ABSTRACT

CHO, YOONJIN. Comparing Predictive Values of Two Diagnostic Tests. (Under the direction of Andrezj Kosinski).

Positive and negative predictive values are important measures of accuracy when one compares medical diagnostic tests. When more than one diagnostic test are available, one may have to choose one of the possible diagnostic tests due to cost, time, or ethical reasons. We consider a paired study design in cohort study where two diagnostic tests are measured on every patient. Our parameter of interest is the log odds of predictive values. In first chapter, we review current methods for comparing diagnostic tests when gold standards are available on every individual. We propose method by series of logistic regressions and derive estimator and test statistics based on likelihood probability. It is often the case that gold standard is not observed on every patient because it may be invasive. If we only consider those who have observed gold standard, the estimator may be biased. In Chapter 2 and 3, we extend the methods to when gold standard is missing. We assume that missing gold standard is missing at random, which means missing pattern only depends on observed data. In Chapter 2, we use semiparametric theory to derive a class of regular and asymptotically normal estimators of our parameter of interest. Out of the class, we derive an estimator which is the most efficient in the class by using the information from available auxiliary covariates which may be associated with the outcome of gold standard. We also use auxiliary covariates in modeling the probability of observing gold standard. In Chapter 3, through M-estimator, we derive another consistent estimator through imputation method.

Comparing Predictive Values of Two Diagnostic Tests

by

Yoonjin Cho

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fullfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

May, 2009

APPROVED BY:

Dr. Daowen Zhang	Dr. Jason Osborne
Dr. Andrzej Kosinski Chair of Advisory Committee	Dr. Anastasios A. Tsiatis Co-Chair of Advisory Committee

DEDICATION

To
My Late Father, Sungho Cho
My Grandmother, Yoon
My Mother, Park
My Brother, Juno
and My Husband, DK

BIOGRAPHY

Yoonjin Cho was born in Seoul, Korea to parents Sungho Cho and Gaeseuk Park on January, 13, 1979. She received B.A. in mathematics and education from Yonsei University in January, 2002. While she was at Yonsei University, she was selected as an exchange student and came to the United States in 1999. She received another B.A in computer science and mathematics from Maryville College in May, 2002. In August, 2003, she entered the program in Statistics at North Carolina State University and received a M.S in Statistics and continued to Ph.D program. On December 29, 2007, she married DK in Seoul, Korea. Upon completion of her doctoral degree, she and her husband will stay at North Carolina and she will start working at Statistical Science in GlaxoSmithKline as a senior statistician.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my adviser, Dr. Andrzej Kosinski for his guidance, and encouragement. I have been very privileged to have him as an adviser who is not only caring, and supportive, but also who has a great sense of humor. I am grateful to Dr. Anastasios A. Tsiatis for the guidance in my research and for his enthusiasm in statistics that I learned from his lectures. I am also very thankful to Dr. Daowen Zhang and Dr. Jason Osborne for the commitment as my committee members and for the valuable classes.

Particular thanks to Dr. Bibhuti Bhattacharyya for his integrity and passion he showed through his classes, and at his office. I would never forget how much he valued and cared for students and he often missed his lunch and free afternoon. My gratitude goes to Dr. Bruce Weir who gave me opportunity to continue my ph.D. with financial support in my first year and to meet with great people. His kindness and his gentleness would remain in spirit to me.

I also appreciate peers who came to the department as the same year as I did for their companies and enjoyable memories. I miss their companies when I had to work on research independently. I also give my thanks to Korean friends who came before and after me to the school. Their shares with food and friendship mean very special to me.

I would like to convey my appreciation to the department for the great programs and classes. I also had great opportunity in working at GSK through the department. I thank my supervisors at GSK: Dr. Richard A. Lewis and Dr. Steven Novick. They inspire me to enjoy statistics even at work and open my eyes to the value of statistics in industry. They also were very helpful in my job search.

Emmanuel Antioch Presbyterian Church members and International Bible Study members make my life in graduate school fruitful. Their prayers and encouragements help me to grow not only spiritually but physically. Many of them taught me how to value friendship. Without their love and support, I would not be able to stand firm where I am now.

This dissertation is dedicated to my family who have been next to me even if we are apart and who have faith in me all the time. I thanks my mother for her support emotionally, financially, and spiritually. Finally, but the most importantly, I thank God, Jesus Christ, and Holy Spirit for His blessings, guidance, peace and grace.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Comparison of Predictive Values of Two Diagnostic Tests Using a Conditional Model

1.1 Introduction

The role of diagnostic test is very important in epidemiology and medicine. It is useful in identifying populations or patients at high risk of disease especially when the gold standard for ascertainment of presence or absence of disease is expensive or invasive. In this paper we will discuss tests with a binary result (positive/negative) and consider situation, in which the gold standard is available for comparison with the test result. The most common measures often used in order to evaluate the diagnostic tests are sensitivity, specificity, positive predictive value, and negative predictive value. Sensitivity is the probability of positive diagnostic test in the diseased population (as identified by the gold standard). Specificity is the probability of negative diagnostic test in the non-diseased population. Positive predictive value (PPV) is the probability of disease when the diagnostic test is positive and negative predictive value (NPV) is the probability of no disease when the diagnostic test is negative. In this paper we will discuss comparison of positive predictive values (PPV_1 versus PPV_2) and negative predictive values (NPV_1 versus NPV_2) for two tests.

The motivating example of coronary artery disease (CAD) data is displayed in

Table 1.1 [1] The data are from a paired study design, in which each subject received both tests. The gold standard for CAD is coronary angiography and the data set considers two diagnostic tests: exercise stress test (Test 1) and clinical history of chest pain (Test 2). A

Table 1.1: Coronary artery disease data.

	CAD		No CAD	
Test 1 result	Test 2 result		Test 2 result	
	Chest pain	No chest pain	Chest pain	No chest pain
Positive stress test	473	29	22	46
Negative stress test	81	25	44	151

general notation is displayed in Table 1.2. Appropriate marginal counts will be denoted by a dot in the location over which marginalization occurs, e.g., $n_{\bullet+D} = n_{++D} + n_{-+D}$, $n_{++\bullet} = n_{++D} + n_{++\overline{D}}$, or $n_{\bullet+\bullet} = n_{\bullet+D} + n_{\bullet+\overline{D}} = n_{++\bullet} + n_{-+\bullet}$. Thus, a direct proportion

Table 1.2: Coronary artery disease data.

	Disease		No Disease	
Test 1 result	Test 2 result		Test 2 result	
	Positive (+)	Negative (−)	Positive (+)	Negative (−)
Positive (+)	n_{++D}	n_{+-D}	$n_{++\overline{D}}$	$n_{+-\overline{D}}$
Negative (−)	n_{-+D}	n_{--D}	$n_{-+\overline{D}}$	$n_{--\overline{D}}$

estimate of PPV for test 1 is $n_{+\bullet D}/n_{+\bullet\bullet}$ and for test 2 is $n_{\bullet+D}/n_{\bullet+\bullet}$. Similarly, NPV can be estimated by $n_{-\bullet\overline{D}}/n_{-\bullet\bullet}$ for test 1 and by $n_{\bullet-\overline{D}}/n_{\bullet-\bullet}$ for test 2.

Since the two tests are measured on the same subject, the correlation between tests is present. To account for this correlation, Leisenring et al. [1] applied marginal model approach [2]. Wang et al. [3] assumed the multinomial distribution and derived the Wald statistics for testing difference in predictive values. In this paper, we focus on estimating and comparing predictive values with a likelihood based conditional regression model framework.

In Section 1.2, we review the current methods for comparing predictive values in a paired design. Section 1.3 presents our approach. In Section 1.4 simulation studies are performed to evaluate the proposed method and compare to the existing methods. Analysis of real data example from a cohort study of coronary artery disease is presented in Section 1.5.

1.2 Review of Current Methods

1.2.1 Marginal Regression

Leisenring et al. [1] construct data so that each subject has a repeated row for each diagnostic test. Variable X_{ij} denotes result for diagnostic test ($X_{ij} = 1$ for positive and $X_{ij} = 0$ for negative test result), with i denoting subject number and j denoting test number ($j = 1, 2$ if two tests are considered). A covariate Z_{ij} is the indicator for the diagnostic test (0 for test 1, 1 for test 2), and the response is a binary variable D_{ij}, with $D_{ij} = 1$ for presence of disease and $D_{ij} = 0$ absence of disease. In general, D_{ij} can have different values for each test. However, to focus on most common situation presented in Table II, in which the same disease status is present within a single patient data, i.e. the situation with $D_{i1} = D_{i2}$.

The diagnostic tests may be correlated conditional on disease status and Leisenring et al. [1] accounted for this correlation by consideration of the Generalized Estimating Equations (GEE) approach. When testing equality of PPVs for test 1 (PPV_1) and test 2 (PPV_2), the marginal model

$$\text{logit } P(D_{ij} = 1 | Z_{ij}, X_{ij} = 1) = \alpha_P + \beta_P Z_{ij}$$

is considered and hypothesis: $H_0 : \beta_P = 0$ equivalent to $H_0 : PPV_1 = PPV_2$ is tested. For testing equality of NPVs, the model

$$\text{logit } P(D_{ij} = 1 | Z_{ij}, X_{ij} = 0) = \alpha_N + \beta_N Z_{ij}$$

is considered and hypothesis: $H_0 : \beta_N = 0$ ($H_0 : NPV_1 = NPV_2$) is tested. Leisenring et al. [1] derived the generalized score statistics and for testing the above two hypotheses. We expressed the score statistics in terms of Table 1.2 cell counts. Score test statistic for comparing PPVs is $T^{GEE}_{PPV} = N_P/D_P$, and for comparing NPVs is $T^{GEE}_{NPV} = N_N/D_N$, where

$$\begin{aligned}
N_P &= \left\{ n_{\bullet +D} - (n_{\bullet +D} + n_{+\bullet D}) \overline{Z}_P \right\}^2 \\
D_P &= (1 - \overline{D}_P)^2 \left\{ n_{++D}(1 - 2\overline{Z}_P)^2 + n_{-+D}(1 - \overline{Z}_P)^2 + n_{+-D}\overline{Z}_P^2 \right\} \\
&\quad + \overline{D}_P^2 \left\{ n_{++\overline{D}}(1 - 2\overline{Z}_P)^2 + n_{-+\overline{D}}(1 - \overline{Z}_P)^2 + n_{+-\overline{D}}\overline{Z}_P^2 \right\} \\
N_N &= \left\{ n_{\bullet -D} - (n_{\bullet -D} + n_{-\bullet D}) \overline{Z}_N \right\}^2 \\
D_N &= (1 - \overline{D}_N)^2 \left\{ n_{-+D}\overline{Z}_N^2 + n_{+-D}(1 - \overline{Z}_N)^2 + n_{--D}(1 - 2\overline{Z}_N)^2 \right\} \\
&\quad + \overline{D}_N^2 \left\{ n_{-+\overline{D}}\overline{Z}_N^2 + n_{+-\overline{D}}(1 - \overline{Z}_N)^2 + n_{--\overline{D}}(1 - 2\overline{Z}_N)^2 \right\}
\end{aligned}$$

and

$$\overline{Z}_P = \frac{n_{\bullet+\bullet}}{n_{\bullet+\bullet}+n_{+\bullet\bullet}},\ \overline{D}_P = \frac{n_{\bullet+D}+n_{+\bullet D}}{n_{\bullet+\bullet}+n_{+\bullet\bullet}}$$
$$\overline{Z}_N = \frac{n_{\bullet-\bullet}}{n_{\bullet-\bullet}+n_{-\bullet\bullet}},\ \overline{D}_N = \frac{n_{\bullet-D}+n_{-\bullet D}}{n_{\bullet-\bullet}+n_{-\bullet\bullet}}.$$

Test statistics T^{GEE}_{PPV} and T^{GEE}_{NPV} are distributed under the corresponding null hypotheses H_0 as the χ^2 distribution with one degree of freedom.

1.2.2 Wald statistics using Weighted Least Squares

When the total number of patients $n_{\bullet\bullet\bullet}$ is fixed, the probabilities for the eight cells in Table II have multinomial distribution, with vector of cell probabilities

$$\mathbf{P} = \left[p_{++D}, p_{-+D}, p_{-+D}, p_{--D}, p_{++\overline{D}}, p_{-+\overline{D}}, p_{+-\overline{D}}, p_{--\overline{D}}\right]^T$$

estimated by

$$\widehat{\mathbf{P}} = \left[n_{++D}, n_{-+D}, n_{-+D}, n_{--D}, n_{++\overline{D}}, n_{-+\overline{D}}, n_{+-\overline{D}}, n_{--\overline{D}}\right]^T / n_{\bullet\bullet\bullet}\ .$$

For large samples, $\sqrt{n_{\bullet\bullet\bullet}}(\mathbf{P}-\widehat{\mathbf{P}})$ is distributed as normal distribution $N(0,\Sigma_P)$, where $\Sigma_P = Diag(\widehat{\mathbf{P}}) - \widehat{\mathbf{P}}\widehat{\mathbf{P}}^T$. Since the measures of interest PPV and NPV are functions of Table II cells, the delta method can be used to estimate variance of PPV and NPV differences. Let $g(\mathbf{P}) = PPV_1 - PPV_2 = p_{+\bullet D}/p_{+\bullet\bullet} - p_{\bullet+D}/p_{\bullet+\bullet}$ and $h(\mathbf{P}) = NPV_1 - NPV_2 = p_{-\bullet D}/p_{-\bullet\bullet} - p_{\bullet-D}/p_{\bullet-\bullet}$. Thus, the Wald statistic for testing $H_0 : PPV_1 = PPV_2$ is

$$T^{WLS}_{PPV} = \frac{(\widehat{PPV}_1 - \widehat{PPV}_2)^2}{Var(\widehat{PPV}_1 - \widehat{PPV}_2)} = \frac{g(\widehat{\mathbf{P}})^2}{n_{\bullet\bullet\bullet}\left[\frac{\partial g(\mathbf{P})}{\partial \mathbf{P}}\right]^T_{\mathbf{P}=\widehat{\mathbf{P}}} \Sigma_P \left[\frac{\partial g(\mathbf{P})}{\partial \mathbf{P}}\right]_{\mathbf{P}=\widehat{\mathbf{P}}}}$$

and the Wald statistic for testing $H_0 : NPV_1 = NPV_2$ is

$$T^{WLS}_{NPV} = \frac{(\widehat{NPV}_1 - \widehat{NPV}_2)^2}{Var(\widehat{NPV}_1 - \widehat{NPV}_2)} = \frac{h(\widehat{\mathbf{P}})^2}{n_{\bullet\bullet\bullet}\left[\frac{\partial h(\mathbf{P})}{\partial \mathbf{P}}\right]^T_{\mathbf{P}=\widehat{\mathbf{P}}} \Sigma_P \left[\frac{\partial h(\mathbf{P})}{\partial \mathbf{P}}\right]_{\mathbf{P}=\widehat{\mathbf{P}}}}.$$

Under corresponding null hypotheses H_0, the test statistics T^{WLS}_{PPV} and T^{WLS}_{NPV} are distributed asymptotically as a as the χ^2 distribution with one degree of freedom. Wang et.al [3] provide extensive formulas for variances as a function of counts from Table 1.2 and advocate use of weighted least squares approach with SAS procedure PROC CATMOD for computations of T^{WLS}_{PPV} and T^{WLS}_{NPV}.

1.3 Conditional Model Regression Framework

In this paper we propose to use likelihood approach and conditional model framework to compare predictive values. For i^{th} individual, let D_i denote disease status (1=disease, 0=no disease), T_{1i} indicate test result (1=positive, 0=negative) of test 1, and T_{2i} indicate result (1=positive, 0=negative) of test 2. The likelihood is as follows:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n\bullet\bullet\bullet} P(D_i, T_{1i}, T_{2i}).$$

We decompose the joint probability $P(D_i, T_{1i}, T_{2i})$ into three conditional probabilities

$$P(D_i, T_{1i}, T_{2i}) = P(D_i|T_{1i}, T_{2i})P(T_{2i}|T_{1i})P(T_{1i})$$

and each component is parametrized with a logistic regression model:

$$\begin{aligned} \text{logit P}(\mathrm{D_i} = 1|\mathrm{T_{1i}}, \mathrm{T_{2i}}) &= \gamma_0 + \gamma_1 T_{1i} + \gamma_2 T_{2i} + \gamma_3 T_{1i} \times T_{2i} \\ \text{logit P}(\mathrm{T_{2i}} = 1|\mathrm{T_{1i}}) &= \beta_0 + \beta_1 T_{1i} \\ \text{logit P}(\mathrm{T_{1i}} = 1) &= \alpha \end{aligned}$$

where logit(p) =log $\{p/(1-p)\}$. The likelihood parameter vector $\boldsymbol{\theta}$ denotes $(\alpha, \beta_0, \beta_1, \gamma_0, \gamma_1, \gamma_2, \gamma_3)^T$.

The predictive values for test 1 (PPV_1 and NPV_1) and test 2 (PPV_2 and NPV_2)

can be expressed in terms of $\boldsymbol{\theta}$ as follows:

$$\begin{aligned} PPV_1 &= P(D=1|T_1=1) = \sum_{t=0}^{1} P(D=1|T_1=1,T_2=t)P(T_2=t|T_1=1) \\ &= \frac{1}{1+e^{\beta_0+\beta_1}}\left\{\frac{e^{\beta_0+\beta_1}}{1+e^{-\gamma_0-\gamma_1-\gamma_2-\gamma_3}}+\frac{1}{1+e^{-\gamma_0-\gamma_1}}\right\} \end{aligned}$$

$$\begin{aligned} NPV_1 &= P(D=0|T_1=0) = \sum_{t=0}^{1} P(D=0|T_1=0,T_2=t)P(T_2=t|T_1=0) \\ &= \frac{1}{1+e^{\beta_0}}\left\{\frac{e^{\beta_0}}{1+e^{\gamma_0+\gamma_2}}+\frac{1}{1+e^{\gamma_0}}\right\} \end{aligned}$$

$$\begin{aligned} PPV_2 &= P(D=1|T_2=1) = \frac{\sum_{t=0}^{1} P(D=1|T_1=t,T_2=1)P(T_2=1|T_1=t)P(T_1=t)}{\sum_{t=0}^{1} P(T_2=1|T_1=t)P(T_1=t)} \\ &= \frac{1}{1+e^{-\beta_0}+e^{-\alpha}+e^{-\alpha-\beta_0-\beta_1}}\left\{\frac{1+e^{-\beta_0}}{1+e^{-\gamma_0-\gamma_1-\gamma_2-\gamma_3}}+\frac{e^{-\alpha}+e^{-\alpha-\beta_0-\beta_1}}{1+e^{-\gamma_0-\gamma_2}}\right\}. \end{aligned}$$

$$\begin{aligned} NPV_2 &= P(D=0|T_2=0) = \frac{\sum_{t=0}^{1} P(D=0|T_1=t,T_2=0)P(T_2=0|T_1=t)P(T_1=t)}{\sum_{t=0}^{1} P(T_2=0|T_1=t)P(T_1=t)} \\ &= \frac{1}{1+e^{\beta_0}+e^{-\alpha}+e^{-\alpha+\beta_0+\beta_1}}\left\{\frac{1+e^{\beta_0}}{1+e^{\gamma_0+\gamma_1}}+\frac{e^{-\alpha}+e^{-\alpha+\beta_0+\beta_1}}{1+e^{\gamma_0}}\right\} . \end{aligned}$$

The statistic for testing $H_0 : PPV_1 = PPV_2$ is

$$T_{PPV}^{cond} = \frac{\{PPV_1(\hat{\boldsymbol{\theta}}) - PPV_2(\hat{\boldsymbol{\theta}})\}^2}{Var\{PPV_1(\hat{\boldsymbol{\theta}})\} + Var\{PPV_2(\hat{\boldsymbol{\theta}})\} - 2\cdot Cov\{PPV_1(\hat{\boldsymbol{\theta}}), PPV_2(\hat{\boldsymbol{\theta}})\}} ,$$

with elements of $\hat{\boldsymbol{\theta}}$ obtained form fitting three logistic regressions, and the variances and covariance for PPV_1 and PPV_2 obtained by the delta method. Namely,

$$\begin{aligned} Var\{PPV_1(\hat{\boldsymbol{\theta}})\} &= \left[\frac{\partial PPV_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} Var(\hat{\boldsymbol{\theta}}) \left[\frac{\partial PPV_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ Var\{PPV_2(\hat{\boldsymbol{\theta}})\} &= \left[\frac{\partial PPV_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} Var(\hat{\boldsymbol{\theta}}) \left[\frac{\partial PPV_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ Cov\{PPV_1(\hat{\boldsymbol{\theta}}), PPV_2(\hat{\boldsymbol{\theta}})\} &= \left[\frac{\partial PPV_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} Var(\hat{\boldsymbol{\theta}}) \left[\frac{\partial PPV_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} . \end{aligned}$$

The variance-covariance matrix $Var\hat{\boldsymbol{\theta}}$ is a 7×7 block diagonal matrix with the first 1×1 block formed by variance of α, the second 2×2 block formed by variance-covariance matrix of β_0 and β_1, and the third 4×4 block formed by variance-covariance matrix of

$\gamma_0, \gamma_1, \gamma_2, \gamma_3$. The three blocks can be obtained from standard statistical software with logistic regression option. Derivatives of predictive values are presented in the Appendix.

