_{4*,i}}{p_{2*,i}^w}\Big]$$
$$\times\begin{pmatrix}p_{1*,i}p_{2*,i}\\ p_{2*,i}^2\\ p_{3*,i}p_{2*,i}\\ p_{4*,i}p_{2*,i}\end{pmatrix}(1\ X_i^{*,T}),$$

$$\begin{aligned}B_2 &= \sum_{i=1}^n\Big[\frac{1}{p_{1*,i}^w}\Big\{(\alpha_{22}^d+\alpha_{32}^d+\alpha_{42}^d)(1-p_{2*,i})-(\alpha_{23}^d+\alpha_{33}^d+\alpha_{43}^d)p_{3*,i}\\ &\quad -(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)p_{4*,i}\Big\}+\frac{\alpha_{32}^d(1-p_{2*,i})-\alpha_{33}^d p_{3*,i}-\alpha_{34}^d p_{4*,i}}{p_{3*,i}^w}\Big]\\ &\quad\times\begin{pmatrix}p_{1*,i}p_{3*,i}\\ p_{2*,i}p_{3*,i}\\ p_{3*,i}^2\\ p_{4*,i}p_{3*,i}\end{pmatrix}(1\ X_i^{*,T}),\end{aligned}$$

$$\begin{aligned}B_3 &= \sum_{i=1}^n\Big[\Big\{\frac{1}{p_{1*,i}^w}(\alpha_{22}^d+\alpha_{32}^d+\alpha_{42}^d)(1-p_{2*,i})-(\alpha_{23}^d+\alpha_{33}^d+\alpha_{43}^d)p_{3*,i}\\ &\quad -(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)p_{4*,i}+\frac{\alpha_{42}^d(1-p_{2*,i})-\alpha_{43}^d p_{3*,i}-\alpha_{44}^d p_{4*,i}}{p_{4*,i}^w}\Big\}\Big]\\ &\quad\times\begin{pmatrix}p_{1*,i}p_{4*,i}\\ p_{2*,i}p_{4*,i}\\ p_{3*,i}p_{4*,i}\\ p_{4*,i}^2\end{pmatrix}(1\ X_i^{*,T}),\end{aligned}$$

$$
\begin{aligned}
C_1 &= \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Big\{-(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})p_{2*,i}+(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})(1-p_{3*,i}) \\
&\quad -(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})p_{4*,i}\Big\}+\frac{-\alpha^d_{22}p_{2*,i}+\alpha^d_{23}(1-p_{3*,i})-\alpha^d_{24}p_{4*,i}}{p^w_{2*,i}}\Bigg] \\
&\quad \times \begin{pmatrix} p_{1*,i}p_{2*,i} \\ p^2_{2*,i} \\ p_{3*,i}p_{2*,i} \\ p_{4*,i}p_{2*,i} \end{pmatrix}(1\ X^{*,T}_i),
\end{aligned}
$$

$$
\begin{aligned}
C_2 &= \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Big\{-(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})p_{2*,i}+(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})(1-p_{3*,i}) \\
&\quad -(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})p_{4*,i}\Big\}+\frac{-\alpha^d_{32}p_{2*,i}+\alpha^d_{33}(1-p_{3*,i})-\alpha^d_{34}p_{4*,i}}{p^w_{3*,i}}\Bigg] \\
&\quad \times \begin{pmatrix} p_{1*,i}p_{3*,i} \\ p_{2*,i}p_{3*,i} \\ p^2_{3*,i} \\ p_{4*,i}p_{3*,i} \end{pmatrix}(1\ X^{*,T}_i),
\end{aligned}
$$

$$
\begin{aligned}
C_3 &= \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Big\{-(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})p_{2*,i}+(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})(1-p_{3*,i}) \\
&\quad -(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})p_{4*,i}\Big\}+\frac{-\alpha^d_{42}p_{2*,i}+\alpha^d_{43}(1-p_{3*,i})-\alpha^d_{44}p_{4*,i}}{p^w_{4*,i}}\Bigg] \\
&\quad \times \begin{pmatrix} p_{1*,i}p_{4*,i} \\ p_{2*,i}p_{4*,i} \\ p_{3*,i}p_{4*,i} \\ p^2_{4*,i} \end{pmatrix}(1\ X^{*,T}_i),
\end{aligned}
$$

$$
\begin{aligned}
D_1 &= \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Big\{-(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})p_{2*,i}-(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})p_{3*,i} \\
&\quad +(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})(1-p_{4*,i})\Big\}+\frac{-\alpha^d_{22}p_{2*,i}-\alpha^d_{23}p_{3*,i}+\alpha^d_{24}(1-p_{4*,i})}{p^w_{2*,i}}\Bigg] \\
&\quad \times \begin{pmatrix} p_{1*,i}p_{2*,i} \\ p^2_{2*,i} \\ p_{3*,i}p_{2*,i} \\ p_{4*,i}p_{2*,i} \end{pmatrix} (1 \; X_i^{*,T}),
\end{aligned}
$$

$$
\begin{aligned}
D_2 &= \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Big\{-(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})p_{2*,i}-(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})(1-p_{3*,i}) \\
&\quad +(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})(1-p_{4*,i})\Big\}+\frac{-\alpha^d_{32}p_{2*,i}-\alpha^d_{33}p_{3*,i}+\alpha^d_{34}(1-p_{4*,i})}{p^w_{3*,i}}\Bigg] \\
&\quad \times \begin{pmatrix} p_{1*,i}p_{3*,i} \\ p_{2*,i}p_{3*,i} \\ p^2_{3*,i} \\ p_{4*,i}p_{3*,i} \end{pmatrix} (1 \; X_i^{*,T}),
\end{aligned}
$$

$$
\begin{aligned}
D_3 &= \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Big\{-(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})p_{2*,i}-(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})p_{3*,i} \\
&\quad +(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})(1-p_{4*,i})\Big\}+\frac{-\alpha^d_{42}p_{2*,i}-\alpha^d_{43}p_{3*,i}+\alpha^d_{44}(1-p_{4*,i})}{p^w_{4*,i}}\Bigg] \\
&\quad \times \begin{pmatrix} p_{1*,i}p_{4*,i} \\ p_{2*,i}p_{4*,i} \\ p_{3*,i}p_{4*,i} \\ p^2_{4*,i} \end{pmatrix} (1 \; X_i^{*,T}),
\end{aligned}
$$

$$
\begin{aligned}
A_5 \;=\; & \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Bigg\{ -(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})(1-p_{2*,i})+(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})p_{3*,i} \\
& +(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})p_{4*,i}\Bigg\}^2 \\
& +\frac{\{\alpha^d_{22}(1-p_{2*,i})-\alpha^d_{23}p_{3*,i}-\alpha^d_{24}p_{4*,i}\}^2}{p^w_{2*,i}} \\
& +\frac{\{\alpha^d_{32}(1-p_{2*,i})-\alpha^d_{33}p_{3*,i}-\alpha^d_{34}p_{4*,i}\}^2}{p^w_{3*,i}} \\
& +\frac{\{\alpha^d_{42}(1-p_{2*,i})-\alpha^d_{43}p_{3*,i}-\alpha^d_{44}p_{4*,i}\}^2}{p^w_{4*,i}}\Bigg]p^2_{2*,i}\begin{pmatrix}1\\ X^*_i\end{pmatrix}^{\otimes 2},
\end{aligned}
$$

$$
\begin{aligned}
A_6 \;=\; & \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Bigg\{ -(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})(1-p_{2*,i})+(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})p_{3*,i} \\
& +(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})p_{4*,i}\Bigg\} \\
& \times\Bigg\{(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})p_{2*,i}-(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})(1-p_{3*,i}) \\
& +(\alpha^d_{24}+\alpha^d_{34}+\alpha^d_{44})p_{4*,i}\Bigg\} \\
& +\frac{\{\alpha^d_{22}(1-p_{2*,i})-\alpha^d_{23}p_{3*,i}-\alpha^d_{24}p_{4*,i}\}\{-\alpha^d_{22}p_{2*,i}+\alpha^d_{23}(1-p_{3*,i})-\alpha^d_{24}p_{4*,i}\}}{p^w_{2*,i}} \\
& +\frac{\{\alpha^d_{32}(1-p_{2*,i})-\alpha^d_{33}p_{3*,i}-\alpha^d_{34}p_{4*,i}\}\{-\alpha^d_{32}p_{2*,i}+\alpha^d_{33}(1-p_{3*,i})-\alpha^d_{34}p_{4*,i}\}}{p^w_{3*,i}} \\
& +\frac{\{\alpha^d_{42}(1-p_{2*,i})-\alpha^d_{43}p_{3*,i}-\alpha^d_{44}p_{4*,i}\}\{-\alpha^d_{42}p_{2*,i}+\alpha^d_{43}(1-p_{3*,i})-\alpha^d_{44}p_{4*,i}\}}{p^w_{4*,i}}\Bigg] \\
& \times p_{2*,i}p_{3*,i}\begin{pmatrix}1\\ X^*_i\end{pmatrix}^{\otimes 2},
\end{aligned}
$$

$$
A_7 \;=\; \sum_{i=1}^{n}\Bigg[\frac{1}{p^w_{1*,i}}\Bigg\{ -(\alpha^d_{22}+\alpha^d_{32}+\alpha^d_{42})(1-p_{2*,i})+(\alpha^d_{23}+\alpha^d_{33}+\alpha^d_{43})p_{3*,i}
$$

$$
\begin{aligned}
&+(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)p_{4*,i}\Big\}\\
&\times\Big\{(\alpha_{22}^d+\alpha_{32}^d+\alpha_{42}^d)p_{2*,i}+(\alpha_{23}^d+\alpha_{33}^d+\alpha_{43}^d)p_{3*,i}\\
&-(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)(1-p_{4*,i})\Big\}\\
&+\frac{\{\alpha_{22}^d(1-p_{2*,i})-\alpha_{23}^d p_{3*,i}-\alpha_{24}^d p_{4*,i}\}\{-\alpha_{22}^d p_{2*,i}-\alpha_{23}^d p_{3*,i}+\alpha_{24}^d(1-p_{4*,i})\}}{p_{2*,i}^w}\\
&+\frac{\{\alpha_{32}^d(1-p_{2*,i})-\alpha_{33}^d p_{3*,i}-\alpha_{34}^d p_{4*,i}\}\{-\alpha_{32}^d p_{2*,i}-\alpha_{33}^d p_{3*,i}+\alpha_{34}^d(1-p_{4*,i})\}}{p_{3*,i}^w}\\
&+\frac{\{\alpha_{42}^d(1-p_{2*,i})-\alpha_{43}^d p_{3*,i}-\alpha_{44}^d p_{4*,i}\}\{-\alpha_{42}^d p_{2*,i}-\alpha_{43}^d p_{3*,i}+\alpha_{44}^d(1-p_{4*,i})\}}{p_{4*,i}^w}\Big]\\
&\times p_{2*,i}p_{4*,i}\begin{pmatrix}1\\ X_i^*\end{pmatrix}^{\otimes 2},
\end{aligned}
$$

$$
\begin{aligned}
A_8 \;=\; &\sum_{i=1}^n\Big[\frac{1}{p_{1*,i}^w}\Big\{(\alpha_{22}^d+\alpha_{32}^d+\alpha_{42}^d)p_{2*,i}-(\alpha_{23}^d+\alpha_{33}^d+\alpha_{43}^d)(1-p_{3*,i})\\
&+(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)p_{4*,i}\Big\}^2\\
&+\frac{\{-\alpha_{22}^d p_{2*,i}+\alpha_{23}^d(1-p_{3*,i})-\alpha_{24}^d p_{4*,i}\}^2}{p_{2*,i}^w}\\
&+\frac{\{-\alpha_{32}^d p_{2*,i}+\alpha_{33}^d(1-p_{3*,i})-\alpha_{34}^d p_{4*,i}\}^2}{p_{3*,i}^w}\\
&+\frac{\{-\alpha_{42}^d p_{2*,i}+\alpha_{43}^d(1-p_{3*,i})-\alpha_{44}^d p_{4*,i}\}^2}{p_{4*,i}^w}\Big]p_{3*,i}^2\begin{pmatrix}1\\ X_i^*\end{pmatrix}^{\otimes 2},
\end{aligned}
$$

$$
\begin{aligned}
A_9 \;=\; &\sum_{i=1}^n\Big[\frac{1}{p_{1*,i}^w}\Big\{(\alpha_{22}^d+\alpha_{32}^d+\alpha_{42}^d)p_{2*,i}-(\alpha_{23}^d+\alpha_{33}^d+\alpha_{43}^d)(1-p_{3*,i})\\
&+(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)p_{4*,i}\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&\times\Big\{(\alpha_{22}^d+\alpha_{32}^d+\alpha_{42}^d)p_{2*,i}+(\alpha_{23}^d+\alpha_{33}^d+\alpha_{43}^d)p_{3*,i}\\
&-(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)(1-p_{4*,i})\Big\}\\
&+\frac{\{-\alpha_{22}^d p_{2*,i}+\alpha_{23}^d(1-p_{3*,i})-\alpha_{24}^d p_{4*,i}\}\{-\alpha_{22}^d p_{2*,i}-\alpha_{23}^d p_{3*,i}+\alpha_{24}^d(1-p_{4*,i})\}}{p_{2*,i}^w}\\
&+\frac{\{-\alpha_{32}^d p_{2*,i}+\alpha_{33}^d(1-p_{3*,i})-\alpha_{34}^d p_{4*,i}\}\{-\alpha_{32}^d p_{2*,i}-\alpha_{33}^d p_{3*,i}+\alpha_{34}^d(1-p_{4*,i})\}}{p_{3*,i}^w}\\
&+\frac{\{-\alpha_{42}^d p_{2*,i}+\alpha_{43}^d(1-p_{3*,i})-\alpha_{44}^d p_{4*,i}\}\{-\alpha_{42}^d p_{2*,i}-\alpha_{43}^d p_{3*,i}+\alpha_{44}^d(1-p_{4*,i})\}}{p_{4*,i}^w}\Big]\\
&\times p_{3*,i}p_{4*,i}\begin{pmatrix}1\\ X_i^*\end{pmatrix}^{\otimes 2},
\end{aligned}
$$

$$
\begin{aligned}
A_{10} \;=\; &\sum_{i=1}^{n}\Big[\frac{1}{p_{1*,i}^w}\Big\{(\alpha_{22}^d+\alpha_{32}^d+\alpha_{42}^d)p_{2*,i}+(\alpha_{23}^d+\alpha_{33}^d+\alpha_{43}^d)p_{3*,i}\\
&-(\alpha_{24}^d+\alpha_{34}^d+\alpha_{44}^d)(1-p_{4*,i})\Big\}^2\\
&+\frac{\{-\alpha_{22}^d p_{2*,i}-\alpha_{23}^d p_{3*,i}+\alpha_{24}^d(1-p_{4*,i})\}^2}{p_{2*,i}^w}\\
&+\frac{\{-\alpha_{32}^d p_{2*,i}-\alpha_{33}^d p_{3*,i}+\alpha_{34}^d(1-p_{4*,i})\}^2}{p_{3*,i}^w}\\
&+\frac{\{-\alpha_{42}^d p_{2*,i}-\alpha_{43}^d p_{3*,i}+\alpha_{44}^d(1-p_{4*,i})\}^2}{p_{4*,i}^w}\Big]p_{4*,i}^2\begin{pmatrix}1\\ X_i^*\end{pmatrix}^{\otimes 2}.
\end{aligned}
$$

Note that when $\alpha_{41}=\alpha_{42}=\alpha_{43}=\alpha_{44}$, $\alpha_{31}=\alpha_{32}=\alpha_{33}=\alpha_{34}$, $\alpha_{21}=\alpha_{22}=\alpha_{23}=\alpha_{24}$, the information matrix is singular. Thus, θ is not a regular point for the information matrix. Now, assume that the misclassification matrix is diagonally dominant. That means $\alpha_{11}>0.5$, $\alpha_{22}>0.5$, $\alpha_{33}>0.5$, and $\alpha_{44}>0.5$. Under this

condition,

$$\boldsymbol{\psi}(\theta) \;=\; \begin{bmatrix} -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & \mathbf{0}^T & \mathbf{0}^T \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0}^T & \mathbf{0}^T \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \mathbf{0}^T & \mathbf{0}^T \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \mathbf{0}^T & \mathbf{0}^T \end{bmatrix}.$$

Observe that in the constrained space, θ is a regular point for $I(\theta)$ and $\boldsymbol{\psi}(\theta)$. Now, the rank of $V(\theta)$ is d_θ. Hence, the parameters are identifiable.

