

Statistical Inference in ROC Curve Analysis

by

Dingding Hu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2025

© Dingding Hu 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Jialiang Li
 Professor
 National University of Singapore

Supervisor: Pengfei Li
 Professor

Internal Member: Changbao Wu
 Professor

 Peijun Sang
 Associate Professor

 Qinglong Tian
 Assistant Professor

Internal-External Member: Ashok Chaurasia
 Associate Professor, School of Public Health Sciences

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

A version of Chapter 2 of this thesis has been published by the *Statistics in Medicine*. A version of Chapter 3 of this thesis has been prepared as a research paper submitted to the *Canadian Journal of Statistics*, which is currently under the second revision. A version of Chapter 4 is under preparation to submit to a journal for publication. My contributions include formulating the research problems, deriving main theorems, implementing the proposed methods through R programming, performing the simulations and real data analysis, and writing the initial drafts.

Abstract

The receiver operating characteristic (ROC) curve is a powerful statistical tool to evaluate the diagnostic abilities of a binary classifier for varied discrimination thresholds. It has been widely applied in various scientific areas. This thesis considers three inference problems in the ROC curve analysis.

In Chapter 1, we introduce the basic concept of the ROC curve, along with some of its summary indices. We then provide an overview of the research problems and outline the structure of the subsequent chapters.

Chapter 2 focuses on improving the ROC curve analysis with a single biomarker by incorporating the assumption that higher biomarker values indicate greater disease severity or likelihood. We interpret “greater severity” as a higher probability of disease, which corresponds to the likelihood ratio ordering between diseased and healthy individuals. Under this assumption, we propose a Bernstein polynomial-based method to model and estimate the biomarker distributions using the maximum empirical likelihood framework. From the estimated distributions, we derive the ROC curve and its summary indices. We establish the asymptotic consistency of our estimators and validate their performance through extensive simulations and compare them with existing methods. A real-data example is used to demonstrate the practical applicability of our approach.

Chapter 3 considers the ROC curve analysis for medical data with non-ignorable missingness in the disease status. In the framework of the logistic regression models for both the disease status and the verification status, we first establish the identifiability of model parameters, and then propose a likelihood method to estimate the model parameters, the ROC curve, and the area under the ROC curve (AUC) for the biomarker. The asymptotic distributions of these estimators are established. Via extensive simulation studies, we com-

pare our method with competing methods in the point estimation and assess the accuracy of confidence interval estimation under various scenarios. To illustrate the application of the proposed method in practical data, we apply our method to the Alzheimer’s disease dataset from the National Alzheimer’s Coordinating Center.

Chapter 4 explores the combination of multiple biomarkers when disease status is determined by an imperfect reference standard, which can lead to misclassification. Previous methods for combining multiple biomarkers typically assume that all disease statuses are determined by a gold standard test, limiting their ability to accurately estimate the ROC curve and AUC in the presence of misclassification. We propose modeling the distributions of biomarkers from truly healthy and diseased individuals using a semiparametric density ratio model. Additionally, we adopt two assumptions from the literature: (1) the biomarkers are conditionally independent of the classification of the imperfect reference standard given the true disease status, and (2) the classification accuracy of the imperfect reference standard is known. Using this framework, we establish the identifiability of model parameters and propose a maximum empirical likelihood method to estimate the ROC curve and AUC for the optimal combination of biomarkers. An Expectation-Maximization algorithm is developed for numerical calculation. Additionally, we propose a bootstrap method to construct the confidence interval for the AUC and the confidence band for the ROC curve. Extensive simulations are conducted to evaluate the robustness of our method with respect to label misclassification. Finally, we demonstrate the effectiveness of our method in a real-data application.

In Chapter 5, we provide a brief summary of Chapters 2–4 and outline several directions for future research.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Pengfei Li, for his invaluable guidance, support, and encouragement throughout my PhD studies. Despite my lack of prior experience in statistical research, he welcomed me into the field and patiently guided me through every stage of my PhD career. I would not have remained as calm, confident, and enthusiastic without his continuous inspiration, unwavering patience, and profound knowledge. His high standards and meticulous approach to academic research will continue to guide me throughout my future career.

I would like to thank all my committee members, Dr. Ashok Chaurasia, Dr. Changbao Wu, Dr. Jialiang Li, Dr. Peijun Sang, and Dr. Qinglong Tian, for dedicating their time to reviewing my thesis and providing constructive feedback.

I would like to thank my parents, who have always been my strongest support since my first arrival in Canada. I would like to thank my friends, Chi, Yuliang, Zijin, and many others, for making my PhD journey more enjoyable and memorable. I would also like to thank all the instructors from my undergraduate and graduate courses, whose teaching provided me with a strong foundation and gradually sparked my interest in statistics. I would like to thank Greg Preston for his consistent support and assistance with administrative matters. I would especially like to thank Mary Lou Dufton for always being a strong support for me and many other graduate students whenever we faced difficulties. I would like to extend my appreciation to all my friends and department staff members not mentioned by name, for their kindness and support.

Last but not least, I would like to express my heartfelt appreciation to Xianwei for her love and support in both life and research. PhD journey has been both long and fleeting. Your companionship has made it meaningful and memorable.

Dedication

*To my beloved parents, grandparents and cousins for their love, support and
companionship.*

*To the friendships I have gained and the many things I have learned throughout this
journey.*

Table of Contents

List of Figures	xiii
List of Tables	xiv
1 Introduction	1
1.1 Receiver operating characteristic curve	1
1.2 Summary indices of the ROC curve	3
1.3 Outline of the thesis	4
2 ROC Curve Analysis under the Two-sample Likelihood Ratio Ordering	7
2.1 Introduction	7
2.2 Main results	10
2.2.1 Bernstein polynomials approach	10
2.2.2 Estimation of the ROC curve and its summary quantities	18
2.2.3 Asymptotic properties	19
2.3 Simulation study	20

2.3.1	Simulation setup	20
2.3.2	Comparison of the ROC curve estimation	23
2.3.3	Comparison of the AUC estimation	23
2.3.4	Comparison of the Youden index and optimal cutoff point estimation	25
2.4	Real data application	27
2.5	Technical details and additional simulation results	33
2.5.1	Additional simulation results under the normal and Gamma distri- bution settings	33
2.5.2	Simulation results under the Beta distribution settings	37
2.5.3	Proof of Proposition 2.1	42
2.5.4	Proof of Theorem 2.1	43
2.5.5	Proof of Theorem 2.2	51
3	ROC Curve Analysis With Non-Ignorable Missing Disease Status	59
3.1	Introduction	59
3.2	Main results	61
3.2.1	Setup	61
3.2.2	Identifiability of model parameters	63
3.2.3	Estimation of model parameters	64
3.2.4	Estimation of the ROC curve and AUC	65
3.2.5	Model verification	67

3.3	Asymptotic properties	68
3.4	Simulation studies	72
3.4.1	Simulation setup	73
3.4.2	AUC point estimation results	75
3.4.3	AUC CIs estimation results	76
3.4.4	ROC curve point estimation results	77
3.4.5	ROC curve CIs estimation results	79
3.5	Real data application	80
3.6	Technical details and additional simulation results	85
3.6.1	Regularity conditions	85
3.6.2	Proof of Theorem 3.1	86
3.6.3	Proof of Proposition 3.1	88
3.6.4	Proof of Theorem 3.2	90
3.6.5	Additional simulation results	103
4	ROC Curve Analysis when Combining Multiple Biomarkers in Presence of Imperfect Reference Standard	112
4.1	Introduction	112
4.2	Main results	116
4.2.1	Problem setup and model identifiability	116
4.2.2	Empirical likelihood method	119

4.3	EM algorithm	122
4.4	Simulation	126
4.4.1	Simulation setup	126
4.4.2	Simulation results for the AUC	127
4.4.3	Simulation results for the ROC curve	128
4.5	Real data application	133
4.6	Technical details	136
4.6.1	Proof of Theorem 4.1	136
4.6.2	Details of the M-step	138
4.6.3	Proof of Theorem 4.2	141
5	Discussion and Future Work	143
5.1	Summary of the thesis	143
5.2	Future work	145
	Bibliography	148

List of Figures

1.1	Two examples of the ROC curves for two different diagnostic tests.	2
2.1	Panel (a) shows $BIC(N)$ versus N ; Panel (b) compares the proposed estimators $(\hat{F}_0, \hat{F}_1)$ and the empirical cdfs $(\tilde{F}_0, \tilde{F}_1)$	30
2.2	Estimated ROC curves based on different methods.	31
3.1	The true ROC curve, the average estimated curve, and its 95% confidence band based on 1000 replications under Scenarios 1–3.	79
3.2	Estimated ROC curves and 95% confidence band.	84
3.3	The true ROC curve, the average estimated curve, and its 95% confidence band based on 1000 replications under Scenarios 4–6.	108
4.1	True ROC curves and average estimated ROC curves using our and DRM methods under Scenario 1.	130
4.2	True ROC curves and average estimated ROC curves using our and DRM methods under Scenario 2.	131
4.3	Estimated ROC curves and 95% CB for DMD dataset.	135

List of Tables

2.1	Simulation settings	22
2.2	Averages of L_1 -distances and L_2 -distances of five methods for estimating the ROC curve	24
2.3	RB (%) and MSE ($\times 1000$) of five methods for estimating the AUC	26
2.4	RB (%) and MSE ($\times 1000$) of five methods for estimating the Youden index	28
2.5	RB (%) and MSE ($\times 1000$) of five methods for estimating the optimal cutoff point	29
2.6	PEs and BPCIs of the AUC, Youden index, and optimal cutoff point by using CK as the biomarker	32
2.7	Frequency of the N values when distributions are normal and Gamma	33
2.8	Averages of L_1 - and L_2 -distances of the ROC curve estimates for other methods	34
2.9	RB (%) and MSE ($\times 1000$) of the AUC estimates for other methods	35
2.10	RB (%) and MSE ($\times 1000$) of the Youden index estimates for other methods	36
2.11	RB (%) and MSE ($\times 1000$) of the optimal cutoff point estimates for other methods	37

2.12	Beta simulation setting	38
2.13	Frequency of the chosen N in Beta distributional setting	38
2.14	Averages of L_1 -distances and L_2 -distances of eight methods for estimating the ROC curve under Beta distribution settings	39
2.15	RB (%) and MSE ($\times 1000$) of eight methods for estimating the AUC under Beta distribution settings	40
2.16	RB (%) and MSE ($\times 1000$) of eight methods for estimating the Youden index under Beta distribution settings	41
2.17	RB (%) and MSE ($\times 1000$) of eight methods for estimating the optimal cutoff point under Beta distribution settings	42
3.1	Summary quantities for Scenarios 1–3	74
3.2	RB (%) and MSE ($\times 1000$) of AUC estimates under Scenarios 1–3	75
3.3	CP and AL of estimated CIs for AUC by IPW and Our methods under Scenarios 1–3	76
3.4	RB (%) and MSE ($\times 1000$) of $ROC(s)$ estimates for $s = 0.1$ and 0.2 under Scenarios 1–3	78
3.5	CP and AL of estimated CIs for $ROC(s)$ with $s = 0.1$ and $s = 0.2$ under Scenarios 1–3	80
3.6	Summary statistics for NACC data on patients aged 65 to 70	81
3.7	Estimated model parameters and their standard errors (Significant coefficients at the 5% level are bolded)	83
3.8	AUC and 95% CI estimates for the NACC data	83

3.9	Summary quantities for Scenarios 4–6	104
3.10	RB (%) and MSE ($\times 1000$) of AUC estimates under Scenarios 4–6	105
3.11	CP and AL of estimated CIs for AUC by IPW and Our methods under Scenarios 4–6	106
3.12	RB (%) and MSE ($\times 1000$) of $ROC(s)$ estimates for $s = 0.1$ under Scenarios 4–6	106
3.13	RB (%) and MSE ($\times 1000$) of $ROC(s)$ estimates for $s = 0.2$ under Scenarios 4–6	107
3.14	CP and AL of estimated CIs for $ROC(s)$ with $s = 0.1$ and $s = 0.2$ under Scenarios 4–6	107
3.15	CP and AL of two types of CIs for AUC under Scenarios 1–3	109
3.16	CP and AL of two types of CIs for AUC under Scenarios 4–6	110
3.17	CP and AL of two types of CIs for $ROC(s)$ at $s = 0.1$ and $s = 0.2$ under Scenarios 1–3	110
3.18	CP and AL of two types of CIs for $ROC(s)$ at $s = 0.1$ and $s = 0.2$ under Scenarios 4–6	111
4.1	Summary of AUC, $ROC(0.1)$, $ROC(0.2)$ for the optimal combination of biomarkers in each scenario	127
4.2	RB (%) and MSE ($\times 1000$) for the AUC estimation under 2 scenarios	128
4.3	CP and AL for the AUC confidence interval estimations under 2 scenarios .	129
4.4	RB (%) and MSE ($\times 1000$) of $ROC(s)$ estimates for $s = 0.1$ and $s = 0.2$ under 2 scenarios	132

4.5	CP and AL of estimated CIs for $ROC(s)$ with $s = 0.1$ and $s = 0.2$ under 2 scenarios	133
4.6	Estimated mean and 95% CI for the AUC based on 1000 shuffled datasets	134

Chapter 1

Introduction

1.1 Receiver operating characteristic curve

This thesis considers several important problems associated with receiver operating characteristic (ROC) curve and its summary indices. In this section, we give a brief introduction to the ROC curve.

The ROC curve is a powerful statistical tool and has been widely applied in many scientific areas; it evaluates the diagnostic abilities of a binary classifier for varied discrimination thresholds. Consider a medical study, where a biomarker serves as a binary classifier for diagnosing the healthy/diseased status of each patient. The ROC curve illustrates the relationship between “sensitivity” and “one minus specificity” across all the possible thresholds of the biomarker. In this context, sensitivity/specificity refers to the probability of correctly identifying the diseased/healthy individuals.

Practically, for an individual, we may assume that the larger biomarker value indicates greater possibility/severity of the disease; therefore, we can classify individuals as diseased

if their corresponding biomarkers exceed a given cutoff point x , and vice versa; see [Yin & Tian \(2014\)](#) for a discussion. Under this assumption, the sensitivity and specificity can be expressed as $1 - F_1(x)$ and $F_0(x)$, where $F_0(\cdot)$ and $F_1(\cdot)$ represent the cumulative distribution functions (cdfs) of the biomarkers for individuals that are healthy and diseased, respectively. The ROC curve is then given by

$$ROC(s) = 1 - F_1(F_0^{-1}(1 - s)), \quad (1.1)$$

for $s \in (0, 1)$.

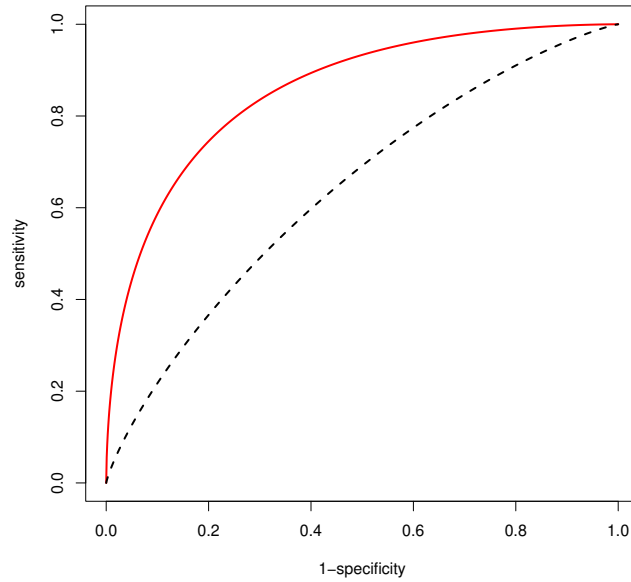


Figure 1.1: Two examples of the ROC curves for two different diagnostic tests.

Two examples of ROC curves are shown in Figure 1.1. From the plot, the diagnostic test for the red ROC curve demonstrates a uniformly better discrimination ability than

the test for the black ROC curve. At each specificity level, the red curve shows a higher sensitivity level. However, this may not always be the case in ROC curve comparisons. We may encounter the situations that one curve does not show advantage at each level over the other, which adds complexity to the ROC curves comparison. Ideally, it is desirable to summarize the accuracy with a single index.

In the next section, we introduce some summary indices of the ROC curve in order to make comparison between ROC curves from various perspectives.

1.2 Summary indices of the ROC curve

One of the most popular summary indices of the ROC curve is the area under the ROC curve (AUC). Mathematically, it can be written as

$$AUC = \int_0^1 ROC(s) ds. \quad (1.2)$$

A direct interpretation of the AUC is the average sensitivity level over all specificity levels. With $ROC(s)$ in (1.1), we can verify that

$$AUC = \int_0^1 \{1 - F_1(F_0^{-1}(1 - s))\} ds = P(X \leq Y),$$

where X and Y have continuous cdfs F_0 and F_1 , respectively. The AUC is then interpreted as the probability that a biomarker of a diseased individual being larger than that of a healthy individual. Note that for a given threshold or cutoff point of the biomarker, it is desirable to have the corresponding value on y -axis (the sensitivity) to be as large as possible. Hence, the larger AUC value indicates that the binary classifier has stronger

classification ability.

Another commonly used summary index of the ROC curve is the Youden index (Youden 1950), which is defined as the maximum value of the sensitivity plus the specificity minus 1, i.e.,

$$J = \max_x \{1 - F_1(x) + F_0(x) - 1\} = \max_x \{F_0(x) - F_1(x)\}. \quad (1.3)$$

One advantage of the Youden index is that it results in a criterion to choose the “optimal” cutoff point, which is the “ x ” where Youden index is achieved. Specifically

$$C = \arg \max_x \{F_0(x) - F_1(x)\}. \quad (1.4)$$

The Youden index lies in $[0, 1]$, with $J = 1$ means that the diagnostic test under the optimal cutoff point can completely separate the diseased and non-diseased individuals, and $J = 0$ means that the test has the predictive accuracy similar to a random guess.

In the literature, the ROC curve, its summary indices, and related inferences have been extensively investigated; we refer to Pepe (2003), Zhou et al. (2011), Qin & Zhang (2003), Chen et al. (2016), Sande et al. (2021), Yuan et al. (2021), and the references therein for a comprehensive review of the existing developments. For convenience of presentation, the literature review for each research problem in this thesis will be provided in the corresponding chapter.

1.3 Outline of the thesis

In this thesis, we focus on statistical inference of the ROC curve and its summary indices under three different problem settings.

In Chapter 2, we consider how to effectively incorporate the assumption that a higher value of the biomarker indicates a greater possibility/severity of the disease to improve the ROC curve analysis with a single biomarker. We mathematically interpret “greater severity of the disease” as “larger probability of being diseased”. This in turn is equivalent to assuming the likelihood ratio ordering of the biomarker between the diseased and healthy individuals. With this assumption, we first propose a Bernstein polynomial method to model the distributions of both samples; we then estimate the distributions by the maximum empirical likelihood principle. The ROC curve estimate and the associated summary indices are obtained subsequently. Theoretically, we establish the consistency of our estimators. Via extensive numerical studies, we compare the performance of our method with competitive methods. The application of our method is illustrated by a real-data example.

In Chapter 3, we consider the ROC curve analysis for medical data with non-ignorable missingness in the disease status. [Yu et al. \(2018\)](#) proposed to model the disease probability (given the biomarker and relative covariates) for the individuals with disease status by a logistic regression model, and the verification probability or the missing probability (given disease status, biomarker, and relative covariates) by another logistic regression model. They further proposed a score method to estimate the unknown model parameters and an inverse probability weighting method to estimate the AUC. Their estimate of model parameters does not maximize the full likelihood, and further their AUC estimate only uses the data with disease status. Both leave the room for improvement. In Chapter 3, we first establish the identifiability of model parameters, and then propose a maximum full likelihood method to estimate the model parameters, the ROC curve, and the AUC for the biomarker. The asymptotic distributions of these estimators are established. Via extensive simulation studies, we compare our method with competing methods in the point estimation and assess the accuracy of confidence interval estimation under various

scenarios. To illustrate the application of the proposed method in practical data, we apply our method to the Alzheimer’s disease dataset from the National Alzheimer’s Coordinating Center (NACC).

In Chapter 4, we address the problem of combining multiple biomarkers in the presence of an imperfect reference standard. Because the reference standard is not perfect, it leads to misclassified disease status. To account for this, we propose linking the distributions of biomarkers for truly healthy and diseased individuals using a density ratio model (DRM) (Qin & Zhang 2010). Under the assumptions of conditional independence, known accuracy of the reference standard, and the DRM, we establish the identifiability of the model parameters. We then propose a maximum empirical likelihood method to estimate the ROC curve and the AUC for the optimal combination of biomarkers. An Expectation-Maximization algorithm is developed for parameter estimation, and a bootstrap method is used to construct confidence intervals for the AUC and confidence bands for the ROC curve. Simulation studies and a real data application are conducted to demonstrate the performance of the proposed method.

In Chapter 5, we provide a brief summary of Chapters 2–4 and outline several directions for future research.

Chapter 2

ROC Curve Analysis under the Two-sample Likelihood Ratio Ordering

2.1 Introduction

In this chapter, we consider statistical inference on the ROC curve and its summary indices under two-sample likelihood ratio ordering. Recall that in Chapter 1 we use $F_0(\cdot)$ and $F_1(\cdot)$ to denote the cdfs of the biomarkers for individuals that are healthy and diseased respectively. Under the assumption that the larger biomarker value indicates greater possibility/severity of the disease, the ROC curve is given by

$$ROC(s) = 1 - F_1(F_0^{-1}(1 - s)), \quad (2.1)$$

for $s \in (0, 1)$.

In the literature, there are two common strategies for mathematically accommodating the assumption that larger biomarker associates with the greater severity of the disease. The first interprets it as the biomarker values of diseased individuals are stochastically greater than those of healthy ones; that is $F_1(x) \leq F_0(x)$ for every $x \in \mathbb{R}$ (Wang et al. 2017). The second approach assumes that a higher biomarker value corresponds to a greater probability of having the disease (Yu et al. 2017). Under the assumption that “a higher biomarker value indicates greater disease severity,” this strategy is effectively equivalent to interpreting “greater disease severity” as “a higher probability of being diseased.” More specifically, let X and D denote the biomarker and indicator of the membership of an individual: $D = 1$ or 0 respectively indicates that the individual is diseased or healthy. Using Bayes’ formula, we have

$$P(D = 1|X = x) = \frac{P(D = 1)f_1(x)}{P(D = 1)f_1(x) + P(D = 0)f_0(x)}; \quad (2.2)$$

where $f_1(x)$ and $f_0(x)$ are the probability density functions (pdfs) of the biomarkers in the diseased population and the healthy population, respectively. Hence, the second interpretation (Yu et al. 2017) equivalently assumes that $f_1(x)/f_0(x)$ is a monotone function of x , which is known as the likelihood ratio ordering (Dykstra et al. 1995). We observe that with the likelihood ratio ordering, the derivative of $ROC(s)$,

$$\frac{d\{ROC(s)\}}{ds} = \frac{f_1(F_0^{-1}(1-s))}{f_0(F_0^{-1}(1-s))},$$

is a nonincreasing function, which implies that $ROC(s)$ is concave for $s \in (0, 1)$. We refer to Gneiting & Vogel (2018) for the importance of concavity in the interpretation and modelling of ROC curves. Furthermore, we observe that as an assumption, likelihood ratio

ordering is also widely adopted in many important applications of the Neyman–Pearson lemma, such as the Karlin–Rubin theorem. In this chapter, we shall incorporate the likelihood ratio ordering assumption to improve the estimation of the ROC curve and its summary indices.

There are three types of estimation methods for the ROC curves: the parametric, the semiparametric (Zhou & Lin 2008, Qin & Zhang 2003, Yuan et al. 2021), and the nonparametric methods (Zhou et al. 2011). We observe that many of these estimation methods did not incorporate the likelihood ratio ordering constraint; ignoring this constraint may result in less accurate ROC estimation. As far as we are aware, there are two estimation methods in the literature that have accommodated this constraint: Dykstra et al. (1995) considered the estimation of F_0 and F_1 with a maximum nonparametric likelihood approach; Yu et al. (2017) proposed a smoothed likelihood approach to estimate f_0 and f_1 . The former leads to a non-smooth estimate of $P(D = 1|X = x)$ in (2.2), resulting in an inefficient estimate of the optimal cutoff point that maximizes the Youden index. The latter relies on a smoothing parameter the optimal choice of which is not easily obtained. For a theoretical comparison of the convergence rates of the optimal cutoff point between the non-smooth and smoothed estimates of $P(D = 1|X = x)$, we refer to Hsieh & Turnbull (1996); the numerical performance can be found in Sections 2.3 and 2.5.

In this chapter, we propose a Bernstein polynomial method that accommodates the likelihood ratio ordering assumption. We further use the empirical likelihood principle to estimate F_0 , F_1 , ROC curve, and two popular summary indices: the AUC and the Youden index. With the help of the existing R function `glmnet` (Friedman et al. 2010), we establish an algorithm to implement our proposed method; we integrate our algorithm as an R package BPLR available at “<https://github.com/Dingding-Hu/BPLR-package>”. We also establish the consistency of our method. It is noteworthy that the resulting estimate

of $P(D = 1|X = x)$ from our method is smooth; through extensive numerical studies, we observe that our ROC curve, Youden index, and the associated optimal cutoff point estimates outperform those from the existing methods in most examples.

The rest of the chapter is organized as follows. Section 2.2 proposes the Bernstein polynomial method with the empirical likelihood principle to establish the estimators for $F_0(\cdot)$ and $F_1(\cdot)$, and subsequently result in the estimators of the ROC curve and its summary indices. The consistency of the proposed method is also studied. Section 2.3 presents the simulation results, and Section 2.4 compares our method with existing methods in a real-data example. For clarity, we put the technical details and some additional simulation results in Section 2.5.

2.2 Main results

2.2.1 Bernstein polynomials approach

Denote by $\{X_1, \dots, X_{n_0}\}$ and $\{Y_1, \dots, Y_{n_1}\}$ the random samples of biomarkers from the healthy and the diseased populations, respectively. The pdf and cdf of X_i 's are f_0 and F_0 ; the pdf and cdf of Y_j 's are f_1 and F_1 . Let $n = n_0 + n_1$ be the total sample size; denote $t_1 < \dots < t_m$ as the distinct values of the combined sample.

We assume that f_0 and f_1 satisfy the likelihood ratio ordering, i.e., $f_1(x)/f_0(x)$ is an increasing function of x . To incorporate this constraint in the estimation procedure, we propose to approximate $\log\{f_1(x)/f_0(x)\}$ with a linear combination of Bernstein polynomials. The definition of the Bernstein polynomials is given below.

Definition 2.1. (*Bernstein polynomials*) Let N be a positive integer. For each $0 \leq l \leq N$,

the Bernstein polynomials are defined to be

$$B_l(x; N) = \binom{N}{l} x^l (1-x)^{N-l}, \quad l = 0, \dots, N, \quad x \in [0, 1].$$

Bernstein polynomials can serve as a set of base functions in the expansion of a nonparametric component. A nice feature of such an expansion is that it is able to incorporate the shape constraints as some condition(s) of the coefficients of polynomials. For a monotonic nonparametric component, we can expand it to be $\sum_{l=0}^N \beta_l B_l(x; N)$, which is an increasing function of x if

$$\beta_0 \leq \beta_1 \leq \dots \leq \beta_N. \quad (2.3)$$

We refer to [Wang & Ghosh \(2012\)](#) for more details of incorporating other types of shape constraints in the Bernstein polynomial expansion.

Next, we apply the Bernstein polynomials to our estimation problem. For presentational convenience, in the development below, we assume that N is given and all biomarkers are ranged in $[0, 1]$. [Remarks 2.1](#) and [2.2](#) give the solutions of how to choose N and transform the biomarkers so that they are contained in $[0, 1]$ in practice. Based on the likelihood ratio ordering assumption, $\log\{f_1(x)/f_0(x)\}$ is an increasing function of x . We propose to approximate it by a linear combination of Bernstein polynomials:

$$\log\{f_1(x)/f_0(x)\} \approx \sum_{l=0}^N \beta_l B_l(x; N), \quad (2.4)$$

with $(\beta_0, \dots, \beta_N)$ satisfying [\(2.3\)](#); and then incorporate the maximum empirical likelihood

to estimate $(\beta_0, \dots, \beta_N)$, F_0 , and F_1 . To this end, let

$$a_i = \sum_{j=1}^{n_0} I(X_j = t_i) \quad \text{and} \quad b_i = \sum_{j=1}^{n_1} I(Y_j = t_i).$$

Based on the observed two-sample data, the full likelihood function is given by

$$L = \prod_{i=1}^m \{f_0(t_i)\}^{a_i} \cdot \{f_1(t_i)\}^{b_i}.$$

For $i = 1, \dots, m$, let

$$p_{i0} = f_0(t_i) \quad \text{and} \quad p_{i1} = f_1(t_i).$$

Then, the empirical likelihood ([Owen 2001](#)) is

$$L = \prod_{i=1}^m p_{i0}^{a_i} \cdot p_{i1}^{b_i}.$$

We assume $n_1/n \rightarrow \lambda \in (0, 1)$ as $n \rightarrow \infty$. For simplicity, hereafter we write $\lambda = n_1/n$ and assume that it is constant, since it does not affect our technical development. We further define

$$\phi_i = (1 - \lambda)f_0(t_i) + \lambda f_1(t_i) = (1 - \lambda)p_{i0} + \lambda p_{i1}.$$

By model [\(2.4\)](#), we have

$$\begin{aligned} p_{i0} &= \frac{\phi_i}{1 - \lambda + \lambda \exp \left\{ \sum_{l=0}^N \beta_l B_l(t_i; N) \right\}}, \\ p_{i1} &= \frac{\exp \left\{ \sum_{l=0}^N \beta_l B_l(t_i; N) \right\} \phi_i}{1 - \lambda + \lambda \exp \left\{ \sum_{l=0}^N \beta_l B_l(t_i; N) \right\}}. \end{aligned}$$

That is, p_{i0} 's and p_{i1} 's are determined by ϕ_i 's and $(\beta_0, \dots, \beta_N)$.

With the above reparameterization, the empirical likelihood function of $(\phi_1, \dots, \phi_m, \beta_0, \dots, \beta_N)$ is then given as

$$L = \prod_{i=1}^m p_{i0}^{a_i} p_{i1}^{b_i} = (\lambda)^{-n_1} (1 - \lambda)^{-n_0} \cdot L_1(\phi_1, \dots, \phi_m) \cdot L_2(\beta_0, \dots, \beta_N),$$

where

$$L_1(\phi_1, \dots, \phi_m) = \prod_{i=1}^m \phi_i^{a_i + b_i} \quad \text{and} \quad L_2(\beta_0, \dots, \beta_N) = \prod_{i=1}^m \left[\{\theta(t_i)\}^{b_i} \{1 - \theta(t_i)\}^{a_i} \right]$$

with

$$\theta(x) = \frac{\lambda \exp \left\{ \sum_{l=0}^N \beta_l B_l(x; N) \right\}}{1 - \lambda + \lambda \exp \left\{ \sum_{l=0}^N \beta_l B_l(x; N) \right\}}.$$

Note that feasible ϕ_i 's satisfy

$$\phi_i \geq 0, \quad \sum_{i=1}^m \phi_i = 1 \tag{2.5}$$

and

$$\sum_{i=1}^m \phi_i \theta(t_i) = \lambda \tag{2.6}$$

to ensure that both F_0 and F_1 are cdfs. The maximum empirical likelihood estimator (MELE) of $(\phi_1, \dots, \phi_m, \beta_0, \dots, \beta_N)$ is then defined to be

$$(\hat{\phi}_1, \dots, \hat{\phi}_m, \hat{\beta}_0, \dots, \hat{\beta}_N) = \arg \max_{\phi_1, \dots, \phi_m, \beta_0, \dots, \beta_N} L$$

subject to constraints (2.5), (2.6), and the inequality constraints in (2.3).

To solve the optimization problem above, we consider the following reparameterization

of β_l 's:

$$\beta_0 = \alpha_0, \quad \beta_1 = \alpha_0 + \alpha_1, \dots, \beta_N = \sum_{l=0}^N \alpha_l.$$

Then (2.3) is equivalent to

$$\alpha_1 \geq 0, \dots, \alpha_N \geq 0.$$

The above reparameterization implies that

$$\sum_{l=0}^N \beta_l B_l(x; N) = \sum_{l=0}^N \alpha_l B_l^*(x; N),$$

where $B_l^*(x; N) = \sum_{k=l}^N B_k(x; N)$ for $l = 1, \dots, N$ and $B_0^*(x; N) = 1$.

With a slight abuse of notation, we write

$$L_2(\alpha_0, \dots, \alpha_N) = \prod_{i=1}^m \left[\{\theta(t_i)\}^{b_i} \{1 - \theta(t_i)\}^{a_i} \right]$$

with

$$\theta(x) = \frac{\lambda \exp \left\{ \alpha_0 + \sum_{l=1}^N \alpha_l B_l^*(x; N) \right\}}{1 - \lambda + \lambda \exp \left\{ \alpha_0 + \sum_{l=1}^N \alpha_l B_l^*(x; N) \right\}}.$$

The following proposition summarizes the results for calculating the MELEs of ϕ_i 's and $(\alpha_0, \alpha_1, \dots, \alpha_N)$. For clarity, the proof is given in Section 2.5.3.