Similarly, to test $H_0 : NPV_1 = NPV_2$ we consider test statistic

$$T^{cond}_{NPV} = \frac{\{NPV_1(\hat{\boldsymbol{\theta}}) - NPV_2(\hat{\boldsymbol{\theta}})\}^2}{Var\{NPV_1(\hat{\boldsymbol{\theta}})\} + Var\{NPV_2(\hat{\boldsymbol{\theta}})\} - 2 \cdot Cov\{NPV_1(\hat{\boldsymbol{\theta}}), NPV_2(\hat{\boldsymbol{\theta}})\}}$$

with

$$\begin{aligned}
Var\{NPV_1(\hat{\boldsymbol{\theta}})\} &= \left[\frac{\partial NPV_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} Var(\hat{\boldsymbol{\theta}}) \left[\frac{\partial NPV_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\
Var\{NPV_2(\hat{\boldsymbol{\theta}})\} &= \left[\frac{\partial NPV_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} Var(\hat{\boldsymbol{\theta}}) \left[\frac{\partial NPV_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\
Cov\{NPV_1(\hat{\boldsymbol{\theta}}), NPV_2(\hat{\boldsymbol{\theta}})\} &= \left[\frac{\partial NPV_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} Var(\hat{\boldsymbol{\theta}}) \left[\frac{\partial NPV_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} .
\end{aligned}$$

Test statistics T^{cond}_{PPV} and T^{cond}_{NPV} are distributed under corresponding null hypotheses H_0 as the χ^2 distribution with one degree of freedom.

In the next section, the simulation study which compares our method to two other methods mentioned in section 1.2.

1.4 Simulation Study

In this section, we perform simulations to evaluate size and power of tests resulting from the conditional model regression approach. We compare performance to the marginal regression and weighted least squares approaches.

Similarly to Leisenring, et.al [1] we generate data from a scenario, in which dependence of the diagnostic tests T_1, and T_2 does not change with disease status. However, we consider generation of data from a multinomial distribution with cells as in Table 1.2 and we specify dependence of T_1 and T_2 conditional on the disease through an odds ratio OR. This odds ratio, 4 predictive values, and prevalence of disease (Prev) specify the multinomial distribution from which generation is performed. We consider various situations with varied disease prevalence, sample size ($n_{\bullet\bullet\bullet}$), and magnitude of dependency of tests conditional on disease status. Under each scenario, we carried out 10,000 simulations.

Simulation results are summarized by proportion (P_{PPV}) of rejections of $H0 : PPV_1 = PPV_2$ at 0.05 level, proportion (P_{NPV}) of rejections of $H0 : NPV_1 = NPV_2$, and following Leisenring, et.al [1] or Wang et.al [3] we also provide average kappa statistic

(Kappa) for comparison of T_1 and T_2 averaged over all simulations. The results are presented for each of the three methods: our approach (cond), marginal regression approach (GEE), and weighted least squares (WLS).

Table 1.4 considered situation with equal PPVs and equal NPVs. The simulation results are similar for all three methods and confirm the type I error nominal value of 0.05.

Table 1.3: Simulation results with PPV_1=0.75, PPV_2=0.75, NPV_1=0.85, NPV_2=0.85.

Scenario				P_{PPV} (type I error)			P_{NPV} (type I error)		
$n_{\bullet\bullet\bullet}$	Prev	OR	Kappa	Cond	GEE	WLS	Cond	GEE	WLS
200	0.25	1.0	0.265	0.059	0.055	0.059	0.050	0.052	0.052
		4.5	0.464	0.055	0.053	0.055	0.050	0.051	0.051
		18.0	0.644	0.057	0.058	0.060	0.041	0.048	0.048
	0.5	1.0	0.348	0.047	0.048	0.048	0.052	0.052	0.053
		4.5	0.539	0.053	0.053	0.053	0.048	0.051	0.051
		18.0	0.709	0.049	0.050	0.050	0.047	0.050	0.050
	0.7	1.0	0.127	0.050	0.051	0.051	0.068	0.058	0.069
		4.5	0.347	0.049	0.049	0.049	0.057	0.050	0.060
		18.0	0.558	0.046	0.048	0.048	0.051	0.047	0.056
500	0.25	1.0	0.266	0.051	0.051	0.052	0.051	0.051	0.051
		4.5	0.466	0.055	0.053	0.053	0.053	0.053	0.053
		18.0	0.646	0.054	0.052	0.053	0.046	0.049	0.049
	0.5	1.0	0.350	0.051	0.051	0.053	0.053	0.052	0.053
		4.5	0.540	0.050	0.050	0.050	0.049	0.049	0.050
		18.0	0.710	0.046	0.049	0.053	0.050	0.053	0.053
	0.7	1.0	0.128	0.050	0.051	0.051	0.054	0.050	0.054
		4.5	0.353	0.053	0.053	0.053	0.056	0.054	0.057
		18.0	0.561	0.050	0.051	0.051	0.055	0.055	0.057

Table 1.4 summarizes simulation results when we considered different PPVs ($PPV_1 = 0.85$, $PPV_2 = 0.75$) and equal NPVs. Again performance of the three methods is similar. As dependency of tests conditional on disease increases, disease prevalence increases, and the sample size increases, the power increases accordingly.

Table 1.4 summarizes simulation results when we considered different NPVs ($NPV_1 = 0.85$, $NPV_2 = 0.80$) and equal PPVs. Performance of the three methods is similar. Power increases as dependency of tests conditional on disease increases and sample size increases. However, as the disease prevalence increases, the power decreases. This results from fewer patients with no disease with increase in diseases prevalence.

Table 1.4: Simulation results with PPV_1=0.85, PPV_2=0.75, NPV_1=0.85, NPV_2=0.85.

Scenario				P_{PPV} (power)			P_{NPV} (type I error)		
$n_{\bullet\bullet\bullet}$	Prev	OR	Kappa	Cond	GEE	WLS	Cond	GEE	WLS
200	0.25	1.0	0.289	0.209	0.202	0.210	0.052	0.053	0.053
		4.5	0.484	0.218	0.216	0.222	0.050	0.052	0.052
		18.0	0.650	0.264	0.265	0.268	0.045	0.053	0.053
	0.5	1.0	0.408	0.622	0.623	0.623	0.054	0.053	0.054
		4.5	0.552	0.764	0.764	0.763	0.055	0.054	0.055
		18.0	0.683	0.899	0.898	0.898	0.050	0.054	0.055
	0.7	1.0	0.193	0.975	0.976	0.976	0.098	0.089	0.099
		4.5	0.330	0.997	0.997	0.997	0.099	0.093	0.101
		18.0	0.414	1.000	1.000	1.000	0.098	0.093	0.099
500	0.25	1.0	0.289	0.401	0.397	0.401	0.051	0.052	0.052
		4.5	0.484	0.454	0.453	0.454	0.051	0.052	0.052
		18.0	0.650	0.548	0.545	0.546	0.053	0.055	0.055
	0.5	1.0	0.408	0.955	0.955	0.955	0.050	0.050	0.050
		4.5	0.552	0.989	0.999	0.989	0.047	0.049	0.049
		18.0	0.683	0.999	0.999	0.999	0.048	0.048	0.048
	0.7	1.0	0.193	1.000	1.000	1.000	0.063	0.059	0.063
		4.5	0.330	1.000	1.000	1.000	0.061	0.059	0.061
		18.0	0.414	1.000	1.000	1.000	0.064	0.062	0.065

Table 1.5: Simulation results with $PPV_1 = 0.85, PPV_2 = 0.85, NPV_1 = 0.85, NPV_2 = 0.80$.

Scenario					P_{PPV} (power)			P_{NPV} (type I error)		
$n_{\bullet\bullet\bullet}$	Prev	OR		Kappa	Cond	GEE	WLS	Cond	GEE	WLS
200	0.25	1.0	1.0	0.176	0.069	0.059	0.069	0.568	0.569	0.569
		4.5	4.5	0.35	0.064	0.058	0.064	0.729	0.733	0.733
		18.0	18.0	0.50	0.067	0.063	0.068	0.864	0.868	0.867
	0.5	1.0	1.0	0.45	0.053	0.053	0.053	0.180	0.179	0.180
		4.5	4.5	0.59	0.051	0.052	0.052	0.213	0.213	0.215
		18.0	18.0	0.73	0.053	0.056	0.056	0.294	0.298	0.298
	0.7	1.0	1.0	0.45	0.047	0.048	0.048	0.085	0.071	0.084
		4.5	4.5	0.59	0.051	0.054	0.054	0.086	0.076	0.086
		18.0	18.0	0.73	0.048	0.050	0.050	0.074	0.074	0.080
500	0.25	1.0	1.0	0.21	0.054	0.052	0.055	0.929	0.930	0.930
		4.5	4.5	0.39	0.056	0.052	0.056	0.983	0.983	0.983
		18.0	18.0	0.52	0.054	0.053	0.055	0.999	0.999	0.999
	0.5	1.0	1.0	0.45	0.052	0.052	0.052	0.355	0.355	0.356
		4.5	4.5	0.59	0.049	0.049	0.049	0.447	0.447	0.448
		18.0	18.0	0.73	0.048	0.049	0.049	0.623	0.623	0.624
	0.7	1.0	1.0	0.45	0.053	0.053	0.053	0.113	0.106	0.113
		4.5	4.5	0.59	0.053	0.054	0.054	0.114	0.109	0.114
		18.0	18.0	0.73	0.049	0.050	0.050	0.119	0.119	0.121

In summary, the simulation results provide evidence that method based on the conditional model approach performs similarly to the GEE based method (score test) and weighted least squares method.

1.5 Example

We illustrate our method with a cardiology data set presented in Table I which consists of 1465 men included in the Coronary Artery Surgery Study (CASS). The gold standard for coronary artery disease (CAD) is coronary angiography and the data set considers two diagnostic tests: exercise stress test (Test 1) and clinical history of chest pain (Test 2). The data are from a paired study design, in which each subject received both tests. The interest here is to compare positive predictive values and negative predictive values of the two tests. The estimated values of predictive values are: $PPV_1 = 0.881$, $PPV_2 = 894$, $NPV_1 = 0.648$, and $NPV_2 = 0.785$. Test statistic $T^{cond}_{PPV} = 0.8020$ (P-value=0.3705) and $T^{cond}_{NPV} = 23.7254$ (P-value < 0.0001). Hence, based on this data set, we conclude that negative predictive value of clinical history of chest pain (Test 2) is superior to that of exercise stress test (Test 1), but negative predictive values are similar. For comparison, $T^{GEE}_{PPV} = 0.8015$ (P-value=0.3706) and $T^{GEE}_{NPV} = 23.5794$ (P-value < 0.0001), and $T^{WLS}_{PPV} = 0.80$ (P-value=0.3705) and $T^{WLS}_{NPV} = 23.73$ (P-value < 0.0001).

1.6 Discussion

We proposed using a likelihood based conditional model regression approach for comparison of positive predictive values and negative predictive values of two binary diagnostic tests. The proposed method performs similar to the marginal model (GEE) based generalized score statistic [1] and the approach based on Weighted Lease Squares proposed by Wang et.al [3].

Although it was not the focus of this paper, our approach yields to a direct extension to a situation with a partially missing disease status by consideration of one more logistic regression modeling probability of missingness, similarly to a one test situation considered by Kosinski and Barnhart [4].

Chapter 2

Improving the efficiency of testing equality of predictive values using auxiliary covariates

2.1 Introduction

As new technologies emerge, it becomes more available to detect disease as early as possible. It is critical to discover the disease early to cure the disease. Developing accurate diagnostic tests in health care is important in early detection of disease. Some of technologies which are used in health care include X-rays, CT scans, MRI, biopsy, Ultrasound, or biochemical tests. It is necessary to evaluate the performance of a diagnostic test to its gold standard before it is used into practice. Gold standard is defined as a test or condition which gives the definite measure of disease. In this article, we use gold standard and disease interchangeably. There are several measures to evaluate the performance of diagnostic tests and those are introduced in the book of Pepe [5]. We focus on predictive values in this article. The positive predictive value is the proportion of diseased individuals among the positive diagnostic tests and the negative predictive value is the proportion of nondiseased individuals among the negative diagnostic tests. When more than one diagnostic test are available, the interest is to determine the diagnostic test which performs the better. Comparing the predictive values of the diagnostic test is one of possible methods and there have been several papers on comparing diagnostic tests when every individual has the outcome

of the gold standard. [3, 1] However, it is often the case that not every patient undergoes the evaluation of gold standard. The reason can be that the risk of going through gold standard can be higher than the disease itself for some patients. Individuals who undergo the gold standard is not selected by random, but it is rather by physicians' decision. If we only use patients who have gold standard outcomes presented in evaluating the accuracy of diagnostic tests, the estimate of the predictive values of the diagnostic test is likely to be biased which is called verification bias [6]. When there is only one diagnostic test which is of interest and missingness of disease only depends on the one diagnostic test, estimating the predictive value from the observed data may not be biased by the definition [7]. However, it is not true when we have more than one diagnostic tests. Our objectives is to compare the accuracy of two diagnostic tests, which is subject to be biased if one only considers those who have the outcome of gold standard. In some situation, baseline auxiliary covariates, such as demographical and physiological records, which may be related to the disease status or the probability of missing disease are available. For example, recent studies show that the prevalence of breast cancer may be related to the average time of sleep and exercise. It is also highly correlated with ages, smoking habits, or family history of breast cancer. Furthermore, when one uses mammography as gold standard of breast cancer, those who are younger have denser breast tissue which makes difficult to interpret the results and may be recorded as missing for gold standard. We call those prognostic factors as covariates and wish to incorporate in testing equality of predictive values.

Wang [3] and Leisenring [1] compare two diagnostic tests when all the gold standard are observed on every patients. Nofuentes et al. [8] accounts for the missing gold standard, and compared two predictive values with EM algorithm. They also use covariate information in predicting predictive values. To use information from covariates, Nofuentes et al.[8] assume distribution on covariates and suggest to integrate out to have marginal predictive value for the whole population or dichotomize the covariates in order to estimate the predictive values in a given population. However, the distribution of observed covariates is unlikely known, and there would be more information if we use available continuous covariates rather than dichotomizing the covariates. In this article, our objective is to compare two diagnostic tests under the condition that gold standards are missing for some individuals and use information from covariates to have more efficient estimator and powerful test for a given hypothesis test. We assume a typical clinical study. Baseline characteristics are collected on patients and diagnostic tests are measured from to all patients, while the invasive or ex-

pensive gold standard may not be available to some patients. We also assume that subjects in this clinical study are randomly selected from the population since the predictive values depend on disease prevalences of populations. In this paper, we expand the idea of semi-paramteric theory in comparing two diagnostic tests with auxiliary covariates. Section 2.2 describes the notation and the model for predictive value. Section 2.3 derives the estimating equation which results in semiparametric estimator under no missingness and missingness, and Section 2.5 shows the property of the estimator. Section 2.6 illustrates the asymptotic property with its variance and give tests statistics to test our hypothesis of interest. Section 2.4 explains the steps to estimate point estimate. We conclude with simulation study in Section 2.7 and application to real data examples in section 2.8 to study the property of our estimator.

2.2 Model Framework

2.2.1 Notation and Model

We denote the full data with three binary random variables had there been no missing gold standard, $V_i = (D_i, T_{1i}, T_{2i})$, $i = 1 \cdots N$, which are independent and identically distributed. For i^{th} individual, D_i denotes the outcome of the gold standard, T_{1i} for test 1 and T_{2i} for test 2; 1 for positive test outcome and 0 for negative test outcome. Our interest, difference between log-odds ratio of predictive values, can be written as the difference of logit of difference in conditional expectations. $\text{logit}\{\text{E}(\text{D}|\text{T}_1 = 1)\} - \text{logit}\{\text{E}(\text{D}|\text{T}_2 = 1)\}$ for logit $\{\text{PPV}_1\}$ − logit $\{\text{PPV}_2\}$, and $\text{logit}\{\text{E}(\text{D} = 0|\text{T}_1 = 0)\} - \text{logit}\{\text{E}(\text{D} = 0|\text{T}_2 = 0)\}$ for logit $\{\text{NPV}_1\}$−logit $\{\text{NPV}_2\}$, where $\text{logit}\{\text{p}\} = \log\{\text{p}/(1-\text{p})\}$. Since modeling the negative predictive values would be parallel to modeling the positive predictive values, we illustrate how one would estimate and test for the difference between logit of positive predictive values of two diagnostic tests relative to its gold standard. To illustrate our hypothesis in a marginal fashion, let us use the notation from [1] for now. Let Z_i denote the test assignment for individual i either (1 for test 1 and 0 for test 2). Let X_i be the outcome of tests, which is also 0 or 1 corresponding to negative, and positive outcome respectively. D_i is a outcome of gold standard which takes 0 or 1. Since we have two measurements for individual i for test 1 and test 2, we would indicate individuals with ID number. For every one ID, Z_i and X_i have two entries, and Z_i always takes 1 and 0, and X_i changes corresponding to the

outcome of T_{1i} and T_{2i}. For example, if we use vector notation, for individuals who have $T_1 = 1$, and $T_2 = 1$, it translates into $\mathbf{Z} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. For individuals who have the outcome of $T_1 = 1$, and $T_2 = 0$, it would be written as $\mathbf{Z} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. We also assume that $\mathbf{D}$ stays the same within individuals. This notation would allow us to express the estimate of predictive values for each tests in a marginal form. The logistic regression can be expressed as

$$\text{logit}\{\text{E}(\text{D}|\text{Z},\text{X})\} = \acute{\beta}_0 + \acute{\beta}_1\text{Z} + \acute{\beta}_2\text{X} + \acute{\beta}_3\text{ZX}, \tag{2.1}$$

,then $\sum_{i=0}^{3}\beta_i$ is the log-odds ratio of PPV_1 and the $\sum_{i=0,2}\beta_i$ is the log-odds of PPV_2. Since we are interested in estimating and testing for positive predictive values, we subset the data where $X = 1$, which means that we only consider individuals who have at least one positive tests outcome. Then model (2.1) is written as

$$\text{logit}\{\text{E}(\text{D}|\text{Z},\text{X}=1)\} = \beta_0 + \beta_1\text{Z}, \tag{2.2}$$

so β_1 is the difference of log-odds ratio of PPV_1 and PPV_2. The hypothesis of our interest can be written as $H_0 : \beta_1 = 0$. However, as we mentioned, we cannot observe gold standard for some of the patients. Thus, we denote the realization of observed data for individual i with possible missingness as $O_i = (R_i, R_iD_i, T_{1i}, T_{2i}, \boldsymbol{C_i})$ for $i = 1, \cdots, N$, with our original notation. R_i denotes the indicator whether gold standard is observed, and $\boldsymbol{C_i}$ denotes the vector of covariates, which may be correlated with the gold standard. When $R_i = 1$, one can observe $(D_i, T_{1i}, T_{2i}, \boldsymbol{C_i})$, and if $R_i = 0$, $(T_{1i}, T_{2i}, \boldsymbol{C_i})$ are observed. Since we cannot observe missing gold standard, we assume Missing at Random, that is, R is independent from D given (T_1, T_2, C). With the assumption of MAR, but no further assumption on distribution, we would like to derive a class of semiparametric estimators which are consistent and asymptotically normal for $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ which is efficient by exploiting the correlation between C and D using the results from the semiparametric theory. We first derive estimating equations with mean zero which leads to consistent and asymptotically normal estimator [2, 9]. We then derive score and Wald statistics from the estimating equations.

2.2.2 Semiparametric Theory

By using semiparametric theory [9], we will develop regular and asymptotically linear estimators (RAL) that have the form,

$$N^{1/2}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta_0}) = N^{-1/2}\sum_{i=1}^{N}\varphi_i + o_p(1) \tag{2.3}$$

for random variable φ_i which satisfies $E(\varphi) = 0$, and $E(\varphi\varphi^T)$ is finite and nonsingular under the truth, $\boldsymbol{\beta_0}$ where $o_p(1)$ means convergence in probability to zero as N goes to infinity. i^{th} function φ_i is referred as an i^{th} influence function of the estimator $\hat{\boldsymbol{\beta}}_N$.
RAL estimators have properties as asymptotically normal which can be shown with the central limit theorem and their variances are equal to the variance of their uniquely defined influence function, i.e., $E(\varphi\varphi^T)$. We find the estimator of the form by deriving the class of all unbiased estimating equations for $\boldsymbol{\beta}$ based on available data. As the theory of Robins and Rotnitzky [10], we find the unbiased estimating equation for the parameter $\boldsymbol{\beta}$ when the gold standard is MAR. The steps are as follows: First, we derive the full data unbiased estimating equations when there are no missing data. Second, when the D is missing under the assumption MAR, we find the Inverse Probability Weighted (IPW) estimating equations to correct the verification bias, and characterize the class of Inverse Probability Weighted estimating equations. Finally, we derive the most efficient estimator in the class using all available data, which is called an augmented estimator.