So, what I have proved is that θ is identifiable under the strictly diagonal dominance constraint on the misclassification probability matrix. This has been proved rigorously for the two categories, three categories, and four categories of X as well for W. Now, let us revisit the definition of observational equivalence given in (3.4). Clearly, two θ and θ^* cannot be observationally equivalent under (C1). The question remains: is it possible to find (θ, β) and (θ, β^*) that are observationally equivalent? First, consider the case when both X and W have two categories. Note that (θ, β) and (θ, β^*) are observationally equivalent implies

$$\text{pr}(Y = 1, W = 2|X^*, Z; \theta, \beta) = \text{pr}(Y = 1, W = 2|X^*, Z; \theta, \beta^*), \tag{3.6}$$

for every X^* and Z. Since,

$$\begin{aligned} & \text{pr}(Y = 1, W = 2|X^*, Z; \theta, \beta) \\ = \;& \text{pr}(Y = 1|X = 1, X^*, Z; \beta)\text{pr}(W = 2|X = 1, X^*, Z, \theta) \\ & + \{\text{pr}(Y = 1|X = 2, X^*, Z; \beta)\text{pr}(W = 2|X = 2, X^*, Z, \theta) \\ & - \text{pr}(Y = 1|X = 1, X^*, Z; \beta)\text{pr}(W = 2|X = 1, X^*, Z, \theta)\}\text{pr}(X = 2|X^*, Z; \theta) \end{aligned}$$

$$
\begin{aligned}
= \quad & \mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z,\theta) \\
& +\{\mathrm{pr}(Y=1|X=2,Z;\beta)\mathrm{pr}(W=2|X=2,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z,\theta)\}\mathrm{pr}(X=2|X^*,Z;\theta),
\end{aligned}
$$

(3.6) implies that

$$
\begin{aligned}
0 \quad = \quad & \mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta^*)\mathrm{pr}(W=2|X=1,Z,\theta) \\
& +\Big[\{\mathrm{pr}(Y=1|X=2,Z;\beta)\mathrm{pr}(W=2|X=2,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z,\theta)\} \\
& -\{\mathrm{pr}(Y=1|X=2,Z;\beta^*)\mathrm{pr}(W=2|X=2,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta^*)\mathrm{pr}(W=2|X=1,Z,\theta)\}\Big] \\
& \times\mathrm{pr}(X=2|X^*,Z;\theta) \qquad (3.7)
\end{aligned}
$$

for every X^* and Z. Let us take two arbitrary values of X^*, X_1^* and X_2^*. Plugging in these two values in (3.7) results in

$$
\begin{aligned}
0 \quad = \quad & \mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z;\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta^*)\mathrm{pr}(W=2|X=1,Z;\theta) \\
& +\Big[\{\mathrm{pr}(Y=1|X=2,Z;\beta)\mathrm{pr}(W=2|X=2,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z,\theta)\} \\
& -\{\mathrm{pr}(Y=1|X=2,Z;\beta^*)\mathrm{pr}(W=2|X=2,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta^*)\mathrm{pr}(W=2|X=1,Z,\theta)\}\Big] \\
& \times\mathrm{pr}(X=2|X_1^*,Z;\theta) \qquad (3.8)
\end{aligned}
$$

and

$$
\begin{aligned}
0 \;=\;& \mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta^*)\mathrm{pr}(W=2|X=1,Z,\theta) \\
& +\Big[\{\mathrm{pr}(Y=1|X=2,Z;\beta)\mathrm{pr}(W=2|X=2,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z,\theta)\} \\
& -\{\mathrm{pr}(Y=1|X=2,Z;\beta^*)\mathrm{pr}(W=2|X=2,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta^*)\mathrm{pr}(W=2|X=1,Z,\theta)\}\Big] \\
& \times \mathrm{pr}(X=2|X_2^*,Z;\theta). \qquad (3.9)
\end{aligned}
$$

Now, subtracting (3.8) from (3.9), I obtain

$$
\begin{aligned}
0 \;=\;& \Big[\{\mathrm{pr}(Y=1|X=2,Z;\beta)\mathrm{pr}(W=2|X=2,Z,\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z;\theta)\} \\
& -\{\mathrm{pr}(Y=1|X=2,Z;\beta^*)\mathrm{pr}(W=2|X=2,Z;\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta^*)\mathrm{pr}(W=2|X=1,Z;\theta)\}\Big] \\
& \times\{\mathrm{pr}(X=2|X_2^*,Z;\theta)-\mathrm{pr}(X=2|X_1^*,Z;\theta)\} \qquad (3.10)
\end{aligned}
$$

for every Z. Now, $\{\mathrm{pr}(X=2|X_2^*,Z;\theta)-\mathrm{pr}(X=2|X_1^*,Z;\theta)\}$ cannot be zero for every Z and any two arbitrary choices X_1^* and X_2^*. Hence, (3.10) implies that

$$
\begin{aligned}
0 \;=\;& \Big[\{\mathrm{pr}(Y=1|X=2,Z;\beta)\mathrm{pr}(W=2|X=2,Z;\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta)\mathrm{pr}(W=2|X=1,Z;\theta)\} \\
& -\{\mathrm{pr}(Y=1|X=2,Z;\beta^*)\mathrm{pr}(W=2|X=2,Z;\theta) \\
& -\mathrm{pr}(Y=1|X=1,Z;\beta^*)\mathrm{pr}(W=2|X=1,Z;\theta)\}\Big] \qquad (3.11)
\end{aligned}
$$

for every Z. So, (3.11) and (3.7) together imply that $\text{pr}(Y = 1|X = 1, Z; \beta)\text{pr}(W = 2|X = 1, Z, \theta) - \text{pr}(Y = 1|X = 1, Z; \beta^*)\text{pr}(W = 2|X = 1, Z, \theta) = 0$ for every Z. Summing up the above arguments, I say (3.7) holds if the two equations

$$\begin{aligned}
&\text{pr}(Y = 1|X = 1, Z; \beta)\text{pr}(W = 2|X = 1, Z, \theta) \\
&-\text{pr}(Y = 1|X = 1, Z; \beta^*)\text{pr}(W = 2|X = 1, Z, \theta) = 0, \\
&\text{pr}(Y = 1|X = 2, Z; \beta)\text{pr}(W = 2|X = 2, Z, \theta) \\
&-\text{pr}(Y = 1|X = 2, Z; \beta^*)\text{pr}(W = 2|X = 2, Z, \theta) = 0
\end{aligned} \tag{3.12}$$

hold for every Z. That means $\text{pr}(Y = 1|X = 1, Z; \beta) = \text{pr}(Y = 1|X = 1, Z; \beta^*)$ and $\text{pr}(Y = 1|X = 2, Z; \beta) = \text{pr}(Y = 1|X = 2, Z; \beta^*)$ for every Z. But this cannot hold because the model for Y given X and Z is identifiable. Hence, the assumption that (θ, β) and (θ, β^*) are observationally equivalent cannot be true. It is seen that for proving result (3.12) from (3.7) it is critical that X^* is independent of 1) Y conditional on X and Z and 2) W conditional on X.

Now, I generalize this idea for the case when both X and W have r categories. In this case,

$$\begin{aligned}
&\text{pr}(Y = 1, W = 2|X^*, Z; \theta, \beta) \\
= \;&\text{pr}(Y = 1|X = 1, Z; \beta)\text{pr}(W = 2|X = 1, Z, \theta) \\
&+\{\text{pr}(Y = 1|X = 2, Z; \beta)\text{pr}(W = 2|X = 2, Z, \theta) \\
&-\text{pr}(Y = 1|X = 1, Z; \beta)\text{pr}(W = 2|X = 1, Z, \theta)\}\text{pr}(X = 2|X^*, Z; \theta) \\
&+\cdots+\{\text{pr}(Y = 1|X = r, Z; \beta)\text{pr}(W = 2|X = r, Z, \theta) \\
&-\text{pr}(Y = 1|X = 1, Z; \beta)\text{pr}(W = 2|X = 1, Z, \theta)\}\text{pr}(X = r|X^*, Z; \theta).
\end{aligned}$$

The observationally equivalent relation $\text{pr}(Y = 1, W = 2|X^*, Z; \theta, \beta) = \text{pr}(Y = 1, W = 2|X^*, Z; \theta, \beta^*)$ implies

$$\begin{aligned}&\text{pr}(Y = 1|X = j, Z; \beta)\text{pr}(W = 2|X = j, Z, \theta)\\ &-\text{pr}(Y = 1|X = j, Z; \beta^*)\text{pr}(W = 2|X = j, Z, \theta) = 0,\end{aligned}$$

for $j = 1, \ldots, r$ and every Z, that in turn implies $\text{pr}(Y = 1|X = j, Z; \beta) - \text{pr}(Y = 1|X = j, Z; \beta^*)$ for $j = 1, \ldots, r$ and every Z. This is in contradiction with the fact that the assumed model for Y given X and Z is identifiable. Hence, model parameters (θ, β) are identifiable under constraint (C1).

3.3 Inference

3.3.1 Bayesian inference

To estimate the model parameters I used a Bayesian procedure. For a Bayesian inference I need two important things, the likelihood of the data that stems from the assumed probability models, and the prior distribution of the parameters. In order to integrate the diagonal dominance constraint I write

$$\alpha_{i,j} = \frac{0.5\exp(\eta_{i,j})}{1 + \sum_{s\neq j}\exp(\eta_{s,j})},\ i \neq j, \tag{3.13}$$

and

$$\alpha_{i,i} = 1 - \sum_{s\neq i}\alpha_{s,i},\ i = 1, \ldots, r. \tag{3.14}$$

The η-parameters are in the real line. Define θ_t as θ with α-parameters replaced by η's. Now the likelihood of the parameter (θ_t, β) given the observed data D =

$\{(W_i, X_i^*, Y_i, Z_i), i = 1, \ldots, n\}$ is

$$\begin{aligned}\mathcal{L}(\theta_t, \beta) &= \prod_{i=1}^{n}\prod_{r'=1}^{r}\Big[\{\mathrm{pr}(Y=0, W=r'|X_i^*, Z_i; \theta_t, \beta)\}^{1-Y_i} \\ &\quad \times\{\mathrm{pr}(Y=1, W=r'|X_i^*, Z_i; \theta_t, \beta)\}^{Y_i}\Big]^{I(W_i=r')}, \end{aligned} \tag{3.15}$$

where for $s = 0$ and 1

$$\begin{aligned}\mathrm{pr}(Y=s, W=r'|X_i^*, Z_i; \theta_t, \beta) &= \sum_x \mathrm{pr}(Y=s|X=x, Z_i; \beta)\mathrm{pr}(W=r'|X=x; \theta_t) \\ &\quad \times\mathrm{pr}(X=x|X_i^*, Z_i; \theta_t).\end{aligned}$$

The expression of $\mathrm{pr}(Y = s|X = x, Z_i; \beta)$, $\mathrm{pr}(X = x|X_i^*, Z_i; \theta_t)$, and $\mathrm{pr}(W = r'|X = x; \theta_t)$ are given in (3.1), (3.2), (3.3), respectively. Suppose that $\pi(\beta)$ and $\pi(\theta_t)$ are the prior distribution on the parameters β and θ_t, respectively. In the standard Bayesian inference, Markov Chain Monte Carlo technique is applied, where each parameter is sampled from its full conditional distribution. When this chain is for a large number of times, the sampled values of the parameters can be considered to be sample from the posterior distribution of the parameters. Then any statistic of the posterior distributions can be estimated with sufficient accuracy. Although this technique is widely used, the computation time is a big challenge, specially when the conditional distributions are not the standard distribution from where samples can be easily drawn. In our specific case, all the conditional distributions are non-standard distributions. Sampling from them requires implementation of the Metropolis-Hastings algorithm with a suitable proposal distribution.

3.3.2 Automatic Differentiation Variational Inference (ADVI)

To avoid slow MCMC procedure one can apply the variational inference (VI) procedure. Although the VI procedure is technically less accurate than the MCMC based approach, it produces fast results. Particularly, in the VI inference, the posterior distributed is approximated by a parametric model. First, a family of parametric models is assumed to approximate the posterior distribution. Then its parameters are estimated by minimizing the Kullback-Lieblier distance between the assumed parametric model and the actual posterior distribution. Thus, the problem reduces to an optimization problem. Though VI procedure is quite fast compared to the MCMC based process, it requires model specific derivations and implementation through computer coding. Therefore, I use the automatic differentiation VI method of [38] that uses a scalable variational inference algorithms. Here is a brief description of the ADVI.