Proposition 2.1. *Let*

$$(\hat{\phi}_1, \dots, \hat{\phi}_m) = \arg \max_{\phi_1, \dots, \phi_m} L_1(\phi_1, \dots, \phi_m) \text{ subject to (2.5)}$$

and

$$(\hat{\alpha}_0, \dots, \hat{\alpha}_N) = \arg \max_{\alpha_0, \dots, \alpha_N} L_2(\alpha_0, \dots, \alpha_N)$$

subject to $\alpha_l \geq 0$ for $l = 1, \dots, N$. Then

(a) $\hat{\phi}_i = (a_i + b_i)/n$ for $i = 1, \dots, m$;

(b) $\sum_{i=1}^m \hat{\phi}_i \hat{\theta}(t_i) = \lambda$, where

$$\hat{\theta}(x) = \frac{\lambda \exp \left\{ \hat{\alpha}_0 + \sum_{l=1}^N \hat{\alpha}_l B_l^*(x; N) \right\}}{1 - \lambda + \lambda \exp \left\{ \hat{\alpha}_0 + \sum_{l=1}^N \hat{\alpha}_l B_l^*(x; N) \right\}}. \quad (2.7)$$

Proposition 2.1 implies that we can maximize L_1 and L_2 separately to obtain the MELEs of ϕ_i 's and $(\alpha_0, \alpha_1, \dots, \alpha_N)$. The MELEs of ϕ_i 's have the closed form in Proposition 2.1 (a). Note that $L_2(\alpha_0, \dots, \alpha_N)$ can be viewed as the likelihood for the standard logistic regression with the intercept being $\alpha_0 + \log\{\lambda/(1 - \lambda)\}$ and covariates being $B_1^*(x; N), \dots, B_N^*(x; N)$. Then $(\hat{\alpha}_0, \dots, \hat{\alpha}_N)$ can be readily calculated by using the existing R function `glmnet`. Once $\hat{\phi}_i$'s and $(\hat{\alpha}_0, \dots, \hat{\alpha}_N)$ are available, the estimates of p_{i0} 's and p_{i1} 's are given by

$$\begin{aligned} \hat{p}_{i0} &= \frac{\hat{\phi}_i}{1 - \lambda + \lambda \exp \left\{ \hat{\alpha}_0 + \sum_{l=1}^N \hat{\alpha}_l B_l^*(t_i; N) \right\}} \\ \hat{p}_{i1} &= \frac{\exp \left\{ \hat{\alpha}_0 + \sum_{l=1}^N \hat{\alpha}_l B_l^*(t_i; N) \right\} \hat{\phi}_i}{1 - \lambda + \lambda \exp \left\{ \hat{\alpha}_0 + \sum_{l=1}^N \hat{\alpha}_l B_l^*(t_i; N) \right\}}, \end{aligned}$$

which leads to the estimates for $F_0(\cdot)$ and $F_1(\cdot)$:

$$\hat{F}_0(x) = \sum_{i=1}^m \hat{p}_{i0} I(t_i \leq x) \quad \text{and} \quad \hat{F}_1(x) = \sum_{i=1}^m \hat{p}_{i1} I(t_i \leq x). \quad (2.8)$$

We make some remarks for the proposed method above.

Remark 2.1. In the development above, we have assumed that N is known. In practice, we can choose it based on the Bayesian information criterion (BIC).

Specifically, since we maximize $L_2(\alpha_0, \dots, \alpha_N)$ to obtain the estimators of $(\alpha_0, \dots, \alpha_N)$, we can use it to establish the BIC criterion. For any N , let

$$(\tilde{\alpha}_0, \tilde{\alpha}_1, \dots, \tilde{\alpha}_N) = \arg \max_{\alpha_0, \dots, \alpha_N} L_2(\alpha_0, \dots, \alpha_N)$$

and

$$\tilde{\theta}(x) = \frac{\lambda \exp \left\{ \tilde{\alpha}_0 + \sum_{l=1}^N \tilde{\alpha}_l B_l^*(x; N) \right\}}{1 - \lambda + \lambda \exp \left\{ \hat{\alpha}_0 + \sum_{l=1}^N \tilde{\alpha}_l B_l^*(x; N) \right\}}.$$

That is, $\tilde{\theta}(x)$ is the MELE of $\theta(x)$ without the monotonicity assumption. Denote by df_N the number of unknown parameters in $\theta(x)$. We define

$$BIC(N) = -2 \log \left[\prod_{i=1}^m \{\tilde{\theta}_N(t_i)\}^{b_i} \{1 - \tilde{\theta}_N(t_i)\}^{a_i} \right] + (\log n) \cdot df_N.$$

Consequently, N is set to be the minimizer of $BIC(N)$. We use $\tilde{\theta}(x)$ instead of $\hat{\theta}(x)$ to construct the BIC criterion because without the monotonicity assumption, the number of unknown parameters in $\theta(x)$ can be clearly counted.

Remark 2.2. Practically, a biomarker, x say, may not be in the range $[0, 1]$. We can consider the transformation

$$x^* = \frac{x - t_{(1)}}{t_{(m)} - t_{(1)}},$$

with $t_{(1)}$ and $t_{(m)}$ being the minimum and the maximum values of t_i 's. Clearly $x^* \in [0, 1]$; we can then apply our method to the transformed biomarkers.

Remark 2.3. We observe that biomarkers may exhibit high variability in practice; applying

a log transformation on them may improve the performance. Furthermore, we may include both the original and transformed biomarkers in the model. Specifically, we may consider

$$\log\{f_1(x)/f_0(x)\} \approx \alpha_0 + \sum_{l=1}^N \alpha_l B_l^*(x^*; N) + \sum_{l=1}^N \alpha_{N+l} B_l^*(z^*; N), \quad (2.9)$$

where

$$z^* = \frac{\log x - \log t_{(1)}}{\log t_{(m)} - \log t_{(1)}},$$

and $\alpha_l \geq 0$ for $l = 1, \dots, (2N)$. All our developments above can be similarly applied to (2.9). This modelling strategy is applied in all the simulation and real data examples.

Remark 2.4. Our method is established on the likelihood ratio ordering assumption. To check the rationale of this assumption in practice, we can plot $(\hat{F}_0, \hat{F}_1)$ in (2.8) and $(\tilde{F}_0, \tilde{F}_1)$, where $\tilde{F}_0$ and $\tilde{F}_1$ are respectively the empirical cdfs of $\{X_1, \dots, X_{n_0}\}$ and $\{Y_1, \dots, Y_{n_1}\}$. If the plots of $\hat{F}_i$ are reasonably close to those of $\tilde{F}_i$ for $i = 0, 1$, we may regard the likelihood ratio ordering as a reasonable assumption. For a more rigorous approach, we can utilize the goodness-of-fit test statistics:

$$\Delta_n = \sup_x |\hat{F}_0(x) - \tilde{F}_0(x)| \quad \text{or} \quad \Delta_n^* = \sup_x \left| \hat{F}_1(x) - \tilde{F}_1(x) \right|,$$

and apply the Bootstrap method to perform the test. It can be demonstrated that $\Delta_n = \lambda \Delta_n^* / (1 - \lambda)$, and thus the testing outcomes based on either Δ_n or Δ_n^* are identical. Therefore, we only need to consider one, for example Δ_n , in practice.

2.2.2 Estimation of the ROC curve and its summary quantities

With the estimates $\hat{F}_0(x)$ and $\hat{F}_1(x)$, the ROC curve can be estimated by

$$\widehat{ROC}(s) = 1 - \hat{F}_1\left(\hat{F}_0^{-1}(1-s)\right). \quad (2.10)$$

Recall that the AUC is given as

$$AUC = \int_0^1 ROC(s) ds$$

and the Youden index (J) is

$$J = \max_x \{F_0(x) - F_1(x)\} = F_0(C) - F_1(C),$$

where

$$C = \arg \max_x \{F_0(x) - F_1(x)\}. \quad (2.11)$$

With $\widehat{ROC}(s)$, AUC can be estimated by

$$\widehat{AUC} = \int_0^1 \widehat{ROC}(s) ds. \quad (2.12)$$

Note that (2.11) implies $f_0(C) = f_1(C)$, and hence $\theta(C) = \lambda$. Therefore we can solve the equation

$$\hat{\theta}(\hat{C}) = \lambda \quad (2.13)$$

to obtain $\hat{C}$ and as a consequence,

$$\hat{J} = \hat{F}_0(\hat{C}) - \hat{F}_1(\hat{C}). \quad (2.14)$$

2.2.3 Asymptotic properties

In this section, we establish the consistency of our estimators. We need the following notation. Let $G(x) = \lambda F_1(x) + (1 - \lambda)F_0(x)$ and denote

$$\theta_0(x) = \frac{\lambda f_1(x)}{(1 - \lambda)f_0(x) + \lambda f_1(x)}. \quad (2.15)$$

The following conditions are needed in our technical developments for Theorems 2.1 and 2.2. They are not necessarily the weakest possible.

A1 The total sample size $n = n_0 + n_1 \rightarrow \infty$ and $n_1/n \rightarrow \lambda$ for $\lambda \in (0, 1)$ being a constant.

A2 The order $N \rightarrow \infty$ as $n \rightarrow \infty$ and $\lim_{n \rightarrow \infty} N/n = 0$.

A3 The likelihood ratio ordering assumption is satisfied, or equivalently, $\theta_0(x)$ in (2.15) is a nondecreasing function of x . Furthermore, $\theta_0(x)$ is a continuous function of x and there exists a $\delta > 0$ such that $\delta \leq \theta_0(x) \leq 1 - \delta$ for all $x \in [0, 1]$.

A4 There exists an $\epsilon_0 > 0$, such that $\theta_0(x)$ is strictly increasing for $x \in [C - \epsilon_0, C + \epsilon_0]$, where C is defined by (2.11).

We first show the L_2 convergence of $\hat{\theta}(\cdot)$ to $\theta_0(\cdot)$, where $\hat{\theta}(\cdot)$ is defined by (2.7).

Theorem 2.1. *Assume Conditions A1–A3. We have*

$$\int_0^1 \left\{ \hat{\theta}(x) - \theta_0(x) \right\}^2 dG(x) = o_p(1).$$

With Theorem 2.1, we are able to establish the consistency of $(\hat{F}_0, \hat{F}_1)$ in (2.8), $\widehat{ROC}(s)$ in (2.10), $\widehat{AUC}$ in (2.12), $\hat{C}$ in (2.13), and $\hat{J}$ in (2.14); the results are given in Theorem 2.2 below.

Theorem 2.2. *Assume Conditions A1–A3. We have*

- (a) $\sup_{x \in [0,1]} \left| \hat{F}_0(x) - F_0(x) \right| = o_p(1)$ and $\sup_{x \in [0,1]} \left| \hat{F}_1(x) - F_1(x) \right| = o_p(1)$,
- (b) $\sup_{s \in [0,1]} \left| \widehat{ROC}(s) - ROC(s) \right| = o_p(1)$,
- (c) $\widehat{AUC} = AUC + o_p(1)$.

Furthermore, if Condition A4 is also satisfied, then

- (d) $\hat{C} = C + o_p(1)$ and $\hat{J} = J + o_p(1)$.

For clarity, the proofs of Theorems 2.1 and 2.2 are provided in Sections 2.5.4 and 2.5.5, respectively.

2.3 Simulation study

2.3.1 Simulation setup

In this section, with simulation examples, we compare the performance of our proposed method (denoted as “BP”) with existing methods in the estimation of the ROC curves and its summary indices. We consider the following seven competitive methods:

- the Box-Cox method in Bantis et al. (2019), denoted as “Box-Cox”;
- the method in Zhou & Lin (2008) under the binormal model, denoted as “ZL”;
- the method in Lin et al. (2012), denoted as “LZL”;

- the ECDF-based method, denoted as “ECDF”, which is also known as the nonparametric estimator in [Pepe \(2003\)](#) and [Zhou et al. \(2011\)](#);
- the maximum nonparametric likelihood method under the likelihood ratio ordering [Dykstra et al. \(1995\)](#), denoted as “MNLE”;
- the kernel-based method in [Bantis et al. \(2019\)](#), denoted as “Kernel”;
- the maximum smoothed likelihood method under the likelihood ratio ordering in [Yu et al. \(2017\)](#), denoted as “MSLE”.

In the Kernel and MSLE methods, we follow [Bantis et al. \(2019\)](#) and [Yu et al. \(2017\)](#). We use the Gaussian kernel with bandwidths h_0 and h_1 for healthy and diseased groups respectively, where

$$h_0 = 0.9 \min \left\{ s_0, \frac{q_0}{1.34} \right\} n_0^{-0.2} \quad \text{and} \quad h_1 = 0.9 \min \left\{ s_1, \frac{q_1}{1.34} \right\} n_1^{-0.2}.$$

Here, s_0 and q_0 are sample standard deviation and sample interquartile range for the sample from the healthy population, and s_1 and q_1 are sample standard deviation and sample interquartile range for the sample from the diseased population.

Throughout our numerical studies, we observe that the MNLE and MSLE methods have similar performance as that of the ECDF and Kernel methods, respectively. Furthermore, the LZL method is designed to accommodate covariates; without covariates, the performance of this method is very similar to that of the ECDF method. For clarity, we present the results for MNLE, MSLE, and LZL methods in Sections [2.5.1](#) and [2.5.2](#).

We consider two distributional settings:

- (1) $f_0 \sim N(\mu_0, \sigma_0^2)$ and $f_1 \sim N(\mu_1, \sigma_1^2)$;

(2) $f_0 \sim \text{Gamma}(a_0, b_0)$ and $f_1 \sim \text{Gamma}(a_1, b_1)$.

Here, $\text{Gamma}(a, b)$ denotes the gamma distribution with shape parameter a and rate parameter b . Note that Setting (1) corresponds to the case that the model assumption for the Box-Cox method is satisfied, whereas Setting (2) corresponds to the case that the model assumption for the Box-Cox method is violated. In both cases, f_0 is fixed. The parameters for f_1 are varied such that the corresponding Youden indices are 0.3, 0.5, and 0.7, respectively. The details of these settings are given in Table 2.1. We consider three different combinations of sample sizes: $(n_0, n_1) = (50, 50)$, $(100, 100)$, and $(150, 50)$. Therefore, for each distributional setting of f_0 and f_1 , we have 9 combinations of parameters and sample sizes. For each combination, we repeat the simulation 2000 times.

Table 2.1: Simulation settings

Distribution	J	AUC	μ_0	σ_0^2	μ_1	σ_1^2
Normal	0.3	0.707	10	1	10.771	1
Normal	0.5	0.830	10	1	11.349	1
Normal	0.7	0.929	10	1	12.073	1
Distribution	J	AUC	a_0	b_0	a_1	b_1
Gamma	0.3	0.708	2	1	3	0.937
Gamma	0.5	0.830	2	1	4	0.944
Gamma	0.7	0.929	2	1	5	0.827

We observe that the proposed method depends on the choice of N . In Section 2.5.1, we give the frequency of the selected N values based on the BIC criterion given in Remark 2.1 from 2000 repetitions. Furthermore, we have considered the distributional setting that f_0 and f_1 follow the Beta distributions; the details and results are given in Section 2.5.2.

2.3.2 Comparison of the ROC curve estimation

In this section, we compare our proposed method with the competitive methods in the ROC curve estimation. The criteria for comparison are the L_1 - and L_2 -distances between the estimated and the true ROC curves. For a generic ROC curve estimate $\overline{ROC}(s)$, the L_1 - and L_2 -distances between $\overline{ROC}(s)$ and the true ROC curve $ROC(s)$ are defined to be:

$$L_1(\overline{ROC}, ROC) = \int_0^1 |\overline{ROC}(s) - ROC(s)| ds$$

and

$$L_2(\overline{ROC}, ROC) = \left[\int_0^1 \{\overline{ROC}(s) - ROC(s)\}^2 ds \right]^{1/2}.$$

Simulation results based on 2000 repetitions are summarized in Table [2.2](#).

From this table, we observe that our BP method results in the smallest average L_1 - and L_2 -distances for all the examples; the improvement is significant. For example, the ratio of the L_2 -distances between our BP method and the ECDF method is ranged between 0.61 and 0.71 with an average of 0.66; the ratio of the L_2 -distances between our BP method and the ZL method is between 0.84 and 0.93 with an average of 0.88.

2.3.3 Comparison of the AUC estimation

In this section, we compare our proposed method with the competitive methods in the AUC estimation. The criteria for comparison are the relative bias (RB) and mean square error (MSE). Suppose we have B point estimates of the AUC $\hat{a}^{(i)}$ for $i = 1, \dots, B$. The RB

Table 2.2: Averages of L_1 -distances and L_2 -distances of five methods for estimating the ROC curve

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		L_1	L_2	L_1	L_2	L_1	L_2
Normal (0.3)	BP	0.040	0.046	0.029	0.033	0.033	0.038
	Box-Cox	0.044	0.053	0.032	0.038	0.037	0.043
	ZL	0.045	0.054	0.032	0.038	0.037	0.044
	ECDF	0.056	0.071	0.040	0.051	0.046	0.058
	Kernel	0.047	0.057	0.035	0.042	0.040	0.048
Normal (0.5)	BP	0.032	0.041	0.023	0.029	0.026	0.034
	Box-Cox	0.036	0.047	0.025	0.034	0.029	0.038
	ZL	0.036	0.049	0.026	0.034	0.030	0.039
	ECDF	0.045	0.064	0.032	0.046	0.037	0.051
	Kernel	0.040	0.053	0.029	0.040	0.033	0.043
Normal (0.7)	BP	0.020	0.033	0.014	0.023	0.016	0.026
	Box-Cox	0.022	0.036	0.015	0.026	0.017	0.028
	ZL	0.023	0.039	0.016	0.027	0.019	0.030
	ECDF	0.029	0.053	0.020	0.038	0.023	0.040
	Kernel	0.027	0.044	0.020	0.033	0.022	0.035
Gamma (0.3)	BP	0.042	0.050	0.030	0.035	0.035	0.041
	Box-Cox	0.045	0.054	0.032	0.038	0.037	0.044
	ZL	0.047	0.056	0.032	0.039	0.038	0.045
	ECDF	0.057	0.072	0.040	0.052	0.046	0.058
	Kernel	0.051	0.062	0.038	0.046	0.044	0.051
Gamma (0.5)	BP	0.034	0.046	0.024	0.032	0.027	0.035
	Box-Cox	0.036	0.049	0.026	0.035	0.028	0.037
	ZL	0.037	0.050	0.026	0.035	0.028	0.038
	ECDF	0.045	0.067	0.032	0.047	0.036	0.050
	Kernel	0.046	0.060	0.035	0.044	0.038	0.046
Gamma (0.7)	BP	0.021	0.036	0.016	0.026	0.016	0.027
	Box-Cox	0.022	0.038	0.016	0.027	0.016	0.028
	ZL	0.023	0.040	0.016	0.028	0.018	0.031
	ECDF	0.029	0.055	0.021	0.040	0.022	0.040
	Kernel	0.033	0.050	0.026	0.038	0.028	0.038

in percentage and the MSE are respectively defined to be:

$$RB(\%) = \frac{1}{B} \sum_{i=1}^B \frac{\hat{a}^{(i)} - a_0}{a_0} \times 100 \quad \text{and} \quad MSE = \frac{1}{B} \sum_{i=1}^B (\hat{a}^{(i)} - a_0)^2,$$

where a_0 is the true value of the AUC. The results are summarized in Table 2.3.

Comparing the RB values in this table, we observe that our method is small in RB values for all examples, though in most cases, the RB values from the ECDF method are the smallest; the Kernel method always results in negative and largest RBs in absolute values. Comparing the MSE values, our BP method has comparable or better performance than competitive methods.

2.3.4 Comparison of the Youden index and optimal cutoff point estimation

In this section, we compare our method with the competitive methods in the estimation of the Youden index and the optimal cutoff point. The criteria for comparison are the RB and MSE, which are similarly defined as those in Section 2.3.3. The simulation results for estimating the Youden index and the optimal cutoff point are summarized in Tables 2.4 and 2.5, respectively.

From Table 2.4, we observe that in the estimation of the Youden index, out of 18 simulation settings, our BP method leads to the smallest RB values in 16 settings, and obtains the second smallest RB values in 2 settings, where the Kernel method produces the smallest RB values. Comparing MSE values, the performance of our BP method and the Box-Cox method is comparable, and is better than the other methods in most of the settings.

Table 2.3: RB (%) and MSE ($\times 1000$) of five methods for estimating the AUC

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Normal (0.3)	BP	0.29	2.39	0.00	1.21	0.20	1.65
	Box-Cox	0.44	2.44	0.06	1.23	0.20	1.68
	ZL	0.44	2.47	0.09	1.24	0.24	1.69
	ECDF	0.09	2.51	-0.10	1.27	-0.01	1.74
	Kernel	-1.67	2.38	-1.54	1.28	-1.53	1.71
Normal (0.5)	BP	0.16	1.50	-0.01	0.75	0.16	1.01
	Box-Cox	0.35	1.50	0.08	0.75	0.17	1.02
	ZL	0.36	1.53	0.10	0.76	0.19	1.03
	ECDF	0.03	1.56	-0.07	0.77	-0.01	1.06
	Kernel	-1.95	1.77	-1.68	0.95	-1.71	1.24
Normal (0.7)	BP	0.05	0.58	-0.01	0.28	0.08	0.37
	Box-Cox	0.10	0.54	0.01	0.27	0.05	0.36
	ZL	0.14	0.58	0.04	0.28	0.10	0.40
	ECDF	-0.02	0.61	-0.04	0.29	-0.01	0.39
	Kernel	-1.56	0.88	-1.28	0.46	-1.33	0.59
Gamma (0.3)	BP	0.69	2.49	0.29	1.26	0.11	1.70
	Box-Cox	0.57	2.58	0.26	1.30	0.23	1.76
	ZL	0.55	2.60	0.24	1.30	0.20	1.75
	ECDF	0.04	2.66	-0.08	1.34	-0.10	1.78
	Kernel	-2.36	2.63	-2.08	1.44	-2.28	1.86
Gamma (0.5)	BP	0.33	1.56	0.06	0.80	0.17	0.94
	Box-Cox	0.40	1.55	0.13	0.80	0.39	0.95
	ZL	0.39	1.58	0.10	0.81	0.36	0.97
	ECDF	-0.01	1.65	-0.14	0.84	0.13	0.99
	Kernel	-2.65	2.12	-2.31	1.20	-2.33	1.37
Gamma (0.7)	BP	0.10	0.59	-0.01	0.30	0.00	0.35
	Box-Cox	0.08	0.55	-0.01	0.29	0.08	0.32
	ZL	0.13	0.59	-0.01	0.30	0.14	0.39
	ECDF	-0.05	0.62	-0.10	0.31	0.00	0.35
	Kernel	-2.19	1.14	-1.82	0.65	-2.03	0.80

From Table 2.5, for estimating the optimal cutoff point, our BP method and the Box-Cox method result in significantly smaller MSE values than other methods. The performance of the BP and the Box-Cox method is mixed. When f_0 and f_1 are simulated from the Gamma distribution with the Youden index $J = 0.7$, Box-Cox method produces smaller MSE values, whereas our BP method leads to smaller MSE values in all other settings.

2.4 Real data application

In this section, we apply our method to a dataset on Duchenne Muscular Dystrophy (DMD). DMD is a type of muscular dystrophy that is genetically transmitted from a mother to her children. Muscle loss occurs at an early age for offspring with the disease. Female offspring with the disease do not suffer from significant symptoms compared to male offspring who die at a young age. Female carriers do not show sign of disease and therefore detection of potential female carrier is of main interest.

Percy et al. (1982) stated that DMD carriers are more likely to have higher measurement of specific biomarkers. Four biomarkers including creatine kinase (CK), hemopexin (H), lactate dehydrogenase (LD), and pyruvate kinase (PK) are measured from the blood serum samples of a healthy group ($n_0 = 127$) and a group of carriers ($n_1 = 67$). The complete dataset was collected by Andrews & Herzberg (2012). Yuan et al. (2021) pointed out that the biomarker CK has the best performance among these four biomarkers since it corresponds to largest estimate of the Youden index. Therefore, we apply our method and other existing methods on the biomarker CK to compare their performance.

Following Remark 2.1, we construct the $BIC(N)$ versus N plot in Panel (a) of Figure 2.1; it suggests to use $N = 1$ in our BP method. In this figure, it can be seen that $BIC(N)$ exhibits an almost linear relationship with N . This is because BIC consists of

Table 2.4: RB (%) and MSE ($\times 1000$) of five methods for estimating the Youden index

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Normal (0.3)	BP	1.90	5.69	0.44	2.87	1.25	3.93
	Box-Cox	4.25	5.86	1.57	2.89	2.50	3.97
	ZL	4.66	5.94	1.80	2.92	2.84	4.01
	ECDF	24.05	11.64	11.57	5.57	19.05	7.75
	Kernel	-0.88	5.61	-2.53	3.08	-1.83	4.05
Normal (0.5)	BP	1.32	4.91	0.40	2.43	0.99	3.29
	Box-Cox	2.64	5.04	1.05	2.42	1.52	3.33
	ZL	2.89	5.24	1.15	2.47	1.70	3.40
	ECDF	11.67	9.15	7.40	4.33	9.28	6.13
	Kernel	-3.53	5.45	-3.73	3.01	-3.54	3.90
Normal (0.7)	BP	1.01	3.51	0.37	1.68	0.74	2.25
	Box-Cox	1.64	3.38	0.70	1.61	0.95	2.20
	ZL	2.09	3.84	0.89	1.77	1.40	2.87
	ECDF	6.29	5.71	4.06	2.79	5.01	3.80
	Kernel	-3.74	4.43	-3.43	2.43	-3.43	3.10
Gamma (0.3)	BP	3.85	6.03	1.74	2.98	1.32	4.03
	Box-Cox	5.27	6.25	2.90	3.08	3.18	4.27
	ZL	5.60	6.32	2.92	3.08	3.19	4.23
	ECDF	23.77	11.66	15.77	5.83	19.08	7.85
	Kernel	-0.67	6.10	-1.91	3.35	-0.95	4.17
Gamma (0.5)	BP	1.92	5.27	0.57	2.56	1.08	3.09
	Box-Cox	3.26	5.33	1.66	2.58	2.62	3.33
	ZL	3.30	5.47	1.54	2.60	2.43	3.35
	ECDF	11.20	9.02	7.25	4.45	9.38	5.94
	Kerel	-3.42	5.91	-3.64	3.21	-2.96	3.55
Gamma (0.7)	BP	1.34	3.66	0.51	1.83	0.71	2.09
	Box-Cox	1.88	3.46	0.95	1.75	1.37	2.08
	ZL	2.28	4.03	1.01	1.89	1.81	3.26
	ECDF	6.17	5.87	3.85	2.86	4.95	3.62
	Kernel	-3.71	4.56	-3.50	2.62	-3.62	2.88

Table 2.5: RB (%) and MSE ($\times 1000$) of five methods for estimating the optimal cutoff point

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Normal (0.3)	BP	-0.19	11.34	-0.18	5.72	-0.17	7.60
	Box-Cox	-0.11	44.14	-0.02	22.61	-0.08	29.02
	ZL	-0.22	58.06	-0.04	27.42	-0.05	35.82
	ECDF	-0.14	137.99	-0.13	95.68	0.15	114.03
	Kernel	0.04	110.52	0.06	64.31	0.28	79.47
Normal (0.5)	BP	-0.15	11.40	-0.13	5.70	-0.11	7.51
	Box-Cox	-0.09	22.44	-0.02	11.15	-0.03	13.51
	ZL	-0.22	33.61	-0.06	15.88	0.00	19.40
	ECDF	-0.22	82.58	-0.10	56.55	-0.14	68.46
	Kernel	-0.03	42.51	-0.04	23.82	0.09	30.02
Normal (0.7)	BP	-0.11	12.96	-0.10	6.37	-0.05	8.44
	Box-Cox	-0.06	15.56	-0.02	7.48	0.01	9.23
	ZL	-0.23	31.01	-0.08	14.66	-0.01	18.19
	ECDF	-0.27	62.74	-0.10	38.93	0.13	51.15
	Kernel	-0.04	27.67	0.01	14.65	0.03	19.03
Gamma (0.3)	BP	4.48	58.33	2.96	31.53	3.67	42.28
	Box-Cox	0.21	88.72	-0.38	43.95	-0.42	52.51
	ZL	0.44	120.92	-0.12	56.77	0.67	71.78
	ECDF	2.86	319.78	1.78	206.42	4.09	274.15
	Kernel	12.12	307.63	8.73	150.92	10.95	189.00
Gamma (0.5)	BP	2.69	45.48	1.98	24.47	2.48	32.88
	Box-Cox	-0.46	52.60	-0.42	27.26	-0.27	30.60
	ZL	-0.39	81.87	-0.07	39.84	0.87	47.04
	ECDF	0.01	210.50	0.62	144.56	2.08	181.49
	Kernel	7.17	141.60	5.72	80.26	6.97	94.56
Gamma (0.7)	BP	1.53	57.16	1.15	31.92	1.53	41.74
	Box-Cox	-0.62	51.57	-0.57	26.03	-0.42	32.98
	ZL	-0.59	106.81	-0.24	53.14	0.39	66.29
	ECDF	-0.59	223.80	-0.12	150.58	1.27	193.70
	Kernel	4.38	140.67	3.56	75.42	4.03	82.75

two terms: the log-likelihood and a penalty term $(\log n) \cdot df_N$. We computed the maximized log-likelihood values and corresponding BIC values for $N = 1, \dots, 10$, and found that the maximized log-likelihood values remain almost constant as N ranges from 1 through 10, indicating that the model with $N = 1$ is sufficient to explain the data; numerical details are omitted for brevity. As a result, the penalty term dominates as N increases, resulting in a linearly increasing trend in BIC as N increases. Based on Remark 2.4, to check whether the likelihood ratio ordering assumption is reasonable, we plot our BP estimates $(\hat{F}_0, \hat{F}_1)$ and the empirical cdf estimates $(\tilde{F}_0, \tilde{F}_1)$ in Panel (b) of Figure 2.1; this plot suggests that the likelihood ratio ordering assumption might be reasonable. We further perform the goodness-of-fit test suggested in Remark 2.4; the p -value based on 1000 bootstrap samples is 0.975. This reinforces the validity of the likelihood ratio ordering assumption.

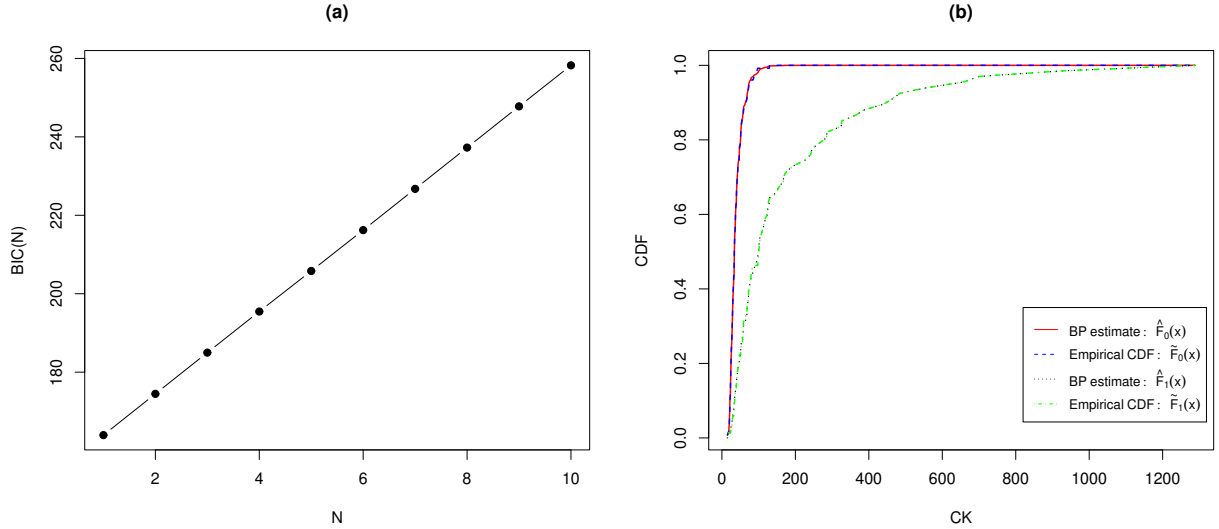


Figure 2.1: Panel (a) shows $BIC(N)$ versus N ; Panel (b) compares the proposed estimators $(\hat{F}_0, \hat{F}_1)$ and the empirical cdfs $(\tilde{F}_0, \tilde{F}_1)$.

In Figure 2.2, we plot the ROC curve estimates from our BP and competitive meth-

ods. The performance of the MNLE and MSLE methods is similar to that of the ECDF and kernel method, respectively; we thus do not include them in this figure. From the figure, we observe that the ROC curve estimate from our BP method is similar to those of Box-Cox, ZL, ECDF, and LZL methods, but different from the kernel method. The difference lies in the fact that the kernel method leads to much smaller sensitivity estimate compared to other methods when 1-specificity is large, greater than 0.5 say. This is mainly because that the CK values for the diseased individuals are very skewed, which inflates the selected bandwidth when the kernel method is used to estimate F_1 ; this in turn makes F_1 overestimated and hence causes the sensitivity underestimated when 1-specificity is large or the value of CK is small.

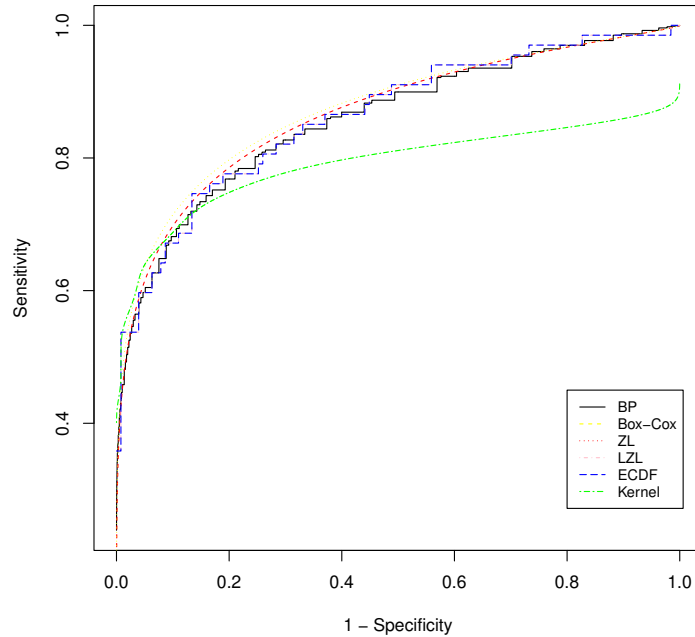


Figure 2.2: Estimated ROC curves based on different methods.