2.3 Estimators

2.3.1 Full data Estimating Equation

In the section, from the logistic regression (2.2) we first derive the estimating equation when there is no missing gold standard, and when the auxiliary covariates are not introduced.

Proposition 2.3.1 *We have estimator $\hat{\boldsymbol{\beta}}_N$ which is defined as the solution of an estimating equation, $\sum_{i=1}^{N} m(D_i, T_{1i}, T_{2i}) = 0^{dim(\boldsymbol{\beta})}$, where the estimating function is written as*

$$m(D, T_1, T_2) = A(T_1, T_2)^{2\times 2}\begin{Bmatrix} D - expit\{\beta_0 + \beta_1\} \\ D - expit\{\beta_0\} \end{Bmatrix}, \tag{2.4}$$

where

$$A(T_1, T_2) = \begin{Bmatrix} I(T_1 = 1) & I(T_2 = 1) \\ I(T_1 = 1) & 0 \end{Bmatrix},$$

and $expit(x) = exp(x)/(1 + exp(x))$.

The proposition can be derived from restricting moments in the model (2.2). Positive predictive values for each test use individuals who have at least one result of diagnostic tests where $X_i = 1$. Those who have at least one positive diagnostic test outcome are grouped into three categories. One is with individuals who result in positive outcome from both of tests ($T_1 = 1, T_2 = 1$), the second group is with only first test positive ($T_1 = 1$), and the last is those who have only the second test positive ($T_2 = 1$). Then from the (2.2), three groups can be represented with three different restricted moments.
For the first group,

$$\text{logit}\{\text{E(D}|\text{Z} = 1, \text{X} = 1)\} = \beta_0 + \beta_1, \ \text{logit}\{\text{E(D}|\text{Z} = 0, \text{X} = 1)\} = \beta_0, \tag{2.5}$$

The second group,

$$\text{logit}\{\text{E(D}|\text{Z} = 1, \text{X} = 1)\} = \beta_0 + \beta_1. \tag{2.6}$$

The third group,

$$\text{logit}\{\text{E(D}|\text{Z} = 0, \text{X} = 1)\} = \beta_0. \tag{2.7}$$

Let $\mu_1 = expit(\beta_0 + \beta_1)$, $\mu_2 = expit(\beta_0)$, and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Then a solution to the estimating equation $\sum_{i=1}^{N} m(D_i, T_{1i}, T_{2i}; \boldsymbol{\beta})$ with an arbitrary choice of $A(T_{1i}, T_{2i})$ is the semiparametric estimator for the restricted moments.

$$m(D, T_{1i}, T_{2i}; \boldsymbol{\beta}) = A(T_{1i}, T_{2i})^{2\times 2} \begin{Bmatrix} D_i - \mu_1 \\ D_i - \mu_2 \end{Bmatrix}.$$

When $A(T_{1i}, T_{2i}) = \left\{ \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} \right\}^T V_i^{-1}$ with working covariance matrix, V_i, gives the optimal estimator when all the data are observed [2, 9]. V_i is a working covariance matrix which characterizes the correlation within individual i. We have set V_i by assuming independent correlation within individual, and derive the proposition 2.3.1. Justification of independent covariance matrix follows from the Pepe-Anderson condition in Pepe [11]. That is,

$$P(D = 1|T_1, T_2) = P(D = 1|T_1).$$

When non-diagonal covariance matrix is used, the condition should be met to lead a consistent estimator. However, (T_1, T_2) together is more informative than T_1 or T_2 itself in prediction of D, so the condition does not hold. For the individuals who have both of T_1 and T_2 positive outcomes, this estimating equation consists of two rows of equations in solving for β_0 and β_1. When only T_1 is positive, only first row in (2.4) is included and when only T_2 is positive, the second equation is included into the estimating equation while none of the equation will be used when neither T_1 nor T_2 are positive. Since the estimating equation is unbiased with its expectation equal to zero, the estimator by solving the equation is consistent, asymptotically normal (CAN) estimator [2]. Furthermore, among the estimators based on (D_i, T_{1i}, T_{2i}), for $i = 1, \cdots, N$, solving the above estimating equation gives not only the efficient estimator which results in the smallest variance, but the most powerful test of $H_0 : \beta_1 = 0$.

2.3.2 Inverse Probability Estimator

However, there are situations when gold standard for some patients is missing. This situation may occur when the physicians do not allow patients to go through diagnostic test because their risk from conducting diagnostic tests are higher than the risk from the disease. In this case, the estimating equation in (2.4) may be biased, so inverse probability weighting (IPW) method which divides the (2.4) by the probability of observing the individual, π, can be used to correct the bias [10].

With the potential missing gold standard for some patients, we denote the data as $O_i = (R_i, RD_i, T_{1i}, T_{2i}, \boldsymbol{C_i})$, for $i = 1, \cdots, N$ where R_i is missing gold standard indicator. The only assumption we have in deriving IPW estimator is that missingness is independent of D conditional on (T_1, T_2, C), which is the Missing at Random (MAR) assumption. The probability of missingness can be written as

$$P(R = 1|D, T_1, T_2, C) = P(R = 1|T_1, T_2, C) = \pi(T_1, T_2, C). \tag{2.8}$$

This means the probability of missingness only depends on observed quantities (T_1, T_2, C). There are some cases that missingness occurrs by design, and then we would know the conditional probability in (2.8), but in most of cases the probability is unknown. Since R is binary, the natural model for missing mechanism is logistic regression with a known

function of g, and finite dimensional parameter $\boldsymbol{\gamma}$,

$$\text{logit P(R}|\text{T}_1, \text{T}_2, \text{C}) = \text{g(T}_1, \text{T}_2, \text{C}; \boldsymbol{\gamma}). \tag{2.9}$$

To have unbiased estimator for $\boldsymbol{\beta}$, we weight the estimating equation with the inverse of the probability that the individual would be verified (i.e. π) and solve the inverse probability weighted estimating equation, which is

$$\sum_{i=1}^{N} \frac{R_i}{\pi(T_{1i}, T_{2i}, \boldsymbol{C_i}, \boldsymbol{\gamma})} m(D_i, T_{1i}, T_{2i}). \tag{2.10}$$

We denote the IPW estimator as $\boldsymbol{\beta}^{IPW}$. The IPW estimating equation is consistent as long as we model the missing data mechanism correctly. We denote the true parameter of $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}_0$ and the true probability of missingness as $\pi(T_1, T_2, C, \boldsymbol{\gamma}_0) = \pi_0$.
The estimator by solving (2.8) is consistent since the expectation of that is equal to zero. It also uses the prognostic factors(covariates) in modeling the probability of observing disease for each individual i, and we denote the probability as π_i. However, this estimator only uses the complete data where gold standard is observed, $R = 1$, and may not be an efficient estimator. Thus, we extend this IPW into the estimator which incorporates all the data and possibly gains efficiency. In the next section, not only we correct the verification bias as it was done for $\boldsymbol{\beta}^{IPW}$, but we also incorporate extra covariates, C, to take advantage of extra information from the correlation between D and C which results in the optimal estimator in the class of estimators.

2.3.3 Augmented Estimating Equation

The observed full data is $O_i = (R_i, RD_i, T_{1i}, T_{2i}, \boldsymbol{C_i})$, and we are to derive semi-parametric model over as large a class as possible which could have generated the data. The only restriction over the class is that the missingness is independent of disease conditional on diagnostic tests and covariates of the patient, which is equivalent to the Missing at Random (MAR) assumption, $R \perp D|\ T_1, T_2, C$. One of the key results in [9] is that the class of all unbiased estimating equations for $\boldsymbol{\beta}$ using all the available data may be written as $\sum_{i=1}^{N} m^* = 0$, where

$$m^* = \frac{R}{\pi(\boldsymbol{\gamma})} m(D, T_1, T_2) - \frac{R - \pi(\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma})} L(T_1, T_2, C), \tag{2.11}$$

and $L(T_1, T_2, C)$ is an arbitrary $dim(\boldsymbol{\beta})$-dimensional function of covariates and diagnostic tests. Recall that $E(m(D, T_1, T_2))$ is zero, and the additional term in (2.11) has expectation zero since we assume $R \perp D|\ T_1, T_2, C$. It follows with iterative conditional expectation,

$$E\left[E\left\{\frac{R-\pi(\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma})}L(T_1,T_2,C)|T_1,T_2,C\right\}\right] = E\left[\frac{P(R=1|T_1,T_2,C)-\pi(\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma})}L(T_1,T_2,C)\right] \tag{2.12}$$

$$= E\left\{\frac{P(R=1|T_1,T_2,C)-P(R=1|T_1,T_2,C)}{P(R=1|T_1,T_2,C)}L(T_1,T_2,C)\right\} = 0.$$

Thus, estimating function, m^* in (2.11) is unbiased estimating equation based on observed data $O = (R, RD, T_1, T_2, C)$. When the arbitrary function, $L(T_1, T_2, C)$ is taken as 0, then the above equation (2.11) is the same as IPW estimating equation (2.10) which does not make use of the cases where the gold standard is not observed. The objective is to find consistent, and the most efficient estimator for β in the class (2.11) for $\boldsymbol{\beta}$ is achieved by selecting the appropriate function of T_1, T_2 and, C, $L(T_1, T_2, C)$. With M-estimator theory [12], the asymptotic variance-covariance estimator from the unbiased estimating equation is $A^{-1}BA^{-1}$, where $A = E\left\{-\partial m^*(D, T_1, T_2)/\partial \boldsymbol{\beta}^T\right\}|_{\boldsymbol{\beta}=\boldsymbol{\beta_0}}$ and $B = E\left\{m^* m^{*T}\right\}|\boldsymbol{\beta} = \boldsymbol{\beta_0}$, and $\boldsymbol{\beta_0}$ is the true value of $\boldsymbol{\beta}$. Thus, the optimal estimating equation in the class (2.11) leads to B_{opt}, where $B - B_{opt}$ is nonnegative definite. It is shown that taking $L(T_1, T_2, C) = E\left\{m(D, T_1, T_2)|T_1, T_2, C\right\}$ gives B_{opt} [9]. Thus, the optimal estimating equation is written as

$$\sum_{i=1}^{N}\left[\frac{R_i}{\pi_i(\boldsymbol{\gamma})}m(D_i, T_{1i}, T_{2i}) - \frac{R_i-\pi_i(\boldsymbol{\gamma})}{\pi_i(\boldsymbol{\gamma})}E\left\{m(D_i, T_{1i}, T_{2i})|T_{1i}, T_{2i}, \boldsymbol{C_i}\right\}\right] = 0. \tag{2.13}$$

The conditional expectation in the augmented term in (2.13) is with respect to the conditional distribution of D given (T_1,T_2,C). Unless we know this distribution, $E(D|T_1, T_2, C)$, the quantity is unknown and we model it with finite dimensional parameter $\boldsymbol{\alpha}$. Since D is a binary random variable, the natural model to fit is logistic regression with some known function h, and finite dimensional parameter, α.

$$\text{logit P(D = 1|T}_1, \text{T}_2, \text{C}) = \text{h(T}_1, \text{T}_2, \text{C}, \boldsymbol{\alpha}), \tag{2.14}$$

and we call it conditional model. Let $P(D = 1|T_1, T_2, C, \boldsymbol{\alpha}) = \eta(\boldsymbol{\alpha})$, then the augmented estimating equation is written as $\sum_{i=1}^{N} m^*(D_i, T_{1i}, T_{2i})$, which we can rewrite the estimating

equations in (2.11) as

$$m^*(D, T_1, T_2) = A(T_1, T_2, \boldsymbol{\beta}^{2\times 2} \left\{ \begin{array}{c} \tilde{D} - expit\{\beta_0 + \beta_1\} \\ \tilde{D} - expit\{\beta_0\} \end{array} \right\},$$

and $\tilde{D} = (1 - R) \times \eta(\boldsymbol{\alpha}) + R \times (\frac{D}{\pi(\boldsymbol{\gamma})} + \eta(\boldsymbol{\alpha}) \times (1 - \frac{1}{\pi(\boldsymbol{\gamma})}))$. Thus, when the disease status is observed, $R = 1$, the weighted average of D and $\eta(\boldsymbol{\alpha})$ with the weight of $\pi(\boldsymbol{\gamma})$ is used, and when the disease is not observed, $R = 0$, $\eta(\boldsymbol{\alpha})$ is used as a response for the augmented estimating equation. We denote the estimator from this estimating equation as $\boldsymbol{\beta}^{Aug}$. In the simulation study, we will show that the estimator, $\boldsymbol{\beta}^{Aug}$, is unbiased and more efficient than $\boldsymbol{\beta}^{IPW}$ while the complete case estimator $\boldsymbol{\beta}^{Com}$ which uses individuals who are observed with D, is biased and the IPW estimator, $\boldsymbol{\beta}^{IPW}$, is unbiased.

2.4 Algorithm to Solve the Augmented Equation

The augmented estimator involves conditional model and the missingness model when they are unknown. Thus, it requires an adaptive algorithm in that we first model conditional and missingness model and use the estimates in the augmented estimating equation. Let X be the regressors in the missingness model in (2.9) and conditional model in (2.14), and use linear model in the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ respectively. The regressor $\mathbf{X}$ can be different for missingness and conditional model, but in simplicity let us use $\mathbf{X}$ for now.

$$\text{logit } P(R = 1|\mathbf{X}) = \boldsymbol{\gamma}^T\mathbf{X},$$
$$\text{logit } P(D = 1|\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$$

1. Find the MLE $\hat{\boldsymbol{\gamma}}_N$ of the $\boldsymbol{\gamma}$, which is the solution to the score vector.

$$S_{\boldsymbol{\gamma}}^{dim(\boldsymbol{\gamma})^T}(\boldsymbol{\gamma}_0) = \frac{\partial logL}{\partial \boldsymbol{\gamma}^T} = \sum (1, X_i^T)^T (R_i - \frac{exp\{\boldsymbol{\gamma}^T X_i\}}{1 + exp\{\boldsymbol{\gamma}^T X_i\}}) = \mathbf{0}$$

2. Find the estimate of $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}}_N$ in the conditional density of D on $\mathbf{X}$ by solving the score vector.

$$S_{\boldsymbol{\alpha}}^{dim(\boldsymbol{\alpha})^T}(\boldsymbol{\alpha}_0) = \frac{\partial logL}{\partial \boldsymbol{\alpha}^T} = \sum (1, X_i^T)^T (D_i - \frac{exp\{\boldsymbol{\alpha}^T \mathbf{X}\}}{1 + exp\{\boldsymbol{\alpha}^T \mathbf{X}\}}) = \mathbf{0}$$

3. The estimate of the parameter of the interest, $\boldsymbol{\beta}$, is obtained by solving the augmented estimating equation, $\sum_{i=1}^{N} m^*(\tilde{D}_i, T_{1i}, T_{2i})$, where

$$m^*(\tilde{D}, T_1, T_2) = A(T_1, T_2, \boldsymbol{\beta})^{2\times 2} \left\{ \begin{array}{c} \tilde{D} - expit\,\{\beta_0 + \beta_1\} \\ \tilde{D} - expit\,\{\beta_0\} \end{array} \right\},$$

and $\tilde{D} = (1 - R) \times \eta(\hat{\boldsymbol{\alpha}}_N) + R \times (\frac{D}{\pi(\hat{\boldsymbol{\gamma}}_N)} + \eta(\hat{\boldsymbol{\alpha}}_N) \times (1 - \frac{1}{\pi(\hat{\boldsymbol{\gamma}}_N)}))$.

2.5 Property of the Augmented Estimator

In previous sections, we considered the class of estimating equations under MAR, and then presented the estimator which is consistent and the most efficient among the class of estimators. We have posited two additional models along with the restriction of the moment of parameters of our interest. In the cases in which the missingness is not by design, we have posited model to estimate the parameter $\boldsymbol{\gamma}$ with logistic regression,

$$\text{logit P(R = 1|T}_1, \text{T}_2, \boldsymbol{C}, \boldsymbol{\gamma}) = \text{g(T}_1, \text{T}_2, \boldsymbol{C}, \boldsymbol{\gamma})$$

The other model, which we call conditional model, attempts to bring efficiency in estimating $\boldsymbol{\beta}$ through the possible dependency between D and (T_1, T_2, C) and it has been posited with logistic regression,

$$\text{logit P(D = 1|T}_1, \text{T}_2, \boldsymbol{C}, \boldsymbol{\alpha}) = \text{h(T}_1, \text{T}_2, \boldsymbol{C}, \boldsymbol{\alpha}).$$

So far, we have taken the stand that missingness model is correctly specified, when the augmented term is zero in (2.12); which means that the proposed model (2.15) contains true parameter, $\boldsymbol{\gamma}_0$, in specifying the probability of missingness,

$$P(R = 1|T_1, T_2, \boldsymbol{C}, \hat{\boldsymbol{\gamma}}_N) = \pi(\hat{\boldsymbol{\gamma}}_N) \xrightarrow{p} \pi(\boldsymbol{\gamma}_0) \tag{2.15}$$

However, in most cases, we do not know the true probability unless the missingness is by design, so $\hat{\boldsymbol{\gamma}}_N$ may converge to some other constant $\boldsymbol{\gamma}^*$ under some regularity conditions,

$$\pi(\hat{\boldsymbol{\gamma}}_N) \xrightarrow{p} \pi(\boldsymbol{\gamma}^*) \neq \pi(\boldsymbol{\gamma}_0).$$

Likewise, conditional model may be incorrectly specified, and the estimate $\hat{\boldsymbol{\alpha}}_N$ may converge to some other value $\boldsymbol{\alpha}^*$ under some regularity conditions,

$$\pi(\hat{\boldsymbol{\alpha}}_N) \xrightarrow{p} \pi(\boldsymbol{\alpha}^*) \neq \pi(\boldsymbol{\alpha}_0).$$

In this section, it will be shown that the estimator, $\hat{\boldsymbol{\beta}}^{Aug}$ has double protection from misspecification of the models, which means that it is robust to the misspecification to the missing model as long as the conditional model is correctly specified. At the same time, it is robust to the misspecification to the conditional model as long as the missing model is correctly specified. This property is called double robustness [9]. The estimator is consistent and asymptotically normal (CAN) if the expectation of the estimating equation is zero, $E(m^*(R, D, RD, T_1, T_2)) = 0$. To prove CAN of the estimator, $\hat{\boldsymbol{\beta}}^{Aug}$ with the parameters in the equation which converge as $\hat{\boldsymbol{\gamma}}_N \xrightarrow{p} \boldsymbol{\gamma}^*$, and $\hat{\boldsymbol{\alpha}}_N \xrightarrow{p} \boldsymbol{\alpha}^*$, we have to show that

$$E\left[\frac{R}{\pi(\boldsymbol{\gamma}^*)} m(D, T_1, T_2, \boldsymbol{\beta_0}) - \left\{\frac{R - \pi(\boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma}^*)}\right\} E(m(D, T_1, T_2, \boldsymbol{\alpha}^*)|T_1, T_2, C)\right] = 0.$$

First, when the missing model is correctly specified, but the conditional model may be misspecified, we have

$$\begin{aligned} \pi(\boldsymbol{\gamma}^*, T_1, T_2, \boldsymbol{C}) = \pi(\boldsymbol{\gamma}_0, T_1, T_2, \boldsymbol{C}) = P(R = 1|T_1, T_2, \boldsymbol{C}), \text{and} \\ \eta(\boldsymbol{\alpha}^*, T_1, T_2, \boldsymbol{C}) \neq \eta(\boldsymbol{\alpha}_0, T_1, T_2, \boldsymbol{C}) = P(D = 1|T_1, T_2, \boldsymbol{C}). \end{aligned} \tag{2.16}$$

After we add and subtract $m(D, T_1, T_2, \boldsymbol{\beta_0})$ to the expectation, we have

$$E\left[m(D, T_1, T_2, \boldsymbol{\beta_0}) + \left\{\frac{R - \pi(\boldsymbol{\gamma}_0)}{\pi(\boldsymbol{\gamma}_0)}\right\} \{m(D, T_1, T_2, \boldsymbol{\beta_0}) - E(m(D, T_1, T_2, \boldsymbol{\alpha}^*)|T_1, T_2, C)\}\right]. \tag{2.17}$$

Since $E\{m(D, T_1, T_2, \boldsymbol{\beta_0})\} = 0$ and if we take conditional expectation on (D, T_1, T_2, C), the above equation (2.17) is equivalent to

$$E\left[\left\{\frac{E(R|D, T_1, T_2, C) - \pi(\boldsymbol{\gamma}_0)}{\pi(\boldsymbol{\gamma}_0)}\right\} \{m(D, T_1, T_2, \boldsymbol{\beta_0}) - E(m(D, T_1, T_2, \boldsymbol{\alpha}^*)|T_1, T_2, C)\}\right]. \tag{2.18}$$

Due to MAR and assumption of correction specification in missingness model, (2.17), $E(R|D, T_1, T_2, C) = E(R|T_1, T_2, C) = \pi(\boldsymbol{\gamma}_0)$, the expectation in (2.18) becomes zero.
Thus, as long as the missingness model is correctly specified, the parameter, $\hat{\boldsymbol{\beta}}^{Aug}$ from the augmented estimating equation in (2.11) is consistent.