First, I transform all the parameter θ, β into unconstrained real parameters, call them $\boldsymbol{\zeta}$. Now, suppose $\pi(\boldsymbol{\zeta}|\mathrm{D})$, the posterior distribution of $\boldsymbol{\zeta}$, will be approximated by a parametric density $f_p(\boldsymbol{\zeta};\vartheta)$. Here ϑ denotes the set of parameters. Then the task is to estimate ϑ by minimizing the Kullback-Liebler divergence between $f_p(\boldsymbol{\zeta};\vartheta)$ and $\pi(\boldsymbol{\zeta}|\mathrm{D})$. The Kullback-Liebler divergence between the two densities is

$$KL(f_p(\boldsymbol{\zeta};\vartheta), \pi(\boldsymbol{\zeta}|\mathrm{D})) = \int \log\left\{\frac{f_p(\boldsymbol{\zeta};\vartheta)}{\pi(\boldsymbol{\zeta}|\mathrm{D})}\right\} f_p(\boldsymbol{\zeta};\vartheta) d\boldsymbol{\zeta}.$$

Let

$$\vartheta^* = \mathrm{argmin}_\vartheta KL(f_p(\boldsymbol{\zeta};\vartheta), \pi(\boldsymbol{\zeta}|\mathrm{D})).$$

Then

$$
\begin{aligned}
\vartheta^* &= \text{argmin}_\vartheta KL(f_p(\boldsymbol{\zeta};\vartheta), \pi(\boldsymbol{\zeta}|\text{D})) \\
&= \text{argmin}_\vartheta KL(f_p(\boldsymbol{\zeta};\vartheta), \pi(\boldsymbol{\zeta})\pi(\text{D}|\boldsymbol{\zeta})) \\
&= \text{argmax}_\vartheta \Big\{ -KL(f_p(\boldsymbol{\zeta};\vartheta), \pi(\boldsymbol{\zeta})\pi(\text{D}|\boldsymbol{\zeta})) \Big\} \\
&= \text{argmax}_\vartheta \int \Big[\log\{\pi(\boldsymbol{\zeta})\} + \log\{\pi(\text{D}|\boldsymbol{\zeta})\} - \log\{f_p(\boldsymbol{\zeta};\vartheta)\} \Big] f_p(\boldsymbol{\zeta};\vartheta) d\boldsymbol{\zeta} \\
&= \text{argmax}_\vartheta \Big(E_{f_p}[\log\{\pi(\boldsymbol{\zeta})\}] + E_{f_p}[\log\{\pi(\text{D}|\boldsymbol{\zeta})\}] - E_{f_p}[\log\{f_p(\boldsymbol{\zeta};\vartheta)\}] \Big) \\
&= \text{argmax}_\vartheta \Big(E_{f_p}[\log\{\pi(\text{D}|\boldsymbol{\zeta})\}] - KL[\log\{f_p(\boldsymbol{\zeta};\vartheta)\}, \pi(\boldsymbol{\zeta})] \Big).
\end{aligned}
$$

Note that $E_{f_p}[\log\{\pi(\text{D}|\boldsymbol{\zeta})\}] - KL[\log\{f_p(\boldsymbol{\zeta};\vartheta)\}, \pi(\boldsymbol{\zeta})]$ is called the evidence lower bound (ELBO). The first term $E_{f_p}[\log\{\pi(\text{D}|\boldsymbol{\zeta})\}]$ is the expected log-likelihood with respect to $f_p(\boldsymbol{\zeta};\vartheta)$, while the second term $KL[\log\{f_p(\boldsymbol{\zeta};\vartheta)\}, \pi(\boldsymbol{\zeta})]$ is the Kullback-Leibler divergence between the approximating posterior and the prior distribution.

In the ADVI approach, the support of the latent variables is first automatically transformed into the real coordinate space. Second, the ELBO is computed using Monte Carlo (MC) integration, which only requires being able to sample from the variational distribution. Third, stochastic gradient ascent is used to maximize the ELBO and automatic differentiation is used to compute gradients without any user input. Thus, I just need to input the model, parameters, and priors. Then one of the following two approaches needs to be applied: 1) the mean-field approach, where the posterior distribution of the transformed parameters assumed to follow independent normal distributions, and 2) full-rank approach where the posterior distribution of the transformed parameters are assumed to follow a multivariate normal distribution. [37] has provided a STAN package to implement this procedure. In order to apply ADVI, the main assumption is that the probability model is differentiable, which

means that the probability model is a continuous and differentiable function of the transformed variable $\boldsymbol{\zeta}$ over the Euclidian space.

3.3.3 ADVI algorithm

Since the parameters in $\boldsymbol{\zeta}$ has support on the real line, this means that transforming the latent parameters is not required and hence I can skip the first step. In the second step I assume the full rank Gaussian approximation, that means $f_p(\boldsymbol{\zeta};\vartheta) \equiv \text{MN}(\boldsymbol{\zeta};\mu, LL^T)$, where L is a lower triangular matrix of real valued entries, and MN stands for multivariate normal. Here ϑ is a vector containing μ and entries of L. Note that μ and LL^T represent the posterior mean and the variance-covariate matrix of the joint posterior distribution of the transformed parameter vector $\boldsymbol{\zeta}$.

Next I apply the elliptical standardization to convert the variational factors to standard normals using the transformation $\xi = S_\vartheta(\boldsymbol{\zeta}) = L^{-1}(\boldsymbol{\zeta} - \mu)$. Hence the solution to transformed objective function is now defined as

$$\vartheta^* \quad = \quad \underset{\vartheta}{\text{argmax}}\mathcal{L}(\vartheta) = \underset{\varphi}{\text{argmax}}\mathbb{E}_{MN(\xi;0,I)}\left[\log\{p(\text{D}, S_\vartheta^{-1}(\xi))\}\right] - E_{f_p(\boldsymbol{\zeta};\vartheta)}[\log\{f_p(\boldsymbol{\zeta};\vartheta)\}],$$

where $S_\vartheta^{-1}(\xi) = L\xi + \mu = \boldsymbol{\zeta}$. In our specific case, $\boldsymbol{\zeta} = (\beta^T, \theta_t^T)^T$ and $p(\text{D}, \boldsymbol{\zeta}) = \mathcal{L}(\theta_t, \beta)$ given in (3.15).

Therefore, the gradients required for optimization are defined as

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}}\mathcal{L}(\theta_t,\beta) &= \mathbb{E}_{MN(\xi;0,I)}\left\{\nabla_{\boldsymbol{\zeta}}\log p(\text{D},\boldsymbol{\zeta})\right\}, \\
\nabla_{\mathbb{L}}\mathcal{L}(\theta_t,\beta) &= \mathbb{E}_{MN(\xi;0,I)}\left\{\nabla_{\boldsymbol{\zeta}}\log p(\text{D},\boldsymbol{\zeta})\right\} + (\mathbf{L}^{-1})^T.
\end{aligned}$$

In particular, for each iteration of the algorithm, I apply the automatic differentiation method to calculate the gradients $\nabla_{\boldsymbol{\zeta}} p(\text{D}, \boldsymbol{\zeta})$ on the inside of the expectations, then

use MC integration to approximate the expectation:

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}(\theta_t, \beta) \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_{\boldsymbol{\zeta}} \log p(\mathrm{D}, \boldsymbol{\zeta}) \tag{3.16}$$

$$\nabla_{\mathbb{L}} \mathcal{L}(\theta_t, \beta) \approx \left(\frac{1}{S} \sum_{s=1}^{S} \nabla_{\boldsymbol{\zeta}} \log p(\mathrm{D}, \boldsymbol{\zeta}) \right) + (\mathbf{L}^{-1})^T, \tag{3.17}$$

where $s = 1, \ldots, S$ represents the sth sample drawn from the distribution of $MN(\xi; 0, I)$. The algorithm is as follows:

Input: Dataset $D = (X^*_{1:n}, W_{1:n}, Y_{1:n}, Z_{1:n})$, model $p(\mathrm{D}, \boldsymbol{\zeta})$

- Set $i = 1$, $\boldsymbol{\mu}^{(1)} = \widehat{\theta} = (\widehat{\beta}, \widehat{\gamma})$, and $\mathbf{L}^{(1)} = \mathbf{I}$.
- Set $\boldsymbol{\zeta}^{(1)} \sim N(0, 2)$.
- While $ELBO(current) - ELBO(previous) > threshold$.
  - Draw M samples $\boldsymbol{\zeta}_m \sim MN(0, I)$.
  - Approximate $\nabla_{\boldsymbol{\mu}} \mathcal{L}(\theta_t, \beta)$ using Equation (3.16).
  - Approximate $\nabla_{\mathbf{L}} \mathcal{L}(\theta_t, \beta)$ using Equation (3.17).
  - Update $\boldsymbol{\mu}^{(i+1)} = \boldsymbol{\mu}^{(i)} + \mathrm{diag}(\boldsymbol{\rho}^{(i)}) \nabla_{\boldsymbol{\mu}} \mathcal{L}(\theta_t, \beta)$, where $\boldsymbol{\rho}^{(i)}$ is the step size whose formula can be found in Kucukelbir et al. (2017).
  - Update $\mathbf{L}^{(i+1)} = \mathbf{L}^{(i)} + \mathrm{diag}(\boldsymbol{\rho}^{(i)}) \nabla_{\mathbf{L}} \mathcal{L}(\theta_t, \beta)$.
  - Set $i = i + 1$.
  - end.
- Return $\boldsymbol{\mu}^* = \boldsymbol{\mu}^{(i)}$.
- Return $\mathbf{L}^* = \mathbf{L}^{(i)}$.

The speed of the algorithm is dependent on the sample size and size of the threshold used to determine convergence. I compared the performance of the ADVI algorithm for estimating the model parameters with the Hamiltonian Monte Carlo (HMC) algorithm, a traditional MCMC based Bayesian approach which is also implemented in RStan. I consider the simulation setup of scenario S1 from 3.3 (detailed in the simulation section). For the ADVI approach, I considered a converegence threshold of 10^{-5}. Because HMC is an MCMC method, I considered 100 and 1000 iterations per chain. As a metric of comparison, I used the average time in minutes to complete ten replications of this scenario.

Table 3.1: Mean time (minutes) to complete 10 replications of Proposed Method using ADVI and HMC
n: Sample Size, ADVI: Automatic Differentiation Variational Inference, HMC: Hamiltonian Monte Carlo, Threshold: Stopping criterion for ADVI method, Iteration: Number of iterations per chain

	ADVI	HMC	
	Threshold	Iteration	
n	10^{-5}	100	1000
5000	24.47	27.51	151.66
10000	51.68	80.70	358.25

The results are shown in Table 3.1. The performance between the ADVI and algorithms is similar when the number of iterations for HMC is small, however as the number of iterations HMC is increased the time to complete ten replications increases exponentially.

3.4 Simulation

Simulation Design: In the simulation study I generated cohorts consisting of n independent observations. The cohort consisted of n iid copies of (X^*, W, Y, Z). In all scenarios, Z was simulated from uniform$(-1, 1)$, and Y was a binary 0-1 response variable. In the first scenario a binary 0-1 instrument X^* was generated from Bernoulli(0.55), and X was generated with success probability $\text{pr}(X = 2|X^*, Z; \gamma) = 1 - \text{pr}(X = 1|X^*, Z; \gamma) = H(\gamma_1 + \gamma_2 X^* + 0.3Z)$, with $\gamma_1 = -1$ and $\gamma_2 = 1$ so that the marginal success probability of $\text{pr}(X = 2)$ was 0.4 and subsequently $\text{pr}(X = 1) = 0.6$. I considered this situation as one where X^* is mildly associated with the true X and refer to this as MA. I also considered a situation where X^* is strongly associated with X through the model $\text{pr}(X = 2|X^*)$ by setting $\gamma_2 = 2$, and it is referred to as SA. The misclassified variable W was generated as follows:

$$W = \begin{cases} 2 & \text{if } B \times I(X = 2) + B^* \times \{1 - I(X = 2)\} = 1, \\ 1 & \text{otherwise,} \end{cases}$$

with $B \sim \text{Bernoulli}(1 - \alpha_{22})$ and $B^* \sim \text{Bernoulli}(1 - \alpha_{11})$.

Here I considered two levels of misclassification: (1) $\alpha_{12} = \alpha_{21} = 0.2$ and (2) $\alpha_{12} = \alpha_{21} = 0.05$. Under the setup where $\alpha_{12} = \alpha_{21} = 0.2$, the marginal probability of $W = 2$ was 44%. Finally, I considered a binary response variable Y which was generated with the success probability

$$\text{pr}(Y = 1|X, Z; \beta) = H\{\beta_0 + \beta_{x,2} I(X = 2) + \beta_z Z\}.$$

with $\beta_z = 0.5$. I set $\beta_0 = -2$ and $\beta_{x,2} = 1$ so that marginally $\text{pr}(Y = 1) = 0.21$. For this scenario I considered sample sizes of $n = 2000$ and $n = 5000$.

In scenario II, I considered W and X as categorical variables with 3 categories. There were three versions of scenario II; scenario IIa is the one in which there is one instrument, while in scenarios IIb and IIc, I considered two instrumental variables. In scenario IIb, both instruments were strongly associated with X while in scenario IIc both instruments were moderately associated with X.

In scenario IIa, the probability for the jth category of X was modeled as

$$\text{pr}(X = j|X^*, Z; \gamma) = \frac{\exp\{\gamma_{j1} + \gamma_{j2} I(X_i^* = 1) + \gamma_{j3} Z\}}{1 + \sum_{j=2}^{3} \exp\{(\gamma_{j1} + \gamma_{j2} I(X_i^* = 1) + \gamma_{j3} Z\}},$$

with $j = 2, 3$ and where $\gamma_{21} = 0.25, \gamma_{22} = 2, \gamma_{23} = 0.3, \gamma_{31} = 0.5, \gamma_{32} = 2.5, \gamma_{33} = -0.3$, and $X^* \sim$ Bernoulli(0.55). In scenario IIb, the probability for the jth category of X was modeled as

$$\text{pr}(X = j|X^*, Z; \gamma) = \frac{\exp\{\gamma_{j1} + \gamma_{j2} I(X_{1i}^* = 2) + \gamma_{j3} Z + \gamma_{j4} X_{i2}^*\}}{1 + \sum_{j=2}^{3} \exp\{(\gamma_{j1} + \gamma_{j2} I(X_{1i}^* = 2) + \gamma_{j3} Z + \gamma_{j4} X_{i2}^*\}},$$

with $j \in (2, 3)$ and where $\gamma_{21} = 0.25, \gamma_{22} = 2, \gamma_{23} = 0.3, \gamma_{24} = 1, \gamma_{31} = 0.5, \gamma_{32} = 2.5, \gamma_{33} = -0.3, \gamma_{34} = -1$, $X_1^* \sim$ Bernoulli(0.55) and $X_2^* \sim$ Uniform$(-1, 1)$.

Finally, in scenario IIc the probability model for the jth category of X was modeled is the same as in scenario IIb, however I took $\gamma_{22} = 0.7$ and $\gamma_{32} = 1$.

In scenarios IIa, IIb, IIc, I calculated $\text{pr}(X = 1|X^*, Z; \gamma) = 1 - \text{pr}(X = 2|X^*, Z; \gamma) - \text{pr}(X = 3|X^*, Z; \gamma)$.

I then generated X from Multinomial$(\text{pr}(X = 1|X^*, Z), \text{pr}(X = 2|X^*, Z; \gamma), \text{pr}(X = 3|X^*, Z; \gamma))$.

The misclassification matrix for the tri-category X and W case was

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix},$$

where the misclassification probabilities were set as $\alpha_{21} = 0.025, \alpha_{31} = 0.015, \alpha_{11} = 0.96$, $\alpha_{22} = 0.85, \alpha_{32} = 0.05, \alpha_{12} = 0.10, \alpha_{23} = 0.1, \alpha_{33} = 0.70, \alpha_{13} = 0.2$. Then to generate W, I sampled η_w from uniform$(0, 1)$ and then set

$$W = \begin{cases} 1, & \text{if } 0 < \eta_W < \text{pr}(W = 1|X = j), \\ 2, & \text{if } \text{pr}(W = 1|X = j) \leq \eta_W < \text{pr}(W \leq 2|X = j), \\ 3, & \text{if } \text{pr}(W \leq 2|X = j) \leq \eta_W, \end{cases}$$

where $\text{pr}(W \leq i|X = j) = \sum_{w=1}^{i} \text{pr}(W = w|X = j)$ represents the cumulative probability. Finally, I generated Y as a binary variable with the success probability $\text{pr}(Y = 1|X, Z) = H(\beta_0 + \beta_{x,2} I(X = 2) + \beta_{x,3} I(X = 3) + \beta_2 Z)$, with $\beta_0 = -2, \beta_1 = 1, \beta_2 = 0.7, \beta_3 = 0.5$.