We evaluate the point estimates (PEs) and bootstrap percentile confidence intervals (BPCIs) of the AUC, Youden index, and optimal cutoff point based on different methods, and display them in Table 2.6. The AUC estimates from all methods are similar, except for the Kernel method, which leads to a smaller AUC estimate. This is consistent with the results in Figure 2.2, and complies with our observation in Table 2.3: the Kernel method has the negative relative biases, and all other methods have small RBs and comparable MSEs. The Youden index estimates are similar for all methods: the maximum difference between any two estimates is less than 0.03. For the optimal cutoff point estimates, the Kernel method leads to a different result from other methods; other methods give similar results. But the ZL, LZL and ECDF methods have resulted in wider BPCIs than our method; we conjecture that this may indicate that our method has given smaller standard error than other methods in the optimal cutoff point estimation. However, unlike with the simulation studies, in this real data example, we are unable to compare the estimates with the true values.

Table 2.6: PEs and BPCIs of the AUC, Youden index, and optimal cutoff point by using CK as the biomarker

Indices Method	AUC		Youden index		Optimal cutoff point	
	PE	BPCI	PE	BPCI	PE	BPCI
BP	0.865	[0.804, 0.914]	0.588	[0.480, 0.688]	58.998	[51.976, 67.311]
Box-Cox	0.874	[0.816, 0.921]	0.616	[0.510, 0.710]	58.017	[51.869, 65.055]
ZL	0.868	[0.805, 0.917]	0.600	[0.498, 0.697]	57.000	[50.975, 69.000]
LZL	0.863	[0.800, 0.914]	0.612	[0.502, 0.727]	56.000	[43.000, 97.500]
ECDF	0.863	[0.800, 0.914]	0.612	[0.502, 0.727]	56.000	[43.000, 97.500]
Kernel	0.790	[0.748, 0.844]	0.591	[0.502, 0.669]	73.356	[54.689, 79.413]

2.5 Technical details and additional simulation results

2.5.1 Additional simulation results under the normal and Gamma distribution settings

In this section, we first present the frequency of the selected N values based on the BIC criterion in Remark 2.1; the results are given in Table 2.7. Based on Remark 2.3, we include both the original and log-transformed biomarkers in the model. Hence, for both normal and Gamma distributions, the true value of N is 1. From Table 2.7, we observe that the proposed BIC criterion can choose $N = 1$ for the majority of repetitions in all the simulation settings.

Tables 2.8–2.11 summarize the results of the MNLE method, the MSLE method, and the LZL method introduced in Section 2.3 in the estimation of the ROC curve, the AUC, the Youden index, and the optimal cutoff point. From these tables, we observe that the performance of MNLE is similar to that of the ECDF based method; the performance of the MSLE is close to that of the Kernel method. The results of LZL are very similar to those of the ECDF method.

Table 2.7: Frequency of the N values when distributions are normal and Gamma

Distribution (J)	$(n_0, n_1) = (50, 50)$			$(n_0, n_1) = (100, 100)$			$(n_0, n_1) = (150, 50)$		
	$N = 1$	$N = 2$	$N = 3$	$N = 1$	$N = 2$	$N = 3$	$N = 1$	$N = 2$	$N = 3$
Normal (0.3)	1994	6	0	1999	1	0	1995	5	0
Normal (0.5)	1999	1	0	1999	1	0	1999	1	0
Normal (0.7)	2000	0	0	2000	0	0	2000	0	0
Gamma (0.3)	1982	18	0	1986	14	0	1995	5	0
Gamma (0.5)	1988	12	0	1991	9	0	1992	8	0
Gamma (0.7)	1996	4	0	2000	0	0	2000	0	0

Table 2.8: Averages of L_1 - and L_2 -distances of the ROC curve estimates for other methods

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		L_1	L_2	L_1	L_2	L_1	L_2
Normal (0.3)	MNLE	0.054	0.066	0.038	0.047	0.045	0.054
	MSLE	0.045	0.054	0.035	0.041	0.039	0.046
	LZL	0.056	0.071	0.040	0.051	0.046	0.058
Normal (0.5)	MNLE	0.042	0.059	0.030	0.043	0.035	0.047
	MSLE	0.039	0.052	0.029	0.039	0.033	0.043
	LZL	0.045	0.064	0.032	0.046	0.037	0.051
Normal (0.7)	MNLE	0.026	0.047	0.019	0.035	0.021	0.036
	MSLE	0.027	0.044	0.020	0.033	0.022	0.035
	LZL	0.029	0.053	0.020	0.038	0.023	0.040
Gamma (0.3)	MNLE	0.054	0.067	0.039	0.048	0.044	0.054
	MSLE	0.049	0.059	0.037	0.044	0.042	0.049
	LZL	0.057	0.072	0.040	0.052	0.046	0.058
Gamma (0.5)	MNLE	0.042	0.061	0.031	0.044	0.034	0.047
	MSLE	0.045	0.058	0.034	0.044	0.037	0.046
	LZL	0.045	0.067	0.032	0.047	0.036	0.050
Gamma (0.7)	MNLE	0.025	0.049	0.019	0.036	0.020	0.036
	MSLE	0.033	0.049	0.025	0.037	0.027	0.038
	LZL	0.029	0.055	0.021	0.040	0.022	0.040

Table 2.9: RB (%) and MSE ($\times 1000$) of the AUC estimates for other methods

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Normal (0.3)	MNLE	5.09	3.39	3.39	1.69	4.17	2.35
	MSLE	-1.35	2.23	-1.41	1.24	-1.33	1.63
	LZL	0.09	2.51	-0.10	1.27	-0.01	1.74
Normal (0.5)	MNLE	2.90	1.88	1.95	0.94	2.37	1.29
	MSLE	-1.88	1.75	-1.66	0.95	-1.67	1.23
	LZL	0.03	1.56	-0.07	0.77	-0.01	1.06
Normal (0.7)	MNLE	1.40	0.63	1.00	0.33	1.18	0.43
	MSLE	-1.54	0.87	-1.27	0.46	1.31	0.59
	LZL	-0.02	0.61	-0.04	0.29	-0.01	0.39
Gamma (0.3)	MNLE	5.04	3.47	3.40	1.77	4.04	2.34
	MSLE	-1.88	2.43	-1.85	1.38	-1.96	1.75
	LZL	0.05	2.66	-0.08	1.35	-0.10	1.78
Gamma (0.5)	MNLE	2.85	1.92	1.88	0.99	2.43	1.25
	MSLE	-2.48	2.04	-2.22	1.17	-2.23	1.33
	LZL	-0.01	1.65	-0.14	0.84	0.13	0.99
Gamma (0.7)	MNLE	1.35	0.63	0.94	0.34	1.14	0.40
	MSLE	-2.10	1.11	-1.77	0.63	-1.98	0.78
	LZL	-0.05	0.62	-0.10	0.31	0.00	0.35

Table 2.10: RB (%) and MSE ($\times 1000$) of the Youden index estimates for other methods

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Normal (0.3)	MNLE	24.05	11.64	11.57	5.57	19.05	7.75
	MSLE	-0.88	5.61	-2.53	3.08	-1.83	4.05
	LZL	24.04	11.63	15.55	5.57	19.03	7.75
Normal (0.5)	MNLE	11.67	9.15	7.40	4.33	9.28	6.13
	MSLE	-3.53	5.45	-3.73	3.01	-3.54	3.90
	LZL	11.66	9.15	7.39	4.33	9.27	6.12
Normal (0.7)	MNLE	6.29	5.71	4.06	2.79	5.01	3.80
	MSLE	-3.74	4.43	-3.43	2.43	-3.43	3.10
	LZL	6.28	5.70	4.06	2.79	5.00	3.79
Gamma (0.3)	MNLE	23.77	11.66	15.77	5.83	19.08	7.85
	MSLE	-0.67	6.10	-1.91	3.35	-0.95	4.17
	LZL	23.76	11.65	15.75	5.82	19.07	7.84
Gamma (0.5)	MNLE	11.20	9.02	7.25	4.45	9.38	5.94
	MSLE	-3.42	5.91	-3.64	3.21	-2.96	3.55
	LZL	11.19	9.01	7.24	4.45	9.37	5.94
Gamma (0.7)	MNLE	6.17	5.87	3.85	2.86	4.95	3.62
	MSLE	-3.71	4.56	-3.50	2.62	-3.62	2.88
	LZL	6.16	5.87	3.85	2.86	4.95	3.62

Table 2.11: RB (%) and MSE ($\times 1000$) of the optimal cutoff point estimates for other methods

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Normal (0.3)	MNLE	-0.13	137.31	-0.13	95.25	0.14	113.75
	MSLE	0.04	110.52	-0.06	64.32	0.28	79.45
	LZL	-0.17	146.41	-0.10	100.75	0.16	118.25
Normal (0.5)	MNLE	-0.22	82.46	-0.10	56.07	-0.14	68.40
	MSLE	-0.04	42.52	-0.04	23.82	0.08	30.02
	LZL	-0.25	87.21	-0.12	59.57	0.13	70.82
Normal (0.7)	MNLE	-0.26	62.73	-0.09	38.80	0.13	51.15
	MSLE	-0.05	27.67	0.00	14.65	0.02	19.03
	LZL	-0.29	64.95	-0.08	40.22	0.15	52.01
Gamma (0.3)	MNLE	2.67	313.56	1.65	204.17	4.00	272.66
	MSLE	12.10	307.44	8.71	150.78	10.93	188.82
	LZL	2.81	341.48	1.94	218.63	4.09	284.06
Gamma (0.5)	MNLE	-0.03	209.13	0.60	144.27	2.03	180.81
	MSLE	7.16	141.44	5.70	80.14	6.95	94.40
	LZL	-0.14	227.38	0.49	149.22	3.86	211.86
Gamma (0.7)	MNLE	-0.62	222.60	-0.12	149.71	1.22	192.81
	MSLE	4.36	140.51	3.54	75.29	4.01	82.60
	LZL	-0.70	228.96	0.08	156.61	1.41	198.65

2.5.2 Simulation results under the Beta distribution settings

In this section, we consider the simulation when $f_0 \sim \text{Beta}(2, 2)$ and $f_1 \sim \text{Beta}(a_1, b_1)$. For (a_1, b_1) , we consider three sets of values as shown in Table 2.12 such that the corresponding J is 0.3, 0.5, and 0.7 respectively. We note that under this simulation setup, the true value of N is 1.

Table 2.13 summarizes the frequency of the selected N values based on the BIC criterion in Remark 2.1; this table shows that the BIC criterion picks the correct order in the

majority of the repetitions. Tables 2.14–2.17 summarize the results of all methods for estimating the ROC curve, the AUC, the Youden index, and the optimal cutoff point. The observations are similar to those when f_0 and f_1 are simulated as normal and Gamma distributions; we omit the details. It is noteworthy that the model assumption for the Box-cox method is violated under this setup. Our BP method performs better than the Box-Cox method for estimating the ROC curve, the Youden index, and the optimal cutoff point; the improvement is significant. For estimating the AUC, our BP method leads to smaller MSEs when $J = 0.3$ and 0.5 , but the Box-Cox method results in smaller MSEs when $J = 0.7$.

Table 2.12: Beta simulation setting

Distribution	J	AUC	a_1	b_1
Beta	0.3	0.702	3.838	2
Beta	0.5	0.822	6.148	2
Beta	0.7	0.919	11.014	2

Table 2.13: Frequency of the chosen N in Beta distributional setting

(n_0, n_1) Distribution (J)	(50, 50)			(100, 100)			(150, 50)		
	$N = 1$	$N = 2$	$N = 3$	$N = 1$	$N = 2$	$N = 3$	$N = 1$	$N = 2$	$N = 3$
Beta (0.3)	1989	11	0	1996	4	0	1998	2	0
Beta (0.5)	1999	1	0	2000	0	0	2000	0	0
Beta (0.7)	2000	0	0	2000	0	0	2000	0	0

Table 2.14: Averages of L_1 -distances and L_2 -distances of eight methods for estimating the ROC curve under Beta distribution settings

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		L_1	L_2	L_1	L_2	L_1	L_2
Beta (0.3)	BP	0.041	0.048	0.029	0.034	0.031	0.037
	Box-Cox	0.045	0.055	0.032	0.039	0.036	0.043
	ZL	0.046	0.057	0.032	0.039	0.035	0.043
	LZL	0.056	0.073	0.040	0.052	0.044	0.057
	ECDF	0.056	0.073	0.040	0.052	0.044	0.057
	MNLE	0.054	0.069	0.039	0.050	0.044	0.055
	Kernel	0.047	0.058	0.034	0.042	0.037	0.046
	MSLE	0.044	0.054	0.034	0.040	0.035	0.044
Beta (0.5)	BP	0.033	0.046	0.024	0.033	0.024	0.033
	Box-Cox	0.039	0.055	0.029	0.041	0.031	0.044
	ZL	0.038	0.054	0.027	0.038	0.027	0.039
	LZL	0.046	0.070	0.032	0.050	0.034	0.051
	ECDF	0.046	0.070	0.032	0.050	0.034	0.051
	MNLE	0.043	0.065	0.032	0.047	0.032	0.047
	Kernel	0.040	0.057	0.029	0.042	0.030	0.042
	MSLE	0.038	0.054	0.028	0.040	0.029	0.041
Beta (0.7)	BP	0.022	0.041	0.015	0.029	0.015	0.028
	Box-Cox	0.028	0.056	0.023	0.046	0.025	0.051
	ZL	0.025	0.050	0.017	0.034	0.020	0.038
	LZL	0.031	0.065	0.021	0.045	0.022	0.044
	ECDF	0.031	0.065	0.021	0.045	0.022	0.044
	MNLE	0.028	0.059	0.020	0.043	0.020	0.041
	Kernel	0.030	0.058	0.022	0.044	0.022	0.042
	MSLE	0.028	0.053	0.021	0.040	0.021	0.039

Table 2.15: RB (%) and MSE ($\times 1000$) of eight methods for estimating the AUC under Beta distribution settings

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Beta (0.3)	BP	1.07	2.42	0.82	1.20	0.82	1.43
	Box-Cox	1.12	2.72	0.99	1.35	1.17	1.61
	ZL	0.62	2.60	0.50	1.27	0.51	1.51
	LZL	0.14	2.65	0.13	1.31	0.18	1.53
	ECDF	0.14	2.65	0.13	1.31	0.18	1.53
	MNLE	5.33	3.57	3.74	1.82	4.35	2.22
	Kernel	-1.53	2.50	-1.19	1.27	-1.18	1.48
	MSLE	-1.08	2.30	-0.98	1.20	-0.93	1.38
Beta (0.5)	BP	0.66	1.58	0.41	0.81	0.31	0.84
	Box-Cox	1.48	1.74	1.30	0.93	1.16	0.90
	ZL	0.56	1.68	0.32	0.85	0.26	0.87
	LZL	0.13	1.73	0.02	0.87	0.02	0.89
	ECDF	0.13	1.73	0.02	0.87	0.02	0.89
	MNLE	3.16	2.08	2.16	1.08	2.37	1.13
	Kernel	-1.79	1.86	-1.51	0.99	-1.46	1.00
	MSLE	-1.57	1.76	-1.38	0.95	-1.36	0.97
Beta (0.7)	BP	0.36	0.69	0.23	0.33	0.16	0.34
	Box-Cox	0.58	0.60	0.50	0.29	0.09	0.25
	ZL	0.36	0.73	0.20	0.36	0.23	0.83
	LZL	0.11	0.77	0.04	0.37	0.06	0.35
	ECDF	0.11	0.77	0.04	0.37	0.06	0.35
	MNLE	1.70	0.83	1.20	0.43	1.23	0.42
	Kernel	-1.79	1.10	-1.48	0.57	-1.35	0.52
	MSLE	1.56	0.98	-1.34	0.52	-1.23	0.48

Table 2.16: RB (%) and MSE ($\times 1000$) of eight methods for estimating the Youden index under Beta distribution settings

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Beta (0.3)	BP	4.59	6.07	3.20	2.96	3.36	3.54
	Box-Cox	7.35	7.10	5.75	3.48	3.08	4.65
	ZL	5.76	6.22	3.69	2.96	4.36	3.63
	LZL	23.60	11.63	15.74	5.58	19.37	7.37
	ECDF	23.61	11.64	15.76	5.58	19.39	7.38
	MNLE	23.61	11.64	15.76	5.58	19.39	7.38
	Kernel	0.51	5.98	0.46	3.04	0.00	3.72
	MSLE	0.51	5.98	0.46	3.04	0.00	3.72
Beta (0.5)	BP	2.35	5.23	1.28	2.59	1.08	2.76
	Box-Cox	8.64	7.60	7.34	4.21	8.66	5.13
	ZL	3.86	5.55	2.16	2.64	2.06	2.89
	LZL	11.26	8.78	7.24	4.29	8.78	5.14
	ECDF	11.27	8.78	7.25	4.30	8.79	5.15
	MNLE	11.27	8.78	7.25	4.30	8.79	5.15
	Kernel	-1.29	5.30	-1.69	2.76	-1.71	3.17
	MSLE	-1.29	5.30	-1.69	2.76	-1.71	3.17
Beta (0.7)	BP	1.31	3.67	0.61	1.72	0.51	1.80
	Box-Cox	8.20	6.27	7.60	4.28	7.84	4.65
	ZL	2.71	4.07	1.44	1.86	1.89	4.71
	LZL	6.13	5.84	4.08	2.78	5.03	3.42
	ECDF	6.14	5.84	4.08	2.79	5.03	3.43
	MNLE	6.14	5.84	4.08	2.79	5.03	3.43
	Kernel	-1.39	3.71	-1.47	1.88	-1.20	2.19
	MSLE	-1.39	3.71	-1.47	1.88	-1.20	2.19

Table 2.17: RB (%) and MSE ($\times 1000$) of eight methods for estimating the optimal cutoff point under Beta distribution settings

Distribution (J)	(n_0, n_1) Method	(50, 50)		(100, 100)		(150, 50)	
		RB	MSE	RB	MSE	RB	MSE
Beta (0.3)	BP	2.18	0.95	2.03	0.61	2.37	0.79
	Box-Cox	-0.41	1.33	-0.30	0.67	-2.03	0.88
	ZL	0.24	2.82	0.71	1.42	0.89	1.81
	LZL	0.97	7.73	1.06	5.59	1.86	6.66
	ECDF	0.97	7.30	1.05	5.25	1.86	6.54
	MNLE	1.10	7.21	1.08	5.22	1.87	6.53
	Kernel	-0.74	5.07	0.44	3.10	-0.14	4.09
	MSLE	-0.74	5.08	-0.45	3.10	-0.14	4.09
Beta (0.5)	BP	1.13	0.62	0.96	0.33	1.10	0.49
	Box-Cox	-1.66	0.72	-1.78	0.42	-2.97	0.75
	ZL	0.08	1.42	0.36	0.68	0.73	0.83
	LZL	0.15	3.84	0.10	2.46	0.94	3.03
	ECDF	0.06	3.68	0.02	2.40	0.96	2.99
	MNLE	0.15	3.68	0.04	2.38	0.96	2.99
	Kernel	-1.24	1.88	-1.07	1.13	-0.89	1.56
	MSLE	-1.24	1.88	-1.07	1.14	-0.90	1.56
Beta (0.7)	BP	0.46	0.44	0.44	0.24	0.62	0.34
	Box-Cox	-1.65	0.53	-1.77	0.36	-2.02	0.55
	ZL	-0.24	0.91	0.06	0.45	0.34	0.63
	LZL	-0.20	2.12	0.01	1.32	0.80	1.50
	ECDF	-0.46	1.93	-0.12	1.26	0.68	1.47
	MNLE	-0.39	1.93	-0.09	1.25	0.69	1.47
	Kernel	-1.16	0.99	-0.82	0.60	-0.78	0.84
	MSLE	-1.17	1.00	-0.82	0.60	-0.78	0.84

2.5.3 Proof of Proposition 2.1

For (a). The proof is straightforward and we omit the details.

For (b). Let $l_2(\alpha_0, \dots, \alpha_N) = \log L_2(\alpha_0, \dots, \alpha_N)$. Since there is no constraint on α_0 ,

and $l_2(\alpha_0, \dots, \alpha_N)$ is a concave function of $(\alpha_0, \dots, \alpha_N)$, $\hat{\alpha}_0$ has to satisfy the first-order condition. That is

$$\frac{\partial l_2(\hat{\alpha}_0, \hat{\alpha}_1, \dots, \hat{\alpha}_N)}{\partial \alpha_0} = 0. \quad (2.16)$$

Note that

$$\frac{\partial}{\partial \alpha_0} l_2(\alpha_0, \alpha_1, \dots, \alpha_N) = \sum_{i=1}^m b_i - \sum_{i=1}^m (a_i + b_i) \theta(t_i).$$

Therefore, (2.16) leads to

$$0 = \sum_{i=1}^m b_i - \sum_{i=1}^m (a_i + b_i) \hat{\theta}(t_i),$$

which, together with the facts that $\sum_{i=1}^m b_i = n_1$ and $\lambda = n_1/n$, imply that $\sum_{i=1}^m \hat{\phi}_i \hat{\theta}(t_i) = \lambda$. This finishes the proof.

2.5.4 Proof of Theorem 2.1

We first define some notation. We denote

$$d^2(\hat{\theta}, \theta_0) = \int_0^1 \left\{ \hat{\theta}(x) - \theta_0(x) \right\}^2 dG(x).$$

Let $R(x) = \log\{f_1(x)/f_0(x)\}$,

$$\theta_N(x) = \frac{\lambda \exp \left\{ \sum_{l=0}^N R(l/N) B_l(x; N) \right\}}{1 - \lambda + \lambda \exp \left\{ \sum_{l=0}^N R(l/N) B_l(x; N) \right\}}, \quad (2.17)$$

and

$$\mathcal{B}_N = \left\{ \theta(x) = \frac{\lambda \exp \left\{ \sum_{l=0}^N \beta_l B_l(x; N) \right\}}{1 - \lambda + \lambda \exp \left\{ \sum_{l=0}^N \beta_l B_l(x; N) \right\}} : \beta_0 \leq \beta_1 \cdots \leq \beta_N \right\}.$$

By Condition A3, $\theta_N(x) \in \mathcal{B}_N$. Further let

$$\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^m \log[\{\theta(t_i)\}^{b_i} \{1 - \theta(t_i)\}^{a_i}].$$

Then

$$\hat{\theta}(x) = \arg \max_{\theta(x) \in \mathcal{B}_N} \ell_n(\theta).$$

By Condition A1, $n_1/n \rightarrow \lambda \in (0, 1)$ as $n \rightarrow \infty$. As we discussed in Section 2.2.1, we write $\lambda = n_1/n$ and assume that it is constant, since it does not affect our technical development. Define

$$\gamma_0(x; \theta) = 4(1 - \lambda) \left\{ \sqrt{\frac{1 - \theta(x)}{1 - \theta_0(x)}} - 1 \right\}$$

and

$$\gamma_1(x; \theta) = 4\lambda \left\{ \sqrt{\frac{\theta(x)}{\theta_0(x)}} - 1 \right\}.$$

The proof of Theorem 1 consists of three steps.

Step 1. We argue that

$$\begin{aligned} d^2(\hat{\theta}, \theta_0) &\leq \int \gamma_0(x; \hat{\theta}) d\{\tilde{F}_0(x) - F_0(x)\} \\ &\quad + \int \gamma_1(x; \hat{\theta}) d\{\tilde{F}_1(x) - F_1(x)\} + 2\{\ell_n(\theta_0) - \ell_n(\theta_N)\}, \end{aligned}$$

where $\tilde{F}_0(x)$ and $\tilde{F}_1(x)$ are the empirical cdfs of $\{X_1, \dots, X_{n_0}\}$ and $\{Y_1, \dots, Y_{n_1}\}$, respectively.

Step 2. We use results from empirical processes to show that

$$\int \gamma_0(x; \hat{\theta}) d\{\tilde{F}_0(x) - dF_0(x)\} = o_p(1) \quad (2.18)$$

and

$$\int \gamma_1(x; \hat{\theta}) d\{\tilde{F}_1(x) - dF_1(x)\} = o_p(1). \quad (2.19)$$

Step 3. We further show that

$$\ell_n(\theta_0) - \ell_n(\theta_N) = o_p(1). \quad (2.20)$$

We start with Step 1. Since $\hat{\theta}(x)$ maximizes $\ell_n(\theta)$ over $\mathcal{B}_N$ and $\theta_N(x) \in \mathcal{B}_N$, we have

$$\begin{aligned} \ell_n(\theta_0) - \ell_n(\theta_N) &\geq \ell_n(\theta_0) - \ell_n(\hat{\theta}) \\ &= -\frac{1}{n} \sum_{i=1}^m \left\{ b_i \log \frac{\hat{\theta}(t_i)}{\theta_0(t_i)} + a_i \log \frac{1 - \hat{\theta}(t_i)}{1 - \theta_0(t_i)} \right\} \\ &= -\frac{1}{n} \left\{ \sum_{i=1}^{n_1} \log \frac{\hat{\theta}(Y_i)}{\theta_0(Y_i)} + \sum_{j=1}^{n_0} \log \frac{1 - \hat{\theta}(X_j)}{1 - \theta_0(X_j)} \right\} \\ &= -\left\{ \lambda \int \log \frac{\hat{\theta}(x)}{\theta_0(x)} d\tilde{F}_1(x) + (1 - \lambda) \int \log \frac{1 - \hat{\theta}(x)}{1 - \theta_0(x)} d\tilde{F}_0(x) \right\}. \end{aligned}$$

Applying the inequality $0.5 \log x \leq \sqrt{x} - 1$ for $x > 0$, we further get

$$\begin{aligned} 2\{\ell_n(\theta_0) - \ell_n(\theta_N)\} &\geq -\int \gamma_1(x; \hat{\theta}) d\tilde{F}_1(x) - \int \gamma_0(x; \hat{\theta}) d\tilde{F}_0(x) \\ &= -\int \gamma_1(x; \hat{\theta}) d\{\tilde{F}_1(x) - dF_1(x)\} - \int \gamma_1(x; \hat{\theta}) dF_1(x) \\ &\quad - \int \gamma_0(x; \hat{\theta}) d\{\tilde{F}_0(x) - dF_0(x)\} - \int \gamma_0(x; \hat{\theta}) dF_0(x), \end{aligned}$$

which leads to

$$\begin{aligned}
-\int \gamma_0(x; \hat{\theta}) dF_0(x) - \int \gamma_1(x; \hat{\theta}) dF_1(x) &\leq \int \gamma_1(x; \hat{\theta}) d\{\tilde{F}_1(x) - dF_1(x)\} \\
&+ \int \gamma_0(x; \hat{\theta}) d\{\tilde{F}_0(x) - F_0(x)\} \\
&+ 2\{\ell_n(\theta_0) - \ell_n(\theta_N)\}.
\end{aligned}$$

To finish the proof of Step 1, it suffices to show that

$$d^2(\hat{\theta}, \theta_0) \leq -\int \gamma_0(x; \hat{\theta}) dF_0(x) - \int \gamma_1(x; \hat{\theta}) dF_1(x). \quad (2.21)$$

By the definition of $\theta_0(x)$, we have that

$$dF_1(x) = \frac{\theta_0(x)/\{1 - \theta_0(x)\}}{\lambda/(1 - \lambda)} dF_0(x) \quad (2.22)$$

and

$$dG(x) = \lambda dF_1(x) + (1 - \lambda) dF_0(x) = \frac{1 - \lambda}{1 - \theta_0(x)} dF_0(x). \quad (2.23)$$

Note that (2.22) and (2.23) imply

$$\begin{aligned}
-\int \gamma_1(x; \hat{\theta}) dF_1(x) &= \int 4\lambda \frac{\theta_0(x)/\{1 - \theta_0(x)\}}{\lambda/(1 - \lambda)} \left\{ 1 - \sqrt{\frac{\hat{\theta}(x)}{\theta_0(x)}} \right\} dF_0(x) \\
&= 4 \int \theta_0(x) \left\{ 1 - \sqrt{\frac{\hat{\theta}(x)}{\theta_0(x)}} \right\} dG(x)
\end{aligned}$$

and

$$-\int \gamma_0(x; \hat{\theta}) dF_0(x) = 4 \int \{1 - \theta_0(x)\} \left\{ 1 - \sqrt{\frac{1 - \hat{\theta}(x)}{1 - \theta_0(x)}} \right\} dG(x).$$

Hence,

$$\begin{aligned} & -\int \gamma_0(x; \hat{\theta}) dF_0(x) - \int \gamma_1(x; \hat{\theta}) dF_1(x) \\ &= 4 \int \left[1 - \sqrt{\hat{\theta}(x)\theta_0(x)} - \sqrt{\{1 - \hat{\theta}(x)\} \{1 - \theta_0(x)\}} \right] dG(x) \\ &= 2 \int \left\{ \sqrt{\hat{\theta}(x)} - \sqrt{\theta_0(x)} \right\}^2 dG(x) + 2 \int \left\{ \sqrt{1 - \hat{\theta}(x)} - \sqrt{1 - \theta_0(x)} \right\}^2 dG(x). \end{aligned}$$

Using the facts that both $\theta_0(x)$ and $\hat{\theta}(x) \in [0, 1]$, we get that

$$\begin{aligned} & -\int \gamma_0(x; \hat{\theta}) dF_0(x) - \int \gamma_1(x; \hat{\theta}) dF_1(x) \\ & \geq 2 \int \frac{1}{4} \left\{ \hat{\theta}(x) - \theta_0(x) \right\}^2 dG(x) + 2 \int \frac{1}{4} \left\{ \hat{\theta}(x) - \theta_0(x) \right\}^2 dG(x) \\ & = d^2(\hat{\theta}, \theta_0). \end{aligned}$$

This finishes the proof of (2.21), and hence that of Step 1.

We now move to Step 2. Let

$$\mathcal{H} = \{h(x) : h(x) \text{ is non-decreasing on } [0, 1] \text{ and } h(x) \in [0, 1]\}.$$

Using Theorems 2.2 and 9.24 of [Kosorok \(2008\)](#), we have that the class $\mathcal{H}$ is P -Glivenko–

Cantelli (P -G-C) for $P = F_0$ and F_1 . That is,

$$\sup_{h(x) \in \mathcal{H}} \left| \int h(x) d\{\tilde{F}_0(x) - F_0(x)\} \right| = o_p(1) \quad \text{and} \quad \sup_{h(x) \in \mathcal{H}} \left| \int h(x) d\{\tilde{F}_1(x) - F_1(x)\} \right| = o_p(1).$$

Since $\mathcal{B}_N \subset \mathcal{H}$, the class $\mathcal{B}_N$ is also P -G-C for $P = F_0$ and F_1 .

Let

$$\Gamma_0 = \{\gamma_0(x; \theta) : \theta(x) \in \mathcal{B}_N\}.$$

We notice that $|\gamma_0(x; \theta)| \leq 4(1 - \lambda)[1/\{1 - \theta_0(x)\} + 1]$ and

$$\begin{aligned} \int 4(1 - \lambda)[1/\{1 - \theta_0(x)\} + 1] dF_0(x) &= 4(1 - \lambda) + \int 4(1 - \lambda)/\{1 - \theta_0(x)\} dF_0(x) \\ &= 4(1 - \lambda) + \int 4 dG(x) \\ &= 4 + 4(1 - \lambda) < \infty. \end{aligned}$$

That is, the class Γ_0 has an integrable envelope. Applying Corollary 9.27 of [Kosorok \(2008\)](#), we then get that Γ_0 is F_0 -G-C, i.e.,

$$\sup_{\theta(x) \in \Gamma_0} \left| \int \gamma_0(x; \theta) d\{\tilde{F}_0(x) - F_0(x)\} \right| = o_p(1),$$

which implies that

$$\int \gamma_0(x; \hat{\theta}) d\{\tilde{F}_0(x) - F_0(x)\} = o_p(1).$$

Similarly,

$$\int \gamma_1(x; \hat{\theta}) d\{\tilde{F}_1(x) - F_1(x)\} = o_p(1).$$

This finishes Step 2.

We consider Step 3. It can be checked that

$$\begin{aligned}\ell_n(\theta_0) - \ell_n(\theta_N) &= -\frac{\lambda}{n_1} \sum_{i=1}^{n_1} \log \frac{\theta_N(Y_i)}{\theta_0(Y_i)} - \frac{1-\lambda}{n_0} \sum_{j=1}^{n_0} \log \frac{1-\theta_N(X_j)}{1-\theta_0(X_j)} \\ &= -\lambda \int \log \frac{\theta_N(x)}{\theta_0(x)} d\tilde{F}_1(x) - (1-\lambda) \int \log \frac{1-\theta_N(x)}{1-\theta_0(x)} d\tilde{F}_0(x).\end{aligned}$$

By Condition A3 and the construction of $\theta_N(x)$ in (2.17), we have

$$\theta_0(x) \in [\delta, 1-\delta] \quad \text{and} \quad \theta_N(x) \in [\delta, 1-\delta],$$

where $\delta > 0$ is provided in Condition A3.

Let

$$\mathcal{H}^* = \{h(x) : h(x) \text{ is non-decreasing on } [0,1] \text{ and } h(x) \in [\delta, 1-\delta]\}.$$

Then $\mathcal{H}^* \subset \mathcal{H}$. Hence, the class $\mathcal{H}^*$ is P -G-C for $P = F_0$ and F_1 . Note that for any $h(x) \in \mathcal{H}^*$, both $|\log\{h(x)/\theta_0(x)\}|$ and $|\log[\{1-h(x)\}/\{1-\theta_0(x)\}]|$ are bounded functions.