Second, let us think about when the conditional density is correctly specified, and missingness mechanism is not, which is

$$\begin{aligned} \pi(\boldsymbol{\gamma}^*, T_1, T_2, C) \neq \pi(\boldsymbol{\gamma}_0, T_1, T_2, C) = P(R = 1|T_1, T_2, C), \\ \eta(\boldsymbol{\alpha}^*, T_1, T_2, C) = \eta(\boldsymbol{\alpha}_0, T_1, T_2, C) = P(D = 1|T_1, T_2, C), \end{aligned} \tag{2.19}$$

and the second assumption in (2.19) is the same as

$$E\left[m(D|T_1,T_2,C,\boldsymbol{\alpha}^*)\right] = E\left[m(D|T_1,T_2,C,\boldsymbol{\alpha}_0)\right].$$

Again, we need to show that the estimating equation is equal to zero, $E(m^*(R,D,RD,T_1,T_2)) = 0$.

$E\left[m^*(R,D,RD,T_1,T_2)\right]$

$$=E\left[m(D,T_1,T_2,\boldsymbol{\beta_0}) + \left\{\frac{R-\pi(\boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma}^*)}\right\}\left\{m(D,T_1,T_2,\boldsymbol{\beta_0}) - E(m(D,T_1,T_2,\boldsymbol{\alpha}^{(0)}))|T_1,T_2,C)\right\}\right].$$

By the law of iterated expectation, we first take conditional expectation on (R,T_1,T_2,C),

$$0+\left\{\frac{R-\pi(\boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma}^*)}\right\}\left[E\left\{m(D,T_1,T_2,\boldsymbol{\alpha}^*)|R,T_1,T_2,C\right\} - E\left\{m(D,T_1,T_2,\boldsymbol{\alpha}^{(0)}))|R,T_1,T_2,C)\right\}\right]$$

$$=\left\{\frac{R-\pi(\boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma}^*)}\right\}\left[E\left\{m(D,T_1,T_2,\boldsymbol{\alpha}_0)|T_1,T_2,C\right\} - E\left\{m(D,T_1,T_2,\boldsymbol{\alpha}^{(0)}))|T_1,T_2,C)\right\}\right] = 0.$$

The equality takes place since we have MAR assumption, D is independent of R given covariates, (T_1,T_2,C).

Thus, in summary, we have shown that the estimator from the augmented estimating equation is doubly robust in that it is robust to the misspecification of the missing model and to the misspecification of the conditional model as long as the other model is correctly specified.

2.6 Variance and Hypothesis testing of the Estimator

In this section, we derive the variance of the estimator of our interest, $\hat{\boldsymbol{\beta}}^{Aug}$. In the augmented equation, not only we have to estimate the parameter $\boldsymbol{\beta}$, but the estimation also contains unknown parameter $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$, which are nuisance parameters. Thus, we account for uncertainty in estimating nuisance parameters in estimating parameter of interest. We denote the estimator as $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ and eliminate the subscript N for simplicity. We will present two estimators of the variance, one is to estimate the variance by deriving influence function using semiparametric theory in section 2.2.2. The other method is to use sandwich variance estimator method for M-estimator [12].

2.6.1 When $\boldsymbol{\alpha}$ is known

Now, we want to derive the influence function for the $\boldsymbol{\beta}^{\hat{Aug}}$ in order to derive the asymptotic variance of it.

Proposition 2.6.1 *The variance of* $\hat{\boldsymbol{\beta}}^{Aug}$ *in a sandwich form is*

$$var_{gg}(\hat{\boldsymbol{\beta}}^{Aug}) = E(\frac{\partial m^*}{\partial \boldsymbol{\beta}^T})^{-1} \times [\frac{1}{N}\sum_{i=1}^{N} gg^T] \left[E(\frac{\partial m^*}{\partial \boldsymbol{\beta}^T})^{-1}\right]^T \tag{2.20}$$

where,

$$g = m^* - E[m^* S_{\boldsymbol{\gamma}}^T(\boldsymbol{\gamma}_0)](E[S_{\boldsymbol{\gamma}} S_{\boldsymbol{\gamma}}^T])^{-1} S_{\boldsymbol{\gamma}}.$$

Recall that the estimator, $\hat{\beta}^{Aug}$ can be obtained from $\sum_{i=1}^{N} m^*(\tilde{D}_i, T_{1i}, T_{2i}) = 0$, where

$$m^*(\tilde{D}_i, T_{1i}, T_{2i}, \boldsymbol{\beta}) = \frac{R_i}{\pi_i(\boldsymbol{\gamma})} m(D_i, T_{1i}, T_{2i}, \boldsymbol{\beta}) - \left[\frac{R_i - \pi_i(\boldsymbol{\gamma})}{\pi_i(\boldsymbol{\gamma})}\right] L(T_{1i}, T_{2i}, \boldsymbol{C_i}).$$

The derivation of variance of the estimator from the augmented estimating equations is as follows. We first derive influence function had we consider $\boldsymbol{\gamma}$ parameter as a fixed value, and then extend the influence function when $\boldsymbol{\gamma}$ is a unknown parameter.
By keeping $\boldsymbol{\gamma}$ as known, we can expand the estimating equation, m^* about the true value $\boldsymbol{\beta_0}$ and have the form

$$\begin{aligned} &N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) = \\ &N^{-1/2}\sum [N^{-1}\sum \frac{\partial m^*(\boldsymbol{\beta}^*, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}^T}]^{-1} m^*(\boldsymbol{\beta_0}, \boldsymbol{\gamma}) + o_p(1), \end{aligned} \tag{2.21}$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta_0}$and $\hat{\boldsymbol{\beta}}$ and with some regularity condition we have

$$[N^{-1}\sum \frac{\partial m^*(\boldsymbol{\beta}^*, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}^T}]^{-1} \xrightarrow{p} [E(\frac{\partial m^*(\boldsymbol{\beta_0}, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}^T})]^{-1}.$$

Let $\tilde{\varphi}(\boldsymbol{\gamma}) = [E(\frac{\partial m^*(\boldsymbol{\beta_0}, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}^T})]^{-1} m^*(\boldsymbol{\beta_0}, \boldsymbol{\gamma})$, then we rewrite (2.21) as

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) = N^{-1/2}\sum_{i=1}^{N} \tilde{\varphi}_i(\boldsymbol{\gamma}) + o_p(1), \tag{2.22}$$

which is the form of (2.3) and $\tilde{\varphi}_i(\boldsymbol{\gamma})$ is an i^{th} influence function of $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\gamma}$ is known.
Next, we consider γ as unknown and we expand $\boldsymbol{\gamma}$ in (2.21) about the true value of the parameter, $\boldsymbol{\gamma}_0$

$$\begin{aligned} N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) &= N^{-1/2}\sum_{i=1}^{N} \{\tilde{\varphi}_i(\boldsymbol{\gamma}_0) + [N^{-1}\sum \frac{\partial \tilde{\varphi}(\gamma_n^*)}{\partial \boldsymbol{\gamma}^T}] N^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\} + o_p(1) \\ &= N^{-1/2}\sum_{i=1}^{N} \{\tilde{\varphi}_i(\boldsymbol{\gamma}_0) + E(\frac{\partial \tilde{\varphi}(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}^T}) N^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\} + o_p(1). \end{aligned} \tag{2.23}$$

Again, the equality follows under some regularity conditions since we have $N^{-1}\sum \frac{\partial\tilde{\varphi}(\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}^T} \xrightarrow{p} E(\frac{\partial\tilde{\varphi}(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}^T})$.

Note that if we know the influence function of $\hat{\boldsymbol{\gamma}}$, then we can replace $N^{1/2}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$ with its influence function. Furthermore, and we know the $E(\frac{\partial\tilde{\varphi}(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}^T})$ in a simpler manner as in lemma 2.6.2 whose proof is in page 203 of the *Semiparametric Theory and Missing Data* [9]

Lemma 2.6.2

$$E(\frac{\partial\tilde{\varphi}(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}^T}) = -E[\tilde{\varphi}(\boldsymbol{\gamma}_0)S_{\boldsymbol{\gamma}_0}^T] \tag{2.24}$$

Let us use the maximum likelihood method in estimating $\boldsymbol{\gamma}$.
The likelihood is where $\boldsymbol{C_i}$ is the vectors of covariates for the individual i, and let $X_i = (T_{1i}, T_{2i}, \boldsymbol{C}_i^T)$.

$$L = \prod_{i=1}^{N} \pi(X_i,\boldsymbol{\gamma})^{R_i}\{1-\pi(X_i,\boldsymbol{\gamma})\}^{1-R_i}$$

Then the score vector with respect to $\boldsymbol{\gamma}$ is

$$S_{\boldsymbol{\gamma}}^{2\times1}(\boldsymbol{\gamma}_0) = \frac{\partial logL}{\partial\boldsymbol{\gamma}^T} = \sum(1, X_i^T)^T(R_i - \frac{exp^{\{(\boldsymbol{\gamma}^T X_i)\}}}{1+exp^{(\boldsymbol{\gamma}^T X_i)}}).$$

By expanding the score vector about $\boldsymbol{\gamma}_0$, the influence function of $\hat{\boldsymbol{\gamma}}$ can be given by $(E[S_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0)S_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0)^T])^{-1}S_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0)$, which is

$$N^{1/2}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0) = \sum_{i=1}^{N}(E[S_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0)S_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0)^T])^{-1}S_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0) + o_p(1). \tag{2.25}$$

Thus, the influence function of $\hat{\boldsymbol{\beta}}$ is equal to the following by plugging lemma 2.6.2 and the right hand side of (2.25) into (2.23)

$$\begin{aligned} &N^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}) = \\ &N^{-1/2}\sum\tilde{\varphi}(\gamma_0) - E[\tilde{\varphi}(\gamma_0)S_{\gamma_0}^T]((E[S_{\gamma_0}S_{\gamma_0}^T])^{-1}S_{\gamma_0}) + o_p(1) \end{aligned} \tag{2.26}$$

Finally, the variance of the $\hat{\boldsymbol{\beta}}$ is the variance of the influence function, whose i^{th} influence function is given as

$$\tilde{\varphi}(\boldsymbol{\beta_0},\boldsymbol{\gamma}_0) - E[\tilde{\varphi}(\boldsymbol{\beta_0},\boldsymbol{\gamma}_0)S_{\boldsymbol{\gamma}}^T(\boldsymbol{\gamma}_0)]((E[S_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0)S_{\boldsymbol{\gamma}}^T(\boldsymbol{\gamma}_0)])^{-1}S_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0)). \tag{2.27}$$

Where,

$$\tilde{\varphi}(\boldsymbol{\beta_0}, \boldsymbol{\gamma}_0) = \left[E(\frac{\partial m^*(\boldsymbol{\beta_0}, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\beta}^T}) \right]^{-1} m^*(\boldsymbol{\beta_0}, \boldsymbol{\gamma}_0),$$

$$m^*(D, T_{1i}, T_{2i}, \boldsymbol{\beta_0}, \boldsymbol{\gamma}_0) = A(T_{1i}, T_{2i}) \left[\frac{R_i}{\pi(\boldsymbol{\gamma}_0)} m(\boldsymbol{\beta_0}) - (\frac{R_i - \pi(\boldsymbol{\gamma}_0)}{\pi(\boldsymbol{\gamma}_0)} \right] L(T_{1i}, T_{2i}, \boldsymbol{C_i})).$$

Thus, variance can be written as the proposition 2.6.1 with its estimator as

$$\widehat{var_{gg}}(\widehat{\boldsymbol{\beta}}) = \widehat{E}(\frac{\partial m^*}{\partial \boldsymbol{\beta}^T})^{-1} \times [\frac{1}{N} \sum_{i=1}^{N} gg^T](\widehat{E}(\frac{\partial m^*}{\partial \boldsymbol{\beta}^T})^{-1})^T \tag{2.28}$$

where,

$$g = m^* - \widehat{E}[m^* S_{\boldsymbol{\gamma}}^T](\widehat{E}[S_{\boldsymbol{\gamma}} S_{\boldsymbol{\gamma}}^T])^{-1} S_{\boldsymbol{\gamma}}$$

$$\widehat{E}(\frac{\partial m^*}{\partial \boldsymbol{\beta}^T}) = N^{-1} \sum_{i=1}^{N} (\frac{\partial m^*}{\partial \boldsymbol{\beta}^T})$$

$$\widehat{E}[m^* S_{\boldsymbol{\gamma}}^T] = N^{-1} \sum_{i=1}^{N} m^* S_{\boldsymbol{\gamma}}^T$$

$$\widehat{E}[S_{\boldsymbol{\gamma}} S_{\boldsymbol{\gamma}}^T] = N^{-1} \sum_{i=1}^{N} S_{\boldsymbol{\gamma}} S_{\boldsymbol{\gamma}}^T,$$

and they are evaluated at the estimator $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\gamma}}$. Note that we have not accounted for the variability in estimating $\boldsymbol{\alpha}$ and we assume that $\boldsymbol{\alpha}$ is known. In the section, we will derive the variance of $\hat{\beta}$ when α is unknown with M-estimator theory. In Theorem 10.3 in the book of *Semiparametric Theory and Missing Data* [9] prove that two estimators for $\boldsymbol{\beta}$ from the augmented estimating equation with $\boldsymbol{\alpha}$ known and $\boldsymbol{\alpha}$ estimated are asymptotically equivalent.

The hypothesis test $H_0 : L^T \beta = 0$ is our interest to test the difference between PPVs where $L^T = (0,\ 1)$, and $\boldsymbol{\beta} = (\beta_0, \beta_1)$. With the variance estimate in 2.28, we can construct the Wald statistics which follows χ_1^2 under the H_0. Wald statistics follows since the asymptotic normality of RAL estimator.

$$T_{gg}^w = L^T \hat{\boldsymbol{\beta}} (\widehat{var_{gg}})^{-1} L \hat{\boldsymbol{\beta}}.$$

The second test which is asymptotically equivalent to the Wald statistics, but which is invariant to any transformations is presented as generalized score statistics. We can derive

generalized score statistics from the estimating equation [13, 14, 15].
Let

$$
\begin{aligned}
S^*_{gg}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0) &= \sum m^*(\tilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0), \\
I_{gg0} &= \sum \partial m^*(\tilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\beta}^T = \sum A(T_1, T_2) \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T}, \\
I_{gg1} &= \sum m^*(\tilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0)^T m(\tilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0) = \sum A(T_1, T_2)(\tilde{D} - \boldsymbol{\mu})(\tilde{D} - \boldsymbol{\mu})^T A(T_1, T_2)^T,
\end{aligned}
$$

and $\tilde{\boldsymbol{\beta}}$ is an estimate from the estimating equation assuming the null hypothesis is true. Then the generalized score statistic is

$$
T^s_{gg} = S^*_{gg}(\tilde{\boldsymbol{\beta}})\ \Sigma^m_{gg}\ L^T\ (L \Sigma^e_{gg} L)^{-1}\ L\ \Sigma^m_{gg}\ S^*_{gg}(\tilde{\boldsymbol{\beta}}),
$$

where $\Sigma^m_{gg} = (I_{gg0})^{-1}$ is a model based covariance matrix of $\widehat{\boldsymbol{\beta}}$. When the working correlation within individual in (2.4) is correctly specified, it is a consistent estimator for the covariance matrix of $\widehat{\boldsymbol{\beta}}$. On the other hands, $\Sigma^e_{gg} = I^{-1}_{gg0} I_{gg1} I^{-1}_{gg0}$ is an empirical estimator of covariance matrix $\hat{\boldsymbol{\beta}}$, and it is still consistent even if the working correlation is not correctly specified. These test statistics from the augmented estimating equation gives the most powerful test any other tests in the class of IPW.

2.6.2 When $\boldsymbol{\alpha}$ is unknown

We derived the influence function of $\hat{\boldsymbol{\beta}}$ by assuming that the parameters from $L(T_1, T_2, C; \boldsymbol{\alpha})$ is known. We derive variance of the parameter of interest and test statistics by accounting for the variability of the parameters both from the missingness model and covariate model in this section. Let $\boldsymbol{\theta}^T = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)$. The composite estimating equation, m^*_{hh}, in estimating all the parameters which are used in augmented estimator can be written as

$$
\sum_{i=1}^{N}
\begin{pmatrix}
\begin{pmatrix} 1 \\ X_i^T \end{pmatrix} & R_i\,(D_i - \eta_i(\boldsymbol{\alpha})) \\
A(T_{1i}, T_{2i}) & \begin{pmatrix} \tilde{D}_i - expit\{\beta_0 + \beta_1\} \\ \tilde{D}_i - expit\{\beta_0\} \end{pmatrix} \\
\begin{pmatrix} 1 \\ X_i^T \end{pmatrix} & (R_i - \pi_i(\boldsymbol{\gamma}))
\end{pmatrix}
=
\begin{pmatrix} 0^{dim(\boldsymbol{\alpha})} \\ 0^{dim(\boldsymbol{\beta})} \\ 0^{dim(\boldsymbol{\gamma})} \end{pmatrix},
$$

where for i^{th} individual $\eta_i(\boldsymbol{\alpha}) = 1/(1+exp^{-\boldsymbol{\alpha}^T X_i})$, $\pi_i(\boldsymbol{\gamma}) = 1/(1+exp^{-\boldsymbol{\gamma}^T X_i})$, and $A(T_1, T_2)$ is defined above. When we write the composite estimating equation, the point estimate of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ does not change from the algorithm described in section (2.4), but the variance for the parameters of interest $\hat{\boldsymbol{\beta}}$ may change. Since (2.29) is the from of M-estimator , we apply sandwich variance estimator. The sandwich variance of the parameters is

$$V_{hh}(\boldsymbol{\theta}_0) = A_{hh}(\boldsymbol{\theta}_0)^{-1} B_{hh}(\boldsymbol{\theta}_0) \{A_{hh}(\boldsymbol{\theta}_0)^{-1}\}^T$$

where

$$\begin{aligned} A_{hh}(\boldsymbol{\theta}_0) &= E_F\left(-\partial m^*_{hh}/\partial \boldsymbol{\theta}^T\right) \\ B_{hh}(\boldsymbol{\theta}_0) &= E_F\left(m^*_{hh}(\boldsymbol{\theta}) m^*_{hh}(\boldsymbol{\theta})^T\right). \end{aligned}$$

The estimator for the variance is estimated as

$$V_{hh}(D, T_1, T_2, C, \hat{\boldsymbol{\theta}}) = A_{hh}(D, T_1, T_2, C, \hat{\boldsymbol{\theta}})^{-1} B_{hh}(D, T_1, T_2, C, \hat{\boldsymbol{\theta}}) \{A_{hh}(D, T_1, T_2, C, \hat{\boldsymbol{\theta}})^{-1}\}^T,$$

where

$$\begin{aligned} A_{hh}(D, T_1, T_2, C, \hat{\boldsymbol{\theta}}) &= \frac{1}{N}\sum_{i=1}^{N}\left(-\partial m^*_{hh}(\hat{\boldsymbol{\theta}})/\partial \theta_N^T\right), \\ B_{hh}(D, T_1, T_2, C, \hat{\boldsymbol{\theta}}) &= \frac{1}{N}\sum_{i=1}^{N}\left(m^*_{hh}(\hat{\boldsymbol{\theta}}) m^*_{hh}(\hat{\boldsymbol{\theta}})^T\right), \end{aligned}$$

where $\hat{\boldsymbol{\theta}_N} = (\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)$ We can write the hypothesis as $H_0 = L_{hh}^T \boldsymbol{\theta} = 0$ where $L_{hh}^T = \left(0^{dim(\boldsymbol{\alpha})},\ 1, 0^{dim(\boldsymbol{\gamma})}\right)$ to compare the difference of log-odds positive predictive values.
The Wald and generalized score statistics is written as

$$\begin{aligned} T^w_{hh} &= L^T_{hh}\hat{\boldsymbol{\theta}}\left\{V_{hh}(\hat{\boldsymbol{\theta}})\right\} L_{hh}\hat{\boldsymbol{\theta}} \\ T^s_{hh} &= S^*_{hh}(\tilde{\boldsymbol{\theta}})\ \Sigma^m_{hh}\ L^T_{hh}\ (L\Sigma^e_{hh}L)^{-1}\ \Sigma^m_{hh}\ S^*_{hh}(\tilde{\boldsymbol{\theta}}), \end{aligned}$$

where $S^*_{hh}(\tilde{\boldsymbol{\theta}})$ is the summation of the estimating equation under the null hypothesis, and the following are also considered under the null hypothesis.

$$\begin{aligned} \Sigma^m_{hh} &= (I_{hh0})^{-1}, \Sigma^e_{hh} = I^{-1}_{hh0} I_{hh1} I^{-1}_{hh0}, \\ I_{hh0} &= N \times A_{hh}(\tilde{\boldsymbol{\theta}}) \\ I_{hh1} &= N \times B_{hh}(\tilde{\boldsymbol{\theta}}) \end{aligned}$$

As it was mention in section (2.6.1), the variance derived from its influence functions $\boldsymbol{\alpha}$ known, and the variance from the sandwich method from M-estimator where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ unknown are asymptotically equivalent.