I considered sample sizes of $n = 5000$ and $n = 10000$. For Table IIc, I also considered an additional sample size of $n = 20000$.

Finally, I considered a simulation, referred to as scenario III, in which I mimicked the black women group of the breast cancer data set analyzed in Section 3.5. Demographic information summarizing this subgroup can be found there.

The instrumental variable X^*, was simulated from the three category multinomial distribution with $\text{pr}(X^* = 1) = 0.74$, $\text{pr}(X^* = 2) = 0.22$, and $\text{pr}(X^* = 3) = 0.04$. There were two binary prognostic factors, Z_1 and Z_2, simulated from Bernoulli

distribution with success probabilities $\text{pr}(Z_1 = 1) = 0.23$ and $\text{pr}(Z_2 = 1) = 0.31$. The model for the true exposure X for the jth category was

$$\text{pr}(X = j|X^*;\gamma) = \frac{\exp\{\gamma_{j1} + \gamma_{j2}I(X^*_{1i} = 2) + \gamma_{j3}X^*_{i2} + \gamma_{j4}Z_1 + \gamma_{j5}Z_2\}}{1 + \sum_{j=2}^{3}\exp\{\gamma_{j1} + \gamma_{j2}I(X^*_{1i} = 2) + \gamma_{j3}X^*_{i2} + \gamma_{j4}Z_1 + \gamma_{j5}Z_2\}},$$

with $j = 2, 3$ and I set $\gamma_{21} = 0.47, \gamma_{22} = 0.80, \gamma_{23} = 0.36, \gamma_{24} = -0.99, \gamma_{25} = -2.00, \gamma_{31} = -0.09, \gamma_{32} = -4.12, \gamma_{33} = 0.87, \gamma_{34} = -0.45, \gamma_{35} = 1.75$. The parameters were the corresponding estimates from the data on black women using the proposed method (case 2).

To generate the misclassified exposure W, I considered the estimated misclassification matrix from the data using the method of case 2:

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.683 & 0.001 & 0.066 \\ 0.221 & 0.500 & 0.270 \\ 0.096 & 0.499 & 0.664 \end{bmatrix},$$

with the elements of the matrix corresponding to $\text{pr}(W = i|X = j)$ for the ith row and j column. For a given level of X that specifies the column of $\boldsymbol{\alpha}$, W was generated by the multinomial distribution with the corresponding column probabilities. Finally, Y was generated as a binary variable with $\text{pr}(Y = 1|X, Z; \beta) = H(\beta_0 + \beta_1 I(X_2 = 1) + \beta_2 I(X_3 = 1) + \beta_3 Z_1 - \beta_4 Z_2)$, with $\beta_0 = 1.13, \beta_1 = 0, \beta_2 = 1.98, \beta_3 = -0.5, \beta_4 = -2.07$. Here I considered the sample size of $n = 10000$. Like the previous scenarios, I considered two versions of the prior standard deviations: (S1) $\sigma = 2$, (S2) $\sigma = 5$.

Method of Analysis: Each data set was analyzed using three methods. First, assuming the true X is observed, I fitted a logistic model to Y with X and Z as the regressor variables, and refer to this method as M1. This is a purely a hypothetical method, because other than the simulation setting the true X is never known.

However, this hypothetical method is included for the sake of comparison. Second, I regressed Y on W and Z in a logistic model, this is referred to as method M2. Note that M2 is the naive method. Third, I considered the proposed method and refer to it as M3. Under all scenarios, I used 200 replications. For methods M1, M2, and M3, I used the ADVI technique with full rank approximation using the program **RStan**. Under M1 and M2, the priors for the parameters were all set as Normal$(0, 2^2)$.

For the proposed method M3, I used the naive estimates as the prior means for the β-parameters. Also, the regression coefficients for the model of W on X^* and Z were used as the prior means for the γ parameters. For the η-parameters prior means were set to zero. For the prior standard deviations, I considered two situations: S1) $\sigma = 2$ and S2) $\sigma = 5$.

For every data set I recorded the posterior mean and standard deviation for each parameter. Then I summarized these results by calculating the median of the posterior means and the median of the posterior standard deviations for each of these parameters. I reported the 95% credible intervals and the median length of the 95% credible intervals for each parameter. Finally, I report the mean square error for each parameter. Finally, I set the threshold of convergence for the ADVI algorithm to 10^{-5}.

Results: Table 3.2 contains the results for Scenario I. Under M1, the bias of the posterior mean for the parameter is negligible for different sample sizes, and the posterior standard deviation decreased with the sample size. Under M2, the estimates tend to be biased regardless of the sample size. For M3, within a fixed sample size, the bias of the estimate and the variance decreased as the strength of association between the instrument γ_2 and the true X increased. Subsequently, the size of the credible intervals also decreased. However, the greatest impact of the change in bias and variance is observed when the level of misclassification is reduced. When considering the effect of an increasing sample size, particularly from $n = 2000$ to $n = 5000$, it is evident that the posterior standard deviations decrease in half for a fixed level of association between the instrument. Also, the bias of M3 decreases with the sample size.

Table 3.2: Results of the simulation study under equal misclassification with 200 generated data sets. MT: method, M: median of posterior means, SD: median of posterior standard deviation $\times 100$, CI: 95% credible interval, CI-L: median length of the 95% credible interval, MSE: Mean Squared Error$\times 100$, MC1: $\alpha_{12} = \alpha_{21} = 0.2$, MC2: $\alpha_{12} = \alpha_{21} = 0.05$, $\text{pr}(X = 1|X^*, Z; \gamma) = H(-1 + \gamma_2 X^* + 0.3Z)$, MA: $\gamma_2 = 1$, SA: $\gamma_2 = 2$, $\text{pr}(Y = 1|X, Z) = H(-2 + X + 0.5Z)$, $\alpha_{21} = \text{pr}(W = 1|X = 0)$, $\alpha_{12} = \text{pr}(W = 0|X = 1)$

		MC1									MC2						
MT		M1			M2			M3			M2			M3			
		β_1	$\beta_{x,2}$	β_z	β_1	$\beta_{x,2}$	β_z	β_1	$\beta_{x,2}$	β_z	β_1	$\beta_{x,2}$	β_z	β_1	$\beta_{x,2}$	β_z
$n = 2000$																	
	M	-2.00	0.99	0.50	-1.82	0.57	0.53	-2.11	1.21	0.48	-1.95	0.88	0.51	-2.12	1.21	0.47
	SD	9.61	12.78	11.19	9.34	12.69	11.01	20.14	26.40	12.07	9.54	12.66	11.08	14.85	19.52	11.55
MA	CI	$(-2.19, -1.82)$	$(0.74, 1.24)$	$(0.28, 0.72)$	$(-2.00, -1.63)$	$(0.33, 0.83)$	$(0.31, 0.74)$	$(-2.62, -1.73)$	$(0.65, 1.78)$	$(0.24, 0.72)$	$(-2.14, -1.77)$	$(0.63, 1.13)$	$(0.29, 0.72)$	$(-2.48, -1.84)$	$(0.84, 1.64)$	$(0.24, 0.70)$
	CI-L	0.38	0.50	0.44	0.37	0.50	0.43	0.79	1.03	0.47	0.37	0.50	0.43	0.58	0.77	0.45
	MSE	0.73	1.45	1.14	4.08	19.30	1.16	10.78	12.90	1.49	0.90	2.77	1.09	6.23	9.84	1.30
	M	-2.00	1.00	0.48	-1.86	0.79	0.50	-2.06	1.06	0.48	-1.93	0.89	0.49	-2.06	1.07	0.48
	SD	10.71	13.00	10.81	10.13	12.70	10.61	14.56	17.40	10.84	10.46	12.85	10.63	13.37	16.19	10.82
SA	CI	$(-2.22, -1.79)$	$(0.74, 1.25)$	$(0.28, 0.71)$	$(-2.06, -1.67)$	$(0.53, 1.03)$	$(0.30, 0.72)$	$(-2.34, -1.76)$	$(0.71, 1.40)$	$(0.28, 0.70)$	$(-2.14, -1.73)$	$(0.63, 1.14)$	$(0.29, 0.71)$	$(-2.33, -1.80)$	$(0.76, 1.40)$	$(0.27, 0.70)$
	CI-L	0.42	0.51	0.42	0.40	0.50	0.42	0.57	0.68	0.43	0.41	0.50	0.42	0.52	0.63	0.42
	MSE	0.89	1.50	1.07	2.55	6.02	1.09	2.03	2.89	1.16	1.29	2.71	1.06	1.99	2.98	1.16
$n = 5000$																	
	M	-2.00	1.00	0.51	-1.81	0.59	0.55	-2.06	1.07	0.51	-1.96	0.89	0.52	-2.08	1.14	0.49
	SD	6.10	8.06	7.10	5.89	7.93	7.03	12.83	16.23	7.49	6.05	7.97	7.12	8.93	12.03	7.22
MA	CI	$(-2.12, -1.88)$	$(0.84, 1.16)$	$(0.37, 0.65)$	$(-1.93, -1.70)$	$(0.43, 0.74)$	$(0.40, 0.68)$	$(-2.36, -1.81)$	$(0.76, 1.43)$	$(0.35, 0.65)$	$(-2.07, -1.83)$	$(0.73, 1.05)$	$(0.38, 0.66)$	$(-2.27, -1.90)$	$(0.89, 1.37)$	$(0.350.64)$
	CI-L	0.24	0.32	0.28	0.23	0.31	0.28	0.50	0.64	0.29	0.24	0.31	0.28	0.35	0.47	0.28
	MSE	0.31	0.51	0.50	3.74	17.60	0.65	4.21	4.17	0.57	0.56	1.81	0.50	1.72	3.21	0.52
	M	-2.00	1.00	0.50	-1.85	0.78	0.51	-2.01	1.03	0.50	-1.92	0.90	0.51	-2.03	1.05	0.50
	SD	6.80	8.25	6.82	6.41	7.98	6.77	9.40	11.13	6.86	6.66	8.16	6.80	8.52	10.29	6.81
SA	CI	$(-2.13, -1.86)$	$(0.83, 1.16)$	$(0.37, 0.64)$	$(-1.98, -1.72)$	$(0.62, 0.93)$	$(0.39, 0.65)$	$(-2.20, -1.83)$	$(0.80, 1.24)$	$(0.37, 0.64)$	$(-2.05, -1.79)$	$(0.72, 1.04)$	$(0.38, 0.65)$	$(-2.20, -1.86)$	$(0.84, 1.24)$	$(0.37, 0.64)$
	CI-L	0.27	0.32	0.27	0.25	0.31	0.27	0.37	0.44	0.27	0.26	0.32	0.27	0.33	0.40	0.27
	MSE	0.41	0.57	0.45	2.59	5.57	0.47	0.98	1.20	0.46	1.06	1.93	0.47	0.77	1.09	0.47

The results for scenario IIa are given in Table 3.3. For all sample sizes, I find that my approach performs better than using the naive estimator in terms of bias reduction. The advantage of M3 is clearly seen as the sample size increases. The results for M3 show that the estimator for the regression coefficient for Z is biased, and the bias seems to decrease with the sample size. It is evident that the MSE is slightly smaller under S1 when the prior $\sigma = 2$.

The results for scenario IIb are given in Table 3.4. As in scenario IIa, there are similar results - M3 performs better than M2 in terms of bias reduction especially as the sample size increases. Moreover, the inclusion of the second instrument appears to reduce the bias greatly, particularly for $\beta_{x,3}$.

The results for scenario IIc are given in Table 3.5. When the instruments are moderately associated with X, M3 performs better than M2 in terms of bias and MSE when the sample size is increased.

Table 3.3: Results of the simulation study for three categories, one instrument

MT: method, M: median of posterior mean, SD: median of posterior standard deviation ×100, CI: 95% credible interval, CI-L: Median length of 95% credible interval, MSE: Mean Squared Error×100, $\text{pr}(Y = 1|X, Z) = H\{-2 + I(X = 2) + 0.7I(X = 3) + 0.5Z\}$, S1: $\sigma = 2$, S2: $\sigma = 5$

MT		M1				M2				M3			
		β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2
$n = 5000$													
	M	−2.00	0.99	0.70	0.50	−1.54	0.50	0.24	0.51	−2.02	1.27	0.61	0.39
	SD	12.91	14.20	13.81	6.52	7.80	9.75	9.74	6.52	21.61	27.29	24.73	8.93
S1	CI	(−2.25, −1.75)	(0.72, 1.27)	(0.42, 0.96)	(0.37, 0.63)	(−1.70, −1.39)	(0.30, 0.68)	(0.06, 0.44)	(0.38, 0.63)	(−2.46, −1.58)	(0.73, 1.84)	(0.12, 1.12)	(0.21, 0.56)
	CI-L	0.51	0.56	0.54	0.26	0.31	0.38	0.38	0.26	0.85	1.07	0.97	0.35
	MSE	1.36	1.78	1.51	0.36	21.08	27.19	21.20	0.37	5.18	15.68	8.28	2.10
	M	−2.01	1.00	0.70	0.50	−1.54	0.50	0.25	0.51	−2.09	1.34	0.65	0.39
	SD	13.05	14.34	13.98	6.54	7.76	9.76	9.79	6.51	24.02	29.62	26.74	9.03
S2	CI	(−2.27, −1.76)	(0.73, 1.29)	(0.43, 0.98)	(0.37, 0.63)	(−1.70, −1.40)	(0.30, 0.68)	(0.06, 0.44)	(0.38, 0.64)	(−2.60, −1.61)	(0.77, 1.98)	(0.13, 1.24)	(0.19, 0.56)
	CI-L	0.51	0.56	0.55	0.26	0.30	0.38	0.38	0.26	0.94	1.16	1.05	0.35
	MSE	1.42	1.85	1.55	0.36	20.89	26.92	20.89	0.36	9.25	24.40	11.00	2.53
$n = 10000$													
	M	−2.01	1.00	0.70	0.50	−1.55	0.50	0.28	0.51	−1.99	1.21	0.63	0.42
	SD	9.18	10.07	9.84	4.67	5.52	6.91	6.93	4.64	15.74	19.59	17.78	6.03
S1	CI	(−2.18, −1.82)	(0.80, 1.19)	(0.51, 0.89)	(0.41, 0.59)	(−1.66, −1.44)	(0.36, 0.63)	(0.13, 0.40)	(0.42, 0.60)	(−2.33, −1.70)	(0.82, 1.61)	(0.29, 1.00)	(0.30, 0.54)
	CI-L	0.36	0.39	0.39	0.18	0.22	0.27	0.27	0.18	0.62	0.77	0.70	0.24
	MSE	0.73	0.87	0.86	0.19	20.35	26.14	19.22	0.19	2.63	8.55	3.82	0.98
	M	−2.01	1.00	0.71	0.50	−1.56	0.50	0.28	0.51	−2.03	1.26	0.67	0.42
	SD	9.24	10.20	9.86	4.67	5.51	6.87	6.91	4.66	16.75	20.44	18.65	6.08
S2	CI	(−2.19, −1.83)	(0.80, 1.20)	(0.51, 0.90)	(0.41, 0.59)	(−1.66, −1.44)	(0.36, 0.63)	(0.13, 0.40)	(0.42, 0.60)	(−2.40, −1.72)	(0.85, 1.67)	(0.30, 1.05)	(0.29, 0.54)
	CI-L	0.36	0.40	0.39	0.18	0.22	0.27	0.27	0.18	0.66	0.80	0.73	0.24
	MSE	0.74	0.87	0.86	0.19	20.28	26.00	19.15	0.20	3.81	11.44	4.31	1.10