Using Corollary 9.27 of Kosorok (2008), we further have the two classes

$$\left\{ \log \frac{h(x)}{\theta_0(x)} : h(x) \in \mathcal{H}^* \right\} \quad \text{and} \quad \left\{ \log \frac{1-h(x)}{1-\theta_0(x)} : h(x) \in \mathcal{H}^* \right\}$$

are also P -G-C for $P = F_0$ and F_1 , which implies that

$$\int \log \frac{\theta_N(x)}{\theta_0(x)} d\{\tilde{F}_1(x) - F_1(x)\} = o_p(1) \quad \text{and} \quad \int \log \frac{1-\theta_N(x)}{1-\theta_0(x)} d\{\tilde{F}_0(x) - F_0(x)\} = o_p(1).$$

Hence,

$$\begin{aligned}\ell_n(\theta_0) - \ell_n(\theta_N) &= -\lambda \int \log \frac{\theta_N(x)}{\theta_0(x)} dF_1(x) - (1 - \lambda) \int \log \frac{1 - \theta_N(x)}{1 - \theta_0(x)} dF_0(x) + o_p(1) \\ &= - \int \left[\theta_0(x) \log \frac{\theta_N(x)}{\theta_0(x)} + \{1 - \theta_0(x)\} \log \frac{1 - \theta_N(x)}{1 - \theta_0(x)} \right] dG(x) + o_p(1).\end{aligned}$$

To finish Step 3, it is sufficient to show that

$$\int \left[\theta_0(x) \log \frac{\theta_N(x)}{\theta_0(x)} + \{1 - \theta_0(x)\} \log \frac{1 - \theta_N(x)}{1 - \theta_0(x)} \right] dG(x) = o(1).$$

Recall that $R(x) = \log\{f_1(x)/f_0(x)\}$. Let

$$R_N(x) = \sum_{l=0}^N R(l/N) B_l(x; N).$$

According to [Lorentz \(1986\)](#), under Condition A3, we have

$$\lim_{N \rightarrow \infty} \sup_{x \in [0,1]} |R(x) - R_N(x)| = 0. \quad (2.24)$$

Further let

$$g(t) = \frac{\lambda \exp(t)}{1 - \lambda + \lambda \exp(t)}.$$

Then, $\theta_0(x) = g(R(x))$ and $\theta_N(x) = g(R_N(x))$. It can be easily verified that $|g'(t)| \leq 1$ for all t , which, together with (2.24), implies that

$$\lim_{N \rightarrow \infty} \sup_{x \in [0,1]} |\theta_N(x) - \theta_0(x)| = 0. \quad (2.25)$$

Using the second-order Taylor expansion, we get

$$\begin{aligned}
& \left| \int \left[\theta_0(x) \log \frac{\theta_N(x)}{\theta_0(x)} + \{1 - \theta_0(x)\} \log \frac{1 - \theta_N(x)}{1 - \theta_0(x)} \right] dG(x) \right| \\
&= 0.5 \int \left[\frac{\theta_0(x)}{\xi_N^2(x)} + \frac{1 - \theta_0(x)}{\{1 - \xi_N(x)\}^2} \right] \{\theta_N(x) - \theta_0(x)\}^2 dG(x) \\
&\leq \frac{0.5}{\delta^2} \int \{\theta_N(x) - \theta_0(x)\}^2 dG(x)
\end{aligned} \tag{2.26}$$

where $\xi_N(x)$ is a function between $\theta_0(x)$ and $\theta_N(x)$, and we have used Condition A3 in the last step. Combining (2.25) and (2.26) leads to

$$\left| \int \left[\theta_0(x) \log \frac{\theta_N(x)}{\theta_0(x)} + \{1 - \theta_0(x)\} \log \frac{1 - \theta_N(x)}{1 - \theta_0(x)} \right] dG(x) \right| \leq \frac{0.5}{\delta^2} \left\{ \sup_{x \in [0,1]} |\theta_N(x) - \theta_0(x)| \right\}^2 \rightarrow 0$$

as $N \rightarrow \infty$. This finishes the proof of Step 3, and hence the proof of Theorem 1.

2.5.5 Proof of Theorem 2.2

Part (a) We concentrate on the consistency of $\hat{F}_1(x)$; the consistency of $\hat{F}_0(x)$ follows similarly.

Let

$$\tilde{G}(x) = \lambda \tilde{F}_1(x) + (1 - \lambda) \tilde{F}_0(x) = \sum_{i=1}^m \frac{a_i + b_i}{n} I(t_i \leq x).$$

Note that for any u ,

$$\begin{aligned}
\hat{F}_1(u) &= \sum_{i=1}^m \hat{p}_{i1} I(t_i \leq u) \\
&= \sum_{i=1}^m \frac{1}{\lambda} \frac{a_i + b_i}{n} \hat{\theta}(t_i) I(t_i \leq u) \\
&= \int \frac{1}{\lambda} \hat{\theta}(x) I(x \leq u) d\tilde{G}(x).
\end{aligned}$$

Recall

$$F_1(u) = \int \frac{1}{\lambda} \theta_0(x) I(x \leq u) dG(x).$$

Then

$$\hat{F}_1(u) - F_1(u) = \frac{1}{\lambda} \int \{\hat{\theta}(x) - \theta_0(x)\} I(x \leq u) dG(x) \quad (2.27)$$

$$+ \frac{1}{\lambda} \int \hat{\theta}(x) I(x \leq u) d\{\tilde{G}(x) - G(x)\}. \quad (2.28)$$

In the following, we show that the two terms in (2.27) and (2.28) are $o_p(1)$.

We first consider (2.27). Note that

$$\begin{aligned}
\left| \frac{1}{\lambda} \int \{\hat{\theta}(x) - \theta_0(x)\} I(x \leq u) dG(x) \right| &\leq \frac{1}{\lambda} \int |\hat{\theta}(x) - \theta_0(x)| I(x \leq u) dG(x) \\
&\leq \frac{1}{\lambda} \int |\hat{\theta}(x) - \theta_0(x)| dG(x) \\
&\leq \frac{1}{\lambda} \left[\int \left\{ \hat{\theta}(x) - \theta_0(x) \right\}^2 dG(x) \right]^{1/2} \\
&= \lambda^{-1} d(\hat{\theta}, \theta_0) = o_p(1), \quad (2.29)
\end{aligned}$$

where we have used the result in Theorem 2.1.

Next, we consider (2.28). According to Lemmas 9.8 and 9.12 of Kosorok (2008),

$$\{I(x \leq u) : u \in \mathcal{R}\}$$

is a Vapnik–Chervonenkis subgraph class with the bounded envelope 1. By Theorems 8.14 and 9.2 of Kosorok (2008), we have that $\{I(x \leq u) : u \in \mathcal{R}\}$ is P -G-C for both F_0 and F_1 . Recall that in the proof of Theorem 2.1, we have argued that $\mathcal{B}_N$ is also P -G-C for both F_0 and F_1 . Then by Corollary 9.27 of Kosorok (2008), $\{I(x \leq u)\theta(x) : \theta \in \mathcal{B}_N, u \in \mathcal{R}\}$ is P -G-C for both F_0 and F_1 . That is,

$$\sup_{\theta(x) \in \mathcal{B}_N, u \in \mathcal{R}} \left| \int I(x \leq u)\theta(x) d\{\tilde{F}_0(x) - F_0(x)\} \right| = o_p(1) \quad (2.30)$$

and

$$\sup_{\theta(x) \in \mathcal{B}_N, u \in \mathcal{R}} \left| \int I(x \leq u)\theta(x) d\{\tilde{F}_1(x) - F_1(x)\} \right| = o_p(1). \quad (2.31)$$

Note that (2.30) and (2.31) together lead to

$$\sup_{\theta(x) \in \mathcal{B}_N, u \in \mathcal{R}} \left| \int I(x \leq u)\theta(x) d\{\tilde{G}(x) - G(x)\} \right| = o_p(1),$$

which implies that

$$\sup_{u \in [0,1]} \left| \int I(x \leq u)\hat{\theta}(x) d\{\tilde{G}(x) - G(x)\} \right| = o_p(1). \quad (2.32)$$

Combining the results in (2.27)–(2.29) and (2.32), we have

$$\sup_{x \in [0,1]} |\hat{F}_1(x) - F_1(x)| = o_p(1).$$

This completes the proof of Part (a).

Part (b) The proof for this part is very similar to the proof of Theorem 1 in [Jokiel-Rokita & Pulit \(2013\)](#). We omit the details here.

Part (c) Note that

$$\begin{aligned} |\widehat{AUC} - AUC| &= \left| \int_0^1 \widehat{ROC}(s) - ROC(s) \, ds \right| \\ &\leq \int_0^1 \left| \widehat{ROC}(s) - ROC(s) \right| \, ds \\ &\leq \int_0^1 \sup_{s \in [0,1]} \left| \widehat{ROC}(s) - ROC(s) \right| \, ds \\ &\leq \sup_{s \in [0,1]} \left| \widehat{ROC}(s) - ROC(s) \right| \\ &= o_p(1). \end{aligned}$$

Part (d) Let $A_n = \left\{ \inf_{x \in [0,1]} \hat{\theta}(x) < \lambda < \sup_{x \in [0,1]} \hat{\theta}(x) \right\}$. We first show that

$$\lim_{n \rightarrow \infty} P(A_n) = 1, \tag{2.33}$$

which implies that the probability that the solution of $\hat{\theta}(x) = \lambda$ exists in $[0, 1]$ approaches to 1 as n goes to infinity. Define

$$A_{n1} = \left\{ \inf_{x \in [0,1]} \hat{\theta}(x) \geq \lambda \right\}$$

and

$$A_{n2} = \left\{ \sup_{x \in [0,1]} \hat{\theta}(x) \leq \lambda \right\}.$$

It suffices to show that

$$\lim_{n \rightarrow \infty} P(A_{n1}) = 0, \quad (2.34)$$

and

$$\lim_{n \rightarrow \infty} P(A_{n2}) = 0. \quad (2.35)$$

We concentrate on (2.34); the proof for (2.35) follows similarly.

By Condition A4, there exists $\epsilon_0 > 0$ such that $\theta_0(x)$ is strictly increasing on $[C - \epsilon_0, C + \epsilon_0]$. Hence, $\theta_0(C) - \theta_0(C - \epsilon_0) > 0$. By the continuity and monotonicity of $\theta_0(x)$, there exists $0 < \delta_0 < \epsilon_0$ such that

$$0 < \theta_0(x) - \theta_0(C - \epsilon_0) \leq \frac{\theta_0(C) - \theta_0(C - \epsilon_0)}{2}$$

for any $x \in [C - \epsilon_0, C - \epsilon_0 + \delta_0]$, which together with the fact $\theta_0(C) = \lambda$ imply that

$$\lambda - \theta_0(x) = \theta_0(C) - \theta_0(x) \geq \frac{\theta_0(C) - \theta_0(C - \epsilon_0)}{2}, \quad \forall x \in [C - \epsilon_0, C - \epsilon_0 + \delta_0]. \quad (2.36)$$

Let $\delta_1 = 0.25\delta_0 \{\theta_0(C) - \theta_0(C - \epsilon_0)\}^2 > 0$. Then (2.36) implies that

$$\begin{aligned} P(A_{n1}) &\leq P\left(\int_{C-\epsilon_0}^{C-\epsilon_0+\delta_0} \left\{\hat{\theta}(x) - \theta_0(x)\right\}^2 dx \geq \delta_1\right) \\ &\leq P\left(\int_0^1 \left\{\hat{\theta}(x) - \theta_0(x)\right\}^2 dx \geq \delta_1\right) \\ &= P\left(d^2(\hat{\theta}, \theta_0) \geq \delta_1\right). \end{aligned}$$

By Theorem 2.1, we get

$$\lim_{n \rightarrow \infty} P(A_{n1}) = 0.$$

This finishes the proof of (2.34) and hence that of (2.33). With the result in (2.33), without loss of generality, we assume that the solution to $\hat{\theta}(x) = \lambda$ exists, i.e.,

$$\hat{\theta}(\hat{C}) = \lambda.$$

We now move to the consistency of $\hat{C}$. Equivalently, we need to show that for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(\hat{C} \geq C + \epsilon) = 0, \tag{2.37}$$

and

$$\lim_{n \rightarrow \infty} P(\hat{C} \leq C - \epsilon) = 0. \tag{2.38}$$

We first consider (2.37). Note that

$$\begin{aligned}
P(\hat{C} \geq C + \epsilon) &\leq P\left(\hat{\theta}(\hat{C}) \geq \hat{\theta}(C + \epsilon)\right) \\
&= P\left(\hat{\theta}(C + \epsilon) - \lambda \leq 0\right) \\
&\leq P\left(\int_C^{C+\epsilon} \left\{\theta_0(x) - \hat{\theta}(x)\right\}^2 dx \geq \int_C^{C+\epsilon} \{\theta_0(x) - \lambda\}^2 dx\right) \\
&\leq P\left(\int_0^1 \left\{\theta_0(x) - \hat{\theta}(x)\right\}^2 dx \geq \int_C^{C+\epsilon} \{\theta_0(x) - \lambda\}^2 dx\right) \\
&= P\left(d^2(\hat{\theta}, \theta_0) \geq \int_C^{C+\epsilon} \{\theta_0(x) - \lambda\}^2 dx\right).
\end{aligned}$$

Condition A4 implies that $\int_C^{C+\epsilon} \{\theta_0(x) - \lambda\}^2 dx > 0$. By Theorem 2.1, we have

$$\lim_{n \rightarrow \infty} P(\hat{C} \geq C + \epsilon) = 0.$$

This completes (2.37). The proof of (2.38) follows similarly.

Finally, we consider the consistency of $\hat{J}$. Note that

$$\begin{aligned}
\hat{J} - J &= \left\{\hat{F}_0(\hat{C}) - \hat{F}_1(\hat{C})\right\} - \{F_0(C) - F_1(C)\} \\
&= \left\{\hat{F}_0(\hat{C}) - \hat{F}_1(\hat{C})\right\} - \left\{F_0(\hat{C}) - F_1(\hat{C})\right\} \\
&\quad + \left\{F_0(\hat{C}) - F_1(\hat{C})\right\} - \{F_0(C) - F_1(C)\}.
\end{aligned}$$

Then

$$|\hat{J} - J| \leq \left|F_0(\hat{C}) - F_0(C)\right| + \left|F_1(\hat{C}) - F_1(C)\right| \quad (2.39)$$

$$+ \sup_{x \in [0,1]} \left|\hat{F}_0(x) - F_0(x)\right| + \sup_{x \in [0,1]} \left|\hat{F}_1(x) - F_1(x)\right|. \quad (2.40)$$

By the consistency of $\hat{C}$ and the continuous mapping theorem, the term in (2.39) is $o_p(1)$. The term in (2.40) is also $o_p(1)$ by the results in Part (a). Hence $\hat{J} - J = o_p(1)$ as claimed.

Chapter 3

ROC Curve Analysis With Non-Ignorable Missing Disease Status

3.1 Introduction

The method on estimating the ROC curve and the corresponding AUC in [Chapter 2](#) depends on having complete observations of the disease status. However, in some applications, verifying disease status for every individual is impractical due to cost, time, or ethical constraints, leading to data with missing disease status for some patients. For these data, a popular assumption is that the missing disease status is missing at random (MAR): the verification status is independent of the true disease status, given the available biomarker and covariate information. Under this assumption, the ROC estimation has been well established in the literature, such as [Begg & Greenes \(1983\)](#), [Zhou \(1998\)](#),

[Fluss et al. \(2009\)](#), [Zhou et al. \(2011\)](#), and the references therein.

However, the MAR assumption is restrictive and is often violated in practice. For example, consider the Alzheimer’s disease (AD) study discussed in [Zhou & Castelluccio \(2004\)](#). In that study, a two-stage design is employed, where the first stage involves a screening test on a large population, and the second consists of a more comprehensive and costly clinical diagnosis on a selected subgroup of individuals from the first stage. The selected patients in the subgroup for the second stage depend on the test results from the first stage and other relevant covariates. The collected data from such a study do not satisfy the MAR assumption, since the diagnostic test in the second stage is applied only to individuals selected based on a mechanism that may relate to their disease status. With this data, the statistical inference for the ROC curve and AUC, based on the MAR assumption, can be biased ([Liu & Zhou 2010](#)).

Assuming the existence of the non-ignorable verification bias, the estimation method for the AUC is available in the literature. [Liu & Zhou \(2010\)](#) proposed a method to model such data with two separate logistic regression models: one regresses the disease status on the biomarker and relevant covariates, and the other regresses verification status on the disease status, biomarker, and relevant covariates. However, as noted by [Liu & Zhou \(2010\)](#), it may not be feasible to validate these models because of the missing disease status in data. To overcome this difficulty, [Yu et al. \(2018\)](#) proposed regressing the disease status on the biomarker and relevant covariates only for individuals with verified disease status, while retaining the same model for verification status as in [Liu & Zhou \(2010\)](#).

We note that [Yu et al. \(2018\)](#)’s method has three limitations. First, their estimation method for the unknown parameters does not make full use of the available data, leaving potential for improvement. Their AUC estimator relies on the inverse probability weighting (IPW) method, which uses data only from individuals with verified disease status,

potentially making it less efficient than methods that utilize all available data (Liu & Zhou 2010, Liu et al. 2022, Li et al. 2023). Second, Yu et al. (2018) assumes the availability of an instrumental variable; in practice, however, it may not be clear of how to identify it. Third, statistical inference for the ROC curve is not considered.

Our contribution in this chapter is threefold. First, under the same framework as Yu et al. (2018), we establish the identifiability of model parameters under mild conditions, without requiring the existence of an instrumental variable. Second, we propose a maximum likelihood method to estimate the unknown model parameters, ROC curve, and AUC. Our approach incorporates all available data, and is therefore expected to be more efficient than the method in Yu et al. (2018); this has been demonstrated in our simulation study. Third, we introduce a two-step verification process for both the disease and verification models.

The rest of this chapter is organized as follows. In Section 3.2, we introduce our proposed method, detailing model identifiability, estimation of the ROC curve and AUC, and model verification. The asymptotic properties of the ROC curve and AUC estimators are discussed in Section 3.3. In Section 3.4, we present simulation studies to compare the performance of our method with existing methods. Section 3.5 illustrates our proposed method using a real data example.

3.2 Main results

3.2.1 Setup

For an individual, let Y denote their disease status ($0 = \text{healthy}$, $1 = \text{diseased}$), X their biomarker, and $\mathbf{V}$ their associated covariates. Let R represent the verification status or

missing indicator for Y , where $R = 0$ indicates that Y is missing, and $R = 1$ indicates that Y is observed. Following [Yu et al. \(2018\)](#), we consider the following disease model

$$\begin{aligned} P(Y = 1|X = x, \mathbf{V} = \mathbf{v}, R = 1) &= P(Y = 1|x, \mathbf{v}, R = 1) \\ &= \frac{1}{1 + \exp(\mu_1 + \mu_2 x + \boldsymbol{\mu}_3^T \mathbf{v})} \end{aligned} \quad (3.1)$$

and the verification model

$$\begin{aligned} P(R = 1|Y = y, X = x, \mathbf{V} = \mathbf{v}) &= P(R = 1|y, x, \mathbf{v}) \\ &= \frac{1}{1 + \exp(\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + \beta y)}. \end{aligned} \quad (3.2)$$

We denote $F_0(x) = P(X \leq x|Y = 0)$ and $F_1(x) = P(X \leq x|Y = 1)$ as the cdfs of biomarkers for healthy and diseased individuals, respectively. Recall that the ROC based on the biomarker X is

$$ROC(s) = 1 - F_1(F_0^{-1}(1 - s)), \quad (3.3)$$

for $s \in (0, 1)$, and the area under the ROC curve is

$$AUC = \int_0^1 ROC(s) ds = \int F_0(x) dF_1(x). \quad (3.4)$$

In this chapter, we consider statistical inference for the ROC curve and the AUC under the models specified in (3.1) and (3.2).

For simplicity of presentation, we denote $\boldsymbol{\mu} = (\mu_1, \mu_2, \boldsymbol{\mu}_3^T)^T$, $\boldsymbol{\phi} = (\psi_1, \psi_2, \boldsymbol{\psi}_3^T, \beta)^T$, and $P_1(x, \mathbf{v}; \boldsymbol{\mu}) = P(Y = 1|x, \mathbf{v}, R = 1)$.

3.2.2 Identifiability of model parameters

We consider the identifiability and the needed conditions for models (3.1) and (3.2). For individuals with $R = 1$, the corresponding $Y, X, \mathbf{V}$ are all observed; the model (3.1) is essentially the classical logistic regression model. Thus, we assume that model parameters $\boldsymbol{\mu}$ are identifiable.

We proceed to consider the identifiability of $\boldsymbol{\phi}$ in (3.2). We have

$$\frac{P(X = x, \mathbf{V} = \mathbf{v}, Y = y, R = 0)}{P(X = x, \mathbf{V} = \mathbf{v}, Y = y, R = 1)} = \exp(\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + \beta y),$$

which leads to

$$\begin{aligned} P(X = x, \mathbf{V} = \mathbf{v}, R = 0) &= \sum_{y=0}^1 P(X = x, \mathbf{V} = \mathbf{v}, Y = y, R = 0) \\ &= P(X = x, \mathbf{V} = \mathbf{v}, R = 1) \exp\{\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + c(x, \mathbf{v}; \boldsymbol{\mu}, \beta)\}, \end{aligned}$$

where

$$c(x, \mathbf{v}; \boldsymbol{\mu}, \beta) = \log E(e^{\beta Y} | x, \mathbf{v}, R = 1) = \log \{\exp(\beta) P_1(x, \mathbf{v}; \boldsymbol{\mu}) + 1 - P_1(x, \mathbf{v}; \boldsymbol{\mu})\}.$$

Applying Bayes' formula, we have

$$\begin{aligned} P(R = 1 | x, \mathbf{v}) &\equiv P(R = 1 | X = x, \mathbf{V} = \mathbf{v}) \\ &= \frac{1}{1 + \exp\{\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + c(x, \mathbf{v}; \boldsymbol{\mu}, \beta)\}}. \end{aligned} \quad (3.5)$$

Because $(R, X, \mathbf{V})$ are observed for all the individuals, $P(R = 1 | x, \mathbf{v})$ is identifiable. As a consequence, according to (3.5), the identifiability of $\boldsymbol{\phi}$ is equivalent to its identifiability in

the expression $\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + c(x, \mathbf{v}; \boldsymbol{\mu}, \beta)$. Theorem 3.1 below ensures the identifiability of the parameters under mild regularity conditions.

Theorem 3.1. *Suppose Condition C1 in Section 3.6.1 is satisfied. Then, the model parameters $\boldsymbol{\mu}$ in (3.1) and $\boldsymbol{\phi}$ in (3.2) are identifiable.*

For presentational continuity, we put the technical details of this theorem in Section 3.6.2. We emphasize that, according to Theorem 3.1, the identifiability of $(\boldsymbol{\mu}, \boldsymbol{\phi})$ does not depend on the existence of an instrumental variable. Therefore, in practice, our method can be applied without the need to identify such a variable.

3.2.3 Estimation of model parameters

Let $\{(Y_i, X_i, \mathbf{V}_i, R_i), i = 1, \dots, n\}$ be independent and identically distributed copies of $(Y, X, \mathbf{V}, R)$, where the biomarker X_i and the covariates $\mathbf{V}_i$ are always observed, and Y_i is observed if and only if $R_i = 1$. Without loss of generality, we assume that $R_i = 1$, for $i = 1, \dots, n_1$, and $R_i = 0$, for $i = n_1 + 1, \dots, n$.

The likelihood for the observed data is given by

$$\begin{aligned} L_n &= \prod_{i=1}^{n_1} P(Y_i, R_i = 1 | X_i, \mathbf{V}_i) \cdot \prod_{i=n_1+1}^n P(R_i = 0 | X_i, \mathbf{V}_i) \cdot \prod_{i=1}^n P(X_i, \mathbf{V}_i) \\ &= \prod_{i=1}^{n_1} P(Y_i | X_i, \mathbf{V}_i, R_i = 1) \cdot \prod_{i=1}^n P(R_i | X_i, \mathbf{V}_i) \cdot \prod_{i=1}^n P(X_i, \mathbf{V}_i). \end{aligned}$$

Thus, after ignoring a constant that does not depend on $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$, the log-likelihood is given by

$$l_n(\boldsymbol{\mu}, \boldsymbol{\phi}) = l_{n1}(\boldsymbol{\mu}) + l_{n2}(\boldsymbol{\mu}, \boldsymbol{\phi}),$$

where

$$l_{n1}(\boldsymbol{\mu}) = \sum_{i=1}^n R_i [Y_i \log \{P_1(X_i, \mathbf{V}_i; \boldsymbol{\mu})\} + (1 - Y_i) \log \{1 - P_1(X_i, \mathbf{V}_i; \boldsymbol{\mu})\}]$$

and

$$l_{n2}(\boldsymbol{\mu}, \boldsymbol{\phi}) = \sum_{i=1}^n [R_i \log \pi(X_i, \mathbf{V}_i; \boldsymbol{\mu}, \boldsymbol{\phi}) + (1 - R_i) \log \{1 - \pi(X_i, \mathbf{V}_i; \boldsymbol{\mu}, \boldsymbol{\phi})\}],$$

where $P_1(x, \mathbf{v}; \boldsymbol{\mu}) = P(Y = 1|x, \mathbf{v}, R = 1)$ is given in (3.1) and $\pi(x, \mathbf{v}; \boldsymbol{\mu}, \boldsymbol{\phi}) = P(R = 1|x, \mathbf{v})$ is given in (3.5). The maximum likelihood estimator of $(\boldsymbol{\mu}, \boldsymbol{\phi})$ is then defined as

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\phi}}) = \arg \max_{\boldsymbol{\mu}, \boldsymbol{\phi}} l_n(\boldsymbol{\mu}, \boldsymbol{\phi}).$$

3.2.4 Estimation of the ROC curve and AUC

Based on the parameter estimators, we proceed to establish estimators for $F_0(\cdot)$ and $F_1(\cdot)$, and then use them to establish estimators for the ROC curve and AUC for the biomarker.

Note (3.2) and (3.5). Applying Bayes's rule, we have

$$\begin{aligned} \frac{P(Y = 1|x, \mathbf{v}, R = 0)}{P(Y = 1|x, \mathbf{v}, R = 1)} &= \frac{P(R = 0|x, \mathbf{v}, Y = 1)/P(R = 0|x, \mathbf{v})}{P(R = 1|x, \mathbf{v}, Y = 1)/P(R = 1|x, \mathbf{v})} \\ &= \frac{\exp(\beta)}{\exp\{c(x, \mathbf{v}; \boldsymbol{\mu}, \beta)\}}, \end{aligned}$$

which, together with model (3.1), leads to

$$\begin{aligned}
P(Y = 1|x, \mathbf{v}, R = 0) &= \frac{\exp(\beta)}{\exp\{c(x, \mathbf{v}; \boldsymbol{\mu}, \beta)\}} P(Y = 1|x, \mathbf{v}, R = 1) \\
&= \frac{P(Y = 1|x, \mathbf{v}, R = 1)e^\beta}{1 - P(Y = 1|x, \mathbf{v}, R = 1) + P(Y = 1|x, \mathbf{v}, R = 1)e^\beta} \\
&= \frac{1}{1 + \exp(\mu_1 + \mu_2 x + \boldsymbol{\mu}_3^T \mathbf{v} - \beta)}.
\end{aligned}$$

Referring to model (3.1), we can write

$$\begin{aligned}
P(Y = 1|x, \mathbf{v}, r) &\equiv P(Y = 1|X = x, \mathbf{V} = \mathbf{v}, R = r) \\
&= \frac{1}{1 + \exp\{\mu_1 + \mu_2 x + \boldsymbol{\mu}_3^T \mathbf{v} + (r - 1)\beta\}}.
\end{aligned} \tag{3.6}$$

Denote $g(x, \mathbf{v}, r; \boldsymbol{\mu}, \beta) = P(Y = 1|x, \mathbf{v}, r)$. The following proposition establishes the connection between $F_0(\cdot)$, $F_1(\cdot)$ and $g(x, \mathbf{v}, r; \boldsymbol{\mu}, \beta)$. The proof is given in Section 3.6.3.

Proposition 3.1. *Let $\boldsymbol{\mu}_0$ and β_0 be the true values of $\boldsymbol{\mu}$ and β , respectively. For any x , we have*

$$E \{g(X, V, R; \boldsymbol{\mu}_0, \beta_0)I(X \leq x)\} = E \{g(X, V, R; \boldsymbol{\mu}_0, \beta_0)\} \cdot F_1(x)$$

and

$$E [\{1 - g(X, V, R; \boldsymbol{\mu}_0, \beta_0)\}I(X \leq x)] = E [\{1 - g(X, V, R; \boldsymbol{\mu}_0, \beta_0)\}] \cdot F_0(x).$$

Based on Proposition 3.1, we can estimate F_0 and F_1 by

$$\hat{F}_1(x) = \frac{\sum_{i=1}^n \hat{g}_i I(X_i \leq x)}{\sum_{i=1}^n \hat{g}_i},$$

and

$$\hat{F}_0(x) = \frac{\sum_{i=1}^n (1 - \hat{g}_i) I(X_i \leq x)}{\sum_{i=1}^n (1 - \hat{g}_i)},$$

where $\hat{g}_i = g(X_i, \mathbf{V}_i, R_i; \hat{\boldsymbol{\mu}}, \hat{\beta})$. Consequently, the ROC curve and AUC are estimated by

$$\widehat{ROC}(s) = 1 - \hat{F}_1(\hat{F}_0^{-1}(1 - s))$$

and

$$\widehat{AUC} = \int \hat{F}_0(x) d\hat{F}_1(x).$$

The asymptotic properties of $\widehat{ROC}(s)$ and $\widehat{AUC}$ will be studied in Section 3.3.

3.2.5 Model verification

The estimation procedures outlined above rely on the model assumptions specified in (3.1) and (3.2). This section describes how to verify these assumptions using the available data $\{(Y_i, X_i, \mathbf{V}_i, R_i), i = 1, \dots, n\}$. We verify models (3.1) and (3.2) separately.

Since model (3.1) relies only on data with responses observed (i.e., $R = 1$), we can validate this model applying existing methods, such as the test based on unweighted sum of squares (Hosmer et al. 1997), available as `resid` in R package `rms`.

For model (3.2), since the responses Y are missed when $R = 0$, we are unable to validate it directly. However, based on the development in Section 3.2.2, models (3.1) and (3.2) lead to model (3.5). Therefore, once model (3.1) is validated, we can verify model (3.2) by performing the goodness of fit test for model (3.5). Motivated by the test based on the

unweighted sum of squares for the standard logistic regression model, we can use

$$T_2 = \sum_{i=1}^n \{(R_i - \hat{\pi}_i)^2 - \hat{\pi}_i(1 - \hat{\pi}_i)\}$$

to construct a test statistic, where $\hat{\pi}_i = \pi(X_i, \mathbf{V}_i; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\phi}})$. Let $se(T_2)$ be the estimated standard error of T_2 obtained from the bootstrap method. Following [Hosmer et al. \(1997\)](#), We use $T_2/se(T_2)$ as the test statistic and $N(0, 1)$ as its reference distribution under the null hypothesis that model (3.2) is correct.

3.3 Asymptotic properties

In this section, we establish the asymptotic distribution of $\widehat{AUC}$ and $\widehat{ROC}(s)$, for each $s \in (0, 1)$. We need some notation. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}^T, \beta)^T$ and $\boldsymbol{\eta} = (\boldsymbol{\mu}^T, \beta, \psi_1, \psi_2, \boldsymbol{\psi}_3^T)^T$, whose dimensions are respectively k_1 and k_2 . We use $\boldsymbol{\theta}_0$ and $\boldsymbol{\eta}_0$ to denote the true values of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Write

$$l_n(\boldsymbol{\eta}) = l_n(\boldsymbol{\mu}, \boldsymbol{\phi}) \quad \text{and} \quad g(x, \mathbf{v}, r; \boldsymbol{\theta}) = g(x, \mathbf{v}, r; \boldsymbol{\mu}, \beta),$$

where $l_n(\boldsymbol{\mu}, \boldsymbol{\phi})$ and $g(x, \mathbf{v}, r; \boldsymbol{\mu}, \beta)$ are defined in Section 3.2.3 and 3.2.4 respectively. Let $\lambda = P(Y = 1)$, $\xi_{1-s} = F_0^{-1}(1 - s)$, and

$$\mathbf{Z}_n = (\mathbf{Z}_{n1}^T, Z_{n2}, Z_{n3}, Z_{n4}, Z_{n5})^T,$$

where

$$\begin{aligned}
\mathbf{Z}_{n1} &= n^{-1} \frac{\partial l_n(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}, \\
Z_{n2} &= n^{-1} \sum_{i=1}^n g_i \{F_0(X_i) - AUC\}, \\
Z_{n3} &= n^{-1} \sum_{i=1}^n (1 - g_i) \{1 - F_1(X_i) - AUC\}, \\
Z_{n4} &= n^{-1} \sum_{i=1}^n (1 - g_i) \{I(X_i \leq \xi_{1-s}) - F_0(\xi_{1-s})\}, \\
Z_{n5} &= n^{-1} \sum_{i=1}^n g_i \{I(X_i \leq \xi_{1-s}) - F_1(\xi_{1-s})\},
\end{aligned}$$

where $g_i = g(X_i, \mathbf{V}_i, R_i; \boldsymbol{\theta}_0)$. It can be verified that

$$E(\mathbf{Z}_n) = 0.$$

Denote the variance-covariance matrix of $\sqrt{n}\mathbf{Z}_n$ by $\boldsymbol{\Sigma}_Z$, and denote the Fisher-information matrix at the true value $\boldsymbol{\eta}_0$ by $\mathbf{J}$. Let $\mathbf{I}_{k_2, k_1}$ be a $k_2 \times k_1$ matrix, where the top $k_1 \times k_1$ block is a identity matrix, and the rest elements are 0's. Define

$$\begin{aligned}
\mathbf{E}_1 &= E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{F_0(X) - AUC\} \right], \\
\mathbf{E}_2 &= E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{1 - F_1(X) - AUC\} \right], \\
\mathbf{E}_3 &= E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(X \leq \xi_{1-s}) - F_0(\xi_{1-s})\} \right], \\
\mathbf{E}_4 &= E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(X \leq \xi_{1-s}) - F_1(\xi_{1-s})\} \right].
\end{aligned}$$

The following theorem summarizes the asymptotic results of $\widehat{AUC}$ and $\widehat{ROC}(s)$. The proof is given in Section 3.6.4.