2.7 Simulation Studies

This simulation is to explore the property of the Doubly Robust estimator compared to the estimators from complete case, and inverse probability weighting when all the models are correct, and one or another models are not correct in estimating and testing the difference of log-odds of PPV for T_1 and T_2. We generate 500 and 1000 individuals with fixed $PPV_1, PPV_2, NPV_1, NPV_2, P(D=1), OR_0, OR_1$. $Pr(D=1)$ is disease prevalence. OR_1 refers to the odds-ratio of T_1 and T_2 when disease is present, and OR_0 refers to the odds-ratio of T_1 and T_2 when disease absent. For each individual $i = 1, \cdots, N$, the covariates $C = (c_{1i}, c_{2i}, c_{3i}, c_{4i})^T$ are generated from independent and identically distributed $N(\mu_d \times d + \mu_{1-d} \times (1-d), 1)$ to have certain correlation with D. The missing indicator, R_i, is generated as Bernoulli with the probability of $P(R=1|T_1, T_2, C) = expit\{\gamma_0 + \gamma_1 T_{1i} + \gamma_2 T_{2i} + \gamma_3 C_{1i} + \gamma_4 C_{2i} + \gamma_5 C_{3i} + \gamma_6 C_{4i}\}$ with known $\boldsymbol{\gamma}$. Missing mechanism is independent of the disease status given covariates which reflexes the assumption of Missing at Random, $R \perp D \mid T_1, T_2, C$. This results in the linear relationship of logit P(D = 1|T_1, T_2, C) with the diagnostic tests and covariates, since

$$\begin{aligned}
log\frac{P(D=1|Z,X,C)}{P(D=0|Z,X,C)} &= log\frac{P(C=c|D=1,Z,X)P(D=1|Z,X)}{P(C=c|D=0,Z,X)P(D=0|Z,X)} \\
&= log\frac{P(C=c|D=1)}{P(C=c|D=0)} + log\frac{P(D=1|Z,X)}{P(D=0|Z,X)} \\
&= log\frac{\frac{1}{\sqrt{2\pi}}e^{-(c-\mu_d)^2/2}}{\frac{1}{\sqrt{2\pi}}e^{-(c-\mu_{1-d})^2/2}} + (\beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX) \\
&= (\mu_d - \mu_{1-d})C + \mu_d^2 - \mu_{1-d}^2 + \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX.
\end{aligned}$$

Recall that Z_i and X_i are vectors, and we change the notation to scalar for individual i as

$$log\frac{P(D=1|T_1,T_2,C)}{P(D=0|T_1,T_2,C)} = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 T_1 T_2 + \alpha_4^T C, \tag{2.29}$$

and C is the vector of covariates. For this simulation, mean of those who do not have disease, μ_{1-d}, is set to 0.1 and those who have disease, μ_d, is set as -0.1 to generate covariates. Miss-

ingness model parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6)$ is set as $(0, -0.6, 0.1, -0.1, -0.1, -0.1, -1)$ and $(0, -0.6, 0.1, -0.1, 0, 0, 0)$ to yield the average probability of missing 40%, but with different number of covariates. To explore the affect of relationship between C and D to efficiency of estimator, we first use 4 covariates which result in the generalized coefficient of determination R^2 [16] of 73.4 % on average. We also use 1 covariate for moderate relationship between D and C, and the generalized coefficient of determination R^2 is 52.3 % on average.

Table 2.1 summaries the analysis under the H_0 where $PPV_1 = 0.85$, $PPV_2 = 0.85$ $NPV_1 = 0.85$, $NPV_2 = 0.85$, $OR_0 = 1$, $OR_1 = 1$, and $P(D = 1) = 0.5$. based on $1,000$ Monte Carlo data set to evaluate the Type I error. β_1^{Com} refers to the analysis in using the complete data. β_1^{IPW} refers to inverse probability weighted estimator. Point estimates of β_1^{gg} and β_1^{hh} uses the same iterative algorithm to solve augmented estimating equations which is described in section 2.4, but the variance are different; β_1^{gg} uses variance estimator when $\boldsymbol{\alpha}$ is known in section 2.6.1, and β_1^{hh} uses variance estimator when $\boldsymbol{\alpha}$ is unknown in section 2.6.2. Although they are expected to perform similarly asymptotically, both of them are included to evaluate the performance with small sample size. The corresponding Wald and score statistics are used. Bias is Monte Carlo bias, AveSE is the average of standard error obtained using appropriate formula described in the sections. MCSE is Monte Carlo standard error, and CP reports how many times the 95% confidence interval contains the true value, $\boldsymbol{\beta}$, out of 1,000 Monte Carlo simulation.

From the Table 2.1, we see that $\hat{\beta}_1^{Com}$ is biased, but the other estimators are unbiased. As we have expected AveSE of β_1^{gg} and β_1^{hh} are not much different, and are smaller than the AveSE of β_1^{IPW}. The gain of efficiency is higher when we use 4 covariates than 1 covariate. CP and type I error from Wald and score are within 1,000 Monte Carlo simulation error. Table 2.2 simulation results from 1,000 Monte Carlo data set under H_a. $PPV_1 = 0.80$ and $PPV_2 = 0.65$ are used and so $\beta_1 = 0.767$. The results shows the similar pattern; β_1^{Com} is biased while the other estimators are not. SE of β_1^{gg} and β_1^{hh} is similar but are more efficient then the SE of β_1^{IPW}. The gain of efficiency is greater when 4 covariates are used than 1 covariate is used in the simulation. The CP is appropriate and the power of β_1^{gg} and β_1^{hh} are bigger than the IPW since augmented estimator gives the most powerful test among the class of inverse probability estimators.

Next, we would like to investigate the doubly robust property when one of the

Table 2.1: Simulation Results when both of models are correct under $\beta_1 = 0$ (1000 Monte-Carlo samples, N=1000, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, Sizes (Type I errors) are shown with Wald and score statistics.)

Estimator	Bias	AveSE	MCSE	CP	Size(Wald)	Size(Score)
			Mild R^2 in the Covariate Model with 1 covariate			
$\hat{\beta}_1^{Com}$	-0.476	0.278	0.282	0.604	0.396	0.419
$\hat{\beta}_1^{gg}$	0	0.160	0.159	0.949	0.051	0.059
$\hat{\beta}_1^{hh}$	0	0.160	0.159	0.950	0.050	0.058
$\hat{\beta}_1^{IPW}$	-0.015	0.250	0.259	0.939	0.061	0.072
			Strong R^2 in the Covariate Model with 4 covariates			
$\hat{\beta}_1^{Com}$	-0.477	0.283	0.274	0.600	0.400	0.423
$\hat{\beta}_1^{gg}$	0	0.127	0.135	0.940	0.060	0.062
$\hat{\beta}_1^{hh}$	0	0.127	0.135	0.939	0.060	0.063
$\hat{\beta}_1^{IPW}$	-0.014	0.256	0.263	0.949	0.051	0.060

proposed models is incorrect. For incorrect model,

$$x_{1i} = -exp\left\{c_{1i} + c_{4i}\right\}/(1 + exp\left\{1 + c_{1i} + c_{4i}\right\})$$
$$x_{2i} = \left\{ \begin{array}{cc} -0.2 & \text{when } c_{2i} + c_{3i} < 0 \\ 0.2 & \text{when } c_{2i} + c_{3i} > 0 \end{array} \right\}.$$

are used as covariates instead of Cs, and missingness model is fitted as

$$\text{logit}\{\text{P}(\text{R} = 1|\text{T}_1, \text{X}_1, \text{X}_2)\} = \gamma_0 + \gamma_1\text{T}_1 + \text{X}_1 + \gamma_4\text{X}_2, \tag{2.30}$$

and the conditional model is used as

$$\text{logit}\{\text{P}(\text{D} = 1|\text{T}_1, \text{C}_1)\} = \alpha_0 + \alpha_1\text{T}_1 + \alpha_2\text{C}_1. \tag{2.31}$$

Table 2.3 is the simulation result when the missing model (2.30) is used instead of using T_1, T_2, C_1, C_2, C_3, and C_4 in the logistic model. β_1^{Com} is still biased and β_1^{IPW} is also biased since it does not have any protection from specifying incorrect missingness model. β_1^{gg} and β_1^{hh} is unbiased and obtain correct CP.

Table 2.4 is when the covariate model (2.31) is used instead of the model $T_1, T_2, C_1, C_2, C_3, C_4, T_1$, and T_2 in the logistic model. IPW does not use the covariate

Table 2.2: Simulation Results when both of models are correct under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, Power in rejecting $H_0 : \beta_1 = 0$ are shown with Wald and score statistics.)

Estimator	Bias	SE	MCSE	CP	Power(Wald)	Power(Score)
			Mild R^2 in the Covariate Model with 1 covariate			
			N=500			
$\hat{\beta}_1^{Com}$	-0.297	0.314	0.327	0.812	0.328	0.347
$\hat{\beta}_1^{gg}$	0.016	0.181	0.195	0.936	0.997	0.997
$\hat{\beta}_1^{hh}$	0.016	0.188	0.195	0.941	0.995	0.999
$\hat{\beta}_1^{IPW}$	-0.020	0.271	0.260	0.948	0.861	0.846
			N=1000			
$\hat{\beta}_1^{Com}$	-0.301	0.218	0.223	0.688	0.596	0.600
$\hat{\beta}_1^{gg}$	0.011	0.127	0.131	0.941	1	1
$\hat{\beta}_1^{hh}$	0.011	0.130	0.131	0.948	1	1
$\hat{\beta}_1^{IPW}$	-0.027	0.187	0.178	0.948	0.993	0.991
			Strong R^2 in the Covariate Model with 4 covariates			
			N=500			
$\hat{\beta}_1^{Com}$	-0.307	0.326	0.314	0.826	0.285	0.307
$\hat{\beta}_1^{gg}$	0	0.143	0.153	0.941	0.999	0.999
$\hat{\beta}_1^{hh}$	0	0.144	0.153	0.942	0.999	0.999
$\hat{\beta}_1^{IPW}$	0.033	0.298	0.293	0.939	0.826	0.856
			N=1000			
$\hat{\beta}_1^{Com}$	-0.319	0.225	0.219	0.694	0.523	0.539
$\hat{\beta}_1^{gg}$	0.001	0.103	0.107	0.944	1	1
$\hat{\beta}_1^{hh}$	0.001	0.103	0.107	0.942	1	1
$\hat{\beta}_1^{IPW}$	0.017	0.207	0.205	0.953	0.987	0.988

Table 2.3: Simulation Results with Wrong Missingness Model, and Correct Covariate Model under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, power in rejecting $H_0 : \beta_1 = 0$ is shown with Wald and score statistics.)

Estimator	Bias	SE	MCSE	CP	Power(Wald)	Power(Score)
			N=500			
$\hat{\beta}_1^{Com}$	-0.307	0.326	0.314	0.826	0.285	0.307
$\hat{\beta}_1^{gg}$	0	0.143	0.153	0.941	0.999	0.999
$\hat{\beta}_1^{hh}$	0	0.143	0.153	0.942	0.999	0.999
$\hat{\beta}_1^{IPW}$	-0.404	0.297	0.284	0.705	0.228	0.243
			N=1000			
$\hat{\beta}_1^{Com}$	-0.319	0.225	0.219	0.694	0.523	0.539
$\hat{\beta}_1^{gg}$	0.001	0.104	0.107	0.946	1	1
$\hat{\beta}_1^{hh}$	0.001	0.103	0.107	0.944	1	1
$\hat{\beta}_1^{IPW}$	-0.410	0.205	0.197	0.460	0.419	0.425

model, so it is not effected from specifying the model incorrectly. β_1^{gg} and β_1^{hh} behave as expected. When both of the models are misspecified by using (2.30) and (2.31), none of the models are expected to perform well as it is summarized in Table 2.5.

2.8 Application

2.8.1 A Cardiology Example

A data set from the area of cardiology has been used. The current gold standard for coronary artery disease (CAD) is coronary angiography. The gold standard is invasive and it has a small risk of morbidity or mortality. Thus, not all patients can go through the gold standard. Single-photon-emission computed tomography (SPECT) stress thallium is used as diagnostic tests. A patients needs to stimulate circulation so that distribution of blood flows in the heart can be ascertained by imaging. We consider two ways of stimulating circulation. The usual way is to use exercise, but some patients cannot exercise and diripidamole is used to stimulate the heart activity. We have the first diagnostic test is to stimulate blood circulation by exercise and the second diagnostic test is to use diripidamole. The gender, age, and weight can be useful physiological factors which may be related to the result of gold standards, and we use them as covariates to increase our efficiency of the estimator.

Table 2.4: Simulation Results with Wrong Covariate Model, and Correct Missingness Model under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, power in rejecting $H_0 : \beta_1 = 0$ is shown with Wald and score statistics.)

Estimator	Bias	SE	MCSE	CP	Power(Wald)	Power(Score)
			N=500			
$\hat{\beta}_1^{Com}$	-0.307	0.326	0.314	0.826	0.285	0.307
$\hat{\beta}_1^{gg}$	0.011	0.177	0.169	0.972	0.995	0.994
$\hat{\beta}_1^{hh}$	0.011	0.177	0.169	0.977	0.994	0.995
$\hat{\beta}_1^{IPW}$	0.033	0.298	0.293	0.939	0.826	0.856
			N=1000			
$\hat{\beta}_1^{Com}$	-0.319	0.225	0.219	0.694	0.523	0.539
$\hat{\beta}_1^{gg}$	0.006	0.122	0.119	0.956	1	1
$\hat{\beta}_1^{hh}$	0.006	0.123	0.119	0.956	1	1
$\hat{\beta}_1^{IPW}$	0.017	0.207	0.205	0.953	0.987	0.988

Table 2.5: Simulation Results with Wrong Covariate Model, and Wrong Missingness Model under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, power in rejecting $H_0 : \beta_1 = 0$ is shown with Wald and score statistics.)

Estimator	Bias	SE	MCSE	CP	Power(Wald)	Power(Score)
			N=500			
$\hat{\beta}_1^{Com}$	-0.307	0.326	0.314	0.826	0.285	0.307
$\hat{\beta}_1^{gg}$	-0.132	0.196	0.186	0.894	0.932	0.930
$\hat{\beta}_1^{hh}$	-0.132	0.196	0.186	0.886	0.935	0.948
$\hat{\beta}_1^{IPW}$	-0.404	0.297	0.284	0.705	0.228	0.243
			N=1000			
$\hat{\beta}_1^{Com}$	-0.319	0.225	0.219	0.694	0.523	0.539
$\hat{\beta}_1^{gg}$	-0.138	0.137	0.131	0.823	0.996	0.996
$\hat{\beta}_1^{hh}$	-0.138	0.136	0.131	0.814	0.996	0.996
$\hat{\beta}_1^{IPW}$	-0.410	0.205	0.197	0.460	0.419	0.425

Table 2.6: Cardiology Data

	$R=1$				$R=0$	
	$D=1$		$D=0$			
T_1	T_2		T_2		T_2	
	(+)	(−)	(+)	(−)	(+)	(−)
(+)	110	2	85	12	434	723
(−)	78	3	138	25	459	348

Table 2.7: Cardiology Example

Estimator	$\widehat{PPV_1}$	$\widehat{PPV_2}$	$\widehat{\beta_1}$	SE($\hat{\beta}_1$)	Wald	Score
Aug.com	0.457	0.536	-0.315	0.104	9.126	9.287
Aug.gg	0.449	0.410	0.162	0.133	1.480	1.486
Aug.hh	0.449	0.410	0.162	0.155	1.093	1.301
IPW	0.455	0.407	0.197	0.145	1.840	1.879

We also use the covariate to model the missingnes probability of patients. The data without covariates are summarized in the Table 2.6.

Those patients(N=271) who have at least one missing covariates has been deleted. Out of total 2417 patients, only $\frac{453}{2417} \times 100 = 19\%$ have observed gold standard. The generalized correlation of coefficient $R^2 = 0.19$ which is weak prediction of D in the covariate model. Our interest of hypothesis is to see if the difference of log-odds of PPVs for T_1 and T_2 are the same.

The naive estimator which eliminates the missing gold standard where $R = 0$ has $\widehat{PPV_1} = 0.457$ and $\widehat{PPV_2} = 0.535$ from the Table 2.7. The logit of the different is $\widehat{\beta_1} = -0.315$. Its Wald and score conclude that there is a significant evidence that we can reject the null hypothesis while the IPW and Aug test statistics cannot reject the H_0.

$$\text{logit}\{\text{P}(\text{R} = 1|\text{T}_1, \text{T}_2, \text{C}_1, \text{C}_2, \text{C}_3)\} = \gamma_0 + \gamma_1\text{T}_1 + \gamma_2\text{T}_2 + \gamma_3\text{C}_1 + \gamma 4\text{C}_2 + \gamma_5\text{C}_3,$$

is used to model the missingness for each individual, and the conditional model is used as

$$\text{logit}\{\text{P}(\text{D} = 1|\text{T}_1, \text{T}_2, \text{C}_1, \text{C}_2, \text{C}_3)\} = \alpha_0 + \alpha_1\text{T}_1 + \alpha_2\text{T}_2 + \alpha_3\text{C}_1 + \alpha_4\text{C}_2 + \alpha_5\text{C}_3 + \alpha_6\text{T}_1\text{T}_2.$$

Our two estimates of variance of augmented method Aug.gg and Aug.hh does not agree well and the reason could be that we only observe 19% of the data and the sample size is not large. We also have only 19% of generalized R^2 for covariate model, and it seems like the covariates are not informative in predicting D and the gain of efficient from IPW for Augmented is not great.

2.8.2 A Coronary Stenosis Study

Table 2.8: Coronary Artery Disease Data.

	$R=1$				$R=0$	
	$D=1$		$D=0$			
T_1	T_2		T_2		T_2	
	(+)	(−)	(+)	(−)	(+)	(−)
			$C=1$			
(+)	224	18	38	6	31	21
(−)	1	1	32	35	24	219
			$C=0$			
(+)	37	0	92	12	16	15
(−)	0	0	79	57	70	522

Coronary stenosis is a heart disease whose patients have symptoms of narrowing or obstruction of heart artery. The prevalence of the disease is higher in men than in women. The current gold standard is coronary angiography. However, it can be dangerous to some patients. The diagnosis which is less invasive than coronary stenosis can be conducted through dynamic trans thoracic echocardiography with effort or dynamic trans thoracic with dobutamine. The risk factors are arterial hypertension, hypercholesterolemia, habitual smoking, diabetes, and family history of coronary heart disease. We were not able to access the whole data, but we use the data which has been dichotomized the risk factors as 1 or 0 in [17]; 1 with patients who are recorded with more than two risk factors, and 0 with patients who are recorded with only one risk factors. Their method is not convenient to incorporate continuous covariate since the distribution of covariate needs to be assumed and integration should be conducted. However, we do not have any distributional assumptions and our covariate model enables us to exploit the correlation between covariate and disease in estimating the difference of odds-ratio for T_1 and T_2. D is the outcome of coronary angiography which is not verified to all of the patients, T_1 is the outcome of dynamic trans thoracic with dobutamine, and T_2 is the outcome of dynamic trans thoracic with effort with one covariate C, which is determined by the number of risk factor. The data with one binary covariate is summarized in table 3.9.

Among 1550 men, only $\frac{632}{1550} \times 100 = 40.8$ % individuals have verified gold standard. There are 281 men who are verified to have positive disease; out of 281, 244 are $C = 1$ which

means that they have more than two risk factors, and 37 of them are $C = 0$ which means that they have less than two risk factors.

$$\text{logit}\{\text{P}(\text{R} = 1|\text{T}_1, \text{T}_2, \text{C}_1, \text{C}_2, \text{C}_3)\} = \gamma_0 + \gamma_1\text{T}_1 + \gamma 2\text{T}_2 + \gamma_3\text{C}_1,$$

is used to model the missingness for each individual, and the conditional model is used as

$$\text{logit}\{\text{P}(\text{D} = 1|\text{T}_1, \text{T}_2, \text{C}_1, \text{C}_2, \text{C}_3)\} = \alpha_0 + \alpha_1\text{T}_1 + \alpha_2\text{T}_2 + \alpha_3\text{C}_1 + \alpha_4\text{T}_1\text{T}_2.$$

The covariate model $\text{logit}(\text{D} = 1|\text{T}_1, \text{T}_2, \text{C}, \text{T}_1\text{T}_2)$ has strong generalized correlation coefficient $R^2 = 50.8\%$.

Table 3.10 is the results using complete case, augmented estimator with two different variance estimates, gg and hh estimators, and Inverse probability weighted estimator. Complete cases estimates PPV_1, PPV_2, and β_1 differently than the other estimator, but Aug.gg, Aug.hh, and IPW have the similar estimates and similiar Wald and score statistics. The variance estimator of Aug are also most efficient and have the more powerful statistics than the statistics of IPW.