Table 3.4: Results of the simulation study for three categories, two instrument with strong association
MT: method, M: median of posterior mean SD: median of posterior standard deviation ×100,CI: 95% credible interval, CI-L: Median length of 95% credible interval, MSE: Mean Squared Error×100, $\text{pr}(Y = 1|X, Z) = H\{-2 + I(X = 2) + 0.7I(X = 3) + 0.5Z\}$, S1: $\sigma = 2$, S2: $\sigma = 5$

MT		M1				M2				M3			
		β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2
$n = 5000$													
	M	−1.98	0.98	0.68	0.51	−1.51	0.47	0.23	0.51	−1.99	1.02	0.70	0.50
	SD	13.63	14.74	14.49	6.47	7.87	9.69	9.83	6.46	21.83	23.94	23.79	6.62
S1	CI	(−2.26, −1.72)	(0.70, 1.28)	(0.40, 0.97)	(0.38, 0.63)	(−1.67, −1.36)	(0.27, 0.65)	(0.03, 0.42)	(0.38, 0.64)	(−2.46, −1.57)	(0.57, 1.52)	(0.24, 1.20)	(0.37, 0.63)
	CI-L	0.53	0.58	0.57	0.25	0.31	0.38	0.39	0.25	0.86	0.94	0.93	0.26
	MSE	1.44	1.65	1.63	0.34	24.13	29.54	23.73	0.35	5.16	5.79	5.83	0.35
	M	−1.99	0.99	0.69	0.50	−1.51	0.47	0.23	0.51	−2.07	1.07	0.77	0.50
	SD	13.79	14.88	14.57	6.50	7.86	9.76	9.92	6.49	24.00	25.83	25.85	6.55
S2	CI	(−2.27, −1.73)	(0.71, 1.29)	(0.41, 0.99)	(0.38, 0.63)	(−1.67, −1.36)	(0.28, 0.66)	(0.03, 0.42)	(0.38, 0.64)	(−2.58, −1.60)	(0.58, 1.63)	(0.26, 1.32)	(0.37, 0.63)
	CI-L	0.54	0.58	0.57	0.25	0.31	0.38	0.39	0.25	0.94	1.01	1.01	0.26
	MSE	1.47	1.71	1.67	0.34	23.89	29.12	23.52	0.36	8.03	8.69	8.46	0.35
$n = 10000$													
	M	−2.01	1.01	0.71	0.50	−1.53	0.48	0.25	0.50	−2.02	1.04	0.73	0.50
	SD	9.71	10.46	10.38	4.65	5.65	6.89	6.94	4.62	16.28	17.38	17.62	4.62
S1	CI	(−2.20, −1.82)	(0.80, 1.21)	(0.51, 0.91)	(0.40, 0.59)	(−1.64, −1.42)	(0.35, 0.61)	(0.11, 0.38)	(0.41, 0.59)	(−2.36, −1.71)	(0.71, 1.40)	(0.40, 1.10)	(0.40, 0.58)
	CI-L	0.38	0.41	0.41	0.18	0.22	0.27	0.27	0.18	0.64	0.68	0.69	0.18
	MSE	0.77	0.84	0.90	0.19	22.18	27.37	20.89	0.19	2.91	3.32	3.37	0.19
	M	−2.01	1.01	0.72	0.50	−1.53	0.48	0.25	0.51	−2.07	1.07	0.78	0.50
	SD	9.82	10.57	10.39	4.64	5.65	6.88	6.95	4.65	17.35	18.48	18.58	4.65
S2	CI	(−2.21, −1.82)	(0.81, 1.22)	(0.51, 0.92)	(0.40, 0.59)	(−1.64, −1.42)	(0.35, 0.62)	(0.11, 0.39)	(0.41, 0.59)	(−2.43, −1.73)	(0.72, 1.46)	(0.42, 1.16)	(0.40, 0.59)
	CI-L	0.38	0.41	0.41	0.18	0.22	0.27	0.27	0.18	0.68	0.72	0.73	0.18
	MSE	0.80	0.87	0.93	0.19	22.11	27.28	20.78	0.18	4.17	4.51	4.60	0.19

Table 3.5: Results of the simulation study for three categories, two instrument with moderate association MT: method, M: median of posterior mean SD: median of posterior standard deviation ×100,CI: 95% credible interval, CI-L: Median length of 95% credible interval, MSE: Mean Squared Error×100, $\text{pr}(Y = 1|X, Z) = H\{-2 + I(X = 2) + 0.7I(X = 3) + 0.5Z\}$, $S1 : \sigma = 2$, $S2 : \sigma = 5$

MT		M1				M2				M3			
		β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2
$n = 5000$													
	M	−2.00	0.99	0.70	0.50	−1.59	0.54	0.30	0.50	−2.39	1.43	1.11	0.50
	SD	11.57	13.00	12.95	6.59	7.58	9.59	9.83	6.58	36.84	37.81	38.23	6.82
S1	CI	(−2.24, −1.78)	(0.75, 1.25)	(0.45, 0.96)	(0.37, 0.63)	(−1.74, −1.44)	(0.35, 0.72)	(0.11, 0.49)	(0.38, 0.64)	(−3.13, −1.71)	(0.70, 2.18)	(0.38, 1.89)	(0.37, 0.64)
	CI-L	0.45	0.51	0.51	0.26	0.30	0.38	0.39	0.26	1.44	1.48	1.50	0.27
	MSE	1.04	1.38	1.19	0.38	17.35	22.17	16.73	0.39	33.75	35.50	35.46	0.40
	M	−2.00	0.99	0.69	0.50	−1.58	0.54	0.30	0.50	−2.28	1.32	1.02	0.50
	SD	11.59	12.97	12.69	6.56	7.58	9.51	9.83	6.58	31.12	32.58	33.17	6.83
S2	CI	(−2.23, −1.77)	(0.74, 1.24)	(0.45, 0.95)	(0.37, 0.63)	(−1.74, −1.44)	(0.35, 0.72)	(0.11, 0.49)	(0.38, 0.64)	(−2.90, −1.68)	(0.68, 1.96)	(0.36, 1.67)	(0.37, 0.63)
	CI-L	0.45	0.51	0.50	0.26	0.30	0.37	0.39	0.26	1.22	1.28	1.30	0.27
	MSE	1.00	1.35	1.14	0.37	17.54	22.30	16.86	0.40	17.69	20.14	19.97	0.40
$n = 10000$													
	M	−2.01	1.01	0.70	0.50	−1.60	0.55	0.31	0.51	−2.26	1.26	0.97	0.50
	SD	8.27	9.20	9.16	4.68	5.38	6.76	7.00	4.65	23.90	24.48	25.14	4.75
S1	CI	(−2.17, −1.85)	(0.83, 1.19)	(0.53, 0.89)	(0.41, 0.60)	(−1.70, −1.49)	(0.42, 0.68)	(0.18, 0.45)	(0.42, 0.60)	(−2.75, −1.80)	(0.80, 1.78)	(0.49, 1.49)	(0.41, 0.60)
	CI-L	0.32	0.36	0.36	0.18	0.21	0.27	0.27	0.18	0.94	0.96	0.99	0.19
	MSE	0.49	0.61	0.59	0.20	16.30	20.44	15.34	0.20	14.39	15.39	15.54	0.22
	M	−2.01	1.01	0.70	0.50	−1.60	0.55	0.31	0.51	−2.20	1.22	0.91	0.50
	SD	8.24	9.16	9.11	4.67	5.37	6.77	6.91	4.67	21.80	22.53	23.04	4.77
S2	CI	(−2.16, −1.84)	(0.82, 1.18)	(0.52, 0.88)	(0.41, 0.60)	(−1.70, −1.49)	(0.42, 0.68)	(0.18, 0.45)	(0.42, 0.60)	(−2.65, −1.78)	(0.79, 1.69)	(0.47, 1.40)	(0.41, 0.60)
	CI-L	0.32	0.36	0.36	0.18	0.21	0.27	0.27	0.18	0.85	0.88	0.90	0.19
	MSE	0.49	0.59	0.57	0.20	16.41	20.61	15.39	0.20	9.12	10.64	10.54	0.22
$n = 20000$													
	M	−2.00	1.00	0.69	0.49	−1.59	0.55	0.30	0.50	−2.10	1.13	0.81	0.49
S1	SD	5.83	6.43	6.37	3.34	3.83	4.79	4.87	3.34	14.94	15.40	15.86	3.34
	CI	(−2.11, −1.88)	(0.87, 1.12)	(0.57, 0.82)	(0.43, 0.56)	(−1.67, −1.52)	(0.45, 0.64)	(0.21, 0.40)	(0.43, 0.57)	(−2.40, −1.81)	(0.82, 1.44)	(0.51, 1.13)	(0.43, 0.56)
	CI-L	0.23	0.25	0.25	0.13	0.15	0.19	0.19	0.13	0.59	0.60	0.62	0.13
	MSE	0.30	0.35	0.35	0.10	16.79	20.85	15.98	0.10	3.74	4.17	4.36	0.11
	M	−2.00	1.00	0.70	0.49	−1.59	0.55	0.30	0.50	−2.12	1.13	0.82	0.49
S2	SD	5.84	6.44	6.39	3.34	3.79	4.80	4.84	3.34	15.32	15.96	16.26	3.32
	CI	(−2.11, −1.88)	(0.87, 1.12)	(0.57, 0.82)	(0.43, 0.56)	(−1.67, −1.52)	(0.45, 0.64)	(0.21, 0.40)	(0.43, 0.57)	(−2.43, −1.82)	(0.82, 1.46)	(0.51, 1.16)	(0.43, 0.56)
	CI-L	0.23	0.25	0.25	0.13	0.15	0.19	0.19	0.13	0.60	0.63	0.64	0.13
	MSE	0.30	0.34	0.34	0.10	16.76	20.75	15.96	0.10	5.05	5.24	5.53	0.11

Finally, I discuss the results of the real data based simulation study, which can be found in Table 3.6. It can be seen that naive method severely underestimates $\beta_{x,2}$ and $\beta_{x,3}$. The proposed method produces lower MSE than the naive method, though not as small as using the true X. However, based on the previous simulations, the MSE would further decrease if the sample size increases. Also, the results seem to be not sensitive to the prior standard deviations that were considered in the simulation study.

Table 3.6: Results of the simulation study based on the real data

MT: method, M: median of posterior mean, SD: median of posterior standard deviation ×100, CI: 95% credible interval, CI-L: Median length of 95% credible interval, MSE: Mean Squared Error×100, $\text{pr}(Y = 1|X, Z) = H\{1.13 + 0I(X = 2) + 1.98I(X = 3) - 0.5Z_1 - 2.07Z_2\}$

MT		M1					M2				M3				
		β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_3	β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_3	β_1	$\beta_{x,2}$	$\beta_{x,3}$	β_2	β_3
$n = 10000$																
	M	1.13	0.00	1.99	−0.50	−2.07	0.95	0.38	0.73	−0.41	−1.11	1.11	0.00	2.00	−0.48	−2.05
S1	SD	5.42	6.33	7.95	6.01	7.28	5.46	6.42	6.32	5.67	5.15	7.60	10.47	13.55	7.19	11.84
	CI	(1.02, 1.23)	(−0.12, 0.13)	(1.82, 2.14)	(−0.61, −0.37)	(−2.21, −1.93)	(0.84, 1.06)	(0.25, 0.50)	(0.60, 0.85)	(−0.53, −0.30)	(−1.21, −1.01)	(0.96, 1.26)	(−0.20, 0.22)	(1.73, 2.27)	(−0.62, −0.34)	(−2.30, −1.83)
	CI-L	0.21	0.25	0.31	0.24	0.29	0.21	0.25	0.25	0.22	0.20	0.30	0.41	0.53	0.28	0.46
	MSE	0.28	0.37	0.52	0.37	0.52	3.48	14.61	157.02	1.06	92.79	0.53	1.02	1.79	0.55	1.26
	M	1.13	0.00	1.99	−0.49	−2.07	0.95	0.38	0.73	−0.41	−1.11	1.12	0.02	1.99	−0.48	−2.02
S2	SD	5.41	6.30	7.95	6.01	7.34	5.46	6.45	6.31	5.66	5.15	7.45	10.46	13.14	7.16	11.39
	CI	(1.02, 1.24)	(−0.12, 0.13)	(1.83, 2.14)	(−0.61, −0.38)	(−2.22, −1.93)	(0.84, 1.06)	(0.25, 0.51)	(0.60, 0.85)	(−0.53, −0.30)	(−1.21, −1.01)	(0.97, 1.26)	(−0.18, 0.23)	(1.73, 2.25)	(−0.62, −0.34)	(−2.27, −1.81)
	CI-L	0.21	0.25	0.31	0.24	0.29	0.21	0.25	0.25	0.22	0.20	0.29	0.41	0.51	0.28	0.45
	MSE	0.29	0.38	0.52	0.37	0.52	3.48	14.64	157.01	1.05	92.57	0.51	1.12	1.85	0.56	1.49

3.5 Real Data Analysis

For the purpose of illustrating our methodology, I applied the methods to analyze the data from the Surveillance, Epidemiology, and End Result (SEER) database provided by the National Cancer Institute [34]. SEER contains information on all cancer incidences starting in 1973 from different cancer registries located in eighteen different states throughout the United States. I restricted the analysis to breast cancer data based on the SEER data, 1975-2016.

The goal is to find association between the 5-year survival and the treatment therapy after adjusting the effect of the age of diagnosis and the stage of the disease.

For the purpose of analysis I considered only female subjects that were diagnosed with the breast cancer disease with stages II and III during 2007, 2008, 2009 and 2010. The reason for excluding subjects who were diagnosed prior to 2007 is due to the unavailability of insurance information from previous years. I do not consider other stages because there is not much of variations in the treatment therapy. The response Y is defined as 0 or 1 if a subject dies before 5 years or survives 5 years or more. I considered only two racial groups, black and white, which includes hispanic black and hispanic white. This is based on the variable number 2: *Race recode (White, Black, Other)* in the SEER database. Summary information about the demographics of the two groups can be found in Table 3.7. I included female patients with only one reported tumor, using the variable numbers 149: *Sequence number* and 150: *First malignant primary indicator*.