Theorem 3.2. *Suppose Conditions C1–C6 in Section 3.6.1 are satisfied. As $n \rightarrow \infty$, we have*

(a) $\sqrt{n}(\widehat{AUC} - AUC) \xrightarrow{d} N(0, \sigma_{AUC}^2)$, where $\sigma_{AUC}^2 = \mathbf{H}_1^T \boldsymbol{\Sigma}_Z \mathbf{H}_1$ with

$$\mathbf{H}_1 = \begin{pmatrix} \mathbf{J}^{-1} \mathbf{I}_{k_2, k_1} \left(\frac{\mathbf{E}_1}{\lambda} - \frac{\mathbf{E}_2}{1-\lambda} \right) \\ \lambda^{-1} \\ (1-\lambda)^{-1} \\ 0 \\ 0 \end{pmatrix};$$

(b) $\sqrt{n} \left\{ \widehat{ROC}(s) - ROC(s) \right\} \xrightarrow{d} N(0, \sigma_s^2)$, where $\sigma_s^2 = \mathbf{H}_2^T \boldsymbol{\Sigma}_Z \mathbf{H}_2$ with

$$\mathbf{H}_2 = \begin{pmatrix} \mathbf{J}^{-1} \mathbf{I}_{k_2, k_1} \left\{ \frac{-\mathbf{E}_4}{\lambda} - \frac{\mathbf{E}_3}{1-\lambda} \frac{f_1(\xi_{1-s})}{f_0(\xi_{1-s})} \right\} \\ 0 \\ 0 \\ (1-\lambda)^{-1} \frac{f_1(\xi_{1-s})}{f_0(\xi_{1-s})} \\ -\lambda^{-1} \end{pmatrix},$$

where $f_1(x)$ and $f_0(x)$ are the probability density functions of $F_1(x)$ and $F_0(x)$, respectively.

The asymptotic variance formulas presented in Theorem 3.2 can be used to construct plug-in variance estimators for $\widehat{AUC}$ and $\widehat{ROC}(s)$. The elements involved in the plug-in

estimators take the form

$$E\{h(X, \mathbf{V}, R; \boldsymbol{\eta}_0, \lambda, F_0, F_1, f_0, f_1)\}$$

for some h function. For example, for $\mathbf{E}_1$, the corresponding h function is given as

$$h(X, \mathbf{V}, R; \boldsymbol{\eta}_0, \lambda, F_0, F_1, f_0, f_1) = \frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{F_0(X) - AUC\}.$$

We can estimate $(\boldsymbol{\eta}_0, F_0, F_1)$ using $(\hat{\boldsymbol{\eta}}, \hat{F}_0, \hat{F}_1)$. From (3.6), we note that

$$\lambda = P(Y = 1) = E\{g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)\}.$$

Thus, an estimator of λ is given by

$$\hat{\lambda} = n^{-1} \sum_{i=1}^n \hat{g}_i.$$

With $\hat{F}_0$ and $\hat{F}_1$, their corresponding densities f_0 and f_1 can be estimated by

$$\hat{f}_0(x) = \int K_{h_0}(u - x) d\hat{F}_0(u), \quad \hat{f}_1(x) = \int K_{h_1}(u - x) d\hat{F}_1(u),$$

where $K_h(x) = K(x/h)/h$ with $K(\cdot)$ being a kernel function. Here, the bandwidths h_0 and h_1 are chosen as

$$h_0 = 1.06n^{-1/5} \min\left(\hat{\sigma}_0, \hat{Q}_0/1.34\right) \quad \text{and} \quad h_1 = 1.06n^{-1/5} \min\left(\hat{\sigma}_1, \hat{Q}_1/1.34\right),$$

where $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are the variance estimators for the biomarker from the healthy and disease group; $\hat{Q}_0$ and $\hat{Q}_1$ are interquartile range estimators for the biomarker from the healthy

and disease groups.

With $(\hat{\boldsymbol{\eta}}_0, \hat{\lambda}, \hat{F}_0, \hat{F}_1, \hat{f}_0, \hat{f}_1)$, we can estimate $E\{h(X, \mathbf{V}, R; \boldsymbol{\eta}_0, \lambda, F_0, F_1, f_0, f_1)\}$ by

$$n^{-1} \sum_{i=1}^n \{h(X_i, \mathbf{V}_i, R_i; \hat{\boldsymbol{\eta}}_0, \hat{\lambda}, \hat{F}_0, \hat{F}_1, \hat{f}_0, \hat{f}_1)\}.$$

For example, we can estimate $\mathbf{E}_1$ as

$$\hat{\mathbf{E}}_1 = \frac{1}{n} \sum_{i=1}^n \frac{\partial g(X_i, \mathbf{V}_i, R_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \{\hat{F}_0(X) - \widehat{AUC}\}.$$

Other terms in σ_{AUC}^2 and σ_s^2 can be similarly estimated, which leads to the plug-in variance estimators $\hat{\sigma}_{AUC}^2$ for σ_{AUC}^2 and $\hat{\sigma}_s^2$ for σ_s^2 in Theorem 3.2. Consequently, the $100(1 - \alpha)\%$ Wald-type confidence interval (CI) for AUC is given by

$$[\widehat{AUC} - z_{1-\alpha/2} \hat{\sigma}_{AUC} / \sqrt{n}, \quad \widehat{AUC} + z_{1-\alpha/2} \hat{\sigma}_{AUC} / \sqrt{n}], \quad (3.7)$$

and likewise, that for $ROC(s)$ is given by

$$[\widehat{ROC}(s) - z_{1-\alpha/2} \hat{\sigma}_s / \sqrt{n}, \quad \widehat{ROC}(s) + z_{1-\alpha/2} \hat{\sigma}_s / \sqrt{n}], \quad (3.8)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ percentile of the standard normal distribution.

3.4 Simulation studies

In this section, we compare the performance of our method with existing methods in the AUC and ROC curve estimation through simulation examples. We also evaluate the accuracy of the Wald-type CIs in (3.7) and (3.8) at the nominal level $1 - \alpha = 0.95$.

3.4.1 Simulation setup

Our method (denoted as “Our”) is compared with four competitive methods:

- the IPW method in [Yu et al. \(2018\)](#), denoted as “IPW”;
- the method that assumes an ignorable mechanism in the verification model, i.e., $\beta = 0$ in (3.2), denoted as “IG”;
- the method based only on the verified data (i.e. $R = 1$), denoted as “VER”;
- the full data method, which assumes all the data are available, denoted as “Full”. This method serves as a benchmark.

We do not include the method of [Liu & Zhou \(2010\)](#) as a competitive method as their model assumptions cannot be verified using the observed data. For the data simulation, the biomarker X is generated from $\text{Uniform}[-1, 1]$. The covariates are simulated as two-dimensional: $\mathbf{V} = (V_1, V_2)^T$, where $V_1 \sim N(0, 1)$ and $V_2 \sim \text{Bernoulli}(0.5)$. Here, X , V_1 , and V_2 are simulated as independent. To simulate Y and R , we consider the following three scenarios.

Scenario 1:

$$\begin{aligned} P(Y = 1|X = x, \mathbf{V} = \mathbf{v}, R = 1) &= \frac{1}{1 + \exp(1.7 - 2.5x - 1.5v_1 - 1.5v_2)}, \\ P(R = 1|X = x, \mathbf{V} = \mathbf{v}, Y = y) &= \frac{1}{1 + \exp(1.3 - 1.5x - 1.2v_1 + v_2)}. \end{aligned}$$

Scenario 2:

$$P(Y = 1|X = x, \mathbf{V} = \mathbf{v}, R = 1) = \frac{1}{1 + \exp(1.7 - 2.5x - 1.5v_1 - 1.5v_2)},$$

$$P(R = 1|X = x, \mathbf{V} = \mathbf{v}, Y = y) = \frac{1}{1 + \exp(1.3 - 1.5x - 1.2v_1 + v_2 - 2y)}.$$

Scenario 3:

$$P(Y = 1|X = x, \mathbf{V} = \mathbf{v}, R = 1) = \frac{1}{1 + \exp(1.7 - 2.5x - 1.5v_1 + 0.5v_1^2 - 1.5v_2)},$$

$$P(R = 1|X = x, \mathbf{V} = \mathbf{v}, Y = y) = \frac{1}{1 + \exp(1.3 - 1.5x - 1.2v_1 + 0.5v_1^2 + v_2 - 2y)}.$$

Scenario 1 corresponds to the case where the missing mechanism is MAR; Scenario 2 corresponds to the case of a non-ignorable missing mechanism. In both scenarios, the models of our method and the IPW method are correctly specified. In Scenario 3, a second order term v_1^2 is added to both the verification and disease models, resulting in model misspecification for both our method and the IPW method.

Summary quantities for these scenarios are displayed in Table 3.1. For each scenario, we consider sample sizes of $n = 5000$ and $n = 10000$ to resemble the sample size of the real data in Section 3.5. For each simulation setting, we repeat 1000 times.

Table 3.1: Summary quantities for Scenarios 1–3

	Scenario 1	Scenario 2	Scenario 3
$P(Y = 1)$	0.370	0.245	0.191
$P(R = 1)$	0.219	0.307	0.242
AUC	0.751	0.776	0.813
$ROC(0.1)$	0.348	0.370	0.419
$ROC(0.2)$	0.549	0.585	0.654

3.4.2 AUC point estimation results

In this section, we compare our method with the competitive methods in the AUC estimation. The comparison is based on the same criteria used in Section 2.3.3: relative bias (RB) and mean squared error (MSE). The results are summarized in Table 3.2.

Table 3.2: RB (%) and MSE ($\times 1000$) of AUC estimates under Scenarios 1–3

n	Method	Scenario 1		Scenario 2		Scenario 3	
		RB	MSE	RB	MSE	RB	MSE
5000	Our	-0.131	0.241	-0.162	0.107	-0.314	0.200
5000	IPW	-0.648	0.869	-0.324	0.328	0.582	0.280
5000	IG	-0.007	0.189	-3.254	0.777	-4.581	1.588
5000	VER	-5.618	2.055	-10.819	7.255	-9.845	6.615
5000	Full	-0.006	0.048	-0.079	0.056	-0.026	0.047
10000	Our	-0.002	0.114	-0.083	0.043	-0.190	0.096
10000	IPW	-0.046	0.358	-0.063	0.132	0.833	0.157
10000	IG	0.003	0.090	-3.237	0.697	-4.524	1.454
10000	VER	-5.579	1.883	-10.776	7.087	-9.809	6.464
10000	Full	0.000	0.025	-0.050	0.025	-0.003	0.024

From the table, we observe that the Full method results in the smallest MSEs across all simulation setups, as expected. Therefore, we exclude the Full method from further discussion. The VER method exhibits the largest absolute RBs compared to the other methods in all scenarios. When compared to the IPW method, our method consistently achieves much smaller MSEs in every scenario. In Scenario 1, where the missingness mechanism for disease status is MAR, the IG method yields the smallest MSE. In this case, our proposed method performs similarly to the IG method, with no significant difference. However, in the other scenarios, our proposed method shows a clear advantage over the IG method in terms of MSE.

3.4.3 AUC CIs estimation results

In this section, we evaluate the performance of the CI estimation for the proposed method at the nominal level of 0.95. Two measures are used to assess the effectiveness of the CI estimation: coverage probability (CP) and average length (AL). Let $\widehat{CI}^{(i)}$ denote the i -th estimated CI for the AUC, where $i = 1, \dots, B$. The CP and AL are defined to be:

$$CP = \frac{1}{B} \sum_{i=1}^B I(a_0 \in \widehat{CI}^{(i)}) \quad \text{and} \quad AL = \frac{1}{B} \sum_{i=1}^B |\widehat{CI}^{(i)}|,$$

where a_0 is the true value of the AUC. We also compute the CPs and ALs of CIs based on the IPW method for comparison. The results are shown in Table 3.3.

Table 3.3: CP and AL of estimated CIs for AUC by IPW and Our methods under Scenarios 1–3

n	Method	Scenario 1		Scenario 2		Scenario 3	
		CP	AL	CP	AL	CP	AL
5000	Our	0.940	0.058	0.950	0.037	0.937	0.050
5000	IPW	0.953	0.111	0.974	0.085	0.924	0.071
10000	Our	0.951	0.041	0.956	0.026	0.936	0.035
10000	IPW	0.944	0.075	0.983	0.055	0.873	0.044

We first focus on the results for the CP. In cases where the models are correctly specified (Scenarios 1 and 2), the CPs for our proposed method are reasonably close to the nominal level. In Scenario 1, the IPW method performs similarly to our proposed method. However, in Scenario 2, the IPW method exhibits an overcoverage problem. In the case of model misspecification (Scenario 3), both methods show an undercoverage problem. The CPs for our proposed method are not significantly smaller than the nominal level, whereas the undercoverage problem is more pronounced for the IPW method. Regarding the AL, our proposed method consistently yields a significantly smaller AL across all scenarios

compared to the IPW method.

3.4.4 ROC curve point estimation results

In this section, we assess the accuracy of the proposed method and compare it with competitive approaches for the ROC curve estimation. Since Yu et al. (2018) does not discuss the estimation of the ROC curve, their method is not included in the comparison.

Figure 3.1 illustrates the accuracy of the proposed ROC curve estimation method. The true ROC curve, $ROC(s)$, is shown alongside the average estimated curve and its 95% confidence band (CB; 2.5% and 97.5% quantiles), based on 1000 replications. The results demonstrate strong agreement between the estimated average curve and the true curve, with the CB consistently covering the true ROC curve.

Next, we compare the estimates of the ROC curve at $s = 0.1$ and $s = 0.2$, corresponding to specificities of 0.9 and 0.8, respectively. Similar to Section 3.4.2, the criteria for comparison are RB and MSE. The simulation results for $ROC(0.1)$ and $ROC(0.2)$ are summarized in Table 3.4.

Similar to the AUC estimation discussed in Section 3.4.2, for both $ROC(0.1)$ and $ROC(0.2)$, the Full method yields the smallest MSE, while the VER method exhibits the largest MSE among all methods across all scenarios.

When the missingness mechanism for disease status is MAR (Scenario 1), the IG method and our proposed method show similar MSEs. However, in other scenarios, our proposed method demonstrates a smaller MSE, with the difference being significant.

Table 3.4: RB (%) and MSE ($\times 1000$) of $ROC(s)$ estimates for $s = 0.1$ and 0.2 under Scenarios 1–3

s	n	Method	Scenario 1		Scenario 2		Scenario 3	
			RB	MSE	RB	MSE	RB	MSE
0.1	5000	Our	-0.165	0.593	-0.278	0.372	3.103	0.824
0.1	5000	IG	0.229	0.569	-5.863	0.895	-7.027	1.553
0.1	5000	VER	-23.707	8.148	-34.804	17.517	-33.923	21.479
0.1	5000	Full	0.156	0.310	-0.294	0.372	-0.102	0.444
0.1	10000	Our	0.125	0.276	-0.163	0.160	3.415	0.528
0.1	10000	IG	0.171	0.263	-5.877	0.677	-6.895	1.176
0.1	10000	VER	-23.801	7.523	-34.549	16.781	-33.847	20.737
0.1	10000	Full	0.134	0.158	-0.263	0.182	-0.010	0.216
0.2	5000	Our	-0.157	0.749	-0.230	0.423	0.206	0.778
0.2	5000	IG	0.122	0.664	-6.113	1.768	-9.028	4.222
0.2	5000	VER	-16.586	9.767	-27.109	26.192	-26.408	31.139
0.2	5000	Full	0.182	0.259	-0.015	0.335	0.050	0.358
0.2	10000	Our	0.112	0.354	-0.031	0.178	0.474	0.393
0.2	10000	IG	0.143	0.315	-6.079	1.501	-8.896	3.758
0.2	10000	VER	-16.470	8.899	-27.151	25.743	-26.548	30.813
0.2	10000	Full	0.113	0.127	0.004	0.149	0.043	0.170

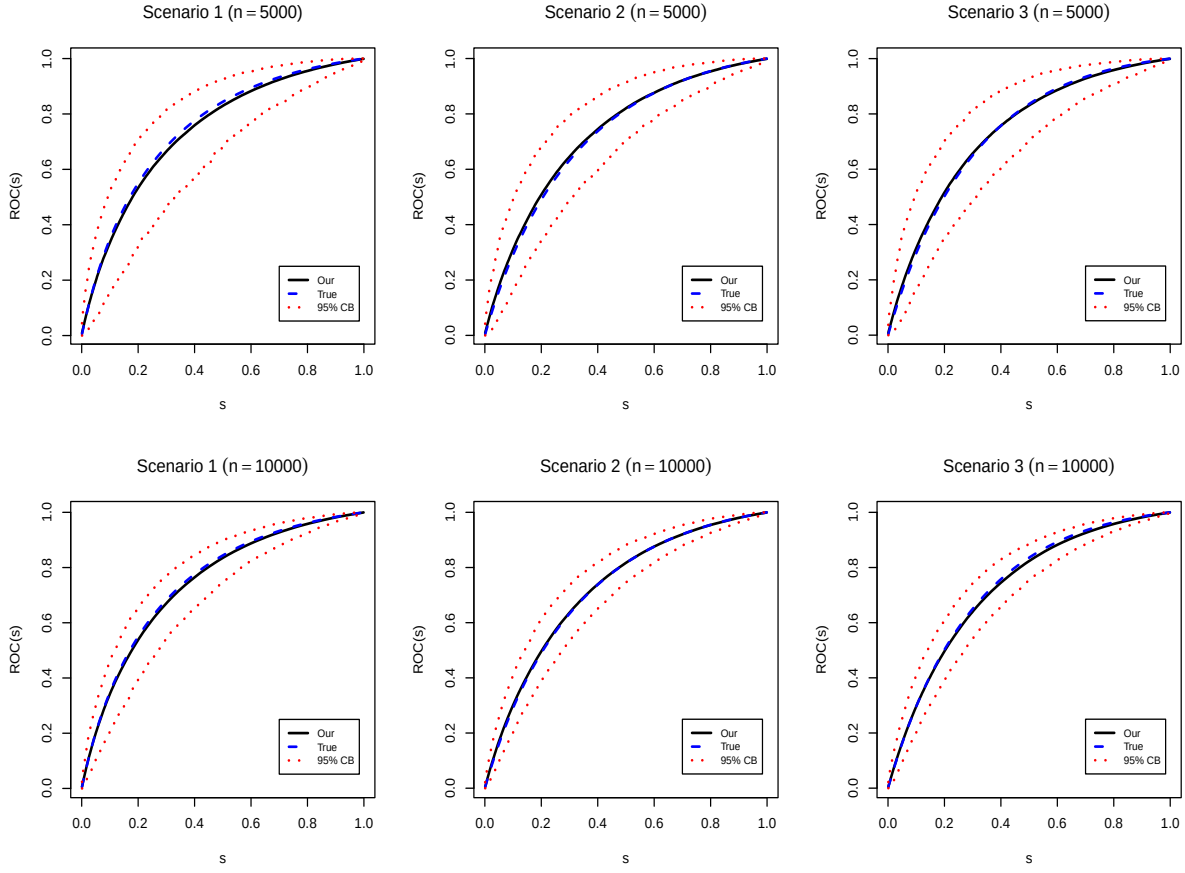


Figure 3.1: The true ROC curve, the average estimated curve, and its 95% confidence band based on 1000 replications under Scenarios 1–3.

3.4.5 ROC curve CIs estimation results

In this section, we assess the performance of the CI estimation of the ROC curve using the proposed method at the nominal level of 0.95. Specifically, we estimate the CIs for $ROC(0.1)$ and $ROC'(0.2)$. Similar to Section 3.4.3, we quantitatively evaluate the effectiveness of the CI estimation using CP and AL. The results are presented in Table 3.5.

Table 3.5: CP and AL of estimated CIs for $ROC(s)$ with $s = 0.1$ and $s = 0.2$ under Scenarios 1–3

s	n	Scenario 1		Scenario 2		Scenario 3	
		CP	AL	CP	AL	CP	AL
0.1	5000	0.932	0.092	0.939	0.074	0.908	0.097
0.1	10000	0.951	0.065	0.963	0.052	0.860	0.068
0.2	5000	0.933	0.103	0.940	0.0777	0.932	0.102
0.2	10000	0.956	0.073	0.951	0.054	0.928	0.071

Under Scenarios 1 and 2, the CPs for our proposed method are reasonably close to the nominal level for both $s = 0.1$ and 0.2 . When the models are misspecified (Scenario 3), the CPs fall below the nominal level for both $s = 0.1$ and 0.2 . In this scenario, the undercoverage problem is more pronounced for $s = 0.1$.

Additional simulations are conducted with sample sizes of $n = 200, 500$, and 1000 . We further evaluate the performance of the CIs estimation of the AUC and $ROC(s)$, with σ_{AUC}^2 and σ_s^2 estimated using the bootstrap method. To save space, these results are presented in Section 3.6.5.

3.5 Real data application

In this section, we apply our proposed method to data collected by the National Alzheimer’s Coordinating Center (NACC). The dataset analyzed is the Uniform Data Set (UDS), which includes all visits for all participants as of the December 2023 data freeze, comprising 185,831 patients. Our focus here is on evaluating the diagnostic ability of the Mini Mental State Examination (MMSE) in detecting Alzheimer’s disease (AD). The MMSE, originally created by [Folstein et al. \(1975\)](#), has since been shown to be effective in quantitatively assessing the severity of cognitive impairment ([Tombaugh & McIntyre 1992](#)). It consists

of a series of questions, with scores ranging from 0 to 30, where lower scores indicate greater cognitive impairment. For this analysis, we use the NIA-AA Alzheimer’s disease neuropathologic change (ADNC) (ABC score) as the gold standard. Since the ABC score requires a brain autopsy, some patients or their families may decline the procedure, leading to missing disease status data. The test results are categorized as “Not AD”, “Low ADNC”, “Intermediate ADNC”, and “High ADNC”. For our study, we define “Not AD” as the healthy group and group the remaining categories under “AD”.

Since missing disease status is strongly correlated with mortality, and individuals who are alive are generally assumed to be in better health, the verification of disease status is likely to be nonignorable. Following [Liu & Zhou \(2010\)](#), we consider age, sex, race, marital status, stroke, Parkinson’s disease, and depression as relevant covariates. For illustration, we focus on patients aged 65 to 70. Patients with missing MMSE results or covariates are excluded from the analysis, resulting in a total of 15,564 patients. Table 3.6 provides a summary of the MMSE scores and relevant covariates for this subgroup.

Table 3.6: Summary statistics for NACC data on patients aged 65 to 70

Variable	Summary statistics
Y	AD: 9.0%; AD-free: 1.2%; missing: 89.8%
MMSE	Mean: 26.3; SD: 5.8; Median: 29
Age	Mean: 67.7; SD: 1.7; Median: 68
Sex	Male: 39.6%; Female: 60.4%
Race	White: 82.5%; other: 17.5%
Marital status	Married: 69.4%; other: 30.6%
Stroke	Yes: 2.0%; No: 98.0%
Parkinson’s disease	Yes: 2.6%; No: 97.4%
Depression	Yes: 19.4%; No: 80.6%

In our data analysis, we define the biomarker as $X = (30 - MMSE)/15$ and transform age to $(age - 65)/5$ for ease of coefficient reporting. Additionally, we denote the relevant

covariates by $\mathbf{V}$. We consider the following disease and verification models:

$$P(Y = 1|X = x, \mathbf{V} = \mathbf{v}, R = 1) = \frac{1}{1 + \exp(\mu_1 + \mu_2 x + \boldsymbol{\mu}_3^T \mathbf{v})}, \quad (3.9)$$

and

$$P(R = 1|X = x, \mathbf{V} = \mathbf{v}, Y = y) = \frac{1}{1 + \exp(\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + \beta y)}. \quad (3.10)$$

The two-step procedure outlined in Section 3.2.5 yields a p-value of 0.557 for the goodness of fit of model (3.9) and a p-value of 0.2 for the goodness of fit of model (3.10). These results indicate that models (3.9) and (3.10) collectively provide a reasonable fit to the data.

Next, we apply our proposed method using models (3.9) and (3.10). The estimated model parameters and their standard errors are summarized in Table 3.7. In the verification model, we find that the non-ignorability parameter β is highly significant, confirming that the missing data mechanism is non-ignorable. Other significant covariates include age, sex, race, and marital status. In the disease model, MMSE, age, and depression are significant at the 5% level. Since our primary focus is on estimating the AUC and ROC curve, and given the relatively small estimated coefficients for the other covariates, we retain the full model without conducting further variable selection in the subsequent analysis.

In terms of AUC estimation, we apply the methods discussed in Section 3.4.1 to the NACC data. We estimate the CI of the AUC using both the IPW method and our proposed method. The AUC and its CI estimates are presented in Table 3.8. Our proposed method yields the largest AUC with a shorter CI estimate compared to the IPW method. Figure 3.2 displays the estimated ROC curve using the IG, VER, and our proposed method. For

Table 3.7: Estimated model parameters and their standard errors (Significant coefficients at the 5% level are bolded)

Variable	Verification model	Disease model
Intercept	5.379 (0.219)	-1.216 (0.277)
(30-MMSE)/15	-0.237 (0.391)	-2.692 (0.139)
(Age-65)/5	0.303 (0.153)	-0.567 (0.196)
Sex	-0.384 (0.103)	-0.266 (0.150)
Race	-1.239 (0.178)	0.248 (0.255)
Marital status	-0.276 (0.132)	0.035 (0.197)
Stroke	-0.070 (0.305)	-0.194 (0.492)
Parkinson's disease	-0.002 (0.246)	-0.166 (0.355)
Depression	0.070 (0.117)	-0.363 (0.168)
Y (gold standard)	-3.099 (0.794)	

our proposed method, we also present the 95% CB. Our proposed method demonstrates a higher sensitivity value at each specificity level compared to the IG and VER methods.

Table 3.8: AUC and 95% CI estimates for the NACC data

Methods	AUC	CI	Length
Our	0.730	[0.704, 0.755]	0.051
IPW	0.697	[-1.012, 2.407]	3.420
IG	0.539	—	—
VER	0.614	—	—

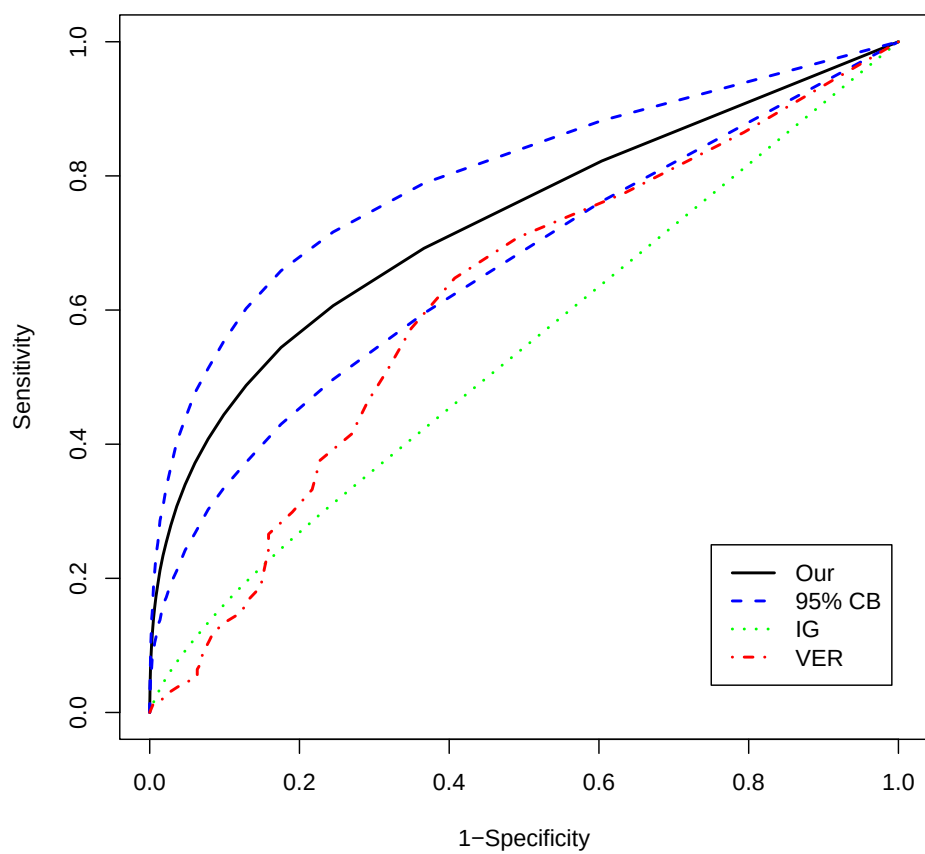


Figure 3.2: Estimated ROC curves and 95% confidence band.

3.6 Technical details and additional simulation results

3.6.1 Regularity conditions

In this section, we list the regularity conditions required in Sections 3.2.2 and 3.3. Let $\mathbf{Z} = (X, \mathbf{V}^T, Y, R)^T$, $\mathbf{z} = (x, \mathbf{v}^T, y, r)^T$, and

$$\begin{aligned} l(\mathbf{z}; \boldsymbol{\eta}) = & r [y \log \{p_1(x, \mathbf{v}; \boldsymbol{\mu})\} + (1 - y) \log \{1 - p_1(x, \mathbf{v}; \boldsymbol{\mu})\}] \\ & + [r \log \{\pi(x, \mathbf{v}; \boldsymbol{\mu}, \boldsymbol{\phi})\} + (1 - r) \log \{1 - \pi(x, \mathbf{v}; \boldsymbol{\mu}, \boldsymbol{\phi})\}]. \end{aligned}$$

The results in Theorems 3.1 and 3.2 rely on the following regularity conditions.

- C1. The biomarker X is continuous, and the coefficient μ_2 in model (3.1) is non-zero. Further, the components of $(1, X, \mathbf{V})$ are stochastically linearly independent. That is, the equation

$$a_1 + a_2 X + \mathbf{a}_3^T \mathbf{V} = 0$$

has only the trivial solution $a_1 = 0$, $a_2 = 0$, $\mathbf{a}_3 = \mathbf{0}$ with probability 1.

- C2. The parameter space Θ of $\boldsymbol{\eta}$ is compact and the true value $\boldsymbol{\eta}_0$ of $\boldsymbol{\eta}$ is the interior point of the parameter space Θ .

- C3. (i) $E(|\log l(\mathbf{Z}; \boldsymbol{\eta}_0)|) < \infty$; (ii) for sufficiently small ρ and for sufficient large t , $E[\log(1 + \log l(\mathbf{Z}; \boldsymbol{\eta}, \rho))] < \infty$ for $\boldsymbol{\eta} \in \Theta$ and $E[\log(1 + \varphi(\mathbf{Z}; t))] < \infty$, where

$$l(\mathbf{Z}; \boldsymbol{\eta}, \rho) = \sup_{\|\boldsymbol{\eta}' - \boldsymbol{\eta}\| < \rho} l(\mathbf{Z}; \boldsymbol{\eta}') \quad \text{and} \quad \varphi(\mathbf{Z}; t) = \sup_{\|\boldsymbol{\eta}\| \geq t} l(\mathbf{Z}; \boldsymbol{\eta});$$

- (iii) $l(\mathbf{Z}; \boldsymbol{\eta}) \rightarrow 0$ in probability as $\|\boldsymbol{\eta}\| \rightarrow \infty$. Here $\|\boldsymbol{\eta}\|$ is the L_2 -norm of $\boldsymbol{\eta}$.

C4. The Fisher information matrix $\mathbf{J}$ exists and is positive definite.

C5. Let η_1 , η_2 , and η_3 be any three elements of $\boldsymbol{\eta}$. There exists a positive ϵ and a function H such that

$$\sup_{\|\boldsymbol{\eta} - \boldsymbol{\eta}_0\| \leq \epsilon} \left| \frac{\partial^3 l^3(\mathbf{Z}; \boldsymbol{\eta})}{\partial \eta_1 \partial \eta_2 \partial \eta_3} \right| \leq H(\mathbf{Z}),$$

where $E\{H(\mathbf{Z})\} < \infty$.

C6. The first and second moments of the biomarker X and covariates $\mathbf{V}$ exist and are finite.

Condition C1 requires that the components of $(1, X, \mathbf{V})$ are stochastically linearly independent. This does not imply that X and $\mathbf{V}$ should be independent. When Condition C6 is also satisfied, i.e., the second moments of X and $\mathbf{V}$ exist and are finite, Condition C1 simply requires that the variance-covariance matrix of $(X, \mathbf{V})$ is positive definite.

3.6.2 Proof of Theorem 3.1

Recall that we consider the following disease model

$$P(Y = 1|x, \mathbf{v}, R = 1) = \frac{1}{1 + \exp(\mu_1 + \mu_2 x + \boldsymbol{\mu}_3^T \mathbf{v})} \quad (3.11)$$

and the verification model

$$P(R = 1|y, x, \mathbf{v}) = \frac{1}{1 + \exp(\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + \beta y)}. \quad (3.12)$$

We need to show that models (3.11) and (3.12) uniquely determine the parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \boldsymbol{\mu}_3^T)^T$ and $\boldsymbol{\phi} = (\psi_1, \psi_2, \boldsymbol{\psi}_3^T, \beta)^T$.

When the components of $(1, X, \mathbf{V})$ are stochastically linearly independent, as assumed in Condition C1, $\boldsymbol{\mu}$ is identifiable in model (3.11).