Table 2.9: Coronary Stenosis

Estimator	$\widehat{PPV_1}$	$\widehat{PPV_2}$	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	Wald	Score
Com	0.521	0.653	-0.550	0.062	78.160	83.208
Aug.gg	0.456	0.639	-0.747	0.069	115.83	125.41
Aug.hh	0.456	0.639	-0.747	0.069	116.113	125.61
IPW	0.454	0.637	-0.748	0.074	102.697	111.86

2.9 Discussion

In this section, we used the semiparametric theory and derived a class of augmented estimators, which are regular and asymptotically linear estimators for $\boldsymbol{\beta}$. We first derived the estimator under no missingness of gold standard and when covariates are not collected from the subjects. In most of cases, the gold standard is missing for some patients. Then the estimator which uses only complete cases with the proposed estimating equation in (2.4) is biased. We, thus, introduce the inverse probability weighted estimator which divides individuals by the probability of observing the patients. We model the probability with logistic model with available covariates and diagnostic tests. Missing at Random assumption $D \perp R|\ (T_1, T_2, C)$ is used to derive and show the asymptotic property of the estimator. The

inverse probability weighted estimator is unbiased, but it does not use individuals whose gold standards are not observed. We extend the estimator to use every possible information in the data set such as the possible association between D and (T_1, T_2, C). In course of that, we use the result from [10] and introduce the class of estimators which contains the IPW and derive the most efficient estimator within the class. This augmented estimator not only models the missingness probability with covariates, but covariate model is used to exploit the information from (T_1, T_2, C) in estimating our parameter of interest. We also derive the most powerful tests; generalized score and Wald with the theory of [13]. The augmented estimator has the double robustness property [9], and simulation studies illustrate the property of the estimator described in previous sections.

Chapter 3

Imputation Method in testing predictive values using auxiliary covariates

3.1 Introduction

Gold standard in health care is used to evaluate the condition of disease for patients. However, the risk from gold standard can be higher on some patients than the risk from the disease itself. One would like to develop non-invasive diagnostic tests which are as accurate as gold standard. The importance of diagnostic tests grows rapidly with the increasing use of biomarkers in health care. Diagnostic biomarkers can be important to detect the early development of disease. Before they are used in practice, it is necessary to evaluate the performance of a diagnostic test to its gold standard. Sensitivity and specificity are two accuracy of classifier of diagnostic tests. They are usually employed on case-control study. On the basis of the disease (gold standard) outcome, study subjects are collected and the diagnostic tests are measured. Such retrospective study requires far less sample size than prospective study, specially when the prevalence of disease is low. Predictive values focus on the accuracy of prediction of disease from diagnostic tests. The positive predictive value is the prediction of diseased individuals among the positive diagnostic tests while the negative predictive value is the proportion of nondiseased individuals among the negative diagnostic tests. Evaluating the accuracy of predictions are possible in cohort study setting. Since

the prevalence of disease in the population is reflected on the predictive values. Statistical methods have been developed in terms of sensitivity and specificity, the predictive values, however, have not been exposed in statistical literature much. We focus on prediction of disease from the diagnostic tests. There are some cases when more than one diagnostic test are available. Specially in the use of biomarkers, hundreds of biomarkers can be available. There have been interest in combing the diagnostic tests in predicting disease [18]. However, due to cost, time, or risk, one has to choose one of the available diagnostic tests in some cases. Our interest is to determine the diagnostic test which performs the better. There are some methods on comparing diagnostic tests when every individual has the outcome of the gold standard. However, it is often the case that not every patient undergoes the gold standard. Moreover, the individual who undergoes the gold standard is not selected by random, but it is rather by physicians' decision. If we only use patients with gold standard outcomes in evaluating the accuracy of the diagnostic test, the estimate of the predictive values of the diagnostic test is likely to be biased unless it is missing at completely random, which is defined by Rubin [19]. In this article, our objective is to compare two diagnostic tests under the condition that gold standards are missing for some individuals. In some situation, baseline auxiliary covariates, such as demographical and physiological records, which may be related to the outcome of diagnostic or disease are available. There have been approaches to weight the individual by the probability of observed for the individual. However, we approach differently. From the available data, we impute the gold standard for those patients who has missing gold standard. We also assume that subjects in this clinical study are randomly selected from the population since the predictive values depend on disease prevalences of populations. In this paper, we use imputation method to estimate the different log-odds ratio of predictive values.

3.2 Model Framework

3.2.1 Notation and Model

When the outcome of gold standard is available on every individual i, for $i = 1 \cdots N$, we let $V_i = (D_i, T_{1i}, T_{2i})$, which is independent and identically distributed. All the three random variables are binary; 1 for positive outcome and 0 for negative outcome. D_i is the outcome of the gold standard, T_{1i} for test 1 and T_{2i} for test 2. We compare the predictive

values in logit transformation. $PPV_1 = Pr(D = 1|T_1 = 1)$, which is defined as a positive predictive value of T_1, denotes the probability of positive prediction of disease among those who have positive diagnostic test outcome using T_1. $NPV_1 = Pr(D = 0|T_1 = 0)$ is a negative predictive value of T_1 and measures the probability of negative gold standard outcome among those who have negative diagnostic outcome using T_1. Diagnostic test with higher PPV and NPV are considered as the more accurate diagnostic test. Our interest, difference between log-odds ratio of predictive values, can be written as the difference in conditional expectation. $\text{logit}\{\text{E(D}|\text{T}_1 = 1)\} - \text{logit}\{\text{E(D}|\text{T}_2 = 1)\}$ for the difference in the positive predictive value, and $\text{logit}\{\text{E(D} = 0|\text{T}_1 = 0)\} - \text{logit}\{\text{E(D} = 0|\text{T}_2 = 0)\}$ for the difference in the negative predictive value, where $\text{logit}\{\text{p}\} = \log\{\text{p}/(1-\text{p})\}$. Since modeling the negative predictive values would be similiar to modeling the positive predictive values, we illustrate how one would estimate and test for the difference between positive predictive values with two diagnostic tests compared to their gold standard. We use the notation from [1] to explicitly express the model in logistic model. Let Z be the indicator for test assignment and it has the value 1 for test 1, and 0 for test 2. Let X be the outcome for the test. Since both of the diagnostic test 1 and 2 are measured on every individual i, i^{th} realization, Z_i and X_i would have two entries for each individual. This notation will allow us to express the predictive values for each tests in logistic regression.

$$\text{logit}\{\text{E(D}|\text{Z, X)}\} = \acute{\beta}_0 + \acute{\beta}_1\text{Z} + \acute{\beta}_2\text{X} + \acute{\beta}_3\text{ZX} \tag{3.1}$$

where $\text{logit(p)} = \log\{\text{p}/(1-\text{p})\}$, and $\text{logit}\,(\text{PPV}_1) = \sum_{\text{i=0}}^{3}\beta_\text{i}$, $\text{logit}\,(\text{PPV}_2) = \sum_{\text{i=0,2}}\beta_\text{i}$, $\text{logit}\,(\text{NPV}_1) = \sum_{\text{i=1}}^{2}\beta_\text{i}$, and $\text{logit}\,(\text{NPV}_2) = \beta_0$. For simplicity, we subset the data where $X = 1$ since we are interested in estimating and testing for positive predictive values. This means that we only include individuals with at least one positive diagnostic test outcome. Then, the simpler logistic regression can be expressed into

$$\text{logit}\{\text{E(D}|\text{Z, X} = 1)\} = \beta_0 + \beta_1\text{Z}. \tag{3.2}$$

$\text{logit}\,(\text{PPV}_1) = \beta_0 + \beta_1$, and $\text{logit}\,(\text{PPV}_2) = \beta_0$,
so β_1 is the difference $\text{logit}\{\text{PPV}_1\}$ and $\text{logit}\{\text{PPV}_2\}$. Our Null hypothesis is $H_0 : \beta_1 = 0$. In order to use the model either (3.1) or (3.2), the disease status, the outcome of diagnostic tests, should be available for all the individuals. However, it is often the case that diagnostic tests are not observed for some individuals. We denote the realization of observed data for individual i as $O_i = (R_i, R_iD_i, T_{1i}, T_{2i}, \boldsymbol{C_i})$ for $i = 1, \cdots, N$. $\boldsymbol{C_i}$ denotes the vector of

covariates, and R_i denotes the indicator for observed gold standard. If $R_i = 1$, one can observe $(D_i, T_{1i}, T_{2i}, \boldsymbol{C_i})$, and if $R_i = 0$, $(T_{1i}, T_{2i}, \boldsymbol{C_i})$ are observed. Since we cannot observe missing gold standard, we need an assumption on the conditional distribution R on the data. The most restrictive assumption is that Missingness is independent of the full data, which is Missing at Completely Random (MCAR). It translates that missingness mechanism does not depend on any of the data. The less restrictive assumption is Missing at Random (MAR) assumption, which means that missingness mechanism only depends on the observed data. That is R is conditionally independent of D given (T_1, T_2, C). When the gold standard is missing for patients, one can imagine that patients with certain covariates or diagnostic tests outcome are more likely to have missing gold standard. In this paper, we assume that missingness only depends on observed data and independent of unobserved data. Thus, R is independent of D conditional on T_1, T_2 and C. With the assumption of MAR and no further assumption on distribution, we would like to derive a class of semiparametric estimators which are consistent and asymptotically normal.

3.2.2 Semiparametric Theory

Let us consider that $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ are the parameters in the statistical model with iid sample $Y_1, \cdots Y_N$ where $\boldsymbol{\beta}$ is 2-dimensional parameter of interest and η is q-dimensional nuisance parameters. We will develop regular and asymptotically linear estimators (RAL) of $\hat{\theta_N}$ that has the form,

$$N^{1/2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta_0}) = N^{-1/2} \sum_{i=1}^{N} \varphi(Y_i) + o_p(1). \tag{3.3}$$

for random variable φ which satisfies $E(\varphi) = 0$, and $E(\varphi\varphi^T)$ is finite and nonsingular under the truth $\boldsymbol{\theta_0}$. $o_p(1)$ is defined as the term which converges in probability in zero as the cluster number, N, goes to infinity. i^{th} function $\varphi(Y_i)$ is referred as influence function of the estimator $\hat{\boldsymbol{\theta}}_N$. RAL estimators have nice properties such as asymptotically normal which can be shown with the central limit theorem and their variances are equal to the variance of their influence function, i.e., $E(\varphi\varphi^T)$. We derive a RAL estimator which is consistent and asymptotically normal through M-estimators [12]. M-estimators are defined as the solution to the sum of vector equations from an independent sample $Y_1, \cdots Y_N$.

$$\sum_{i=1}^{N} m(Y_i, \boldsymbol{\theta}) = 0, \tag{3.4}$$

m is a known function which does not depend on i or N, and it satisfies that $E\{m(Y,\boldsymbol{\theta})\}=0$, and $E\{m(Y,\boldsymbol{\theta})^T m(Y,\boldsymbol{\theta})\}<\infty$. With some regularity conditions and some are discussed in the book [20], M-estimator can be written as the form of RAL in (3.3) and the consistency and normality of the M-estimators follows from semiparametric theory. We expand the (3.4) around the true value $\hat{\theta}_0$,

$$0=\sum_{i=1}^{N} m(Y_i,\boldsymbol{\theta_0})+\{\sum_{i=1}^{N}\frac{\partial m(Y_i,\boldsymbol{\theta}_N^*)}{\partial \boldsymbol{\theta}^T}\}(\boldsymbol{\theta}_N^*-\boldsymbol{\theta})+o_p(1), \tag{3.5}$$

where $\boldsymbol{\theta}_N^*$ is an intermediate value between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_N^*$. With some regularity conditions, the (3.5) is written as

$$\begin{aligned} N^{1/2}(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta_0}) &= -\left[N^{-1}\sum_{i=1}^{N}\frac{\partial m(Y_i,\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^T}\right]^{-1}\left\{N^{-1/2}\sum_{i=1}^{N} m(Y_i,\boldsymbol{\theta_0})\right\}+o_p(1) \\ &= -N^{-1/2}\left[E\left\{\frac{\partial m(Y_i,\boldsymbol{\theta_0})}{\partial\boldsymbol{\theta}^T}\right\}\right]^{-1}\left\{\sum_{i=1}^{N} m(Y_i,\boldsymbol{\theta_0})\right\}+o_p(1). \end{aligned} \tag{3.6}$$

Then $\hat{\boldsymbol{\theta}}_N$ is a regular and asymptotically linear estimator for $\boldsymbol{\theta}$ and its i-th influence function is given as $-\left[E\left\{\frac{\partial m(Y_i,\boldsymbol{\theta_0})}{\partial\boldsymbol{\theta}^T}\right\}\right]^{-1} m(Y_i,\boldsymbol{\theta_0})$. We first construct a M-estimator when there are no missing data, and expand the estimator when some patients have missing D under the Missing at Random assumption.

3.3 Estimators

3.3.1 Estimators under no missingness

In this section, we first derive a M-estimator when there is no missing gold standard, and the observed data are (D_i,T_{1i},T_{2i}) from the model (3.2). We use an estimating equation, $\sum_{i=1}^{N} m(D_i,T_{1i},T_{2i})=0^{2\times 1}$ to estimate the $\boldsymbol{\beta}$'s, where the estimating function is written as

$$m(D,T_1,T_2)=A(T_1,T_2)^{2\times 2}\begin{Bmatrix} D-expit\{\beta_0+\beta_1\} \\ D-expit\{\beta_0\} \end{Bmatrix}, \tag{3.7}$$

where

$$A(T_1,T_2)=\begin{Bmatrix} I(T_1=1) & I(T_2=1) \\ I(T_1=1) & 0 \end{Bmatrix},$$

and $expit(x) = exp(x)/(1 + exp(x))$.

When both of T_1 and T_2 have positive outcomes, this estimating equation translates into two equations solving for β_0 and β_1. When only T_1 is positive, the first equation is included and when only T_2 is positive, the second equation is included into the estimating equation while none of the equation will be used when neither T_1 nor T_2 is positive. Since our estimating equations is unbiased since expectation is equal to zero, the estimator from the estimating equation is consistent, and asymptotically normal estimators. Hypothesis test of $H_0 : \beta_1 = 0$ is available using Wald test and score test.

3.3.2 Estimators under missingness

Now, we estimate the difference in log-odds ratio of positive predictive values for T_1 and T_2 under missing gold standard. When $R = 1$, we can observe D and the above estimating is sufficient to estimate $\boldsymbol{\beta}$. However, when $R = 0$, D cannot be observed. We replace D as $\tilde{D} = RD + (1 - R)\hat{D}$, where $\hat{D}$ is estimated from the conditional distribution given observed data (T_1, T_2, C). Then, (3.7) would be written as

$$m^*(\tilde{D}, T_1, T_2) = A(T_1, T_2)^{2\times 2} \left\{ \begin{array}{c} \tilde{D} - expit\{\beta_0 + \beta_1\} \\ \tilde{D} - expit\{\beta_0\} \end{array} \right\}. \tag{3.8}$$

Since we have binary diagnostic outcomes, the conditional distribution of D given T_1, T_2, and C is Bernoulli with the probability of observing $D = 1$, which is equal to $Pr(D = 1|T_1, T_2, C)$. Since $D \perp R \mid (T_1, T_2, C)$, it is the same as $Pr(D = 1|R = 1, T_1, T_2, C)$ and we let $Pr(D = 1|R = 1, T_1, T_2, C) = \eta(T_1, T_2, C)$. The natural model we can fit for $\eta(T_1, T_2, C)$ is logit $\eta(\mathrm{T}_1, \mathrm{T}_2, \mathrm{C}, \boldsymbol{\alpha})$ with unknown finite dimensional parameters $\boldsymbol{\alpha}$. Under the parametric model, unknown parameters $\boldsymbol{\alpha}$ can be estimated by maximum likelihood method with observed data. For example, the MLE of $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}}_N$ from the logistic model can be obtained by maximizing the likelihood with respect to $\boldsymbol{\alpha}$,

$$\prod_{i=1}^{N} \eta(X_i, \boldsymbol{\alpha})^{R_i} (1 - \eta(X_i, \boldsymbol{\alpha}))^{(1-R_i)}, \tag{3.9}$$

where we let $X = (T_1, T_2, C^T)^T$, and C is a vector of covariates. If the estimates $\hat{\boldsymbol{\alpha}}_N$ converges to true parameter $\boldsymbol{\alpha}_0$ as N goes to infinity, the estimator by solving the equation $\sum_{i=1}^{N} m^*$ where m^* is defined in (3.8) will be an asymptotically linear estimator since $E(m^*(\boldsymbol{\beta})) = 0$.

Let us recall that we have

$$m^*(R,D,T_1,T_2,C) = A(T_1,T_2)^{2\times 2}\left\{\begin{array}{c} RD+(1-R)\eta(T_1,T_2,C,\hat{\boldsymbol{\alpha}}_N) - expit\{\beta_0+\beta_1\} \\ RD+(1-R)\eta(T_1,T_2,C,\hat{\boldsymbol{\alpha}}_N) - expit\{\beta_0\} \end{array}\right\}. \tag{3.10}$$

$E(m^*)$ follows from taking iterative expectation on (T_1,T_2,C).

$$\begin{aligned} &E\left[E\left\{m^*|T_1,T_2,C\right\}\right] \\ &= E\left[A(T_1,T_2)\left\{\begin{array}{c} E\left\{RD+(1-R)\eta(T_1,T_2,C,\hat{\boldsymbol{\alpha}}_N)|T_1,T_2,C\right\} - expit\{\beta_0+\beta_1\} \\ E\left\{RD+(1-R)\eta(T_1,T_2,C,\hat{\boldsymbol{\alpha}}_N)|T_1,T_2,C\right\} - expit\{\beta_0\} \end{array}\right\}\right] \\ &= E\left[A(T_1,T_2)\left\{\begin{array}{c} E\left(RD|T_1,T_2,C\right) + E\left\{(1-R)\eta(T_1,T_2,C,\hat{\boldsymbol{\alpha}}_N)|T_1,T_2,C\right\} - expit\{\beta_0+\beta_1\} \\ E\left(RD|T_1,T_2,C\right) + E\left\{(1-R)\eta(T_1,T_2,C,\hat{\boldsymbol{\alpha}}_N)|T_1,T_2,C\right\} - expit\{\beta_0\} \end{array}\right\}\right] \end{aligned}$$

Since R is independent of D given T_1,T_2,C, and $\eta(T_1,T_2,C,\hat{\boldsymbol{\alpha}}_N) \xrightarrow{p} \eta(T_1,T_2,C,\boldsymbol{\alpha}_0)$, the expectation becomes

$$E\left\{m^*\right\} = E\left[A(T_1,T_2)\left\{\begin{array}{c} E(D|T_1,T_2,C) - expit\{\beta_0+\beta_1\} \\ E(D|T_1,T_2,C) - expit\{\beta_0\} \end{array}\right\}\right].$$

If we take another iterative expectation on (T_1,T_2,C), then it will be the same as the taking expectation on m in (3.7), and thus, the estimator by solving the equation $\sum_{i=1}^N m^*$ is a M-estimator. If we take conditional expectation on (T_1,T_2,C) in 3.7, we have $E(m^*) = E(m)$ and we have shown before that $E(m) = 0$

Furthermore, if we take derivative with respect to $\boldsymbol{\alpha}$ from the likelihood (3.9), the estimating equation for the parameter $\boldsymbol{\alpha}$ can be written as an M-estimator.

$$\sum_{i=1}^{N} \left(\begin{array}{c} 1 \\ X_i^T \end{array}\right) R_i\left(D_i - \eta_i(\boldsymbol{\alpha})\right) = 0^{dim(\boldsymbol{\alpha})} \tag{3.11}$$

When the model for the parameters are correctly specified, we have $\hat{\boldsymbol{\alpha}}_N \xrightarrow{p} \boldsymbol{\alpha}_0$ along with $\eta(X,\hat{\boldsymbol{\alpha}}_N) \xrightarrow{p} \eta(X,\boldsymbol{\alpha}_0)$, where $\boldsymbol{\alpha}_0$ is a true parameter value for the conditional probability in specifying $Pr(D=1|X,\boldsymbol{\alpha}_0)$. To account for variability of parameter, $\boldsymbol{\alpha}$, and for the parameter of our interest, $\boldsymbol{\beta}$, we write the composite estimating equation for the parameter $\theta = (\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T$. For i^{th} individual the composite estimating equation, m_{imp}, in estimating

all the parameters can be written as

$$\sum_{i=1}^{N} \begin{pmatrix} \begin{pmatrix} 1 \\ X_i^T \end{pmatrix} & R_i\left(D_i - \eta_i(\boldsymbol{\alpha})\right) \\ A(T_{1i}, T_{2i}) & \begin{pmatrix} \tilde{D}_i - expit\{\beta_0 + \beta_1\} \\ \tilde{D}_i - expit\{\beta_0\} \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 0^{dim(\boldsymbol{\alpha})} \\ 0^{dim(\boldsymbol{\beta})} \end{pmatrix},$$

where $X = (1, T_1, T_2, C^T)^T$ and C is a vector of covariates.

3.4 Algorithm

The imputation estimator involves conditional model of D on (T_1, T_2, C). Thus, it requires an adaptive algorithm in that we first model conditional and plug in the estimates in the composite estimating equation.