The treatment therapy is considered to be the main exposure variable X. I considered only surgical treatment, surgical treatment and chemotherapy, and a combination of surgical treatment, chemotherapy and radiation treatment, as the three treatment options. [2] suggested that these type of treatments are the most common

for stage II cancer, therefore I focused only on these options. The reported treatment therapy in SEER is considered to be misclassified [50], and according to my notation it is W. The dichotomized age of diagnosis 1 ($\geq$ 65 years) or 0 ($<$ 65 years), and dichotomized stage variable 1 (stage III) and 0 (stage II) were used as two prognostic factors, and they were denoted by $Z = (Z_1, Z_2)^T$. I excluded subjects whose age of diagnosis was less than 35 years. I took a three category insurance status as the instrument X^*, from the variable based on the variable field number 201: *Insurance Recode (2007+)*. The basis for using the insurance status as an instrument comes from different sources. For example, [12] found that medicaid status was associated with late stage diagnosis and treatment utilization, [22] found lower use of chemotherapy receipt for those on medicare aged 65 and up, and finally [13] found a non-zero association between insurance type and receipt of surgical treatment. The categories of X^* are insured/insured no specifics, any medicaid, and not insured. After all the exclusions, there were 43,453 and white and 7,069 black women left in the two datasets, respectively.

I took the following model for the response Y

$$\text{pr}(Y = 1|X, Z; \beta) = H\{\beta_0 + \beta_{x,2} I(X = 2) + \beta_{x,3} I(X = 3) + \beta_{z,1} Z_1 + \beta_{z,2} Z_2\},$$

where $\beta_{x,2}$ and $\beta_{x,3}$ are the log-odds ratio parameter when the therapy is surgical treatment and chemotherapy and a combination of surgical treatment, chemotherapy and radiation treatment, respectively. Here, only surgical treatment was considered as the reference category. Also, $\beta_{z,1}$ is the effect of age greater than equal to 65 years while less than 65 years is considered as the reference category. Finally, $\beta_{z,2}$ is the effect (regression parameter) corresponding to stage III with reference to stage II. Once again, the goal is statistical inference of these parameters. Unlike the

Table 3.7: Summary statistics of black and white women in SEER cohort that were analyzed.

Variable	Category	Black	White
Response (Y)	Survival ≥ 5 years	72%	78%
Misclassified Treatment (W)	Surgical Only	23%	27%
	Surgical and Chemotherapy	31%	29%
	Surgical, Chemotherapy, and Radiation Therapy	46%	44%
Insurance Status (X^*)	Insured/Insured No Specifics	74%	86%
	Any Medicaid	22%	12%
	Not Insured	4%	2%
Age (Z_1)	< 65	77%	71%
	≥ 65	23%	29%
Stage of cancer (Z_2)	Stage II	69%	73%
	Stage III	31%	27%
Age (Actual data)	Minimum	35	35
	25% Percentile	47	48
	Median	54	57
	75% Percentile	64	67
	Max	103	107

simulation study, the true X is not observed. So, first I analyzed the data using the naive method by regressing Y on W and Z using the ADVI method of Bayesian inference. I refer to this approach as M2, and set the priors for all components of β as Normal$(0, 2^2)$. Second, I analyzed the data using the proposed method and refer to it as M3. For the proposed method, I used three different sets of prior distributions. In the first case, the priors for all components of $\boldsymbol{\zeta} = (\beta^T, \gamma^T, \eta^T)^T$ were set to Normal$(0, 2^2)$. In the second case, the priors for all components of $\boldsymbol{\zeta}$ were now set to Normal$(0, 5^2)$. Finally, in case 3, I gathered prior information from previous work. [50] looked at the radiation and chemotherapy information from the SEER database and validated it with the medicare claim data. They found, for chemotherapy $\text{pr}(C_R = 0|C_T = 1) = 0.017$ and $\text{pr}(C_R = 1|C_T = 0) = 0.32$, and for radiation therapy, $\text{pr}(R_R = 0|R_T = 1) = 0.18$ and $\text{pr}(R_R = 1|R_T = 0) = 0.03$, where C_R, R_R, S_R denote reported status of chemotherapy, radiation and surgery, while

C_T, R_T, S_T denote the true status of chemotherapy, radiation and surgery. Also, 0 and 1 stand for no and yes, respectively. Following [14], I set the misclassification probability for surgery, $\text{pr}(S_R = 0|S_T = 1) = \text{pr}(S_R = 1|S_T = 0) = 0.2$. Under the conditional independence assumption, I calculated $\text{pr}(S_R = s_1, C_R = s_2, R_R = s_3|S_T = s_4, C_T = s_5, R_T = s_6) = \text{pr}(S_R = s_1|S_T = s_4)\text{pr}(C_R = s_2|C_T = s_5)\text{pr}(R_R = s_3|R_T = s_6)$. The above prior probabilities helped to compute the following prior misclassification probability table:

	(S_T, C_T, R_T)		
(S_R, C_R, R_R)	(1,0,0)	(1,1,0)	(1,1,1)
(1,0,0)	0.53	0.01	0.00
(1,1,0)	0.25	0.76	0.14
(1,1,1)	0.01	0.02	0.64

The (1, 1)th entry of the above table is $\text{pr}(S_R = 1, C_R = 0, R_R = 0|S_T = 1, C_T = 0, R_T = 0) = \text{pr}(S_R = 1|S_T = 1)\text{pr}(C_R = 0|C_T = 0)\text{pr}(R_R = 0|R_T = 0) = 0.8 \times 0.68 \times 0.97 = 0.53$, the (2, 2)th entry is $\text{pr}(S_R = 1, C_R = 1, R_R = 0|S_T = 1, C_T = 1, R_T = 0) = \text{pr}(S_R = 1|S_T = 1)\text{pr}(C_R = 1|C_T = 1)\text{pr}(R_R = 0|R_T = 0) = 0.8 \times 0.983 \times 0.97 = 0.76$ and the (3, 3)th entry is $\text{pr}(S_R = 1, C_R = 1, R_R = 1|S_T = 1, C_T = 1, R_T = 1) = \text{pr}(S_R = 1|S_T = 1)\text{pr}(C_R = 1|C_T = 1)\text{pr}(R_R = 0|R_T = 1) = 0.8 \times 0.983 \times 0.82 = 0.64$. The other entries were obtained in the similar way. However, before using this approach I had to standardize the table since the elements in each column do not add to 1. This was done by dividing each element of a column by its column sum. For example, the sum of the first column is equal to 0.79, so dividing the first element in the first column by 0.79 leads to the new standardize estimate of 0.67. This approach lead to the following misclassification probability matrix found in Table 3.8.

Table 3.8: Modeling misclassifications for the treatment types

	(S_T, C_T, R_T)		
(S_R, C_R, R_R)	(1,0,0)	(1,1,0)	(1,1,1)
(1,0,0)	0.67	0.02	0.00
(1,1,0)	0.32	0.95	0.18
(1,1,1)	0.01	0.03	0.82

Now, with this standardized misclassification probability matrix in hand, I solved the η-parameters of (3.13) and (3.14) where $\alpha_{i,j}$'s are the (i, j)th entry of the above misclassification matrix. I took these solutions of η as the mean of the normal prior of the η-parameters, and used 2 as the standard deviation. This prior for the η parameters was used for both the black and white women analysis. For the β-parameters, I used the naive estimate of β as the mean of the normal prior and the prior standard deviation was set to 2. Similarly for the γ-parameters, I used naive estimates of γ from regressing W on X^* and Z as the mean of a normal prior distribution with the prior standard deviation set to 2.

Results: The results can be found in Table 3.9. For each method, I present the posterior mean, standard deviation and the 95% credible interval. For both white and black women, M2 and M3 indicate statistically significant effects (association) of the third treatment category (having surgical, chemotherapy, and radiation therapy) when compared to just surgery alone on the 5-year survival. However, the effect of treatment improves the chances of survival for white women more than black women. For the white women group, treatment categories 2 and 3 seem to show statistically significant association with the survival when compared with the reference category 1.

In the naive approach, the odds ratio of survival for surgical therapy and chemotherapy for black women is $\exp(0.37) = 1.45$ while for white women this is $\exp(0.55) =$

Table 3.9: Analysis of the SEER data.
The upper panel is for black and the lower panel is for white women. Cases 1, 2, 3 correspond to three different sets of prior. The prior for the naive method was similar to that of case 1. M2: Naive, M3: proposed method, PM: posterior mean; CI: 95% credible interval

		M2	M3		
			Case 1	Case 2	Case 3
β_0	PM	1.14	1.10	1.13	1.10
	CI	$(1.00, 1.28)$	$(0.86, 1.34)$	$(0.90, 1.35)$	$(0.86, 1.33)$
$\beta_{x,2}$	PM	0.37	-0.14	-0.20	-0.13
	CI	$(0.20, 0.53)$	$(-0.51, 0.24)$	$(-0.58, 0.17)$	$(-0.50, 0.25)$
$\beta_{x,3}$	PM	0.68	1.93	1.98	1.88
	CI	$(0.52, 0.84)$	$(1.39, 2.48)$	$(1.40, 2.57)$	$(1.33, 2.44)$
$\beta_{z,1}$	PM	-0.52	-0.49	-0.53	-0.49
	CI	$(-0.65, -0.38)$	$(-0.73, -0.24)$	$(-0.77, -0.29)$	$(-0.73, -0.25)$
$\beta_{z,2}$	PM	-1.21	-2.03	-2.07	-1.98
	CI	$(-1.33, -1.08)$	$(-2.42, -1.63)$	$(-2.49, -1.65)$	$(-2.38, -1.59)$
β_0	PM	1.33	0.89	0.68	0.61
	CI	$(1.27, 1.39)$	$(0.72, 1.06)$	$(0.55, 0.80)$	$(0.45, 0.77)$
$\beta_{x,2}$	PM	0.55	0.75	1.01	1.16
	CI	$(0.48, 0.62)$	$(0.56, 0.95)$	$(0.85, 1.17)$	$(0.96, 1.37)$
$\beta_{x,3}$	PM	0.83	1.92	1.75	1.70
	CI	$(0.76, 0.89)$	$(1.66, 2.18)$	$(1.58, 1.92)$	$(1.53, 1.87)$
$\beta_{z,1}$	PM	-0.64	-0.34	-0.27	-0.21
	CI	$(-0.70, -0.59)$	$(-0.45, -0.23)$	$(-0.36, -0.19)$	$(-0.34, -0.07)$
$\beta_{z,2}$	PM	-0.93	-1.47	-1.20	-1.13
	CI	$(-0.98, -0.87)$	$(-1.69, -1.25)$	$(-1.33, -1.07)$	$(-1.23, -1.02)$

1.73, which corresponds to a 45% and 73% increase in the odds of 5-year survival for black women and white women, respectively, compared with the only surgical groups. The odds ratio of survival for having surgery, chemotherapy, and radiation therapy to the only surgery group for black women is $\exp(0.68) = 1.97$ while that for white women group is $\exp(0.83) = 2.29$. These represent a 97% and 129% increase in the odds of 5-year survival for the black women and white women, respectively. One thing to note is that for white women, both M2 and M3 show that increasing the variety of treatment improves the chances of 5-year survival, since $\widehat{\beta}_{x,3} > \widehat{\beta}_{x,2}$. However, for black women, this does not appear to be the situation for M3, cases 1, 2 and 3.

Because I used a Bayesian approach for model estimation, I determined the best models for the black and white women by estimating Bayes factors using the **bridgesampling** package. For black women, the Bayes factor for Case 1 vs Case 3 was < 0.01, while the Bayes factor for Case 2 vs Case 3 was > 100. For white women, the Bayes factor for Case 1 vs Case 3 was 6, while the Bayes factor for Case 2 vs Case 3 was > 100. I concluded that Case 2 was the preferred model under the white and black women analysis.

When compared to the naive approach, the proposed method indicates that having more types of treatment has a very profound impact on survival. Under case 2 for black women, the odds ratio of survival when undergoing all three treatments is $\exp(1.98) = 7.24$, indicating a 624% increase in the odds of 5-year survival compared to that having surgery alone. In case 3 for white women, this is odds ratio is $\exp(1.75) = 5.52$.

Additionally, for both white and black women, both Z_1 and Z_2 have statistically significant effect on the survival probability in M3 (case 2). Overall the effect of having Stage III cancer has a more severe impact on black women's survival than white. For example in case 2, black women have an $\exp(-2.07) = 0.13$ odds ratio when they have Stage III cancer, whereas for white women it is $\exp(-1.20) = 0.30$. In words, for black women, the odds of survival decreases by 87% for stage II to stage III, while for the white women group that decrease is about 70%.

Finally, I point out some limitations of this analysis. First, X^* should not have any influence on Y given X, if there is any association it will cause bias. However, this assumption is difficult to verify since X is never observed. Second, misclassification probabilities should not depend on the instrumental variable, but once again, in the absence of true X this is difficult to verify.

3.6 Discussion

In this work, I provided conditions for identification when the exposure is misclassified using a fast Bayesian computational algorithm. Though parameter identification for misclassified exposure has been considered in a nonparametric setup, to the best of our knowledge no one has considered the parametric setup with the logistic model which is widely used in epidemiological studies. The proposed estimation strategy is novel in the epidemiological context, and this fast computational algorithm can be applied to a large dataset. The simulation results are quite encouraging showing the effectiveness of the proposed approach in hugely reducing bias that is seen in the naive method. Finally I used this method to analyze the SEER data.

One limitation of the method is that I assume that the misclassification probabilities do not depend on covariate Z and the response Y. In future, I will consider relaxing this assumption which is not a trivial extension of this current work. Furthermore, I will check if my identification result holds for any parametric model, or a class of parametric models that possesses any specific characteristic.

4. FUTURE WORK

In both chapters 2 and 3, I considered non-differential misclassification of the exposure variable. However, differential misclassification can occur in epidemiological studies [21] . Therefore, in future I will develop methods that can allow differential misclassification of the exposure variable. This new methodology will depend on the observable variables in the study, and consequently I will examine what things are identifiable from the observable data. Secondly, throughout this dissertation the misclassification probabilities are assumed to be independent of the other covariates (prognostic factors) and the instrumental variable conditional on the true covariate. In my future research I will study in what extent these assumptions can be relaxed.

The proposed methods in this dissertation do not require any validation data that is often difficult to obtain. Rather the proposed methods make use of the instrumental variables to learn the misclassification probabilities and consequently the disease-exposure association. Although, it is a big jump in terms of relaxing the requirements, finding a good/strong instrument is often a difficult task. Also, there are certain assumptions that need to be satisfied by instrumental variables, but these set of assumptions are not easy to verify when the true value of the exposure variable is never known, not even for a subset of the data. Thus, part of my future research is to study the possibility of reducing the misclassification bias without requiring any validation data or instrumental variables [10].

Another direction of my future research includes the use of these bias reduction methods into mediation analysis models. The purpose of mediation analysis is to study the direct effect of a treatment or exposure on a response, and the indirect effect of a treatment or exposure through some mediating variable. A good review

of mediation analysis is provided by [61], who explains how mediation analysis can be used as a tool for causal inference using the language of the potential outcome framework. This frame posits that under a binary treatment regime, subjects will have two potential outcomes for each treatment but due to the nature of the study only one of those outcomes will be realized or observed. In order to make causal inferences from an observational study, one has to make certain assumptions about the study, such as *unmeasured confounders assumption*, the *exchangeability assumption*, and *consistency*, and the details can be found in [58].