Next, we argue that $\boldsymbol{\phi}$ is identifiable. Recall that identifying $\boldsymbol{\phi}$ is equivalent to identifying it in

$$\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + c(x, \mathbf{v}; \boldsymbol{\mu}, \beta),$$

where

$$c(x, \mathbf{v}; \boldsymbol{\mu}, \beta) = \log E(e^{\beta Y} | x, \mathbf{v}, R = 1) = \log \{ \exp(\beta) P_1(x, \mathbf{v}; \boldsymbol{\mu}) + 1 - P_1(x, \mathbf{v}; \boldsymbol{\mu}) \} \quad (3.13)$$

with $P_1(x, \mathbf{v}; \boldsymbol{\mu}) = P(Y = 1 | x, \mathbf{v}, R = 1)$.

Suppose that two sets of parameters, $(\psi_1, \psi_2, \boldsymbol{\psi}_3, \beta)$ and $(\psi_1^*, \psi_2^*, \boldsymbol{\psi}_3^*, \beta^*)$, satisfy

$$\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} + c(x, \mathbf{v}; \boldsymbol{\mu}, \beta) = \psi_1^* + \psi_2^* x + \boldsymbol{\psi}_3^{*T} \mathbf{v} + c(x, \mathbf{v}; \boldsymbol{\mu}, \beta^*) \quad (3.14)$$

for all $(x, \mathbf{v})$. It suffices to show that $(\psi_1, \psi_2, \boldsymbol{\psi}_3, \beta) = (\psi_1^*, \psi_2^*, \boldsymbol{\psi}_3^*, \beta^*)$.

We start by arguing that $\beta = \beta^*$. Condition C1 assumes that X is a continuous biomarker and $\mu_2 \neq 0$. Without loss of generality, we assume that $\mathbf{0}$ lies within the support of $\mathbf{V}$.

Taking the first derivative of both sides of (3.14) and setting $\mathbf{v} = \mathbf{0}$, we obtain, after some simplification:

$$\psi_2 + \frac{\mu_2 \exp(\mu_1 + \mu_2 x)}{\exp(\beta) + \exp(\mu_1 + \mu_2 x)} = \psi_2^* + \frac{\mu_2 \exp(\mu_1 + \mu_2 x)}{\exp(\beta^*) + \exp(\mu_1 + \mu_2 x)}. \quad (3.15)$$

Taking another derivative of both sides of (3.15), we further obtain

$$\frac{\mu_2^2 \exp(\beta) \exp(\mu_1 + \mu_2 x)}{\{\exp(\beta) + \exp(\mu_1 + \mu_2 x)\}^2} = \frac{\mu_2^2 \exp(\beta^*) \exp(\mu_1 + \mu_2 x)}{\{\exp(\beta^*) + \exp(\mu_1 + \mu_2 x)\}^2}. \quad (3.16)$$

Since $\mu_2 \neq 0$, after some algebraic manipulation, we get

$$\exp(\beta) - \exp\left(\frac{\beta + \beta^*}{2}\right) + \left\{1 - \exp\left(\frac{\beta - \beta^*}{2}\right)\right\} \exp(\mu_1 + \mu_2 x) = 0$$

for all x . This implies that $\beta = \beta^*$.

Substituting $\beta = \beta^*$ into (3.14) gives

$$\psi_1 + \psi_2 x + \boldsymbol{\psi}_3^T \mathbf{v} = \psi_1^* + \psi_2^* x + \boldsymbol{\psi}_3^{*T} \mathbf{v}. \quad (3.17)$$

When the components of $(1, X, \mathbf{V})$ are stochastically linearly independent, as assumed in Condition C1, (3.17) implies that

$$(\psi_1, \psi_2, \boldsymbol{\psi}_3) = (\psi_1^*, \psi_2^*, \boldsymbol{\psi}_3^*).$$

This completes the proof of Theorem 3.1.

3.6.3 Proof of Proposition 3.1

Recall that

$$\begin{aligned} g(x, \mathbf{v}, r; \boldsymbol{\mu}, \beta) &= P(Y = 1 | X = x, \mathbf{V} = \mathbf{v}, R = r) \\ &= \frac{1}{1 + \exp\{\mu_1 + \mu_2 x + \boldsymbol{\mu}_3^T \mathbf{v} + (r - 1)\beta\}}. \end{aligned}$$

Throughout this section, we assume that the values of $\boldsymbol{\mu}$ and β are equal to their respective true values $\boldsymbol{\mu}_0$ and β_0 .

By the law of total expectation, we have

$$P(Y = 1) = E\{P(Y = 1|X, \mathbf{V}, R)\} = E\{g(X, \mathbf{V}, R; \boldsymbol{\mu}_0, \beta_0)\} \quad (3.18)$$

and

$$\begin{aligned} E\{I(Y = 1) \cdot I(X \leq x)\} &= E\{P(Y = 1|X, \mathbf{V}, R) \cdot I(X \leq x)\} \\ &= E\{g(X, \mathbf{V}, R; \boldsymbol{\mu}_0, \beta_0) \cdot I(X \leq x)\}. \end{aligned} \quad (3.19)$$

Note that $E\{I(Y = 1) \cdot I(X \leq x)\}$ can be rewritten as

$$\begin{aligned} E\{I(Y = 1) \cdot I(X \leq x)\} &= P(Y = 1, X \leq x) \\ &= P(X \leq x|Y = 1)P(Y = 1) = F_1(x) \cdot P(Y = 1). \end{aligned} \quad (3.20)$$

Combining (3.18)–(3.20), we conclude that

$$E\{g(X, V, R; \boldsymbol{\mu}_0, \beta_0)I(X \leq x)\} = E\{g(X, V, R; \boldsymbol{\mu}_0, \beta_0)\} \cdot F_1(x).$$

Similarly, we have

$$E\{[1 - g(X, V, R; \boldsymbol{\mu}_0, \beta_0)]I(X \leq x)\} = E\{[1 - g(X, V, R; \boldsymbol{\mu}_0, \beta_0)]\} \cdot F_0(x).$$

This completes the proof of Proposition 3.1.

3.6.4 Proof of Theorem 3.2

Some preparation

We first review some notation and present useful results in preparation for the proofs of Theorem 3.2.

Let $\lambda = P(Y = 1)$, $g_i = g(X_i, \mathbf{V}_i, R_i; \boldsymbol{\mu}_0, \beta_0) = g(X_i, \mathbf{V}_i, R_i; \boldsymbol{\theta}_0)$, $\xi_{1-s} = F_0^{-1}(1 - s)$, and

$$\mathbf{Z}_n = (\mathbf{Z}_{n1}^T, Z_{n2}, Z_{n3}, Z_{n4}, Z_{n5})^T$$

with

$$\begin{aligned} \mathbf{Z}_{n1} &= n^{-1} \frac{\partial l_n(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}, \\ Z_{n2} &= n^{-1} \sum_{i=1}^n g_i \{F_0(X_i) - AUC\}, \\ Z_{n3} &= n^{-1} \sum_{i=1}^n (1 - g_i) \{1 - F_1(X_i) - AUC\}, \\ Z_{n4} &= n^{-1} \sum_{i=1}^n (1 - g_i) \{I(X_i \leq \xi_{1-s}) - F_0(\xi_{1-s})\}, \\ Z_{n5} &= n^{-1} \sum_{i=1}^n g_i \{I(X_i \leq \xi_{1-s}) - F_1(\xi_{1-s})\}. \end{aligned}$$

It can be verified that

$$E(\mathbf{Z}_n) = 0.$$

Note that $\mathbf{Z}_n$ are the average of independent and identically distributed (iid) random

vectors. By the central limit theorem, we have

$$\sqrt{n}\mathbf{Z}_n \rightarrow N(0, \boldsymbol{\Sigma}_Z), \quad (3.21)$$

where $\boldsymbol{\Sigma}_Z$ is the variance-covariance matrix of $\sqrt{n}\mathbf{Z}_n$.

Denote $\mathbf{J}$ as the Fisher-information matrix at the true value $\boldsymbol{\eta}_0$. Let $\mathbf{I}_{k_2, k_1}$ be a $k_2 \times k_1$ matrix where the top $k_1 \times k_1$ block is an identity matrix, and all other elements are zero. Similarly, let $\mathbf{I}_{k_1, k_2}$ be a $k_1 \times k_2$ matrix where the leftmost $k_1 \times k_1$ block is an identity matrix, with the remaining elements being zero. It can be verified that $\mathbf{I}_{k_2, k_1} = \mathbf{I}_{k_1, k_2}^T$. Further, define

$$\mathbf{E}_1 = E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{F_0(X) - AUC\} \right], \quad (3.22)$$

$$\mathbf{E}_2 = E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{1 - F_1(X) - AUC\} \right], \quad (3.23)$$

$$\mathbf{E}_3 = E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(X \leq \xi_{1-s}) - F_0(\xi_{1-s})\} \right], \quad (3.24)$$

$$\mathbf{E}_4 = E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(X \leq \xi_{1-s}) - F_1(\xi_{1-s})\} \right]. \quad (3.25)$$

Let $\hat{g}_i = g(X_i, \mathbf{V}_i, R_i; \hat{\boldsymbol{\mu}}, \hat{\beta})$ and $\hat{\lambda} = \sum_{i=1}^n \hat{g}_i / n$. Then,

$$\hat{F}_1(x) = \frac{1}{n\hat{\lambda}} \sum_{i=1}^n \hat{g}_i I(X_i \leq x) \quad (3.26)$$

and

$$\hat{F}_0(x) = \frac{1}{n(1 - \hat{\lambda})} \sum_{i=1}^n (1 - \hat{g}_i) I(X_i \leq x). \quad (3.27)$$

Lemma 3.1. *Suppose Conditions C1–C6 are satisfied. We have*

- (a) $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1} + o_p(n^{-1/2}) = O_p(n^{-1/2})$, where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$;
- (b) $\hat{\lambda} = \lambda + O_p(n^{-1/2})$;
- (c) the processes $\sqrt{n}\{\hat{F}_0(t) - F_0(t)\}$ and $\sqrt{n}\{\hat{F}_1(s) - F_1(s)\}$ converge jointly to a bivariate, tight, mean-zero Gaussian process.

Proof. (a) When the model parameters are identifiable, models (3.11) and (3.12) are regular. The classical expansion gives

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = \mathbf{J}^{-1} \mathbf{Z}_{n1} + o_p(n^{-1/2}).$$

Since $\boldsymbol{\theta}$ comprises the first k_1 elements of $\boldsymbol{\eta}$, the first equation in (a) then follows directly. By (3.21), we further obtain

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_p(n^{-1/2}). \quad (3.28)$$

(b) Let $\mathbf{W}_i = (1, X_i, \mathbf{V}_i^T, (R_i - 1))^T$. By mean value theorem, we have

$$\hat{\lambda} - \lambda = n^{-1} \sum_{i=1}^n (g_i - \lambda) - n^{-1} \sum_{i=1}^n g(X_i, \mathbf{V}_i, R_i; \tilde{\boldsymbol{\theta}}) \{1 - g(X_i, \mathbf{V}_i, R_i; \tilde{\boldsymbol{\theta}})\} \mathbf{W}_i^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.

The central limit theorem implies that

$$n^{-1} \sum_{i=1}^n (g_i - \lambda) = O_p(n^{-1/2}).$$

Note that

$$0 \leq g(X_i, \mathbf{V}_i, R_i; \tilde{\boldsymbol{\theta}}) \{1 - g(X_i, \mathbf{V}_i, R_i; \tilde{\boldsymbol{\theta}})\} \leq \frac{1}{4}.$$

Using Condition C6, the law of large numbers, and (3.28), we obtain

$$n^{-1} \sum_{i=1}^n g(X_i, \mathbf{V}_i, R_i; \tilde{\boldsymbol{\theta}}) \{1 - g(X_i, \mathbf{V}_i, R_i; \tilde{\boldsymbol{\theta}})\} \mathbf{W}_i^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = O_p(n^{-1/2}). \quad (3.29)$$

Thus, we conclude that $\hat{\lambda} - \lambda = O_p(n^{-1/2})$ as claimed.

(c) Note that

$$\hat{F}_0(t) - F_0(t) = \frac{D_{n0}(t)}{1 - \hat{\lambda}}, \quad (3.30)$$

where

$$D_{n0}(t) = n^{-1} \sum_{i=1}^n (1 - \hat{g}_i) \{I(X_i \leq t) - F_0(t)\}.$$

Next, we find an approximation of $D_{n0}(t)$. By applying the first-order Taylor expansion, we have

$$\begin{aligned} D_{n0}(t) &= n^{-1} \sum_{i=1}^n (1 - g_i) \{I(X_i \leq t) - F_0(t)\} \\ &\quad - n^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \sum_{i=1}^n \frac{\partial g(X_i, \mathbf{V}_i, R_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(X_i \leq t) - F_0(t)\} + \epsilon_{n1}(t) \\ &= n^{-1} \sum_{i=1}^n (1 - g_i) \{I(X_i \leq t) - F_0(t)\} - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{E}_3(t) + \epsilon_n(t), \end{aligned} \quad (3.31)$$

where $\epsilon_{n1}(t)$ is the reminder term after the first-order Taylor expansion, and

$$\epsilon_n(t) = \epsilon_{n1}(t) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \epsilon_{n2}(t)$$

with

$$\epsilon_{n2}(t) = n^{-1} \sum_{i=1}^n \frac{\partial g(X_i, \mathbf{V}_i, R_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(X_i \leq t) - F_0(t)\} - \mathbf{E}_3(t)$$

and

$$\mathbf{E}_3(t) = E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(X \leq t) - F_0(t)\} \right]. \quad (3.32)$$

Similar to (3.29), it can be verified that

$$\sup_t |\epsilon_{n1}(t)| = o_p(n^{-1/2}). \quad (3.33)$$

Let $\mathcal{H}_1$ be a function class indexed by t , defined as:

$$\mathcal{H}_1 = \left\{ \frac{\partial g(x, \mathbf{v}, r; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(x \leq t) - F_0(t)\} : t \in (-\infty, \infty) \right\}$$

By Condition C6, the elements of

$$\frac{\partial g(x, \mathbf{v}, r; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$$

have finite second-moment. By Theorem 2.10.6 of [van der Vaart & Wellner \(1996\)](#), we have $\mathcal{H}_1$ is a Donsker class. Consequently,

$$\sup_t |\epsilon_{n2}(t)| = O_p(n^{-1/2}).$$

Combining this with (3.28) and (3.33), we obtain

$$\sup_t |\epsilon_n(t)| = o_p(n^{-1/2}). \quad (3.34)$$

Plugging (3.34) into (3.31) yields

$$D_{n0}(t) = n^{-1} \sum_{i=1}^n (1 - g_i) \{I(X_i \leq t) - F_0(t)\} - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{E}_3(t) + o_p(n^{-1/2}), \quad (3.35)$$

where the remainder term is uniformly in t .

Note that

$$0 \leq 1 - g(x, \mathbf{v}, r; \boldsymbol{\theta}_0) \leq 1.$$

By Example 2.10.10 of van der Vaart & Wellner (1996), the function class

$$\mathcal{H}_2 = \left\{ \{1 - g(x, \mathbf{v}, r; \boldsymbol{\theta}_0)\} \{I(x \leq t) - F_0(t)\} : t \in (-\infty, \infty) \right\}$$

is Donsker. Hence,

$$\sup_t |n^{-1} \sum_{i=1}^n (1 - g_i) \{I(X_i \leq t) - F_0(t)\}| = O_p(n^{-1/2}). \quad (3.36)$$

Under Conditions C1 and C6, it is straightforward to verify that $\mathbf{E}_3(t)$ is continuous and

$$\sup_t \|\mathbf{E}_3(t)\| < \infty.$$

Consequently,

$$\sup_t |(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{E}_3(t)| = O_p(n^{-1/2}). \quad (3.37)$$

Combining (3.35)–(3.37), we have

$$\sup_t |D_{n0}(t)| = O_p(n^{-1/2}).$$

This, together with, (3.30), (3.35), and the results in part (b), leads to

$$\hat{F}_0(t) - F_0(t) = \frac{1}{1-\lambda} \left[n^{-1} \sum_{i=1}^n (1-g_i) \{I(X_i \leq t) - F_0(t)\} - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{E}_3(t) \right] + o_p(n^{-1/2}).$$

Let $\mathbf{K}(x, \mathbf{v}, y, r; \boldsymbol{\eta}_0)$ be the influence function of $\hat{\boldsymbol{\theta}}$. That is,

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \frac{1}{n} \sum_{i=1}^n \mathbf{K}(X_i, \mathbf{V}_i, Y_i, R_i; \boldsymbol{\eta}_0) + o_p(n^{-1/2}).$$

Then,

$$\begin{aligned} & \hat{F}_0(t) - F_0(t) \\ &= \frac{1}{n(1-\lambda)} \sum_{i=1}^n \left[(1-g_i) \{I(X_i \leq t) - F_0(t)\} - \mathbf{K}(X_i, \mathbf{V}_i, Y_i, R_i; \boldsymbol{\eta}_0)^T \mathbf{E}_3(t) \right] \\ & \quad + o_p(n^{-1/2}). \end{aligned} \tag{3.38}$$

Following the similar procedures to (3.38), we also have

$$\begin{aligned} & \hat{F}_1(s) - F_1(s) \\ &= \frac{1}{n\lambda} \sum_{i=1}^n \left[g_i \{I(X_i \leq s) - F_1(s)\} + \mathbf{K}(X_i, \mathbf{V}_i, Y_i, R_i; \boldsymbol{\eta}_0)^T \mathbf{E}_4(s) \right] + o_p(n^{-1/2}), \end{aligned} \tag{3.39}$$

where

$$\mathbf{E}_4(s) = E \left[\frac{\partial g(X, \mathbf{V}, R; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{I(X \leq s) - F_1(s)\} \right]. \tag{3.40}$$

With Conditions C1 and C6, we can easily verify that $\mathbf{E}_4(s)$ is continuous and

$$\sup_s ||\mathbf{E}_4(s)|| < \infty.$$

Let

$$\begin{aligned} L_0(x, \mathbf{v}, y, r; t) &= \frac{1}{1-\lambda} \{1 - g(x, \mathbf{v}, r; \boldsymbol{\theta}_0)\} \{I(x \leq t) - F_0(t)\} \\ &\quad - \frac{1}{1-\lambda} \mathbf{K}(x, \mathbf{v}, y, r; \boldsymbol{\eta}_0)^T \mathbf{E}_3(t) \end{aligned} \quad (3.41)$$

and

$$\begin{aligned} L_1(x, \mathbf{v}, y, r; s) &= \frac{1}{\lambda} g(x, \mathbf{v}, r; \boldsymbol{\theta}_0) \{I(x \leq s) - F_1(s)\} \\ &\quad + \frac{1}{\lambda} \mathbf{K}(x, \mathbf{v}, y, r; \boldsymbol{\eta}_0)^T \mathbf{E}_4(s). \end{aligned} \quad (3.42)$$

The approximations in (3.38) and (3.39) imply that both $\hat{F}_0(t)$ and $\hat{F}_1(s)$ are asymptotically linear with influence functions being $L_0(x, \mathbf{v}, y, r; t)$ and $L_1(x, \mathbf{v}, y, r; s)$, respectively.

With the above discussion, the class of $L_0(x, \mathbf{v}, y, r; t)$ (indexed by t) and the class of $L_1(x, \mathbf{v}, y, r; s)$ (indexed by s) are both Donsker classes. Hence, the processes $\sqrt{n}\{\hat{F}_0(t) - F_0(t)\}$ and $\sqrt{n}\{\hat{F}_1(s) - F_1(s)\}$ converge jointly to a bivariate, right, mean-zero, Gaussian process. $\square$

Proof of Part (a) of Theorem 3.2

Recall that

$$\widehat{AUC} = \int \hat{F}_0(x) d\hat{F}_1(x).$$

By Part (c) of Lemma 3.1 and the functional delta method (see, for example, Theorem 20.8, Lemma 20.10, and Example 20.11 in van der Vaart (2000)), we have

$$\begin{aligned}\widehat{AUC} - AUC &= - \int \{\hat{F}_1(x) - F_1(x)\} dF_0(x) \\ &\quad + \int \{\hat{F}_0(x) - F_0(x)\} dF_1(x) + o_p(n^{-1/2}).\end{aligned}$$

With (3.26)–(3.27) and some manipulation, we further obtain

$$\begin{aligned}\widehat{AUC} - AUC &= \hat{\lambda}^{-1} \left[\frac{1}{n} \sum_{i=1}^n \hat{g}_i \{F_0(X_i) - AUC\} \right] \\ &\quad + \left(1 - \hat{\lambda}\right)^{-1} \left[\frac{1}{n} \sum_{i=1}^n (1 - \hat{g}_i) \{1 - F_1(X_i) - AUC\} \right] + o_p(n^{-1/2}).\end{aligned}\quad (3.43)$$

Using the first-order Taylor expansion and Condition C6, we get that

$$\begin{aligned}&\frac{1}{n} \sum_{i=1}^n \hat{g}_i \{F_0(X_i) - AUC\} \\ &= \frac{1}{n} \sum_{i=1}^n g_i \{F_0(X_i) - AUC\} \\ &\quad + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \frac{1}{n} \sum_{i=1}^n \frac{\partial g(X_i, \mathbf{V}_i, R_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \{F_0(X_i) - AUC\} + o_p(n^{-\frac{1}{2}}) \\ &= Z_{n2} + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{E}_1 + o_p(n^{-\frac{1}{2}}) \\ &= Z_{n2} + \mathbf{E}_1^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1} + o_p(n^{-\frac{1}{2}}),\end{aligned}\quad (3.44)$$

where we use the law of large numbers in the second step, with $\mathbf{E}_1$ defined in (3.22), and

Part (a) of Lemma 3.1. Similarly,

$$\frac{1}{n} \sum_{i=1}^n (1 - \hat{g}_i) \{1 - F_1(X_i) - AUC\} = Z_{n3} - \mathbf{E}_2^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1} + o_p(n^{-\frac{1}{2}}). \quad (3.45)$$

Recall that

$$\mathbf{H}_1 = \begin{pmatrix} \mathbf{J}^{-1} \mathbf{I}_{k_2, k_1} \left(\frac{\mathbf{E}_1}{\lambda} - \frac{\mathbf{E}_2}{1-\lambda} \right) \\ \lambda^{-1} \\ (1-\lambda)^{-1} \\ 0 \\ 0 \end{pmatrix}.$$

Plugging (3.44)–(3.45) into (3.43) leads to

$$\begin{aligned} \widehat{AUC} - AUC &= \frac{Z_{n2} + \mathbf{E}_1^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1}}{\hat{\lambda}} + \frac{Z_{n3} - \mathbf{E}_2^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1}}{1 - \hat{\lambda}} + o_p(n^{-1/2}) \\ &= \frac{Z_{n2} + \mathbf{E}_1^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1}}{\lambda} + \frac{Z_{n3} - \mathbf{E}_2^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1}}{1 - \lambda} + o_p(n^{-1/2}) \\ &= \mathbf{H}_1^T \mathbf{Z}_n + o_p(n^{-1/2}), \end{aligned}$$

where we have used Part (b) of Lemma 3.1, (3.21), and Slutsky's theorem to replace $\hat{\lambda}$ by λ . Applying Slutsky's Theorem, we further have

$$\sqrt{n}(\widehat{AUC} - AUC) \xrightarrow{d} N(0, \sigma_{AUC}^2),$$

with $\sigma_{AUC}^2 = \mathbf{H}_1^T \Sigma_Z \mathbf{H}_1$.

Proof of Part (b) of Theorem 3.2

Proof. For any $s \in (0, 1)$, let $\hat{\xi}_s = \hat{F}_0^{-1}(s)$. Then, we have

$$\begin{aligned}\widehat{ROC}(s) - ROC(s) &= \left\{1 - \hat{F}_1(\hat{\xi}_{1-s})\right\} - \left\{1 - F_1(\xi_{1-s})\right\} \\ &= \left\{F_1(\xi_{1-s}) - F_1(\hat{\xi}_{1-s})\right\} + \left\{F_1(\xi_{1-s}) - \hat{F}_1(\xi_{1-s})\right\} + e_n,\end{aligned}\quad (3.46)$$

where

$$e_n = \left\{F_1(\hat{\xi}_{1-s}) - \hat{F}_1(\hat{\xi}_{1-s})\right\} - \left\{F_1(\xi_{1-s}) - \hat{F}_1(\xi_{1-s})\right\}.$$

We first argue that the term $e_n = o_p(n^{-1/2})$.

By Part (c) of Lemma 3.1, we have

$$\sup_x \left| \hat{F}_1(x) - F_1(x) \right| = O_p(n^{-1/2}),$$

which implies that

$$\hat{\xi}_{1-s} - \xi_{1-s} = O_p(n^{-1/2}).$$

Using similar arguments to Lemma A.2 of [Chen & Liu \(2013\)](#) and Theorem 4 in the supplementary document of [Chen et al. \(2021\)](#), we have

$$|e_n| = \left| \left\{F_1(\hat{\xi}_{1-s}) - \hat{F}_1(\hat{\xi}_{1-s})\right\} - \left\{F_1(\xi_{1-s}) - \hat{F}_1(\xi_{1-s})\right\} \right| = o_p(n^{-1/2}). \quad (3.47)$$

Plugging (3.47) into (3.46), we get

$$\widehat{ROC}(s) - ROC(s) = \left\{F_1(\xi_{1-s}) - F_1(\hat{\xi}_{1-s})\right\} + \left\{F_1(\xi_{1-s}) - \hat{F}_1(\xi_{1-s})\right\} + o_p(n^{-1/2}). \quad (3.48)$$

Following the similar arguments to Theorem 3.1 in [Chen & Liu \(2013\)](#), we have the following Bahadur representation

$$\hat{\xi}_{1-s} - \xi_{1-s} = \frac{1 - s - \hat{F}_0(1 - s)}{f_0(\xi_{1-s})} + o_p(n^{-1/2}).$$

Using the Delta method, we further obtain

$$\begin{aligned} F_1(\xi_{1-s}) - F_1(\hat{\xi}_{1-s}) &= f_1(\xi_{1-s}) \left(\xi_{1-s} - \hat{\xi}_{1-s} \right) + o_p(n^{-1/2}) \\ &= f_1(\xi_{1-s}) \frac{\hat{F}_0(\xi_{1-s}) - (1 - s)}{f_0(\xi_{1-s})} + o_p(n^{-1/2}) \\ &= \frac{f_1(\xi_{1-s})}{f_0(\xi_{1-s})} \left\{ \hat{F}_0(\xi_{1-s}) - 1 + s \right\} + o_p(n^{-1/2}), \end{aligned}$$

which, together with [\(3.48\)](#), results in

$$\begin{aligned} \widehat{ROC}(s) - ROC(s) &= \frac{f_1(\xi_{1-s})}{f_0(\xi_{1-s})} \left\{ \hat{F}_0(\xi_{1-s}) - 1 + s \right\} \\ &\quad + \left\{ F_1(\xi_{1-s}) - \hat{F}_1(\xi_{1-s}) \right\} + o_p(n^{-1/2}). \end{aligned} \quad (3.49)$$

With [\(3.26\)](#)–[\(3.27\)](#), [\(3.49\)](#) implies that

$$\begin{aligned} \widehat{ROC}(s) - ROC(s) &= \frac{f_1(\xi_{1-s})}{(1 - \hat{\lambda})f_0(\xi_{1-s})} \left[\frac{1}{n} \sum_{i=1}^n (1 - \hat{g}_i) \{I(X_i \leq \xi_{1-s}) - F_0(\xi_{1-s})\} \right] \\ &\quad - \frac{1}{\hat{\lambda}} \left[\frac{1}{n} \sum_{i=1}^n \hat{g}_i \{I(X_i \leq \xi_{1-s}) - F_1(\xi_{1-s})\} \right] + o_p(n^{-1/2}). \end{aligned} \quad (3.50)$$

Recall $\mathbf{E}_3$ and $\mathbf{E}_4$ defined in (3.24) and (3.25), respectively. Similar to (3.44), we have

$$\frac{1}{n} \sum_{i=1}^n (1 - \hat{g}_i) \{I(X_i \leq \xi_{1-s}) - F_0(\xi_{1-s})\} = Z_{n4} - \mathbf{E}_3^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1} + o_p(n^{-\frac{1}{2}}) \quad (3.51)$$

and

$$\frac{1}{n} \sum_{i=1}^n \hat{g}_i \{I(X_i \leq \xi_{1-s}) - F_1(\xi_{1-s})\} = Z_{n5} + \mathbf{E}_4^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1} + o_p(n^{-\frac{1}{2}}). \quad (3.52)$$

Plugging (3.51) and (3.52) into (3.50), we have

$$\begin{aligned} \widehat{ROC}(s) - ROC(s) &= \frac{f_1(\xi_{1-s})}{(1 - \hat{\lambda})f_0(\xi_{1-s})} (Z_{n4} - \mathbf{E}_3^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1}) \\ &\quad - \frac{1}{\hat{\lambda}} (Z_{n5} + \mathbf{E}_4^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1}) + o_p(n^{-1/2}). \end{aligned} \quad (3.53)$$

By Part (b) of Lemma 3.1, we further get

$$\begin{aligned} \widehat{ROC}(s) - ROC(s) &= \frac{f_1(\xi_{1-s})}{(1 - \lambda)f_0(\xi_{1-s})} (Z_{n4} - \mathbf{E}_3^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1}) \\ &\quad - \frac{1}{\lambda} (Z_{n5} + \mathbf{E}_4^T \mathbf{I}_{k_1, k_2} \mathbf{J}^{-1} \mathbf{Z}_{n1}) + o_p(n^{-1/2}) \\ &= \mathbf{H}_2^T \mathbf{Z}_n + o_p(n^{-1/2}), \end{aligned}$$

where

$$\mathbf{H}_2 = \begin{pmatrix} \mathbf{J}^{-1} \mathbf{I}_{k_2, k_1} \left\{ \frac{-\mathbf{E}_4}{\lambda} - \frac{\mathbf{E}_3}{1-\lambda} \frac{f_1(\xi_{1-s})}{f_0(\xi_{1-s})} \right\} \\ 0 \\ 0 \\ (1-\lambda)^{-1} \frac{f_1(\xi_{1-s})}{f_0(\xi_{1-s})} \\ -\lambda^{-1} \end{pmatrix}.$$

Hence,

$$\sqrt{n} \left\{ \widehat{ROC}(s) - ROC(s) \right\} \xrightarrow{d} N(0, \sigma_s^2),$$

where $\sigma_s^2 = \mathbf{H}_2^T \Sigma_Z \mathbf{H}_2$. This finishes the proof of Theorem 3.2. $\square$

3.6.5 Additional simulation results

Simulation results for three additional scenarios

We first compare the performance of our methods with existing approaches under three additional scenarios (Scenarios 4–6), considering sample sizes of $n = 200$, $n = 500$, and $n = 1000$. As in Section 3.4, the biomarker X is generated from $\text{Uniform}[-1, 1]$, and the covariates are simulated as two-dimensional: $\mathbf{V} = (V_1, V_2)^T$, where $V_1 \sim N(0, 1)$ and $V_2 \sim \text{Bernoulli}(0.5)$. The variables X , V_1 , and V_2 are independently simulated. The response Y and the missing indicator R for Scenarios 4–6 are generated as follows.

Scenario 4:

$$\begin{aligned} P(Y = 1 | X = x, \mathbf{V} = \mathbf{v}, R = 1) &= \frac{1}{1 + \exp(1.7 - 2.5x - 1.5v_1 - 1.5v_2)}, \\ P(R = 1 | X = x, \mathbf{V} = \mathbf{v}, Y = y) &= \frac{1}{1 + \exp(-1.3 + 1.5x + 1.5v_1 - v_2)}. \end{aligned}$$

Scenario 5:

$$P(Y = 1|X = x, \mathbf{V} = \mathbf{v}, R = 1) = \frac{1}{1 + \exp(1.7 - 3x - 1.5v_1 - 1.5v_2)},$$

$$P(R = 1|X = x, \mathbf{V} = \mathbf{v}, Y = y) = \frac{1}{1 + \exp(-0.5 + 1.5x + 1.5v_1 - v_2 - 3y)}.$$

Scenario 6:

$$P(Y = 1|X = x, \mathbf{V} = \mathbf{v}, R = 1) = \frac{1}{1 + \exp(1.7 - 3x - 1.5v_1 + 0.5v_1^2 - 1.5v_2)},$$

$$P(R = 1|X = x, \mathbf{V} = \mathbf{v}, Y = y) = \frac{1}{1 + \exp(-0.5 + 1.5x + 1.5v_1 - 0.5v_1^2 - v_2 - 3y)}.$$

Scenario 4 represents the case where the missing data mechanism is MAR, while Scenario 5 represents the case of a non-ignorable missing data mechanism. In both scenarios, the models for our method and the IPW method are correctly specified. Scenario 6, on the other hand, corresponds to model misspecification for both our method and the IPW method. Summary quantities for these three scenarios are presented in Table 3.9. For each simulation setting, we performed 1000 replications.

Table 3.9: Summary quantities for Scenarios 4–6

	Scenario 4	Scenario 5	Scenario 6
$P(Y = 1)$	0.370	0.298	0.265
$P(R = 1)$	0.764	0.773	0.811
AUC	0.751	0.769	0.774
$ROC(0.1)$	0.348	0.364	0.360
$ROC(0.2)$	0.549	0.575	0.574

The results for the point estimates of the AUC, the CI estimation of AUC, the point estimates of $ROC(s)$ at $s = 0.1$ and $s = 0.2$, as well as the CI estimation of $ROC(s)$ at

$s = 0.1$ and $s = 0.2$, are provided in Tables 3.10–3.14, respectively. Figure 3.3 displays the true ROC curve $ROC(s)$, the average estimated curve, and its 95% confidence band (CB) based on 1000 replications. The overall trends for the proposed method and its comparison with competitive methods align with those reported in Section 3.4.