1. Find the estimate of $\boldsymbol{\alpha}$, $\hat{\alpha_N}$ in the conditional density of D on T_1, T_2, and C from the following equations.

$$\sum_{i=1}^{N} \begin{pmatrix} 1 \\ X_i^T \end{pmatrix} R_i\left(D_i - \eta_i(\boldsymbol{\alpha})\right) = 0^{dim(\boldsymbol{\alpha})}$$

2. Let $\tilde{D}_i = R_i D_i + (1 - R_i)\eta_i(\hat{\boldsymbol{\alpha}})$ for individual i.

3. The estimate of the parameter of the interest, $\boldsymbol{\beta}$, is obtained by solving the augmented estimating equation, $\sum_{i=1}^{N} m^*$, where

$$m^* = A(T_1, T_2)^{2\times 2} \begin{Bmatrix} \tilde{D}_i - expit\left\{\beta_0 + \beta_1\right\} \\ \tilde{D}_i - expit\left\{\beta_0\right\} \end{Bmatrix},$$

and

$$A(T_1, T_2) = \begin{Bmatrix} I(T_1 = 1) & I(T_2 = 1) \\ I(T_1 = 1) & 0 \end{Bmatrix}.$$

3.5 Variance and Test Statistics of the Estimator

M-estimators are written as asymptotically linear estimator as in (3.6) with its influence function as

$$-\left[E\left\{\frac{\partial m(D_i, T_{i1}, T_{2i}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right\}\right]^{-1} m(D_i, T_{i1}, T_{2i}, \boldsymbol{\beta})$$

for i-th individual, when

$$m(D, T_1, T_2, \boldsymbol{\beta}) = A(T_1, T_2)^{2\times 2} \left\{ \begin{array}{c} D - expit\left\{\beta_0 + \beta_1\right\} \\ D - expit\left\{\beta_0\right\}, \end{array} \right\}. \tag{3.12}$$

When one observes D on every subjects, the variance of the the estimator is the variance of the influence function. However, it is often the case that D is missing for some patients and we assume that $D \perp R|\ (T_1, T_2, C)$. Then one more M-estimator is needed since we do not know the conditional probability of D given (T_1, T_2, C). We call the composite estimating equation as Σm_{imp} which is,

$$\sum_{i=1}^{N} \left(\begin{array}{cc} \left(\begin{array}{c} 1 \\ X_i^T \end{array} \right) & R_i\left(D_i - \eta_i(\boldsymbol{\alpha})\right) \\ A(T_{1i}, T_{2i}) & \left(\begin{array}{c} \tilde{D}_i - expit\{\beta_0 + \beta_1\} \\ \tilde{D}_i - expit\{\beta_0\} \end{array} \right) \end{array} \right) = \left(\begin{array}{c} 0^{dim(\boldsymbol{\alpha})} \\ 0^{dim(\boldsymbol{\beta})} \end{array} \right),$$

where $X^T = (T_1, T_2, C^T)$ and C is a vector of covariates.

As the examples in Stefanski and Boos [12], the above estimating equation is partial M-estimator in the sense that it may not be M-estimators separately, but had we estimate the parameters $\boldsymbol{\alpha}$ and plug the estimate on the top equation then it would be a M-estimators. Then by Taylor expansion and suitable regularity conditions, it would be asymptotically normal with consistent and asymptotically normal with the asymptotic covariance matrix below.

$$V_{imp}(\boldsymbol{\theta_0}) = A_{imp}(\boldsymbol{\theta_0})^{-1} B_{imp}(\boldsymbol{\theta_0}) \{A_{imp}(\boldsymbol{\theta_0})^{-1}\}^T \tag{3.13}$$

where

$$\begin{aligned} A_{imp}(\boldsymbol{\theta_0}) &= E_F\left\{-\partial m^* / \partial \boldsymbol{\theta}^T\right\} \\ B_{imp}(\boldsymbol{\theta_0}) &= E_F\left\{m^*(\boldsymbol{\theta}) m^*(\boldsymbol{\theta})^T\right\}. \end{aligned}$$

The estimator for the variance is

$$\hat{V}_{imp}(\tilde{D}, T_1, T_2, \hat{\boldsymbol{\theta}}) = \left\{A_{imp}(\tilde{D}, T_1, T_2, \hat{\boldsymbol{\theta}})^{-1}\right\}^{-1} B_{imp}(\tilde{D}, T_1, T_2, \hat{\boldsymbol{\theta}}) \left\{A_{imp}(\tilde{D}, T_1, T_2, \hat{\boldsymbol{\theta}})^{-1}\right\}^{T},$$

where

$$\begin{aligned} A_{imp}(\tilde{D}, T_1, T_2, \hat{\boldsymbol{\theta}}) &= \frac{1}{N} \sum_{i=1}^{N} \left(-\partial m^*(\hat{\boldsymbol{\theta}}_N) / \partial \boldsymbol{\theta}^T\right) \\ B_{imp}(\tilde{D}, T_1, T_2, \hat{\boldsymbol{\theta}}) &= \frac{1}{N} \sum_{i=1}^{N} \left(m^*(\hat{\boldsymbol{\theta}}_N) m^*(\hat{\boldsymbol{\theta}})^T\right), \end{aligned}$$

and $\hat{\boldsymbol{\theta}}_N$ is the estimator by the adaptive algorithm in (3.4). The parameters are $\boldsymbol{\theta}^T = (\boldsymbol{\alpha}^T, \beta_0, \beta_1)$. Our hypothesis is whether $\beta_1 = 0$ and we can write the hypothesis as $H_0 = L_{imp}^T \boldsymbol{\theta} = 0$ where $L_{imp}^T = \left(0^{dim(\boldsymbol{\alpha})},\ 1\right)$.
The Wald and score statistics which follows χ_1^2 are written as

$$T_{imp}^w = L_{imp}^T \hat{\boldsymbol{\theta}} (\hat{var}(\hat{\boldsymbol{\theta}})) L_{imp} \hat{\boldsymbol{\theta}},$$
$$T_{imp}^s = S_{imp}^*(\tilde{\boldsymbol{\theta}})\ \Sigma_{imp}^m\ L_{imp}^T\ (L_{imp} \Sigma_{imp}^e L_{imp})^{-1}\ L\ \Sigma_{imp}^m\ S_{imp}^*(\tilde{\boldsymbol{\theta}})$$

where $S_{imp}^*(\tilde{\boldsymbol{\theta}})$ is the summation of the estimating equation under the null hypothesis, and the following quantities are considered under the null hypothesis.

$$\Sigma_{imp}^m = (I_{imp0})^{-1}, \Sigma_{imp}^e = I_{imp0}^{-1} I_{imp1} I_{imp0}^{-1},$$
$$I_{imp0} = N \times A_{imp},$$
$$I_{imp1} = N \times B_{imp}.$$

3.6 Simulation Studies

The purpose of the simulations is to explore the property of the imputation method and compare its property to the estimators from complete case, inverse probability weighting, and augmented estimator which are described in Chapter 2. The generation of data are the same as Chapter 2. We generate 500 and 1000 subjects with fixed $PPV_1, PPV_2, NPV_1, NPV_2,\ Pr(D = 1), OR_0, OR_1$. Then, we have data $V = (D, T_1, T_2)$. For each individual $i = 1, \cdots, N$, the covariates $\boldsymbol{C_i} = (c_{1i}, c_{2i}, c_{3i}, c_{4i})^T$ are generated from independent and identically distributed $N(\mu_d \times d + \mu_{1-d} \times (1-d), 1)$ to have certain correlation with D. The missing indicator, R_i, is generated as Bernoulli with the probability of $P(R = 1 | T_1, T_2, C) = expit\{\gamma_0 + \gamma_1 T_{1i} + \gamma_2 T_{2i} + \gamma_3 C_{1i} + \gamma_4 C_{2i} + \gamma_5 C_{3i} + \gamma_6 C_{4i}\}$ with fixed γ. Missing mechanism is independent of the D given covariates which reflexes the assumption of Missing at Random, $R \perp D \mid T_1, T_2, C$. This results in the linear relationship

of logit$\{\Pr(\mathrm{D}=1|\mathrm{T}_1,\mathrm{T}_2,\mathrm{C})\}$ on the diagnostic tests and covariates, since

$$\begin{aligned} log\frac{Pr(D=1|Z,X,C)}{Pr(D=0|Z,X,C)} &= log\frac{Pr(C=c|D=1,Z,X)Pr(D=1|Z,X)}{Pr(C=c|D=0,Z,X)Pr(D=0|Z,X)} \\ &= log\frac{Pr(C=c|D=1)}{Pr(C=c|D=0)} + log\frac{Pr(D=1|Z,X)}{Pr(D=0|Z,X)} \\ &= log\frac{\frac{1}{\sqrt{2\pi}}e^{-(c-\mu_d)^2/2}}{\frac{1}{\sqrt{2\pi}}e^{-(c-\mu_{1-d})^2/2}} + (\beta_0+\beta_1 Z+\beta_2 X+\beta_3 ZX) \\ &= (\mu_d-\mu_{1-d})C+\mu_d^2-\mu_{1-d}^2+\beta_0+\beta_1 Z+\beta_2 X+\beta_3 ZX. \end{aligned}$$

Recall that Z_i and X_i are vectors, and we change the notation to scalar for individual i as

$$log\frac{Pr(D=1|T_1,T_2,C)}{Pr(D=0|T_1,T_2,C)} = \alpha_0+\alpha_1 T_1+\alpha_2 T_2+\alpha_3 T_1 T_2+\alpha_4^T C.$$

For this simulation, means of those who do not have disease positive μ_{1-d} is set to 0.1 and those who have disease μ_d is set as -0.1 to generate covariates. $(\gamma_0,\gamma_1,\gamma_2,\gamma_3,\gamma_4,\gamma_5,\gamma_6)$ is set as $(0,-0.6,0.1,-0.1,-0.1,-0.1,-1)$ to yield the average probability of missing 40%, with 4 covariates, and $(0,-0.6,0.1,-0.1,0,0,0)$ to yield the average probability of missing 40%, with 1 covariate.

To explore the affect of relationship between C and D to the Type-I error of imputation method, we use 1, 2, 3 and 4 covariates. Table 3.1 summaries the result under the H_0 where $PPV_1 = 0.85$ and $PPV_2 = 0.85$ based on $1,000$ Monte Carlo data set to evaluate the Type I error of imputation method. Bias is Monte Carlo bias, AveSE is the average of standard error obtained using appropriate formula described in the sections. MCSE is Monte Carlo standard error. 95% confidence interval are contructed and CP reports the proportiont that the constructed confidence intervals contain the true value of β_1 out of 1,000 Monte Carlo data set. Wald and score statistics with χ_1^2 are used to test the $H_0 : \beta_1 = 0$ and the number of rejecting H_0 is recorded which represent Type-I error. To our surprise, when one covariate is used, the score statistic has inflated type-I error. Table 3.2 results from 1,000 Monte Carlo data set under H_a.

$PPV_1 = 0.80$ and $PPV_2 = 0.65$ are used and so $\beta_1 = 0.767$. The results shows that $\hat{\beta}_1^{Com}$ is biased as we have expected while the other estimators are not. The estimator from the imputation method described in Chapter 3 ($\hat{\beta_1}^{Imp}$) behaves very similarly as the augmented method($\hat{\beta_1}^{gg}$, and $\hat{\beta_1}^{hh}$) when both of the models are correct. $\hat{\beta_1}^{gg}$,$\hat{\beta_1}^{hh}$ are

Table 3.1: Simulation Results of Type I Error of Imputation Method under $\beta_1 = 0$ (1000 Monte-Carlo samples, N=1000, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence interval, Sizes (Type I errors) are shown with Wald and score statistics.)

Bias	SE	MCSE	CP	Size(Wald)	Size(Score)
			4 Covariates		
0	0.127	0.134	0.940	0.060	0.067
			3 Covariates		
0	0.134	0.140	0.943	0.057	0.084
			2 Covariates		
-0.013	0.147	0.149	0.952	0.048	0.089
			1 Covariate		
-0.014	0.161	0.158	0.955	0.045	0.116

described in Chapter 2. AveSE of $\hat{\beta}_1^{Imp}$, $\hat{\beta}_1^{gg}$ and $\hat{\beta}_1^{hh}$ are similar but they are more efficient than the SE of $\hat{\beta}_1^{IPW}$, which is inverse probability weighted estimator. The gain of efficiency is greater when 4 covariates are used than 1 covariate is used in the simulation. The CP is appropriate and the power of $\hat{\beta}_1^{Imp}$, $\hat{\beta}_1^{gg}$, and $\hat{\beta}_1^{hh}$ are greater than the $\hat{\beta}_1^{IPW}$ since the covariation information is incorporated through the covariate model.

To investigate the performance of estimators under the misspecification of some of models, we introduce different covariates than we had originally introduced, let

$$x_{1i} = -exp\left\{c_{1i} + c_{4i}\right\}/(1 + exp\left\{1 + c_{1i} + c_{4i}\right\})$$
$$x_{2i} = \left\{ \begin{array}{ll} -0.2 & \text{when } c_{2i} + c_{3i} < 0 \\ 0.2 & \text{when } c_{2i} + c_{3i} > 0 \end{array} \right\}.$$

are used as covariates instead of C, and missingness model is fitted as

$$\text{logit}\{\Pr(\mathrm{R} = 1|\mathrm{T}_1, \mathrm{X}_1, \mathrm{X}_2)\} = \gamma_0 + \gamma_1 \mathrm{T}_1 + \mathrm{X}_1 + \gamma_4 \mathrm{X}_2, \tag{3.14}$$

and the conditional model is used as

$$\text{logit}\{\Pr(\mathrm{D} = 1|\mathrm{T}_1, \mathrm{T}_2, \mathrm{X}_2)\} = \alpha_0 + \alpha_1 \mathrm{T}_1 + \alpha_2 \mathrm{T}_2 + \alpha_3 \mathrm{X}_2. \tag{3.15}$$

Table 3.3 is the simulation result when the missing model (3.14) is used instead of using T_1, T_2, C_1, C_2, C_3, and C_4 in the logistic model. $\hat{\beta}_1^{Com}$ is still biased and $\hat{\beta}_1^{IPW}$ is also

Table 3.2: Simulation Results when both of models are correct under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, Power in rejecting $H_0 : \beta_1 = 0$ are shown with Wald and score statistics.)

Estimator	Bias	SE	MCSE	CP	Power(Wald)	Power(Score)
			Mild R^2 in the Covariate Model with 1 covariate			
			N=500			
$\hat{\beta}_1^{Com}$	-0.297	0.314	0.327	0.812	0.328	0.347
$\hat{\beta}_1^{gg}$	0.016	0.181	0.195	0.936	0.997	0.997
$\hat{\beta}_1^{hh}$	0.016	0.188	0.195	0.941	0.995	0.999
$\hat{\beta}_1^{IPW}$	-0.020	0.271	0.260	0.948	0.861	0.846
$\hat{\beta}_1^{IMP}$	0.016	0.186	0.194	0.939	0.995	1
			N=1000			
$\hat{\beta}_1^{Com}$	-0.301	0.218	0.223	0.688	0.596	0.600
$\hat{\beta}_1^{gg}$	0.011	0.127	0.131	0.941	1	1
$\hat{\beta}_1^{hh}$	0.011	0.130	0.131	0.948	1	1
$\hat{\beta}_1^{IPW}$	-0.027	0.187	0.178	0.948	0.993	0.991
$\hat{\beta}_1^{IMP}$	0.011	0.130	0.131	0.948	1	1
			Strong R^2 in the Covariate Model with 4 covariates			
			N=500			
$\hat{\beta}_1^{Com}$	-0.307	0.326	0.314	0.826	0.285	0.307
$\hat{\beta}_1^{gg}$	0	0.143	0.153	0.941	0.999	0.999
$\hat{\beta}_1^{hh}$	0	0.144	0.153	0.942	0.999	0.999
$\hat{\beta}_1^{IPW}$	0.033	0.298	0.293	0.939	0.826	0.856
$\hat{\beta}_1^{IMP}$	0	0.143	0.152	0.943	0.999	0.999
			N=1000			
$\hat{\beta}_1^{Com}$	-0.319	0.225	0.219	0.694	0.523	0.539
$\hat{\beta}_1^{gg}$	0.001	0.103	0.107	0.944	1	1
$\hat{\beta}_1^{hh}$	0.001	0.103	0.107	0.942	1	1
$\hat{\beta}_1^{IPW}$	0.017	0.207	0.205	0.953	0.987	0.988
$\hat{\beta}_1^{IMP}$	0.002	0.103	0.106	0.943	1	1

biased since it does not have any protections from specifying incorrect missingness model. However, $\hat{\beta}_1^{IMP}$ does not have any effects since missingness model was not included in the modeling of $\hat{\beta}_1{}^{IMP}$. $\hat{\beta}_1{}^{gg}$, and $\hat{\beta}_1{}^{hh}$ are unbiased and obtain correct CP.

Table 3.4 is when the covariate model (3.15) is used instead of the model condition on T_1, T_2, C_1, C_2, C_3, C_4, and T_1T_2 in the logistic model. $\hat{\beta}_1{}^{IPW}$ does not use the covariate model, so it is not affected by specifying the covariate model incorrectly, and $\hat{\beta}_1^{gg}$ and $\hat{\beta}_1^{hh}$ behave as expected which is the Doubly Robust property. However, imputation method has bias since the covariate model was not correctly specified.

When both of the models are misspecified by using (3.14) and (3.15), none of the models are expected to perform well as it is summarized in Table 3.5. Augmented method performs not badly in our simulation setting. However, we can find some other scenario when the augmented method is biased such as simulation result in Chapter 2. We also investigate how the methods would perform under the violation of MAR (nonmissing at random). Instead of generating R under the assumption that $R \perp D \mid (T_1, T_2, \boldsymbol{C})$, we generate missing data through the probability depending on unobserved D, $P(R = 1|T_1, T_2, C, D) = expit\{\gamma_0 + \gamma_1 T_{1i} + \gamma_2 T_{2i} + \gamma_3 C_{1i} + \gamma_4 C_{2i} + \gamma_5 C_{3i} + \gamma_6 C_{4i} + \phi D_i\}$ with fixed γ which is $(0, -0.6, 0.1, -0.1, -0.1, -0.1, -1, 2)$. After the generation, we still analyze the data with the two proposed models as follows under the assumption of MAR since we do not have all available data where $R = 0$,

$$log\frac{Pr(D = 1|T_1, T_2, C)}{Pr(D = 0|T_1, T_2, C)} = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 T_1 T_2 + \alpha_4^T \mathbf{C},$$
$$log\frac{Pr(R = 1|T_1, T_2, C)}{Pr(R = 0|T_1, T_2, C)} = \gamma_0 + \gamma_1 T_1 + \gamma_2 T_2 + \gamma_4^T \mathbf{C}.$$

Then now only the missingness model is incorrect, but the covariate model is also misspecified, since

$$Pr(D = 1|T_1, T_2, C, R = 1) \neq Pr(D = 1|T_1, T_2, C, R = 0).$$

Table 3.6 shows the performance under NMAR. As it is expected, none of the proposed method perform with desirable coverage proportion. Since missingness may depend on unobserved data, it cause some nonidentifiability issues. Some people have done sensitivity analysis under NMAR which is to change the value of ϕ in the generation of R and investigate the effect to the parameter of interest [21].

Table 3.3: Simulation Results with Wrong Missingness Model, and Correct Covariate Model under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, power in rejecting $H_0 : \beta_1 = 0$ is shown with Wald and score statistics.)

Estimator	Bias	SE	MCSE	CP	Power(Wald)	Power(Score)
			N=500			
$\hat{\beta}_1^{Com}$	-0.307	0.326	0.314	0.826	0.285	0.307
$\hat{\beta}_1^{gg}$	0	0.143	0.153	0.941	0.999	0.999
$\hat{\beta}_1^{hh}$	0	0.143	0.153	0.942	0.999	0.999
$\hat{\beta}_1^{IPW}$	-0.404	0.297	0.284	0.705	0.228	0.243
$\hat{\beta}_1^{IMP}$	0	0.143	0.153	0.943	0.999	0.999
			N=1000			
$\hat{\beta}_1^{Com}$	-0.319	0.225	0.219	0.694	0.523	0.539
$\hat{\beta}_1^{gg}$	0.001	0.104	0.107	0.946	1	1
$\hat{\beta}_1^{hh}$	0.001	0.103	0.107	0.944	1	1
$\hat{\beta}_1^{IPW}$	-0.410	0.205	0.197	0.460	0.419	0.425
$\hat{\beta}_1^{IMP}$	0.002	0.103	0.106	0.943	1	1

Table 3.4: Simulation Results with Wrong Covariate Model, and Correct Missingness Model under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, power in rejecting $H_0 : \beta_1 = 0$ is shown with Wald and score statistics.)