The typical cause of concern in mediation analysis are biases induced by confounding variables which may or may not be observed, and so there are verifiable and unverifiable assumptions to validate this issue. However, misclassification or mismeasured mediators are also a cause of concern. [62] provided some analytical results of the effect of mismeasured or misclassified mediators in a mediation analysis. [4] conducted simulations on the effect of misclassified mediators and misclassified exposure and demonstrated that misclassified mediators has a greater influence on biasing indirect effects estimates than a misclassified exposure. In future, I will develop the bias reduction methods to a mediation model where the mediator is a binary or categorical variable.

5. SUMMARY

Addressing bias in analysis of observational data due to misclassified exposure variables is studied in this dissertation from two aspects. I first consider this issue in the matched case control setting, using instrumental variables in two proposed methods to adjust for misclassification in a binary exposure. Second, I consider a more general setting, detailing sufficient conditions for identification for a general categorical exposure, and implementing a novel bayesian algorithm for estimation. The major progresses of this dissertation are summarized as follows:

In the second chapter two consistent methods for bias reduction when estimating the association parameters in a matched case-control study were proposed. The novelty of the methods are the application of instrumental variables to obtain the measurement uncertainty when there is no validation data. Although the use of instrumental variable to reduce bias in when the binary exposure variable is misclassified is not new, adopting this general idea in the matched case-control studies is indeed new. The methodology is accompanied with an uncertainty measure, and contains a theoretical justification of the large sample properties. The simulation studies provide insight on the proposed method's performance. In particular, the simulation results indicate satisfactory performance of the proposed methods when there is a large sample size, strong association between the true exposure and the instruments, and moderate misclassification probabilities.

In the third chapter, I demonstrated the effect of a misclassified multicategory exposure/covariate on a binary regression model. It is shown that under some constraints on the misclassification probability matrix, the model parameters are identifiable. Additionally, the real data example and simulations demonstrate how using

instrumental variables aids in reducing the bias from the misclassified covariate. Estimation of these parameters utilizes the recently developed ADVI method, which is implemented in the Stan language. This computational method avoids time consuming MCMC method, and uses an optimization technique to do a fast Bayesian inference for large datasets. The simulation results show that larger sample sizes are required for bias reduction when there are more categories in the misclassified exposure. Finally, further bias reduction is possible by incorporating more accurate prior knowledge on the parameters.

REFERENCES

[1] A. Agresti. *Categorical data analysis*. Wiley, New York, 2013.

[2] J.R. Bellon, S.E. Come, R.S. Gelman, I.C. Henderson, L.N. Shulman, B.J. Silver, J.R. Harris, and A. Recht. Sequencing of chemotherapy and radiation therapy in early-stage breast cancer: Updated results of a prospective randomized trial. *Journal of Clinical Oncology*, 23(9):1934–1940, 2005.

[3] N. J. Birkett. Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure. *American Journal of Epidemiology*, 136(3):356–362, 1992.

[4] T. Blakely, S. McKenzie, and K. Carter. Misclassification of the mediator matters when estimating indirect effects. *Journal of Epidemiology and Community Health*, 67:458–466, 2013.

[5] J. Bound and A. Krueger. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right. *Journal of Labor Economics*, 12:1–24, 1991.

[6] IDJ Bross. Misclassification in 22 tables. *Biometrics*, 10:478–486, 1954.

[7] J.P. Buonaccorsi, P. Laake, and M. B. Veierød. On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, 61:831–836, 2005.

[8] R.J. Carroll, D. Ruppert, L.A. Stefanski, and C. Crainiceanu. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* Chapman and Hall, CRC Press, New York, 2003.

[9] N Chatterjee and S Wacholder. Validation studies: Bias, efficiency, and exposure assessment. *Epidemiology*, 13(5):503–506, 2002.

[10] X. Chen, Y. Hu, and A Lewbel. Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments. *Economics Letters*, 100(3):381–384, 2008.

[11] R. Chu, P. Gustafson, and N. Le. Bayesian adjustment for exposure misclassification in case-control studies. *Statistics in Medicine*, 29:994–1003, 2010.

[12] Bradley C.J., C.W. Given, and C. Roberts. Race, socioeconomic status, and breast cancer treatment and survival. *Journal of the National Cancer Institute*, 94(7):490–496, 2002.

[13] N. Coburn, J. Fulton, D.N. Pearlman, C. Law, B. DiPaolo, and B. Cady. Treatment variation by insurance status for breast cancer patients. *The Breast Journal*, 14:128–134, 2008.

[14] G. S. Cooper, B. Virnig, C.N. Klabunde, N. Schussler, J. Freeman, and J.L. Warren. Use of seer-medicare data for measuring cancer surgery. *Medical Care*, 40(8):43–48, 2002.

[15] O. Davidov, D. Faraggi, and B. Reiser. Misclassification in logistic regression with discrete covariates. *Biometrical Journal*, 45:541–553, 2003.

[16] N.E. Day and D.P. Byar. Testing hypotheses in case-control studies-equivalence of mantel-haenszel statistics and logit score tests. *Biometrics*, 35:623–630, 1979.

[17] S.W. Duffy, T.E. Rohan, R. Kandel, T.C. Prevost, K. Rice, and J.P. Myles. Misclassification in a matched case-control study with variable matching ratio: application to a study of c-erbb-2 overexpression and breast cancer. *Statistics in Medicine*, 22:2459–2468, 2003.

[18] B.L. Egleston, S.M. Miller, and N.J. Meropol. The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. *Statistics in Medicine*, 30:3560–3572, 2011.

[19] W.N. Evans and J.S. Ringel. Can higher cigarette taxes improve birth outcomes.

Journal of Public Economics, 72:135–154, 1999.

[20] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38, 1993.

[21] K. Flegal, P. Keyl, and F. Nieto. Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134:1233–1244, 1991.

[22] R.A. Freedman, K.S. Virgo, Y. He, A.L. Pavluck, E.P. Winer, E.M. Ward, and N.L Keating. The association of race/ethnicity, insurance status, and socioeconomic factors with breast cancer care. *Cancer*, 117:180–189, 2011.

[23] V. Godambe. Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63:277–284, 1976.

[24] S.L. Gomez and S.L. Glaser. Misclassification of race/ethnicity in a population-based cancer registry (united states). *Cancer Causes Control*, 58:771–781, 2006.

[25] C. Gourierroux and A. Monfort. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Preventing Chronic Disease*, 17:83–97, 1981.

[26] S. Greenland. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29:722–729, 2000.

[27] S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14:29–46, 1999.

[28] P. Gufstaffson. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* Chapman and Hall, CRC Press, New York, 2003.

[29] J. A. Hausman, J. Abrevaya, and F.M. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87:239–269, 1998.

[30] J.W. Hausman, W. Newey, H. Ichimura, and J. Powell. Measurement errors in polynomial regression models. *Journal of Econometrics*, 50:273–295, 1991.

[31] M.A. Hernán and J. M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

[32] M.A. Hernán and J.M. Robins. *Causal Inference*. Chapman and Hall, CRC, Boca Raton, 2008.

[33] Y. Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144:27–61, 2008.

[34] National Cancer Institute. Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) seer*stat database: Incidence - seer 9 regs research data, nov 2018 sub (1975-2016) (katrina/rita population adjustment) - linked to county attributes - total u.s., 1969-2017 countie, 2019. based on the November 2018 submission.

[35] R. Jagsi, P. Abrahamse, S.T. Hawley, J.J. Graff, A.S. Hamilton, and S.J. Katz. Underascertainment of radiotherapy receipt in surveillance, epidemiology, and end results registry data. *Cancer*, 118:333–341, 2012.

[36] H. Kuchenhoff, S.M. Mwalili, and E Lesaffre. A general method for dealing with misclassification in regression: The misclassification simex. *Biometrics*, 62(95):85–96, 2006.

[37] A. Kucukelbir, R. Ranganath, A. Gelman, and D. Blei. Automatic variational inference in stan. *Neural Information Processing Systems*, pages 568–576, 2017.

[38] A. Kucukelbir, T. Tran, R. Ranganath, A. Gelman, and A. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:1–45, 2017.

[39] S. Kuritz, J.R. Landis, and G.G. Koch. A general overview of mantel-haenszel

methods: Applications and recent development. *Annual Review of Public Health*, 9:123–160, 1983.

[40] T.L. Lash, M. Schmidt, A.Ø. Jensen, and M.C. Engebjerg. Methods to apply probabilistic bias analysis to summary estimates of association. *Pharmacoepidemiology and Drug Safety*, 19:638–644, 2010.

[41] J. Liu, P. Gustafson, N. Cherry, and I. Burstyn. Bayesian analysis of a matched case-control study with expert prior information on both the misclassification of exposure and the exposure-disease association. *Statistics in Medicine*, 28:3411–3423, 2009.

[42] Y. Liu, J. Liu, and F. Zhang. Bias analysis for misclassification in a multicategorical exposure in a logistic regression model. *Statistics and Probability Letters*, 83(12):2621–2626, 2013.

[43] R.H. Lyles and J. Lin. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in medicine*, 29(22):2297–2309, 2010.

[44] A. Mahajan. Identification and estimation of regression models with misclassification. *Econometrica*, 74:631–665, 2006.

[45] C.M. Manuel, S. Sinha, and S. Wang. Matched casecontrol data with a misclassified exposure: what can be done with instrumental variables? *Biostatistics*, kxz012, 2019. Online at https://doi.org/10.1093/biostatistics/kxz012.

[46] L. Martin, M. McNamara, A.S. Milot, T. Halle, and E. Hair. The effects of father involvement during pregnancy on receipt of prenatal care and maternal smoking. *Maternal and Child Health Journal*, 11:595–602, 2007.

[47] M.J. Morrissey and D. Spiegelman. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*, 55:338–344, 1999.

[48] National Center for Health Statistics. Data file documentations, natality,1989, 1992. data retrieved from NCHS' Vital Statistics Natality Birth Data, http://www.nber.org/data/vital-statistics-natality-data.html.

[49] R Njai, P.Z. Siegel, Miller J.W., and Y Liao. Misclassification of survey responses and black-white disparity in mammography use, behavioral risk factor surveillance system, 1995-2006. *Preventing Chronic Disease*, 8(3):A59, 2011.

[50] A.M. Noone, J.L. Lund, A. Mariotto, K. Cronin, T. McNeel, D. Deapen, and J.L. Warren. Comparison of seer treatment data with medicare claims. *Medical Care*, 54:e55–e64, 2016.

[51] G.J. Prescott and P.H. Garthwaite. Bayesian analysis of misclassified binary data from a matched casecontrol study with a validation sub-study. *Statistics in Medicine*, 24:379–401, 2005.

[52] P. Rathouz, G. Satten, and R. Carroll. Semiparametric inference in matched case-control studies with missing covariate data. *Biometrika*, 89:905–916, 2002.

[53] Kenneth Rice. Full-likelihood approaches to misclassification of a binary exposure in matched case-control studies. *Statistics in Medicine*, 22:3177–3194, 2003.

[54] T. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

[55] J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958.

[56] S. Schennach. Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica*, 75:201–239, 2007.

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

449(95):51–61, 2000.

[58] VanderWeele T.J. Mediation analysis: A practitioners guide. *Annual Review of Public Health*, 37:1732, 2016.

[59] J. Townsend, P. Roderick, and J. Cooper. Cigarette smoking by socioeconomic group, sex, and age: effects of price, income, and health publicity. *BMJ*, 309:923–927, 1994.

[60] M. Van Den Brink, E. BandellHoekstra, and H.H. AbuSaad. The occurrence of recall bias in pediatric headache: A comparison of questionnaire and diary data. *The Journal of Head and Face Pain*, 41:11–20, 2001.

[61] T.J. VanderWeele and S. Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468, 2009.

[62] Ogburn E.L. VanderWeele T.J., Valeri L. The role of measurement error and misclassification in mediation analysis: mediation and measurement error. *Epidemiology*, 23(4):561–564, 2009.

[63] M.B. Veierød and P. Laake. Exposure misclassification: bias in category specific poisson regression coefficients. *Statistics in Medicine*, 20:771–784, 2001.

[64] C.A. Weinberg, D.M. Umbach, and S. Greenland. When will nondifferential misclassification of an exposure preserve the direction of a trend? *American Journal of Epidemiology,*, 140(6):565–571, 1994.

[65] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 2011.

APPENDIX A

PROOF OF THEORETICAL RESULTS[1]

A.1 Identification of the parameters of the model

$$\text{pr}(W = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$$

The identification comes from the assumed non-linear structure for $\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$. Had $\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$ been linear, the parameters would not be identifiable. In short I write $H(\gamma_0 + \boldsymbol{\gamma}_1^T\boldsymbol{S} + \boldsymbol{\gamma}_2^T\boldsymbol{X}^* + \boldsymbol{\gamma}_3^T\boldsymbol{Z})$ as $H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})$. In our case $H(\cdot)$ is the logistic function, which is nonlinear.

To see the identifiability issue, I need to show that for every given parameter set $(\boldsymbol{\gamma}, \alpha_0, \alpha_1)$ if another parameter set $(\boldsymbol{\gamma}^*, \alpha_0^*, \alpha_1^*)$ satisfies $\text{pr}(W = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}; \alpha_0, \alpha_1, \boldsymbol{\gamma}) = \text{pr}(W = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}; \alpha_0^*, \alpha_1^*, \boldsymbol{\gamma}^*)$ for every choice of $\boldsymbol{S}$, $\boldsymbol{X}^*$ and $\boldsymbol{Z}$, then $(\boldsymbol{\gamma}^*, \alpha_0^*, \alpha_1^*) = (\boldsymbol{\gamma}, \alpha_0, \alpha_1)$. To see this, by Equation (3.3) I start with

$$\alpha_0 + (1 - \alpha_0 - \alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) = \alpha_0^* + (1 - \alpha_0^* - \alpha_1^*)H(\boldsymbol{\gamma}^*, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) \quad \text{(A.1)}$$

for every choice of $(\boldsymbol{S}^T, \boldsymbol{X}^{*,T}, \boldsymbol{Z}^{*,T})^T$. Let $\boldsymbol{\gamma}^* = -\boldsymbol{\gamma}$, $\alpha_0^* = 1 - \alpha_1$ and $\alpha_1^* = 1 - \alpha_0$. Then $H(\boldsymbol{\gamma}^*, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) = H(-\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) = 1 - H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})$ and

$$\begin{aligned}
\alpha_0^* + (1 - \alpha_0^* - \alpha_1^*)H(\boldsymbol{\gamma}^*, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) &= (1 - \alpha_1) + (1 - 1 + \alpha_1 - 1 + \alpha_0) \\
&\quad \times H(-\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) \\
&= (1 - \alpha_1) + (-1 + \alpha_0 + \alpha_1) \\
&\quad \times \{1 - H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})\}
\end{aligned}$$

[1]Portions of this work reprinted with permission from Manuel, C.M. and Wang, S. and Sinha, S., "Matched Case-Control Data with a Misclassified Exposure: What can be done with Instrumental Variables?", Biostatistics, 2019, kxz012, by permission of Oxford University Press

$$
\begin{aligned}
&= (1-\alpha_1) + (-1+\alpha_0+\alpha_1) \\
&\quad -(-1+\alpha_0+\alpha_1) H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) \\
&= \alpha_0 + (1-\alpha_0-\alpha_1) H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}).
\end{aligned}
$$

On the other hand, under the monotonicity restriction $\alpha_0 + \alpha_1 < 1$, if $\alpha_1^* = 1 - \alpha_0$ and $\alpha_0^* = 1 - \alpha_1$, then $\alpha_0^* + \alpha_1^* = (1 - \alpha_1 + 1 - \alpha_0) = 1 + (1 - \alpha_0 - \alpha_1) > 1$. Hence, this particular choice of α_0^*, α_1^* does not satisfy the restriction, and is not a cause of concern anymore.