Table 3.10: RB (%) and MSE ($\times 1000$) of AUC estimates under Scenarios 4–6

n	Method	Scenario 4		Scenario 5		Scenario 6	
		RB	MSE	RB	MSE	RB	MSE
200	Our	-1.151	2.192	0.252	1.645	0.107	1.810
200	IPW	-2.657	3.348	-0.292	2.146	0.349	2.203
200	IG	-0.141	1.678	2.261	1.704	3.054	1.981
200	VER	-1.880	2.198	3.351	2.148	3.661	2.270
200	Full	-0.105	1.326	-0.211	1.369	0.046	1.450
500	Our	-0.911	0.909	-0.004	0.617	-0.717	0.702
500	IPW	-1.851	1.407	-0.058	0.875	0.012	0.999
500	IG	-0.132	0.642	2.360	0.858	3.031	1.041
500	VER	-1.831	0.964	3.439	1.231	3.612	1.276
500	Full	-0.090	0.505	-0.081	0.485	0.054	0.483
1000	Our	-0.755	0.455	0.015	0.365	-1.181	0.460
1000	IPW	-1.379	0.758	-0.026	0.538	-0.141	0.625
1000	IG	-0.108	0.305	2.485	0.647	3.069	0.818
1000	VER	-1.794	0.550	3.606	1.057	3.678	1.075
1000	Full	-0.104	0.241	0.018	0.273	0.058	0.263

Table 3.11: CP and AL of estimated CIs for AUC by IPW and Our methods under Scenarios 4–6

Method	n	Scenario 4		Scenario 5		Scenario 6	
		CP	AL	CP	AL	CP	AL
Our	200	0.946	0.202	0.935	0.160	0.920	0.157
IPW	200	0.927	0.225	0.945	0.231	0.921	0.221
Our	500	0.959	0.119	0.949	0.102	0.937	0.100
IPW	500	0.952	0.161	0.968	0.153	0.938	0.163
Our	1000	0.954	0.083	0.938	0.072	0.907	0.071
IPW	1000	0.968	0.114	0.965	0.115	0.962	0.129

Table 3.12: RB (%) and MSE ($\times 1000$) of $ROC(s)$ estimates for $s = 0.1$ under Scenarios 4–6

s	n	Method	Scenario 4		Scenario 5		Scenario 6	
			RB	MSE	RB	MSE	RB	MSE
0.1	200	Our	-4.035	7.927	3.384	7.898	4.446	9.340
0.1	200	IG	0.046	6.480	13.055	9.497	18.068	12.021
0.1	200	VER	-4.680	10.671	17.615	15.370	17.794	16.699
0.1	200	Full	0.571	7.696	-0.616	8.318	0.235	9.329
0.1	500	Our	-2.640	3.318	1.509	2.947	0.318	3.120
0.1	500	IG	-0.025	2.338	13.894	5.167	18.363	6.999
0.1	500	VER	-4.438	4.189	18.680	8.694	18.468	8.798
0.1	500	Full	0.900	3.000	0.725	3.228	0.715	3.397
0.1	1000	Our	-2.440	1.664	0.651	1.632	-2.419	1.758
0.1	1000	IG	-0.094	1.126	14.660	4.186	18.586	5.798
0.1	1000	VER	-4.527	2.105	18.899	6.883	18.152	6.451
0.1	1000	Full	0.043	1.417	0.338	1.805	-0.156	1.637

Table 3.13: RB (%) and MSE ($\times 1000$) of $ROC(s)$ estimates for $s = 0.2$ under Scenarios 4–6

s	n	Method	Scenario 4		Scenario 5		Scenario 6	
			RB	MSE	RB	MSE	RB	MSE
0.2	200	Our	-2.800	8.336	1.275	7.261	1.793	8.425
0.2	200	IG	-0.132	6.483	7.145	7.809	9.871	9.744
0.2	200	VER	-4.244	9.953	10.086	11.953	10.689	12.530
0.2	200	Full	-0.018	6.698	-0.492	7.726	-0.289	8.575
0.2	500	Our	-2.465	3.549	0.094	2.693	-0.911	3.057
0.2	500	IG	-0.424	2.525	7.302	4.046	9.769	5.406
0.2	500	VER	-3.981	4.174	10.519	6.737	11.139	7.309
0.2	500	Full	-0.102	2.652	-0.269	2.734	-0.059	2.910
0.2	1000	Our	-1.960	1.798	0.157	1.585	-2.211	1.913
0.2	1000	IG	-0.294	1.157	7.671	3.111	9.901	4.398
0.2	1000	VER	-3.678	2.120	11.121	5.650	11.440	6.061
0.2	1000	Full	-0.080	1.184	0.191	1.506	0.119	1.672

Table 3.14: CP and AL of estimated CIs for $ROC(s)$ with $s = 0.1$ and $s = 0.2$ under Scenarios 4–6

s	n	Scenario 4		Scenario 5		Scenario 6	
		CP	AL	CP	AL	CP	AL
0.1	200	0.928	0.387	0.935	0.348	0.904	0.346
0.1	500	0.934	0.228	0.944	0.218	0.925	0.217
0.1	1000	0.936	0.162	0.948	0.156	0.917	0.153
0.2	200	0.940	0.405	0.933	0.341	0.920	0.340
0.2	500	0.949	0.240	0.954	0.217	0.946	0.217
0.2	1000	0.951	0.167	0.949	0.154	0.918	0.154

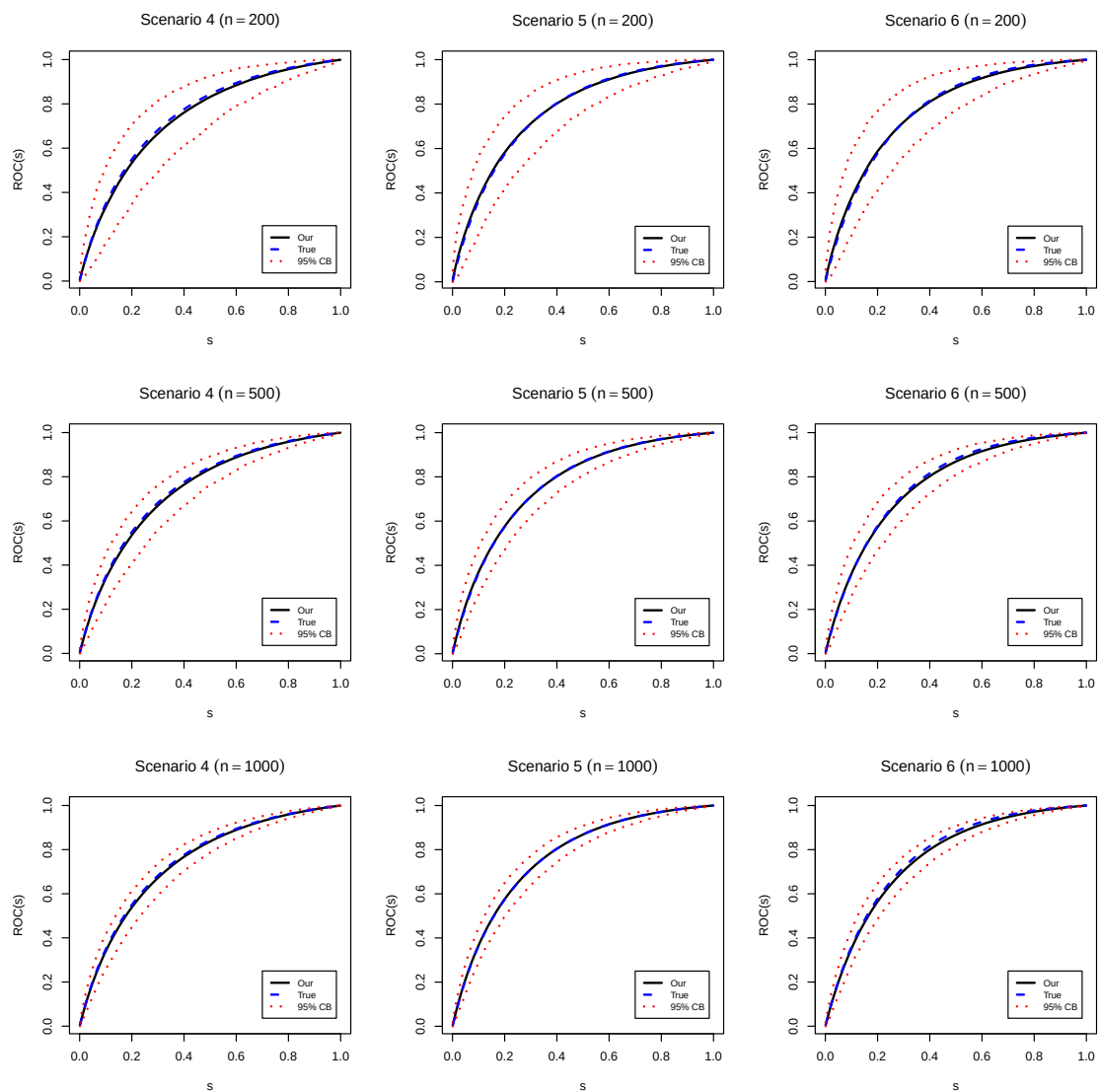


Figure 3.3: The true ROC curve, the average estimated curve, and its 95% confidence band based on 1000 replications under Scenarios 4–6.

Comparison of two types of CIs for AUC and ROC curve

To construct Wald-type CIs for AUC and $ROC(s)$, we require consistent estimators for σ_{AUC}^2 and σ_s^2 , respectively. In Section 3.3, we employ the plug-in method to obtain these estimators. Alternatively, the bootstrap method can be used.

In this section, we apply the nonparametric bootstrap method with 200 bootstrap samples to estimate σ_{AUC}^2 and σ_s^2 . The resulting CIs (referred to as bootstrap CIs) are then compared with those obtained using the plug-in method (referred to as plug-in CIs). The simulation results for these two types of CIs for AUC under Scenarios 1–3 and 4–6 are summarized in Tables 3.15 and 3.16, respectively. Similarly, the simulation results for the two types of CIs for $ROC(s)$ at $s = 0.1$ and $s = 0.2$ under Scenarios 1–3 and 4–6 are summarized in Tables 3.17 and 3.18, respectively.

We observe that both CIs achieve satisfactory coverage in Scenarios 1–2 and 4–5, where the models are correctly specified, though the bootstrap CIs may exhibit over-coverage in certain settings. In cases where the models are misspecified (Scenarios 3 and 6), the bootstrap CIs generally provide better coverage than the plug-in CIs in some settings.

Table 3.15: CP and AL of two types of CIs for AUC under Scenarios 1–3

CI	n	Scenario 1		Scenario 2		Scenario 3	
		CP	AL	CP	AL	CP	AL
Plug-in	5000	0.940	0.058	0.950	0.037	0.937	0.050
Bootstrap	5000	0.950	0.061	0.970	0.054	0.960	0.072
Plug-in	10000	0.951	0.041	0.956	0.026	0.936	0.035
Bootstrap	10000	0.954	0.043	0.962	0.030	0.959	0.041

Table 3.16: CP and AL of two types of CIs for AUC under Scenarios 4–6

CI	n	Scenario 4		Scenario 5		Scenario 6	
		CP	AL	CP	AL	CP	AL
Plug-in	200	0.946	0.202	0.935	0.160	0.920	0.157
Bootstrap	200	0.935	0.175	0.939	0.157	0.936	0.157
Plug-in	500	0.959	0.119	0.949	0.102	0.937	0.100
Bootstrap	500	0.955	0.115	0.956	0.102	0.943	0.105
Plug-in	1000	0.954	0.083	0.938	0.072	0.907	0.071
Bootstrap	1000	0.951	0.082	0.945	0.073	0.925	0.076

Table 3.17: CP and AL of two types of CIs for $ROC(s)$ at $s = 0.1$ and $s = 0.2$ under Scenarios 1–3

CI	s	n	Scenario 1		Scenario 2		Scenario 3	
			CP	AL	CP	AL	CP	AL
Plug-in	0.1	5000	0.932	0.092	0.939	0.074	0.908	0.097
Bootstrap	0.1	5000	0.937	0.095	0.966	0.089	0.942	0.121
Plug-in	0.1	10000	0.951	0.065	0.963	0.052	0.860	0.068
Bootstrap	0.1	10000	0.960	0.066	0.966	0.055	0.883	0.075
Plug-in	0.2	5000	0.933	0.103	0.940	0.077	0.932	0.102
Bootstrap	0.2	5000	0.945	0.107	0.960	0.101	0.955	0.137
Plug-in	0.2	10000	0.956	0.073	0.951	0.054	0.928	0.071
Bootstrap	0.2	10000	0.956	0.075	0.955	0.059	0.950	0.082

Table 3.18: CP and AL of two types of CIs for $ROC(s)$ at $s = 0.1$ and $s = 0.2$ under Scenarios 4–6

Method	s	n	Scenario 4		Scenario 5		Scenario 6	
			CP	AL	CP	AL	CP	AL
Plug-in	0.1	200	0.928	0.387	0.935	0.348	0.904	0.346
Bootstrap	0.1	200	0.925	0.342	0.952	0.360	0.946	0.368
Plug-in	0.1	500	0.934	0.228	0.944	0.218	0.925	0.217
Bootstrap	0.1	500	0.942	0.222	0.958	0.228	0.955	0.237
Plug-in	0.1	1000	0.936	0.162	0.948	0.156	0.917	0.153
Bootstrap	0.1	1000	0.945	0.161	0.956	0.163	0.961	0.171
Plug-in	0.2	200	0.940	0.405	0.933	0.341	0.920	0.340
Bootstrap	0.2	200	0.932	0.353	0.945	0.340	0.940	0.348
Plug-in	0.2	500	0.949	0.240	0.954	0.217	0.946	0.217
Bootstrap	0.2	500	0.947	0.233	0.956	0.220	0.959	0.231
Plug-in	0.2	1000	0.951	0.167	0.949	0.154	0.918	0.154
Bootstrap	0.2	1000	0.951	0.167	0.953	0.157	0.942	0.166

Chapter 4

ROC Curve Analysis when Combining Multiple Biomarkers in Presence of Imperfect Reference Standard

4.1 Introduction

The methods presented in Chapters 2 and 3 assume the availability of a gold standard test that determines the true disease status for all or part of the dataset, thereby implying no misclassification in the observed disease status. In practice, however, a gold standard may be unavailable or impractical to implement at scale due to cost, time, or ethical constraints. In such cases, an imperfect reference standard is often used instead ([Sun et al. 2024](#)). For instance, in brain tumor diagnosis, Magnetic Resonance Imaging (MRI) is commonly used

as the standard diagnostic tool, despite its sensitivity and specificity being less than perfect. Biopsies, which involve direct tissue examination, offer superior sensitivity and significantly higher specificity ([Taghipour Zahir et al. 2011](#)), but their invasive nature and associated risks limit their routine clinical use. Consequently, although MRI is an imperfect reference standard, its accessibility and feasibility make it more widely adopted than the preferable but less practical biopsy.

In this chapter, we consider ROC curve analysis in the presence of an imperfect reference standard, where a gold standard is not available. The reference standard is imperfect in the sense that it may misclassify diseased individuals as healthy and healthy individuals as diseased. As a result, the nominal diseased group (denoted by $R = 1$) may include some truly healthy individuals, while the nominal healthy group (denoted by $R = 0$) may include some truly diseased individuals.

Naively treating R as the true disease status (denoted by D , where $D = 1$ indicates diseased and $D = 0$ indicates healthy) and applying the existing ROC analysis methods may lead to an underestimated ROC curve and biased summary indices. For further discussion, see [Sun et al. \(2024\)](#), [Sun et al. \(2025\)](#), and Section 4.4. Under the two-sample framework, with one sample of biomarkers from the $R = 1$ group, and the other from the $R = 0$ group, [Sun et al. \(2024\)](#) and [Sun et al. \(2025\)](#) proposed fully nonparametric methods to estimate the ROC curve and related summary indices for a univariate biomarker. These methods rely on the conditional independence assumption (Assumption 1 in Section 4.2.1) and known accuracy of the reference standard R (Assumption 2 in Section 4.2.1). Despite the presence of misclassification in both the $R = 0$ and $R = 1$ groups, they showed that their approaches can still consistently estimate the ROC curve and its associated summary indices.

In some practical settings, a single biomarker may lack sufficient sensitivity and speci-

ficity. Consequently, combining multiple biomarkers is often pursued to develop more effective diagnostic tools. The literature offers a variety of methods for combining multiple biomarkers and constructing ROC curves for the resulting combined biomarkers; see [Ma & Huang \(2005\)](#), [Lin et al. \(2011\)](#), [Qin & Zhang \(2010\)](#), [Chen et al. \(2016\)](#), [Meisner et al. \(2021\)](#), [Huang & Sanda \(2022\)](#), [Du et al. \(2024\)](#), and references therein. However, when these methods are applied in the presence of an imperfect reference standard, the resulting ROC curves pertain only to the nominal diseased and nominal healthy groups. As such, they cannot be directly used to distinguish truly diseased from truly healthy individuals. See Remark 4.1 for more details. As previously discussed, the methods proposed in [Sun et al. \(2024\)](#) and [Sun et al. \(2025\)](#) are applicable under an imperfect reference standard. However, their approaches are limited to univariate biomarkers and cannot be directly extended to combine multiple biomarkers. In summary, there is currently no available method for combining multiple biomarkers and constructing an ROC curve that can directly distinguish truly diseased individuals from truly healthy ones in the presence of an imperfect reference standard. This chapter aims to shed light on this gap.

Let $\mathbf{T}$ denote the p -dimensional vector of biomarkers for an individual. We use F_0 and F_1 to denote the cdfs of $\mathbf{T}$ for truly healthy and truly diseased individuals, respectively. Following [Qin & Zhang \(2010\)](#), we propose modeling F_0 and F_1 using a density ratio model (DRM):

$$\frac{dF_1(\mathbf{x})}{dF_0(\mathbf{x})} = \exp\{\alpha + \beta^T \mathbf{q}(\mathbf{x})\}, \quad (4.1)$$

where $\alpha \in \mathcal{R}$, $\beta \in \mathcal{R}^q$, and $\mathbf{q}(\cdot)$ is a q -dimensional vector of known functions. The most common choices for $\mathbf{q}(\cdot)$ are $\mathbf{x}$ and $\log(\mathbf{x})$. Alternatively, other basis functions, such as Bernstein polynomials, polynomials of $\log(\mathbf{x})$ and others, can also be considered. When $q = 1$, the likelihood ratio ordering discussed in Chapter 2 is satisfied given that $\beta q(x)$ is an increasing function of x . In model (4.1), the baseline cdf $F_0(\mathbf{x})$ is unspecified, making the

DRM a semiparametric model. The DRM has been widely applied in various fields such as medical research, survival analysis, covariate shift adaptation, and machine learning. For recent developments and applications of the DRM, see [Qin \(2017\)](#), [Wang \(2017\)](#), and [Yuan \(2021\)](#).

According to the celebrated Neyman–Pearson lemma, the optimal combination of multiple biomarkers is given by the likelihood ratio ([Eguchi & Copas 2002](#), [McIntosh & Pepe 2002](#)). Due to the invariance property of the ROC curve, under model (4.1), the optimal ROC curve can be constructed based on the combination $\beta^T \mathbf{q}(\mathbf{T})$. In this chapter, under model (4.1) and assuming the conditional independence assumption (Assumption 1 in Section 4.2.1) and known accuracy of the reference standard R (Assumption 2 in Section 4.2.1), we first demonstrate that the model parameters $(\alpha, \beta, F_0, F_1)$ in (4.1) are identifiable under a mild condition. We then propose a maximum empirical likelihood method to estimate $(\alpha, \beta, F_0, F_1)$, the ROC curve corresponding to the combination $\beta^T \mathbf{q}(\mathbf{T})$, and the associated AUC. An Expectation-Maximization (EM) algorithm is developed for numerical implementation, and a bootstrap method is introduced to construct confidence intervals for the AUC and confidence bands for the ROC curve.

The remainder of this chapter is organized as follows. Section 4.2 presents the research problem, discusses model identifiability, and introduces the proposed estimation procedure. Section 4.3 details the EM algorithm used for implementation. In Section 4.4, simulation studies are conducted to evaluate the performance of the proposed method and assess the accuracy of the confidence interval estimates. Finally, Section 4.5 illustrates the application of our method using a real data example.

4.2 Main results

4.2.1 Problem setup and model identifiability

Recall that we use $\mathbf{T}$ to denote the p -dimensional biomarker, D to represent the true disease status, and R to represent the nominal (observed) disease status. Define

$$\lambda_0 = P(R = 0|D = 0) \text{ and } \lambda_1 = P(R = 1|D = 1),$$

which quantify the classification accuracy of R . Note that R is imperfect in the sense that at least one of $\lambda_0 < 1$ or $\lambda_1 < 1$ must hold.

To facilitate the development of our method, we adopt the following assumptions from [Sun et al. \(2024\)](#) and [Sun et al. \(2025\)](#):

Assumption 1. $R \perp \mathbf{T} \mid D$; that is the nominal disease status R and the biomarkers $\mathbf{T}$ are conditionally independent given the true disease status D ;

Assumption 2. The classification accuracies λ_0 and λ_1 are known, and $\lambda_0 + \lambda_1 > 1$.

We also adopt the two-sample framework proposed in [Sun et al. \(2024\)](#) and [Sun et al. \(2025\)](#), in which one sample of biomarkers is collected from individuals with $R = 0$ and another from those from $R = 1$. Under model (4.1) and Assumptions 1–2, we establish the identifiability of the parameters $(\alpha, \boldsymbol{\beta}, F_0, F_1)$ under a mild condition. Recall the notion of stochastic linear independence introduced in Section 3.6.1.

Theorem 4.1. *Assume model (4.1) and Assumptions 1–2 hold. If the components of $(1, \mathbf{q}(\mathbf{x}))$ are stochastically linearly independent, then the model parameters $(\alpha, \boldsymbol{\beta}, F_0, F_1)$ in model (4.1) are identifiable.*

For continuity of presentation, the proof of Theorem 4.1 is provided in Section 4.6.1.

The validity of Theorem 4.1 relies on Assumptions 1–2 and the correctness of model (4.1). As noted in Sun et al. (2024), Assumption 1 typically holds when the reference standard and the biomarkers are derived from distinct sources of information. For example, this can happen when the reference test uses blood samples, while the biomarkers are based on imaging results, since these methods provide different kinds of information (Sun et al. 2024). Assumption 2 requires that both λ_0 and λ_1 are known. This condition can be satisfied if the accuracy of the imperfect reference standard has been established in prior studies. For instance, rapid COVID-19 tests are widely used due to their speed and affordability, and their sensitivity and specificity have been extensively evaluated; see Yamayoshi et al. (2020), Okoye et al. (2022), and related works. Additionally, we assume that $\lambda_0 + \lambda_1 > 1$, ensuring that the reference test is more informative than random guessing. It is important to emphasize that the validity of Theorem 4.1 also depends on the correctness of model (4.1). Developing diagnostic tools to assess this assumption remains a valuable direction for future research. In what follows, we proceed under the assumption that Assumptions 1–2 and model (4.1) are valid.

With the above setup, we now proceed to formally define the research problem of interest. Recall that F_0 and F_1 denote the cdfs of $\mathbf{T}$ the truly healthy and truly diseased individuals, respectively; that is,

$$F_0(\mathbf{t}) = P(\mathbf{T} \leq \mathbf{t} | D = 0) \quad \text{and} \quad F_1(\mathbf{t}) = P(\mathbf{T} \leq \mathbf{t} | D = 1).$$

As discussed earlier, biomarker data are collected from individuals with $R = 0$ and $R = 1$. Based on Assumption 1, the distributions of $\mathbf{T}$ conditional on R are given by

$$\mathbf{T} | R = 0 \sim \lambda_0 F_0 + (1 - \lambda_0) F_1 \quad \text{and} \quad \mathbf{T} | R = 1 \sim \lambda_1 F_1 + (1 - \lambda_1) F_0. \quad (4.2)$$

Recall that under model (4.1), $\beta^T \mathbf{q}(\mathbf{T})$ is the optimal combination of the biomarkers. Let F_{c0} and F_{c1} denote the cdfs of $\beta_0^T \mathbf{q}(\mathbf{T})$ for the truly healthy and truly diseased individuals, respectively, where β_0 is the true value of β . That is,

$$F_{c0}(x) = P(\beta_0^T \mathbf{q}(\mathbf{T}) \leq x | D = 0) \quad \text{and} \quad F_{c1}(x) = P(\beta_0^T \mathbf{q}(\mathbf{T}) \leq x | D = 1). \quad (4.3)$$

Our objective is to estimate the optimal ROC curve defined by

$$ROC(s) = 1 - F_{c1}(F_{c0}^{-1}(1 - s)), \quad \text{for } s \in (0, 1), \quad (4.4)$$

and the corresponding AUC,

$$AUC = \int_0^1 ROC(s) \, ds. \quad (4.5)$$

Remark 4.1. We note that under model (4.1) and Assumptions 1–2, the methods proposed in [Ma & Huang \(2005\)](#), [Lin et al. \(2011\)](#), and [Chen et al. \(2016\)](#) for combining multiple biomarkers yield consistent estimators of $\beta/||\beta||$, where $||\beta||$ denotes the L_2 -norm of β . However, the primary goal of these methods is to estimate the ROC curve

$$ROC^*(s) = 1 - G_{c1}(G_{c0}^{-1}(1 - s)),$$

where G_{c0} and G_{c1} are the cdfs of $\beta_0^T \mathbf{q}(\mathbf{T})$ for the nominal healthy and nominal diseased individuals, respectively; that is,

$$G_{c0} = \lambda_0 F_{c0} + (1 - \lambda_0) F_{c1} \quad \text{and} \quad G_{c1} = \lambda_1 F_{c1} + (1 - \lambda_1) F_{c0}.$$

Note that the ROC curve $ROC^*(s)$ characterizes the ability to distinguish nominally healthy

individuals from nominally diseased individuals, rather than truly healthy from truly diseased.

4.2.2 Empirical likelihood method

Let $\{\mathbf{X}_1, \dots, \mathbf{X}_{n_0}\}$ and $\{\mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}\}$ denote the biomarkers from the nominally healthy and nominally diseased groups, respectively. We denote the combined sample by

$$\mathcal{T} = \{\mathbf{T}_1, \dots, \mathbf{T}_n\} = \{\mathbf{X}_1, \dots, \mathbf{X}_{n_0}, \mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}\},$$

where $n = n_0 + n_1$.

Based on the two-sample data, the full likelihood function is given by

$$\prod_{i=1}^{n_0} \{\lambda_0 dF_0(\mathbf{X}_i) + (1 - \lambda_0) dF_1(\mathbf{X}_i)\} \times \prod_{j=1}^{n_1} \{\lambda_1 dF_1(\mathbf{Y}_j) + (1 - \lambda_1) dF_0(\mathbf{Y}_j)\}.$$

For $k = 1, \dots, n$, define

$$p_k = dF_0(\mathbf{T}_k).$$

Under model (4.1), we have

$$q_k = dF_1(\mathbf{T}_k) = p_k \exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k) \}.$$

Let $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, p_1, \dots, p_n)$. Then, the empirical likelihood function of $\boldsymbol{\theta}$ is given by (Owen

2001):

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n_0} [\lambda_0 + (1 - \lambda_0) \exp\{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{X}_i)\}] \cdot \prod_{j=1}^{n_1} [\lambda_1 \exp\{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{Y}_j)\} + (1 - \lambda_1)] \prod_{k=1}^n p_k,$$

subject to the constraints

$$p_k \geq 0, \quad \sum_{k=1}^n p_k = 1, \quad \sum_{k=1}^n p_k \exp\{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k)\} = 1. \quad (4.6)$$

Note that the first two constraints in (4.6) ensures that F_0 is a valid cdf, while the third constraint in (4.6) ensures that F_1 is also a valid cdf.

The maximum empirical likelihood estimator (MELE) of $\boldsymbol{\theta}$ is then defined as

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}), \quad (4.7)$$

subject to the constraints in (4.6). In general, $\hat{\boldsymbol{\theta}}$ is not available in closed form. In Section 4.3, we introduces an EM algorithm for numerically computing $\hat{\boldsymbol{\theta}}$, which comes with a theoretical guarantee of convergence.

Recall the definitions of F_{c0} and F_{c1} in (4.3). Once $\hat{\boldsymbol{\theta}}$ is obtained, the estimators of F_{c1} and F_{c0} are given by

$$\hat{F}_{c1}(x) = \sum_{k=1}^n \hat{p}_k \exp\{\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{q}(\mathbf{T}_k)\} I \left\{ \hat{\boldsymbol{\beta}}^T \mathbf{q}(\mathbf{T}_k) \leq x \right\}, \quad (4.8)$$

and

$$\hat{F}_{c0}(x) = \sum_{k=1}^n \hat{p}_k I \left\{ \hat{\boldsymbol{\beta}}^T \mathbf{q}(\mathbf{T}_k) \leq x \right\}, \quad (4.9)$$

respectively.

The optimal ROC curve in (4.4) and the corresponding AUC in (4.5) can then be estimated by

$$\widehat{ROC}(s) = 1 - \hat{F}_{c1} \left\{ \hat{F}_{c0}^{-1}(1 - s) \right\} \quad (4.10)$$

and

$$\widehat{AUC} = \int_0^1 \widehat{ROC}(s) ds, \quad (4.11)$$

respectively.

At the current stage, we are unable to establish the asymptotic properties of $\widehat{ROC}(s)$ and $\widehat{AUC}$. Therefore, we propose a bootstrap procedure to construct a confidence interval for the AUC and a confidence band for the ROC curve. In what follows, let B denote the number of bootstrap samples.

Step 1. Using $\hat{\boldsymbol{\theta}}$, estimate F_0 and F_1 as follows:

$$\hat{F}_0(\mathbf{x}) = \sum_{k=1}^n \hat{p}_k I(\mathbf{T}_k \leq \mathbf{x}),$$

and

$$\hat{F}_1(\mathbf{x}) = \sum_{k=1}^n \hat{p}_k \exp\{\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{q}(T_k)\} I(\mathbf{T}_k \leq \mathbf{x}).$$

Step 2. For each $b = 1, \dots, B$, we generate a bootstrap two-sample dataset as follows:

$$\left\{ \mathbf{X}_1^{(b)}, \dots, \mathbf{X}_{n_0}^{(b)} \right\} \sim \lambda_0 \hat{F}_0 + (1 - \lambda_0) \hat{F}_1, \quad \left\{ \mathbf{Y}_1^{(b)}, \dots, \mathbf{Y}_{n_1}^{(b)} \right\} \sim \lambda_1 \hat{F}_1 + (1 - \lambda_1) \hat{F}_0.$$

For each bootstrap sample, apply the proposed maximum empirical likelihood method described above to estimate the ROC curve and AUC. Denote the resulting estimates by $\widehat{ROC}^{(b)}(s)$ and $\widehat{AUC}^{(b)}$, respectively.

Step 3. The $1 - \alpha$ confidence interval for the AUC is given by

$$\left[\widehat{AUC}_{\alpha/2}, \widehat{AUC}_{1-\alpha/2} \right],$$

where $\widehat{AUC}_{\alpha/2}$ and $\widehat{AUC}_{1-\alpha/2}$ are the $\alpha/2$ and $1 - \alpha/2$ empirical quantiles of the bootstrap replicates $\left\{ \widehat{AUC}^{(1)}, \dots, \widehat{AUC}^{(B)} \right\}$.

Similarly, construct a pointwise $1 - \alpha$ confidence band for the ROC curve, we estimate the confidence interval for each $ROC(s)$ by

$$\left[\widehat{ROC}_{\alpha/2}(s), \widehat{ROC}_{1-\alpha/2}(s) \right], \quad \text{for } s \in (0, 1),$$

where $\widehat{ROC}_{\alpha/2}(s)$ and $\widehat{ROC}_{1-\alpha/2}(s)$ are the corresponding empirical quantiles of the bootstrap replicates $\left\{ \widehat{ROC}^{(1)}(s), \dots, \widehat{ROC}^{(B)}(s) \right\}$.

In our numerical studies, we set $B = 500$. The performance of the proposed bootstrap method is evaluated in Section 4.4.

4.3 EM algorithm

In this section, we apply the EM algorithm to numerically calculate $\hat{\boldsymbol{\theta}}$. For $i = 1, \dots, n_0$ and $j = 1, \dots, n_1$, we define latent variables Z_{i0} and Z_{j1} , where $Z_{i0} = 1$ if the i th individual in the $R = 0$ group is truly healthy, and $Z_{i0} = 0$ if truly diseased; $Z_{j1} = 1$ if the j th individual in the $R = 1$ group is truly diseased, and $Z_{j1} = 0$ if truly healthy.

With Z_{i0} and Z_{j1} , the complete log-likelihood function is given by

$$\begin{aligned}
l_c(\boldsymbol{\theta}) = & \sum_{i=1}^{n_0} [Z_{i0} \log(\lambda_0) + (1 - Z_{i0}) \log(1 - \lambda_0) + (1 - Z_{i0}) \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{X}_i)\}] \\
& + \sum_{j=1}^{n_1} [(1 - Z_{j1}) \log(1 - \lambda_1) + Z_{j1} \log(\lambda_1) + Z_{j1} \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{Y}_j)\}] \\
& + \sum_{k=1}^n \log(p_k). \tag{4.12}
\end{aligned}$$

The core of the EM algorithm is the EM iteration, which contains a E-step and M-step. We use $\boldsymbol{\theta}^{(m-1)} = (\alpha^{(m-1)}, \boldsymbol{\beta}^{(m-1)}, p_1^{(m-1)}, \dots, p_n^{(m-1)})$ to denote the parameter estimates after the $(m-1)$ th iteration. When $m = 1$, $\boldsymbol{\theta}^{(0)} = (\alpha^{(0)}, \boldsymbol{\beta}^{(0)}, p_1^{(0)}, \dots, p_n^{(0)})$ denote the initial value of $\boldsymbol{\theta}$.