Estimator	Bias	AveSE	MCSE	CP	Power(Wald)	Power(Score)
			N=500			
$\hat{\beta}_1^{Com}$	-0.307	0.326	0.314	0.826	0.285	0.307
$\hat{\beta}_1^{gg}$	0.019	0.198	0.198	0.946	0.991	0.992
$\hat{\beta}_1^{hh}$	0.019	0.177	0.169	0.977	0.994	0.995
$\hat{\beta}_1^{IPW}$	0.033	0.298	0.293	0.939	0.826	0.856
$\hat{\beta}_1^{IMP}$	0.106	0.221	0.204	0.969	1	1
			N=1000			
$\hat{\beta}_1^{Com}$	-0.319	0.225	0.219	0.694	0.523	0.539
$\hat{\beta}_1^{gg}$	0.013	0.140	0.147	0.938	0.999	0.999
$\hat{\beta}_1^{hh}$	0.013	0.142	0.147	0.941	0.999	0.999
$\hat{\beta}_1^{IPW}$	0.017	0.207	0.205	0.953	0.987	0.988
$\hat{\beta}_1^{IMP}$	0.101	0.153	0.148	0.933	1	1

Table 3.5: Simulation Results with Wrong Covariate Model, and Wrong Missingness Model under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, power in rejecting $H_0 : \beta_1 = 0$ is shown with Wald and score statistics.)

Estimator	Bias	AveSE	MCSE	CP	Power(Wald)	Power(Score)
			N=500			
$\hat{\beta}_1^{Com}$	-0.307	0.326	0.314	0.826	0.285	0.307
$\hat{\beta}_1^{gg}$	0.020	0.194	0.191	0.949	0.978	0.981
$\hat{\beta}_1^{hh}$	0.020	0.197	0.191	0.953	0.981	0.998
$\hat{\beta}_1^{IPW}$	-0.404	0.297	0.284	0.705	0.228	0.243
$\hat{\beta}_1^{IMP}$	0.106	0.221	0.204	0.969	1	1
			N=1000			
$\hat{\beta}_1^{Com}$	-0.319	0.225	0.219	0.694	0.523	0.539
$\hat{\beta}_1^{gg}$	-0.012	0.137	0.139	0.938	1	1
$\hat{\beta}_1^{hh}$	-0.012	0.138	0.139	0.940	1	1
$\hat{\beta}_1^{IPW}$	-0.404	0.297	0.284	0.705	0.228	0.243
$\hat{\beta}_1^{IMP}$	0.101	0.153	0.148	0.933	1	1

Table 3.6: Simulation Results with Non-Missing at Ranodom under $\beta_1 = 0.767$ (1000 Monte-Carlo samples, AveSE is average SE over Monte-Carlo samples, MCSE is Monte-Carlo SE, CP is coverage probability of β_1 in 95% confidence Interval, power in rejecting $H_0 : \beta_1 = 0$ is shown with Wald and score statistics.)

Estimator	Bias	AveSE	MCSE	CP	Power(Wald)	Power(Score)
			N=500			
$\hat{\beta}_1^{Com}$	-0.22	0.373	0.363	0.888	0.298	0.299
$\hat{\beta}_1^{gg}$	-0.063	0.125	0.177	0.796	0.998	0.998
$\hat{\beta}_1^{hh}$	-0.063	0.127	0.177	0.802	0.998	0.998
$\hat{\beta}_1^{IPW}$	-0.092	0.269	0.271	0.916	0.738	0.778
$\hat{\beta}_1^{IMP}$	-0.063	0.127	0.177	0.802	0.998	0.998
			N=1000			
$\hat{\beta}_1^{Com}$	-0.242	0.257	0.257	0.825	0.541	0.547
$\hat{\beta}_1^{gg}$	-0.060	0.091	0.127	0.775	1	1
$\hat{\beta}_1^{hh}$	-0.060	0.092	0.127	0.780	1	1
$\hat{\beta}_1^{IPW}$	-0.108	0.184	0.194	0.883	0.948	0.956
$\hat{\beta}_1^{IMP}$	-0.061	0.092	0.127	0.779	1	1

3.7 Application

3.7.1 A Cardiology Example

Table 3.7: Cardiology Data

	$R=1$				$R=0$	
	$D=1$		$D=0$			
T_1	T_2		T_2		T_2	
	(+)	(−)	(+)	(−)	(+)	(−)
(+)	110	2	85	12	434	723
(−)	78	3	138	25	459	348

To investigate the performance of the estimators, we apply the methodologies in cardiology dataset. The current gold standard for Coronary artery disease (CAD) is coronary angiography. Coronary angiography is invasive and it has a small risk of morbidity or mortality. Thus, gold standard is not observed in every patient in the dataset. Single-photon-emission computed tomography (SPECT) stress thallium is used as a diagnostic test. Patients needs to stimulate circulation so that distribution of blood flows in the heart can be ascertained by imaging. We consider two ways of stimulating circulation. The usual way is to use exercise, but some patients cannot exercise and diripidamole is used to stimulate the heart activity. Thus, we can think of having two diagnostic tests. The first diagnostic test is to stimulate blood circulation by exercise and the second diagnostic test is to use diripidamole. There are several prognostic factors which can be useful in predicting disease such as gender, age, and weight. The data without covariates are summarized in the Table 3.7.

There are 271 patients who have at least one of the covariates is missing, and we did not include the patients in our analysis. Out of total 2417 patients, only 453 patients, 19% have observed gold standard. The naive estimator which eliminates the missing gold standard where $R = 0$ has $\widehat{PPV_1} = 0.457$ and $\widehat{PPV_2} = 0.535$. The logit of the different is -0.315. Its Wald and score lead to the conclusion that there is a significant evidence that we can reject the null hypothesis. IPW which divides the estimating equation by the probability of observing the gold standard has the estimate of β_1, 0.197, and its score and Wald statistics result in the different conclusion from the complete case analysis. Similarly, augmented test statistics does not lead to reject the H_0. The R^2 is 0.19 which is weak association in

Table 3.8: Cardiology Example

Estimator	$\widehat{PPV_1}$	$\widehat{PPV_2}$	$\widehat{\beta}_1$	SE	Wald	Score
Com	0.457	0.536	-0.315	0.104	9.126	9.287
Aug.gg	0.449	0.410	0.162	0.133	1.480	1.486
Aug.hh	0.449	0.410	0.162	0.155	1.093	1.301
IPW	0.455	0.407	0.197	0.145	1.840	1.879
IMP	0.449	0.416	0.133	0.168	0.624	6.116

the covariate model logit $\Pr(\mathrm{D} = 1|\mathrm{T}_1, \mathrm{T}_2, \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \mathrm{C}_4, \mathrm{T}_1\mathrm{T}_2)$. With the similar spirit in the table 3.2, score statistic of imputation was quite different from Wald statistics. In our observation, when the data are performing in the NULL, where $\text{logit}\{\mathrm{PPV}_1\}$ and $\text{logit}\{\mathrm{PPV}_2\}$ are not different, and the covariates are not good indicators of disease, the score statistics behaves not reasonably. This is something that we need to look into more detail in further research.

3.7.2 A Coronary Stenosis Study

Table 3.9: Coronary Artery Disease Data.

	$R=1$				$R=0$	
	$D=1$		$D=0$			
T_1	T_2		T_2		T_2	
	(+)	(−)	(+)	(−)	(+)	(−)
			$C=1$			
(+)	224	18	38	6	31	21
(−)	1	1	32	35	24	219
			$C=0$			
(+)	37	0	92	12	16	15
(−)	0	0	79	57	70	522

Coronary stenosis is a heart disease whose patients have symptoms of narrowing or obstruction of heart arteries. The prevalence of disease is higher in men than women. The current gold standard is coronary angiography. However, it can give dangerous reactions to some patients such as arrhythmia, apoplexies et al. The diagnosis which is less invasive than coronary angiography is through dynamic trans thoracic echocardiography with effort or dynamic trans thoracic with dobutamine. The risk factors are arterial hypertension,

hypercholesterolemia, habitual smoking, diabetes, and family history of coronary heart disease. We were not able to access the whole data set, but we use the data which has been dichotomized the risk factors as 1 or 0 in [17]; 1 with patients who are recorded with more than two risk factors, and 0 with patients who are recorded with only one or zero risk factor. Their method is not convenient to incorporate continuous covariate since the distribution of covariate needs to be assumed and integration should be conducted. However, we do not have any distributional assumptions and our covariate model enables us to exploit the covariate and disease correlation. The data uses random variables as described. D is the outcome of coronary angiography which is not verified to all of the patients, T_1 is the outcome of dynamic trans thoracic with dobutamine, and T_2 is the outcome of dynamic trans thoracic with effort with one covariate C, which is determined by the number of risk factor. The data with one binary covariate is summarized in table 3.9. Among 1550 men, only 632, 40.8% individuals have verified gold standard. There are 281 men who are verified to have positive disease; out of 281, 244 are $C = 1$ which means that they have more than two risk factors, and 37 of them are $C = 0$. The covariate model $\text{logit}(\text{D} = 1|\text{T}_1, \text{T}_2, \text{C}, \text{T}_1\text{T}_2)$ has the generalized $R^2 = 50.8\%$, which is strong. Table 3.10 is the result using Complete case, augmented with gg and hh estimators, Inverse probability weighted estimator, and imputation method. Complete cases estimates PPV_1, PPV_2, and β_1 differently than the other estimators, but Aug.gg, Aug.hh, IPW, IMP have the similar estimates and the close Wald and score statistics. The variance estimates of Aug are also most efficient and have more powerful statistics than the IPW.

Table 3.10: Coronary Stenosis

Estimator	$\widehat{PPV_1}$	$\widehat{PPV_2}$	$\widehat{\beta}_1$	SE	Wald	Score
Com	0.521	0.653	-0.550	0.062	78.160	83.208
Aug.gg	0.456	0.639	-0.747	0.069	115.83	125.41
Aug.hh	0.456	0.639	-0.747	0.069	116.113	125.61
IPW	0.454	0.637	-0.748	0.074	102.697	111.86
IMP	0.456	0.639	-0.750	0.070	115.838	148.203

3.8 Discussion

In this section, we have used an imputation method for estimating positive predictive values and β_1 which is the logit of different of predictive values of T_1 and T_2. We

also derive the Wald and generalized score statistics using the M-estimation theory [12]. Through the simulation studies, we compared to the estimators described in Chapter 2. The imputation method is as powerful as the augmented method, and required one of the model to be correct. However, the covariate model may not be valid, and the imputation estimator does not perform well under incorrect covariate model. Furthermore, simulation Table 3.1 and through cardiology result Table 3.8 shows that. generalized score statistics does not perform correctly when the positive predictive values for T_1 and T_2 are the same. The misbehvarior of generalized score statistic should be investigated further.

Bibliography

[1] W Leisenring, T Alonzo, and MS Pepe. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics*, 56(2):345–351, 2000.

[2] KY LIANG and SL ZEGER. Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika*, 73(1):13–22, APR 1986.

[3] W Wang, CS Davis, and SJ Soong. Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine*, 25(13):2215–2229, 2006.

[4] AS Kosinski and HX Barnhart. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*, 59(1):163–171, MAR 2003.

[5] Margaret Sullivan Pepe. *The Statistical Evaluation Of Medical Tests For Classification And Prediction*. Oxford University Press, 2003.

[6] CB Begg and RA Greens. Assessment Of Diagnostic-Tests When Disease Verification Is Subject To Selection Bias. *Biometrics*, 39(1):207–215, 1983.

[7] Xh Zhou. Effect Of Verification Bias On Positive And Negative Predictive Values. *Statistics In Medicine*, 13(17):1737–1745, SEP 15 1994.

[8] José Antonio Roldán Nofuentes and Juan de Dios Luna del Castillo. Comparing two binary diagnostic tests in the presence of verification bias. *Computational Statistics & Data Analysis*, 50(6):1551–1564, 2006.

[9] Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

[10] JM ROBINS, A ROTNITZKY, and LP ZHAO. Estimation Of Regression-Coefficients When Some Regressors Are Not Always Observed. *Journal Of The American Statistical Association*, 89(427):846–866, SEP 1994.

[11] Ms Pepe and Gl Anderson. A Cautionary Note On Inference For Marginal Regression-Models With Longitudinal Data And General Correlated Response Data. *Communications In Statistics-Simulation And Computation*, 23(4):939–951, 1994.

[12] LA Stefanski and DD Boos. The calculus of M-estimation. *American Statistician*, 56(1):29–38, FEB 2002.

[13] A ROTNITZKY and NP JEWELL. Hypothesis-Testing Of Regression Parameters In Semiparametric Generalized Linear-Models For Cluster Correlated Data. *Biometrika*, 77(3):485–497, SEP 1990.

[14] DD BOOS. On Generalized Score Tests. *American Statistician*, 46(4):327–333, NOV 1992.

[15] SAS online Documentation. `http://support.sas.com/onlinedoc/913/docMainpage.jsp`.

[16] E. J. Snell David Roxbee Cox. *Analysis of Binary Data*. CRC Press, 1989.

[17] J. A. Roldan Nofuentes and J. D. Luna del Castillo. The effect of verification bias on the comparison of predictive values of two binary diagnostic tests. *Journal Of Statistical Planning And Inference*, 138(4):950–963, APR 1 2008.

[18] MS Pepe, TX Cai, and G Longton. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1):221–229, MAR 2006.

[19] D. B. Rubin. *Multiple imputation for nonresponse in survey*. Wiley, 1987.

[20] Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden, editors, *Handbook of Econometrics*.

[21] Andrea Rotnitzky, David Faraggi, and Enrique Schisterman. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verifica-

tion bias. *Journal Of The American Statistical Association*, 101(475):1276–1288, SEP 2006.

Appendices

APPENDIX A

Derivatives in Chapter 1

Let A be

$$
\begin{aligned}
&P(T_1 = 1, D = 1) \\
&= \frac{1}{((1 + exp(-\gamma_0 - \gamma_1 - \gamma_2 - \gamma_3)) \times (1 + exp(-\beta_0 - \beta_1)) \times (1 + exp(-\alpha_0)))} \\
&+ \frac{1}{((1 + exp(-\gamma_0 - \gamma_2)) \times (1 + exp(-\beta_0)) \times (1 + exp(\alpha_0)))}
\end{aligned}
$$

Let B be

$$
\begin{aligned}
P(T_1 = 0, D = 0) =& \frac{1}{((1 + exp(\gamma_0 + \gamma_1)) \times (1 + exp(\beta_0 + \beta_1)) \times (1 + exp(-\alpha_0)))} \\
&+ \frac{1}{((1 + exp(\gamma_0)) \times (1 + exp(\beta_0)) \times (1 + exp(\alpha_0)))}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial PPV_1}{\partial \alpha_0} = \frac{1}{P(T_1 = 1)}(\frac{\partial A}{\partial \alpha_0} - \frac{PPV_1}{P(T_1 = 1)} \times \frac{\partial P(T_1 = 1)}{\partial \alpha_0}), \\
&\frac{\partial PPV_1}{\partial \beta_i} = \frac{1}{P(T_1 = 1)}(\frac{\partial A}{\partial \beta_i} - \frac{PPV_1}{P(T_1 = 1)} \times \frac{\partial P(T_1 = 1)}{\partial \beta_i}), \quad \text{where} \quad i = 0, 1, \\
&\frac{\partial PPV_1}{\partial \gamma_i} = \frac{1}{P(T_1 = 1)}(\frac{\partial A}{\partial \gamma_i} - \frac{PPV_1}{P(T_1 = 1)} \times \frac{\partial P(T_1 = 1)}{\partial \gamma_i}), \quad \text{where} \quad i = 0, 1, 2, 3,
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial NPV_1}{\partial \alpha_0} = \frac{1}{P(T_1 = 0)}(\frac{\partial B}{\partial \alpha_0} + \frac{NPV_1}{P(T_1 = 1)} \times \frac{\partial P(T_1 = 1)}{\partial \alpha_0}), \\
&\frac{\partial NPV_1}{\partial \beta_i} = \frac{1}{P(T_1 = 0)}(\frac{\partial B}{\partial \beta_i} + \frac{NPV_1}{P(T_1 = 1)} \times \frac{\partial P(T_1 = 1)}{\partial \beta_i}) \quad \text{where} \quad i = 0, 1, \\
&\frac{\partial NPV_1}{\partial \gamma_i} = \frac{1}{P(T_1 = 0)}(\frac{\partial B}{\partial \gamma_i} + \frac{NPV_1}{P(T_1 = 1)} \times \frac{\partial P(T_1 = 1)}{\partial \gamma_i}) \quad \text{where} \quad i = 0, 1, 2, 3,
\end{aligned}
$$

where

$$\frac{\partial A}{\partial \alpha_0} = \frac{exp(-\alpha_0)}{((1+exp(-\gamma_0-\gamma_1-\gamma_2-\gamma_3))\times(1+exp(-\beta_0-\beta_1))\times(1+exp(-\alpha_0))^2)} - \frac{exp(\alpha_0)}{((1+exp(-\gamma_0-\gamma_2))\times(1+exp(-\beta_0))\times(1+exp(\alpha_0))^2)}$$

$$\frac{\partial B}{\partial \alpha_0} = \frac{exp(-\alpha_0)}{((1+exp(\gamma_0+\gamma_1))\times(1+exp(\beta_0+\beta_1))\times(1+exp(-\alpha_0))^2)} - \frac{exp(\alpha_0)}{((1+exp(\gamma_0))\times(1+exp(\beta_0))\times(1+exp(\alpha_0))^2)}$$

$$\frac{\partial A}{\partial \beta_0} = \frac{exp(-\beta_0-\beta_1)}{((1+exp(-\gamma_0-\gamma_1-\gamma_2-\gamma_3))\times(1+exp(-\beta_0-\beta_1))^2\times(1+exp(-\alpha_0)))} + \frac{exp(-\beta_0)}{((1+exp(-\gamma_0-\gamma_2))\times(1+exp(\alpha_0))\times(1+exp(-\beta_0))^2)}$$

$$\frac{\partial B}{\partial \beta_0} = \frac{-exp(\beta_0+\beta_1)}{((1+exp(\gamma_0+\gamma_1))\times(1+exp(\beta_0+\beta_1))^2\times(1+exp(-\alpha_0)))} - \frac{exp(\beta_0)}{((1+exp(\gamma_0))\times(1+exp(\alpha_0))\times(1+exp(\beta_0))^2)}$$

$$\frac{\partial A}{\partial \beta_1} = \frac{exp(-\beta_0-\beta_1)}{((1+exp(-\gamma_0-\gamma_1-\gamma_2-\gamma_3))\times(1+exp(-\alpha_0))\times(1+exp(-\beta_0-\beta_1))^2)}$$

$$\frac{\partial B}{\partial \beta_1} = \frac{-exp(\beta_0+\beta_1)}{((1+exp(\gamma_0+\gamma_1))\times(1+exp(-\alpha_0))\times(1+exp(\beta_0+\beta_1))^2)}$$

$$\frac{\partial A}{\partial \gamma_0} = \frac{exp(-\gamma_0-\gamma_1-\gamma_2-\gamma_3)}{((1+exp(-\beta_0-\beta_1))\times(1+exp(-\alpha_0))\times(1+exp(-\gamma_0-\gamma_1-\gamma_2-\gamma_3))^2)} + \frac{exp(-\gamma_0-\gamma_2)}{((1+exp(-\beta_0))\times(1+exp(\alpha_0))\times(1+exp(-\gamma_0-\gamma_2))^2)}$$

$$\frac{\partial B}{\partial \gamma_0} = \frac{-exp(\gamma_0+\gamma_1)}{((1+exp(\beta_0+\beta_1))\times(1+exp(-\alpha_0))\times(1+exp(\gamma_0+\gamma_1))^2)} - \frac{exp(\gamma_0)}{(1+exp(\beta_0))\times(1+exp(\alpha_0))\times(1+exp(\gamma_0))^2)}$$

$$\frac{\partial A}{\partial \gamma_1} = \frac{exp(-\gamma_0 - \gamma_1 - \gamma_2 - \gamma_3)}{((1 + exp(-\beta_0 - \beta_1)) \times (1 + exp(-\alpha_0)) \times (1 + exp(-\gamma_0 - \gamma_1 - \gamma_2 - \gamma_3))^2)}$$

$$\frac{\partial B}{\partial \gamma_1} = \frac{-exp(\gamma_0 + \gamma_1)}{((1 + exp(\beta_0 + \beta_1)) \times (1 + exp(-\alpha_0)) \times (1 + exp(\gamma_0 + \gamma_1))^2)}$$

$$\frac{\partial A}{\partial \gamma_2} = \frac{exp(-\gamma_0 - \gamma_1 - \gamma_2 - \gamma_3)}{((1 + exp(-\beta_0 - \beta_1)) \times (1 + exp(-\alpha_0)) \times (1 + exp(-\gamma_0 - \gamma_1 - \gamma_2 - \gamma_3))^2)} + \frac{exp(-\gamma_0 - \gamma_2)}{((1 + exp(-\gamma_0 - \gamma_2)) \times (1 + exp(-\beta_0))^2 \times (1 + exp(\alpha_0)))}$$

$$\frac{\partial B}{\partial \gamma_2} = 0$$

$$\frac{\partial A}{\partial \gamma_3} = \frac{exp(-\gamma_0 - \gamma_1 - \gamma_2 - \gamma_3)}{((1 + exp(-\beta_0 - \beta_1)) \times (1 + exp(-\alpha_0)) \times (1 + exp(-\gamma_0 - \gamma_1 - \gamma_2 - \gamma_3))^2)}$$

$$\frac{\partial B}{\partial \gamma_3} = 0$$