Finally, I check if there is any other choice of $(\alpha_0^*, \alpha_1^*, \boldsymbol{\gamma}^*)$ that satisfies (A.1). Suppose that there exists $(\alpha_0^*, \alpha_1^*, \boldsymbol{\gamma}^*)$ that satisfies (A.1) for every choice of $\boldsymbol{S}$, $\boldsymbol{X}^*$ and $\boldsymbol{Z}$. This implies that for every $(\boldsymbol{S}_k, \boldsymbol{X}_k^*, \boldsymbol{Z}_k)$, $k = 1, 2, \ldots,$

$$
\alpha_0^* + (1 - \alpha_0^* - \alpha_1^*) H(\boldsymbol{\gamma}^*, \boldsymbol{S}_k, \boldsymbol{X}_k^*, \boldsymbol{Z}_k) = \alpha_0 + (1 - \alpha_0 - \alpha_1) H(\boldsymbol{\gamma}, \boldsymbol{S}_k, \boldsymbol{X}_k^*, \boldsymbol{Z}_k).
$$

Since $1 - \alpha_0^* - \alpha_1^* > 0$ and $1 - \alpha_0 - \alpha_1 > 0$, it is readily seen that each element of $(\boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*)$ must have the same sign as the corresponding element of $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$. By letting $T = \gamma_0 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z} \to -\infty$ (and then $T^* = \gamma_0^* + \boldsymbol{\gamma}_1^{*T} \boldsymbol{S} + \boldsymbol{\gamma}_2^{*T} \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z} \to -\infty$ also), it is clear that $\alpha_0^* = \alpha_0$. Likewise, due to the nonlinearity of $H(\cdot)$, $\alpha_1^* = \alpha_1$. This leads to $T^* = T$ and thus $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}$, showing the identifiability of these parameters.

A.2 Proof of Lemma 1

Because of the logistic model assumption and the assumption on W and $\boldsymbol{X}^*$, I can write

$$
1 - \text{pr}(Y = 0 | \boldsymbol{S}, W, X, \boldsymbol{X}^*, \boldsymbol{Z}) = \text{pr}(Y = 1 | \boldsymbol{S}, W, X, \boldsymbol{X}^*, \boldsymbol{Z})
$$

$$
\begin{aligned}
&= \quad \mathrm{pr}(Y=1|\boldsymbol{S},X,\boldsymbol{Z}) \\
&= \quad \exp\{g_0(\boldsymbol{S})+\beta_1 X+\boldsymbol{\beta}_2^T\boldsymbol{Z}\}\mathrm{pr}(Y=0|\boldsymbol{S},X,\boldsymbol{Z}),
\end{aligned}
$$

where $g_0(\cdot)$ is given in model (2.1). Next, consider

$$
\begin{aligned}
&\mathrm{pr}(Y=1|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z}) \\
= &\sum_{x=0,1}\mathrm{pr}(Y=1|\boldsymbol{S},W,X=x,\boldsymbol{X}^*,\boldsymbol{Z})\mathrm{pr}(X=x|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z}) \\
= &\sum_{x=0,1}\mathrm{pr}(Y=1|\boldsymbol{S},X=x,\boldsymbol{Z})\mathrm{pr}(X=x|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z}) \\
= &\sum_{x=0,1}\exp\{g_0(\boldsymbol{S}_i)+\beta_1 x+\boldsymbol{\beta}_2^T\boldsymbol{Z}\}\mathrm{pr}(Y=0|\boldsymbol{S},X=x,\boldsymbol{Z})\mathrm{pr}(X=x|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z}) \\
= &\sum_{x=0,1}\exp\{g_0(\boldsymbol{S}_i)+\beta_1 x+\boldsymbol{\beta}_2^T\boldsymbol{Z}\}\mathrm{pr}(X=x|\boldsymbol{S},W,\boldsymbol{X}^*,Y=0,\boldsymbol{Z}) \\
&\times\mathrm{pr}(Y=0|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z}) \\
= &\ \mathrm{pr}(Y=0|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z})\sum_{x=0,1}\exp\{g_0(\boldsymbol{S})+\beta_1 x+\boldsymbol{\beta}_2^T\boldsymbol{Z}\} \\
&\times\mathrm{pr}(X=x|\boldsymbol{S},W,\boldsymbol{X}^*,Y=0,\boldsymbol{Z}) \\
= &\ \mathrm{pr}(Y=0|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z})\exp\{g_0(\boldsymbol{S})+\boldsymbol{\beta}_2^T\boldsymbol{Z}\}\{\exp(\beta_1) \\
&\times\mathrm{pr}(X=1|\boldsymbol{S},W,\boldsymbol{X}^*,Y=0,\boldsymbol{Z})+\mathrm{pr}(X=0|\boldsymbol{S},W,\boldsymbol{X}^*,Y=0,\boldsymbol{Z})\} \\
\equiv &\ \mathrm{pr}(Y=0|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z})\exp\{g_0(\boldsymbol{S})+\boldsymbol{\beta}_2^T\boldsymbol{Z}+g_1(\beta_1,\boldsymbol{S}_i,W,\boldsymbol{X}^*,\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\eta})\},
\end{aligned}
$$

where the expression of $g_1(\beta_1,\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\eta})$ is obtained after plugging the expression for $\mathrm{pr}(X=1|\boldsymbol{S},W,\boldsymbol{X}^*,Y=0,\boldsymbol{Z})$ and $\mathrm{pr}(X=0|\boldsymbol{S},W,\boldsymbol{X}^*,Y=0,\boldsymbol{Z})$ from Equations (3.4) and (3.5). In particular,

$$
\begin{aligned}
&\exp\{g_1(\beta_1,\boldsymbol{S},W=1,\boldsymbol{X}^*,\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\eta})\} \\
= &\ \exp(\beta_1)\mathrm{pr}(X=1|\boldsymbol{S},W=1,\boldsymbol{X}^*,Y=0,\boldsymbol{Z}) \\
&+\mathrm{pr}(X=0|\boldsymbol{S},W=1,\boldsymbol{X}^*,Y=0,\boldsymbol{Z})
\end{aligned}
$$

$$
\begin{aligned}
&= \exp(\beta_1)\frac{(1-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}{\alpha_0+(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})} \\
&\quad +1-\frac{(1-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}{\alpha_0+(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})} \\
&= \frac{\exp(\beta_1)(1-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})+\alpha_0\{1-H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\}}{\alpha_0+(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}, \qquad \text{(A.2)}
\end{aligned}
$$

$$
\begin{aligned}
&\exp\{g_1(\beta_1,\boldsymbol{S},W=0,\boldsymbol{X}^*,\boldsymbol{\gamma},\boldsymbol{\eta})\} \\
&= \exp(\beta_1)\mathrm{pr}(X=1|\boldsymbol{S},W=0,\boldsymbol{X}^*,Y=0,\boldsymbol{Z}) \\
&\quad +\mathrm{pr}(X=0|\boldsymbol{S},W=0,\boldsymbol{X}^*,Y=0,\boldsymbol{Z}) \\
&= \exp(\beta_1)\frac{\alpha_1 H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}{1-\alpha_0-(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}+1 \\
&\quad -\frac{\alpha_1 H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}{1-\alpha_0-(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})} \\
&= \frac{\exp(\beta_1)\alpha_1 H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})+(1-\alpha_0)\{1-H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\}}{1-\alpha_0-(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}. \qquad \text{(A.3)}
\end{aligned}
$$

A.3 Proof of Theorem 1

Collecting $\boldsymbol{S}_\gamma(\boldsymbol{\gamma},\boldsymbol{\eta}), \boldsymbol{S}_\eta(\boldsymbol{\gamma},\boldsymbol{\eta}), S_{\beta_1}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta}), \boldsymbol{S}_{\boldsymbol{\beta}_2}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta})$ together and letting $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T,\boldsymbol{\eta}^T,\beta_1,\boldsymbol{\beta}_2^T)^T$ and $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\gamma}}^T,\widehat{\boldsymbol{\eta}}^T,\widehat{\beta}_1,\widehat{\boldsymbol{\beta}}_2^T)^T$, I can write

$$
\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = A^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i + o_p(1),
$$

where $\boldsymbol{U}_i's$ are iid and mean zero and finite variance random vectors. $A = -E(\partial\boldsymbol{U}_i/\partial\boldsymbol{\theta})$. By the Central Limit Theorem I obtain the asymptotic normality of $\widehat{\boldsymbol{\theta}}$, and the asymptotic variance of $\sqrt{n}\widehat{\boldsymbol{\theta}}$ is $A^{-1}\mathrm{var}(\boldsymbol{U}_1)A^{-T}$. This asymptotic variance can be consistently estimated by $\widehat{A}^{-1}(\sum_{i=1}^n \widehat{\boldsymbol{U}}_i\widehat{\boldsymbol{U}}_i^T/n)\widehat{A}^{-T}$ with $\widehat{A} = -(1/n)\sum_{i=1}^n \partial\widehat{\boldsymbol{U}}_i/\partial\boldsymbol{\theta}$ and $\widehat{\boldsymbol{U}}_i$ being $\boldsymbol{U}_i$ with $\boldsymbol{\theta}$ replaced by $\widehat{\boldsymbol{\theta}}$.

A.4 Proof of Lemma 2

Part i) of Lemma 2

$$\begin{aligned}
\mathrm{pr}(Y=1|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) &= \sum_x \mathrm{pr}(Y=1|\boldsymbol{S},X=x,\boldsymbol{X}^*,\boldsymbol{Z})\mathrm{pr}(X=x|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) \\
&= \sum_x \exp\{g_0(\boldsymbol{S})+\beta_1 x+\boldsymbol{\beta}_2^T\boldsymbol{Z}\}\mathrm{pr}(Y=0|\boldsymbol{S},X=x,\boldsymbol{X}^*,\boldsymbol{Z}) \\
&\quad \times \mathrm{pr}(X=x|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) \\
&= \sum_x \exp\{g_0(\boldsymbol{S})+\beta_1 x+\boldsymbol{\beta}_2^T\boldsymbol{Z}\}\mathrm{pr}(X=x|\boldsymbol{S},\boldsymbol{X}^*,Y=0,\boldsymbol{Z}) \\
&\quad \times \mathrm{pr}(Y=0|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) \\
&= \mathrm{pr}(Y=0|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})[\exp\{g_0(\boldsymbol{S})+\boldsymbol{\beta}_2^T\boldsymbol{Z}\} \\
&\quad \times\{1-H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\} \\
&\quad +\exp\{g_0(\boldsymbol{S})+\beta_1+\boldsymbol{\beta}_2^T\boldsymbol{Z}\}H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})] \\
&= \mathrm{pr}(Y=0|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\exp\{g_0(\boldsymbol{S})+\boldsymbol{\beta}_2^T\boldsymbol{Z}\} \\
&\quad \times\{1-H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})+\exp(\beta_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\}.
\end{aligned}$$

This implies

$$\mathrm{pr}(Y=1|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) = H\{g_0(\boldsymbol{S})+\boldsymbol{\beta}_2^T\boldsymbol{Z}+g_2(\boldsymbol{\gamma},\beta_1,\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\},$$

where

$$g_2(\boldsymbol{\gamma},\beta_1,\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) = \log\{1-H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})+\exp(\beta_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\}.$$

Part ii) of Lemma 2

$$\mathrm{pr}(X=1|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z},Y=1) = \frac{\mathrm{pr}(Y=1|\boldsymbol{S},X=1,\boldsymbol{X}^*,\boldsymbol{Z})\mathrm{pr}(X=1|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}{\mathrm{pr}(Y=1|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}$$

$$
\begin{aligned}
&= \frac{\exp\{g_0(\boldsymbol{S}) + \beta_1 + \boldsymbol{\beta}_2^T \boldsymbol{Z}\}}{\mathrm{pr}(Y=1|\boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})} \\
&\quad \times \mathrm{pr}(Y=0|\boldsymbol{S}, X=1, \boldsymbol{X}^*, \boldsymbol{Z}) \mathrm{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) \\
&= \frac{\exp\{g_0(\boldsymbol{S}) + \beta_1 + \boldsymbol{\beta}_2^T \boldsymbol{Z}\}}{\mathrm{pr}(Y=1|\boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})} \\
&\quad \times \mathrm{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=0, \boldsymbol{Z}) \mathrm{pr}(Y=0|\boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) \\
&= \frac{\exp\{g_0(\boldsymbol{S}) + \beta_1 + \boldsymbol{\beta}_2^T \boldsymbol{Z}\} H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}{\exp\{g_0(\boldsymbol{S}) + \boldsymbol{\beta}_2^T \boldsymbol{Z} + g_2(\boldsymbol{\gamma}, \beta_1, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})\}} \\
&= \frac{\exp(\beta_1) H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}{\exp\{g_2(\boldsymbol{\gamma}, \beta_1, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})\}} \\
&= \frac{\exp(\beta_1) H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})}{1 - H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}) + \exp(\beta_1) H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})} \\
&= \frac{\exp(\gamma_0 + \beta_1 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z})}{1 + \exp(\gamma_0 + \beta_1 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z})} \\
&= H(\gamma_0 + \beta_1 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z}).
\end{aligned}
$$

Part iii) of Lemma 2

$$
\begin{aligned}
&\mathrm{pr}(W=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \\
&= \mathrm{pr}(W=1|\boldsymbol{S}, X=0, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \mathrm{pr}(X=0|\boldsymbol{S}, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \\
&\quad + \mathrm{pr}(W=1|\boldsymbol{S}, X=1, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \mathrm{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \\
&= \mathrm{pr}(W=1|X=0) \mathrm{pr}(X=0|\boldsymbol{S}, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \\
&\quad + \mathrm{pr}(W=1|X=1) \mathrm{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \\
&= \alpha_0 \{1 - \mathrm{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=1, \boldsymbol{Z})\} + (1-\alpha_1) \mathrm{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \\
&= \alpha_0 + (1 - \alpha_0 - \alpha_1) \mathrm{pr}(X=1|\boldsymbol{S}, \boldsymbol{X}^*, Y=1, \boldsymbol{Z}) \\
&= \alpha_0 + (1 - \alpha_0 - \alpha_1) H(\gamma_0 + \beta_1 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z}).
\end{aligned}
$$