E-step. We compute the posterior probabilities:

$$\begin{aligned}
w_{i,0}^{(m)} &= E(Z_{i0} | \mathcal{T}; \boldsymbol{\theta}^{(m-1)}) \\
&= \frac{\lambda_0}{\lambda_0 + (1 - \lambda_0) \exp [\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(X_i)]},
\end{aligned}$$

and

$$\begin{aligned}
w_{j,1}^{(m)} &= E(Z_{j1} | \mathcal{T}; \boldsymbol{\theta}^{(m-1)}) \\
&= \frac{\lambda_1 \exp [\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(Y_j)]}{\lambda_1 \exp [\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(Y_j)] + 1 - \lambda_1},
\end{aligned}$$

for $i = 1, \dots, n_0$ and $j = 1, \dots, n_1$. In the above, the expectations are with respect to the conditional distributions of Z_{i0} 's and Z_{j1} 's given $\mathcal{T}$, using $\boldsymbol{\theta}^{(m-1)}$ as the parameters of conditional distribution.

After that, we obtain $Q^{(m)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)}) = E\{l_c(\boldsymbol{\theta})|\mathcal{T}; \boldsymbol{\theta}^{(m-1)}\}$ as follows:

$$\begin{aligned}
& Q^{(m)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)}) \\
&= \sum_{i=1}^{n_0} \left[w_{i,0}^{(m)} \log(\lambda_0) + (1 - w_{i,0}^{(m)}) \log(1 - \lambda_0) + (1 - w_{i,0}^{(m)}) \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{X}_i)\} \right] \\
&+ \sum_{j=1}^{n_1} \left[(1 - w_{j,1}^{(m)}) \log(1 - \lambda_1) + w_{j,1}^{(m)} \log(\lambda_1) + w_{j,1}^{(m)} \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{Y}_j)\} \right] \\
&+ \sum_{k=1}^n \log(p_k).
\end{aligned}$$

M-step. We update $\boldsymbol{\theta}$ from $\boldsymbol{\theta}^{(m-1)}$ to $\boldsymbol{\theta}^{(m)}$ by maximizing $Q^{(m)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)})$ subject to (4.6). Let

$$\alpha^* = \alpha + \log \left[\frac{\sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} + \sum_{j=1}^{n_1} w_{j,1}^{(m)}}{n - \sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} - \sum_{j=1}^{n_1} w_{j,1}^{(m)}} \right].$$

In the Section 4.6.2, we show that

$$\{\alpha^{*(m)}, \boldsymbol{\beta}^{(m)}\} = \arg \max_{\alpha^*, \boldsymbol{\beta}} h^{(m)}(\alpha^*, \boldsymbol{\beta}),$$

where

$$\begin{aligned}
h^{(m)}(\alpha^*, \boldsymbol{\beta}) &= \sum_{i=1}^{n_0} (1 - w_{i,0}^{(m)}) \{\alpha^* + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{X}_i)\} \\
&+ \sum_{j=1}^{n_1} w_{j,1}^{(m)} \{\alpha^* + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{Y}_j)\} \\
&- \sum_{k=1}^n \log[1 + \exp\{\alpha^* + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k)\}].
\end{aligned} \tag{4.13}$$

Note that $h^{(m)}(\alpha^*, \boldsymbol{\beta})$ is proportional to a weighted log-likelihood function of a logistic

regression model. Hence, we can easily obtain $\{\alpha^{*(m)}, \beta^{(m)}\}$ by fitting a logistic regression, which can be done by the R function `glmnet`, for instance.

Once $h^{(m)}(\alpha^*, \beta)$ is obtained, we then update $\alpha^{(m)}$ with

$$\alpha^{(m)} = \alpha^{*(m)} - \log \left[\frac{\sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} + \sum_{j=1}^{n_1} w_{j,1}^{(m)}}{n - \sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} - \sum_{j=1}^{n_1} w_{j,1}^{(m)}} \right]$$

and update the p_k 's with

$$p_k^{(m)} = \frac{1}{n + \left[\sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} + \sum_{j=1}^{n_1} w_{j,1}^{(m)} \right] (\exp[\alpha^{(m)} + \{\beta^{(m)}\}^T \mathbf{q}(\mathbf{T}_k)] - 1)}.$$

With the EM algorithm established above, we also show that our algorithm eventually converges. The following theorem states that the empirical likelihood function $L(\boldsymbol{\theta})$ does not decrease after each iteration.

Theorem 4.2. *With the EM algorithm described above, for $m \geq 1$, we have*

$$L(\boldsymbol{\theta}^{(m)}) \geq L(\boldsymbol{\theta}^{(m-1)}).$$

The proof of Theorem 4.2 is provided in Section 4.6.3. Note that by definition, $L(\boldsymbol{\theta}) \leq 1$, which, together with Theorem 4.2, implies that the EM algorithm eventually converges for the given initial value $\boldsymbol{\theta}^{(0)}$. Our numerical experience indicates that the EM-algorithm, under the current setup, is robust to the choice of initial values. That is, different initial values always lead to the same estimate.

4.4 Simulation

In this section, we compare the performance of our proposed method with that of [Qin & Zhang \(2010\)](#) in terms of the ROC curve and AUC estimation. We also evaluate the accuracy of the bootstrap confidence interval estimation procedure, as discussed in Section 4.2.2, at the nominal level of 95%. The method of [Qin & Zhang \(2010\)](#) is essentially equivalent to our proposed method with $\lambda_0 = \lambda_1 = 1$. In the following, we refer to their method as the “DRM” method and our method as “Our”.

4.4.1 Simulation setup

In the simulation study, we consider two scenarios for the underlying distributions F_0 and F_1 .

Scenario 1: the biomarkers $\mathbf{T} = (T_1, T_2)^T$ are two-dimensional and F_0 and F_1 are cdfs of $N(\boldsymbol{\mu}_{10}, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_{11}, \boldsymbol{\Sigma}_1)$, respectively, where

$$\boldsymbol{\mu}_{10} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_{11} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}.$$

Scenario 2: the biomarkers $\mathbf{T} = (T_1, T_2, T_3)^T$ are three-dimensional and F_0 and F_1 are cdfs of $N(\boldsymbol{\mu}_{20}, \boldsymbol{\Sigma}_2)$ and $N(\boldsymbol{\mu}_{21}, \boldsymbol{\Sigma}_2)$, respectively, where

$$\boldsymbol{\mu}_{20} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_{21} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 1 \end{pmatrix}.$$

Note that in both scenarios, the model assumption in (4.1) is satisfied with $\mathbf{q}(\mathbf{x}) = \mathbf{x}$.

In Scenario 1, the optimal combination of biomarkers is $\frac{4}{5}T_1 + \frac{4}{5}T_2$, while in Scenario 2, the optimal combination of biomarker is $\frac{2}{3}T_1 + \frac{2}{3}T_2 + \frac{2}{3}T_3$. The true values of the AUC, $ROC(0.1)$ and $ROC(0.2)$ for the optimal combination of biomarkers in each scenario are summarized in Table 4.1.

Table 4.1: Summary of AUC, $ROC(0.1)$, $ROC(0.2)$ for the optimal combination of biomarkers in each scenario

Scenario	ROC(0.1)	ROC(0.2)	AUC
1	0.493	0.664	0.815
2	0.553	0.717	0.841

For each of distributional scenario, we consider two sample sizes settings, $(n_0, n_1) = (100, 100)$ and $(200, 200)$, and two settings for (λ_0, λ_1) : $\lambda_0 = \lambda_1 = \lambda$, where $\lambda = 0.75$ and 0.9 . This results in a total of 8 combinations. For each combination, the number of replications is 1000.

4.4.2 Simulation results for the AUC

In this section, we compare our proposed method with the DRM method in the AUC estimation and assess the performance of the confidence interval (CI) estimation for the AUC at the nominal level of 0.95.

We begin by comparing the AUC estimates using relative bias (RB) and mean squared error (MSE), as defined in Section 2.3.3. The results are summarized in Table 4.2. From the table, we can see that our method has much lower RB and MSE than the DRM method in all cases. The difference becomes more noticeable when the misclassification rate, $1 - \lambda$, is higher. As the sample size increases, the RB for our method gets closer to zero. In

contrast, the RB for the DRM method stays about the same. This is expected, since the DRM method does not account for label misclassification.

Table 4.2: RB (%) and MSE ($\times 1000$) for the *AUC* estimation under 2 scenarios

(n_0, n_1, λ)	Method	Scenario 1		Scenario 2	
		RB	MSE	RB	MSE
(100, 100, 0.9)	Our	-0.219	1.901	0.551	1.563
	DRM	-7.970	5.502	-7.807	5.404
(200, 200, 0.9)	Our	0.006	0.930	0.271	0.794
	DRM	-8.000	4.877	-8.241	5.383
(100, 100, 0.75)	Our	-0.502	4.825	0.899	4.156
	DRM	-19.347	26.381	-19.675	28.765
(200, 200, 0.75)	Our	-0.239	2.471	0.332	2.497
	DRM	-19.668	26.404	-20.381	30.116

Next, we present the results for the AUC CI estimation. As in Section 3.4.3, we use coverage probability (CP) and average length (AL) to evaluate CI performance. The results are summarized in Table 4.3. From the table, we observe that the CPs are reasonably close to the nominal level. In all scenarios with $\lambda = 0.75$ and $(n_0, n_1) = (100, 100)$, the estimated CIs show a slight overcoverage issue. In other settings, the CPs are very close to 0.95. Within each scenario, the ALs decrease as the sample size increases and the proportion of mislabeled data decreases.

4.4.3 Simulation results for the ROC curve

In this section, we evaluate the accuracy of the proposed method and compare it to the DRM method for the ROC curve estimation. Additionally, we assess the performance of the ROC curve CI estimation using the proposed method at a nominal level of 0.95.

Table 4.3: CP and AL for the AUC confidence interval estimations under 2 scenarios

(n_0, n_1, λ)	Scenario 1		Scenario 2	
	CP	AL	CP	AL
(100, 100, 0.9)	0.954	0.157	0.938	0.146
(200, 200, 0.9)	0.936	0.113	0.941	0.107
(100, 100, 0.75)	0.970	0.246	0.968	0.222
(200, 200, 0.75)	0.952	0.191	0.942	0.178

Figures 4.1 and 4.2 show the average estimated ROC curves, alongside the true ROC curves for Scenarios 1 and 2. In both figures, we observe that the average estimated ROC curves under our method closely align with the true ROC curves across all settings of λ and (n_0, n_1) . The DRM method consistently underestimates the ROC curves for all $s \in (0, 1)$. This demonstrates the advantage of the proposed method over the DRM method in terms of the ROC curve estimation in the presence of data mislabeling.

We also compare the point estimates of the ROC curve at $s = 0.1$ and 0.2 . For both the estimates of $ROC(0.1)$ and $ROC(0.2)$, we observe that our method has a significant advantage over the DRM method in terms of RB and MSE. Similar to the patterns observed in the AUC estimation, the RB for our method decreases to 0 as (n_0, n_1) increase, while the RB for the DRM method remains at a certain level.

Next, we present the results for the ROC CI estimation using Our method at a nominal level of 0.95. Similar to Section 4.4.2, we use CP and AL to assess the effectiveness of the CI estimation. The results are summarized in Table 4.5. Our results show that, for both $s = 0.1$ and $s = 0.2$, the CPs are very close to the nominal level. Similar to the AUC CI estimation, an overcoverage issue arises when the sample sizes are small ($n_0 = n_1 = 100$) and mislabeling proportion is high ($\lambda = 0.75$). The ALs decrease as the sample sizes increase across all scenarios.

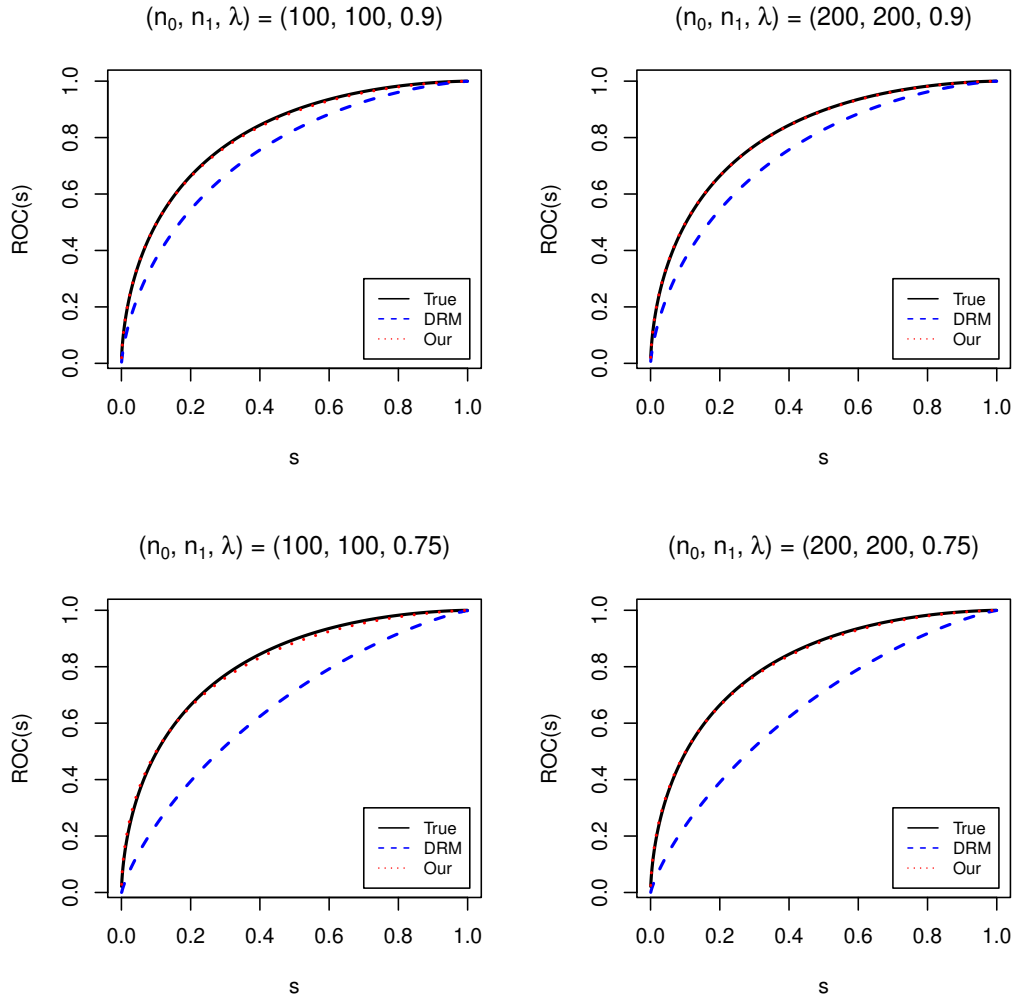


Figure 4.1: True ROC curves and average estimated ROC curves using our and DRM methods under Scenario 1.

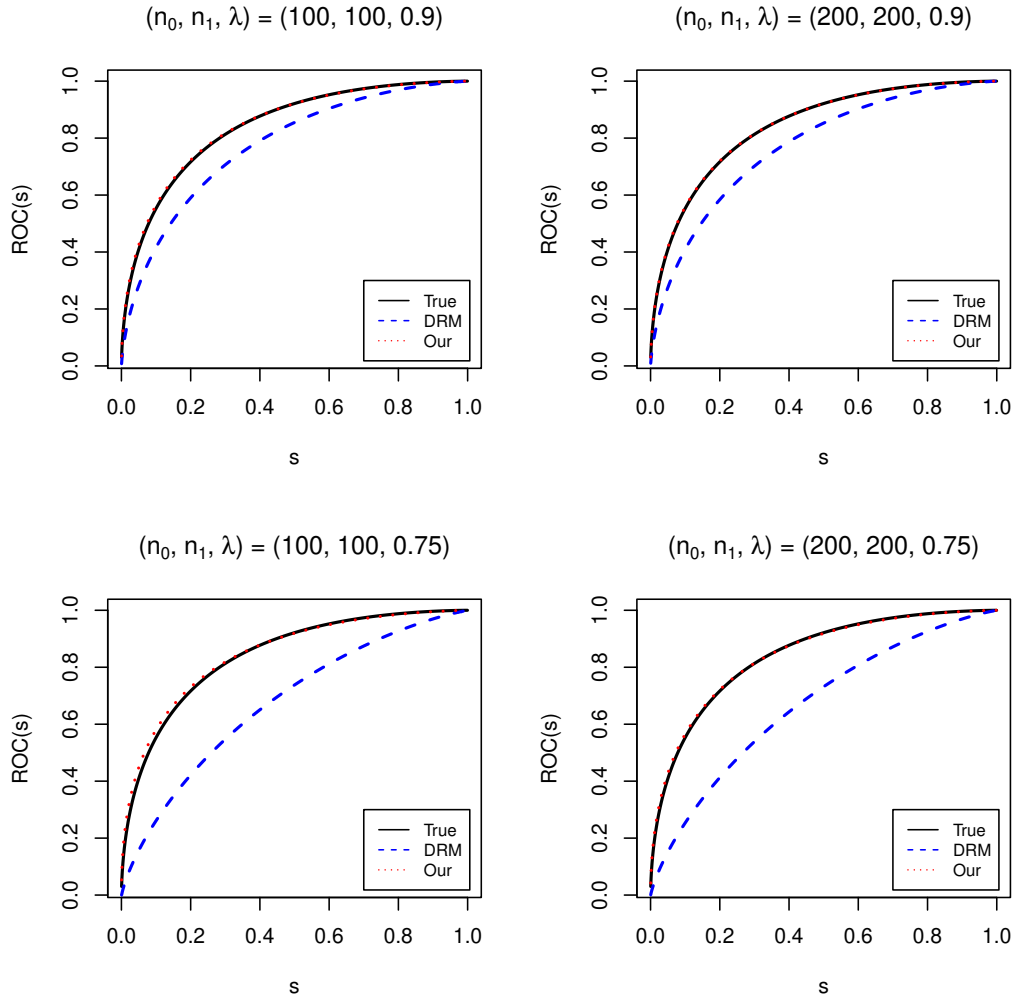


Figure 4.2: True ROC curves and average estimated ROC curves using our and DRM methods under Scenario 2.

Table 4.4: RB (%) and MSE ($\times 1000$) of $ROC(s)$ estimates for $s = 0.1$ and $s = 0.2$ under 2 scenarios

$ROC(s)$ s	(n_0, n_1, λ)	Scenario 1		Scenario 2		MSE
		Method	RB	MSE	RB	
0.1	(100, 100, 0.9)	Our	0.248	8.519	2.052	8.608
		DRM	-24.631	18.479	-24.795	22.726
0.1	(200, 200, 0.9)	Our	0.658	4.292	0.898	4.291
		DRM	-24.698	16.668	-26.032	22.709
0.1	(100, 100, 0.75)	Our	1.182	20.300	4.937	21.710
		DRM	-51.579	66.871	-52.816	87.663
0.1	(200, 200, 0.75)	Our	0.755	10.782	2.201	13.166
		DRM	-52.266	67.396	-54.184	90.967
$ROC(s)$ s	(n_0, n_1, λ)	Scenario 1		Scenario 2		MSE
		Method	RB	MSE	RB	
0.2	(100, 100, 0.9)	Our	-0.450	6.875	0.853	5.835
		DRM	-17.945	18.222	-17.707	19.858
0.2	(200, 200, 0.9)	Our	0.071	3.414	0.312	3.047
		DRM	-17.993	16.255	-18.556	19.675
0.2	(100, 100, 0.75)	Our	-0.769	16.525	1.814	14.957
		DRM	-40.647	76.229	-41.210	90.670
0.2	(200, 200, 0.75)	Our	-0.322	8.728	0.569	9.165
		DRM	-41.246	76.578	-42.434	94.302

Table 4.5: CP and AL of estimated CIs for $ROC(s)$ with $s = 0.1$ and $s = 0.2$ under 2 scenarios

$ROC(s)$		Scenario 1		Scenario 2	
s	(n_0, n_1, λ)	CP	AL	CP	AL
0.1	(100, 100, 0.9)	0.957	0.331	0.940	0.341
0.1	(200, 200, 0.9)	0.941	0.241	0.944	0.250
0.1	(100, 100, 0.75)	0.973	0.492	0.969	0.498
0.1	(200, 200, 0.75)	0.953	0.396	0.944	0.408
$ROC(s)$		Scenario 1		Scenario 2	
s	(n_0, n_1, λ)	CP	AL	CP	AL
0.2	(100, 100, 0.9)	0.953	0.296	0.947	0.283
0.2	(200, 200, 0.9)	0.938	0.216	0.946	0.209
0.2	(100, 100, 0.75)	0.971	0.446	0.969	0.418
0.2	(200, 200, 0.75)	0.952	0.354	0.943	0.339

4.5 Real data application

Recall the dataset on Duchenne Muscular Dystrophy (DMD) analyzed in Section 2.4, which includes four biomarkers: CK, LD, PK, and H. We assume that the disease statuses in the original data are correctly identified. While Section 2.4 focuses on ROC curve analysis for the single biomarker CK, in this section we consider all four biomarkers jointly and estimate the ROC curve and AUC based on their combination.

We first apply the DRM method of Qin & Zhang (2010) to the original dataset to estimate the ROC curve and AUC, which serve as benchmarks. To evaluate the effectiveness of our method in adjusting for disease status misclassification, we randomly switch the labels of n_s observations from each of the truly diseased and truly healthy groups. We consider two scenarios, where n_s can take value 5 or 10. Recall from Section 2.4 that the sample sizes for the healthy and diseased group are $n_0 = 127$ and $n_1 = 67$, respectively.

Since n_s is known, λ_0 and λ_1 can be computed as $1 - n_s/n_0$ and $1 - n_s/n_1$, respectively. This process is repeated 1,000 times, generating 1,000 shuffled datasets. For each shuffled dataset, we apply both our method and the DRM method.

The mean and 95% CIs for the AUC, based on 1,000 shuffled datasets, are summarized in Table 4.6. As shown in the table, the benchmark method yields the highest estimated AUC for both $n_s = 5$ and 10. The average AUCs estimated by our method are slightly lower than the benchmark AUC, with only minor differences. Importantly, the 95% CIs from our method successfully cover the benchmark AUC in both cases. As expected, the lengths of the CIs increase with larger values of n_s , reflecting the greater uncertainty introduced by more label switching. In contrast, the average AUCs estimated by the DRM method are noticeably lower than the benchmark AUC, and the corresponding CIs fail to include the benchmark value in either scenario.

Table 4.6: Estimated mean and 95% CI for the AUC based on 1000 shuffled datasets

Method	$n_s = 5$		$n_s = 10$	
	Mean	95% CI	Mean	95% CI
Benchmark	0.951		0.951	
Our	0.944	[0.926,0.963]	0.944	[0.919,0.973]
DRM	0.893	[0.858,0.924]	0.838	[0.798,0.878]

Figure 4.3 presents the estimated ROC curves obtained using both our method and the DRM method, alongside the benchmark ROC curve, for $n_s = 5$ and 10. In the left panel ($n_s = 5$), the ROC curve estimated by our method closely aligns with the benchmark curve, and the associated confidence band (CB) successfully covers the benchmark. In contrast, the DRM method yields a curve with noticeably lower sensitivity across all specificity levels. A similar pattern is observed in the right panel ($n_s = 10$); however, the estimated 95% CB is wider, reflecting greater uncertainty. The discrepancy between the DRM and

benchmark ROC curves is also more pronounced when $n_s = 10$, further highlighting the robustness of our method to label misclassification.

In summary, the proposed method consistently estimates the ROC curve and AUC even in the presence of disease status misclassification, as demonstrated in both simulation studies and real data analysis.

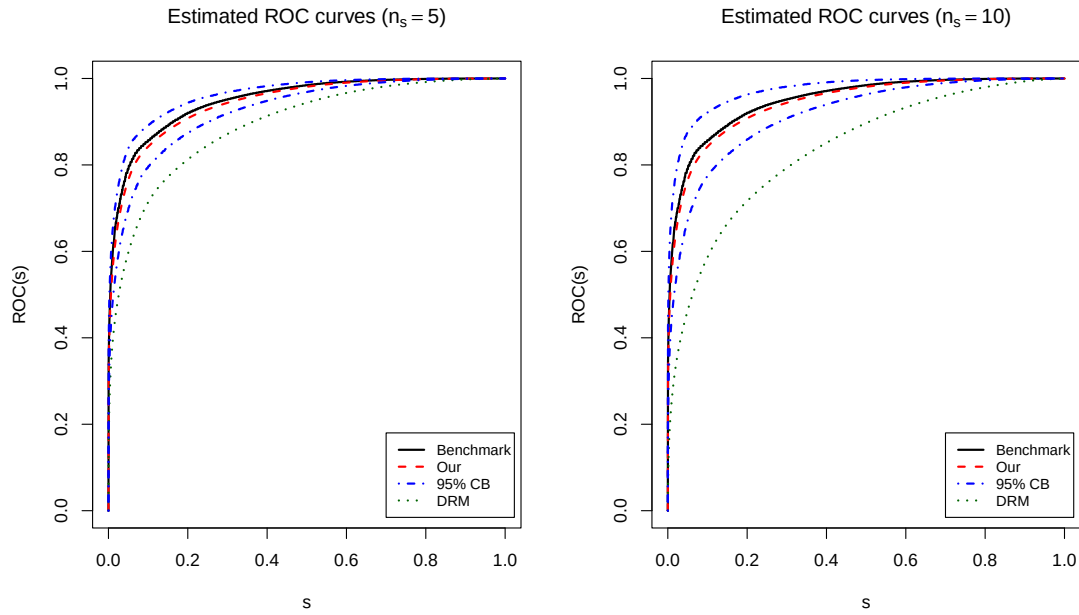


Figure 4.3: Estimated ROC curves and 95% CB for DMD dataset.

4.6 Technical details

4.6.1 Proof of Theorem 4.1

Let $g_0 = \lambda_0 f_0 + (1 - \lambda_0) f_1$ and $g_1 = \lambda_1 f_1 + (1 - \lambda_1) f_0$ denote the probability density functions (pdfs) of biomarkers for individuals with $R = 0$ and $R = 1$, respectively, where f_0 and f_1 are the pdfs of biomarkers for truly healthy and truly diseased individuals, respectively. Since we have random samples from both the $R = 0$ and $R = 1$ groups, g_0 and g_1 are identifiable. Therefore, the ratio $g_1(\mathbf{x})/g_0(\mathbf{x})$ is also identifiable.

Suppose we have two pairs of parameters, $(\alpha, \boldsymbol{\beta})$ and $(\alpha^*, \boldsymbol{\beta}^*)$, that yield the same value of $g_1(\mathbf{x})/g_0(\mathbf{x})$ for all $\mathbf{x}$. That is, for all $\mathbf{x}$, we have

$$\frac{\lambda_1 \exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{x}) \} + (1 - \lambda_1)}{\lambda_0 + (1 - \lambda_0) \exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{x}) \}} = \frac{\lambda_1 \exp \{ \alpha^* + \boldsymbol{\beta}^{*T} \mathbf{q}(\mathbf{x}) \} + (1 - \lambda_1)}{\lambda_0 + (1 - \lambda_0) \exp \{ \alpha^* + \boldsymbol{\beta}^{*T} \mathbf{q}(\mathbf{x}) \}}. \quad (4.14)$$

After some algebraic manipulation, we obtain

$$\begin{aligned} & \lambda_0 \lambda_1 \exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{x}) \} + (1 - \lambda_0) \lambda_1 + (1 - \lambda_0)(1 - \lambda_1) \exp \{ \alpha^* + \boldsymbol{\beta}^{*T} \mathbf{q}(\mathbf{x}) \} \\ &= \lambda_0 \lambda_1 \exp \{ \alpha^* + \boldsymbol{\beta}^{*T} \mathbf{q}(\mathbf{x}) \} + \lambda_0(1 - \lambda_1) + (1 - \lambda_0)(1 - \lambda_1) \exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{x}) \}, \end{aligned}$$

which implies that

$$\begin{aligned} & \{ (1 - \lambda_0)(1 - \lambda_1) - \lambda_0 \lambda_1 \} \exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{x}) \} \\ &= \{ (1 - \lambda_0)(1 - \lambda_1) - \lambda_0 \lambda_1 \} \exp \{ \alpha^* + \boldsymbol{\beta}^{*T} \mathbf{q}(\mathbf{x}) \}. \end{aligned}$$

Under Assumption 2,

$$\{(1 - \lambda_0)(1 - \lambda_1) - \lambda_0\lambda_1\} = 1 - \lambda_0 - \lambda_1 < 0.$$

Therefore, it follows that

$$\exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{x}) \} = \exp \{ \alpha^* + \boldsymbol{\beta}^{*T} \mathbf{q}(\mathbf{x}) \},$$

for all $\mathbf{x}$.

As assumed, the components in $(1, \mathbf{q}(\mathbf{x}))$ are stochastically linearly independent. We must have $\alpha = \alpha^*$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^*$; that is, $(\alpha, \boldsymbol{\beta})$ are identifiable in model (4.1).

Since g_0 and $(\alpha, \boldsymbol{\beta})$ are identifiable, the pdfs

$$f_0 = \frac{g_0(\mathbf{x})}{\lambda_0 + (1 - \lambda_0) \exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{x}) \}}$$

and

$$f_1(\mathbf{x}) = \exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{x}) \} f_0(\mathbf{x})$$

are also identifiable. Therefore, F_0 and F_1 are identifiable, which completes the proof.

4.6.2 Details of the M-step

Recall that

$$\begin{aligned}
& Q^{(m)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)}) \\
&= \sum_{i=1}^{n_0} \left[w_{i,0}^{(m)} \log(\lambda_0) + (1 - w_{i,0}^{(m)}) \log(1 - \lambda_0) + (1 - w_{i,0}^{(m)}) \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{X}_i)\} \right] \\
&\quad + \sum_{j=1}^{n_1} \left[(1 - w_{j,1}^{(m)}) \log(1 - \lambda_1) + w_{j,1}^{(m)} \log(\lambda_1) + w_{j,1}^{(m)} \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{Y}_j)\} \right] \\
&\quad + \sum_{k=1}^n \log(p_k),
\end{aligned}$$

and

$$\boldsymbol{\theta}^{(m)} = \arg \max_{\boldsymbol{\theta}} Q^{(m)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)}),$$

subject to the constraint (4.6). We apply the Lagrange multiplier method for the above optimization problem.

Define

$$\begin{aligned}
& \mathcal{L}^{(m)}(\boldsymbol{\theta}, \xi_1, \xi_2) \\
&= Q^{(m)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)}) + \xi_1 \left(\sum_{k=1}^n p_k - 1 \right) + \xi_2 \left(\sum_{k=1}^n p_k [\exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k) \} - 1] \right) \\
&= \sum_{i=1}^{n_0} \left[w_{i,0}^{(m)} \log(\lambda_0) + \{1 - w_{i,0}^{(m)}\} \log(1 - \lambda_0) + \{1 - w_{i,0}^{(m)}\} \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{X}_i)\} \right] \\
&\quad + \sum_{j=1}^{n_1} \left[\{1 - w_{j,1}^{(m)}\} \log(1 - \lambda_1) + w_{j,1}^{(m)} \log(\lambda_1) + w_{j,1}^{(m)} \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{Y}_j)\} \right] \\
&\quad + \sum_{k=1}^n \log(p_k) + \xi_1 \left(\sum_{k=1}^n p_k - 1 \right) + \xi_2 \left(\sum_{k=1}^n p_k [\exp \{ \alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k) \} - 1] \right),
\end{aligned}$$

where ξ_1 and ξ_2 are the Lagrange multipliers.

Note that $\boldsymbol{\theta}^{(m)}$ and the corresponding Lagrange multipliers $(\xi_1^{(m)}, \xi_2^{(m)})$ need to solve the following equations:

$$\frac{\partial \mathcal{L}^{(m)}}{\partial \xi_1} = \sum_{k=1}^n p_k - 1 = 0, \quad (4.15)$$

$$\frac{\partial \mathcal{L}^{(m)}}{\partial \xi_2} = \sum_{k=1}^n p_k [\exp \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k)\} - 1] = 0, \quad (4.16)$$

$$\frac{\partial \mathcal{L}^{(m)}}{\partial p_k} = \xi_1 + \xi_2 [\exp \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k)\} - 1] + \frac{1}{p_k} = 0, \quad (4.17)$$

$$\frac{\partial \mathcal{L}^{(m)}}{\partial \alpha} = \sum_{i=1}^{n_0} (1 - w_{i,0}^{(m)}) + \sum_{j=1}^{n_1} w_{j,1}^{(m)} + \xi_2 \sum_{k=1}^n p_k \exp \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k)\} = 0, \quad (4.18)$$

and

$$\frac{\partial \mathcal{L}^{(m)}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n_0} (1 - w_{i,0}^{(m)}) \mathbf{q}(\mathbf{X}_i) + \sum_{j=1}^{n_1} w_{j,1}^{(m)} \mathbf{q}(\mathbf{Y}_j) + \xi_2 \sum_{k=1}^n p_k \exp \{\alpha + \boldsymbol{\beta}^T \mathbf{q}(\mathbf{T}_k)\} \mathbf{q}(\mathbf{T}_k) = \mathbf{0}. \quad (4.19)$$

Combining (4.15)–(4.18), we obtain

$$\xi_1^{(m)} = -n \quad \text{and} \quad \xi_2^{(m)} = - \sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} - \sum_{j=1}^{n_1} w_{j,1}^{(m)}, \quad (4.20)$$

which, together with (4.17), implies that

$$p_k^{(m)} = \frac{1}{n + \left[\sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} + \sum_{j=1}^{n_1} w_{j,1}^{(m)} \right] \left[\exp \{\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(\mathbf{T}_k)\} - 1 \right]}. \quad (4.21)$$

Recall

$$\alpha^* = \alpha + \log \left[\frac{\sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} + \sum_{j=1}^{n_1} w_{j,1}^{(m)}}{n - \sum_{i=1}^{n_0} \{1 - w_{i,0}^{(m)}\} - \sum_{j=1}^{n_1} w_{j,1}^{(m)}} \right]$$

and

$$\begin{aligned} h^{(m)}(\alpha^*, \beta) &= \sum_{i=1}^{n_0} (1 - w_{i,0}^{(m)}) \{\alpha^* + \beta^T \mathbf{q}(\mathbf{X}_i)\} \\ &\quad + \sum_{j=1}^{n_1} w_{j,1}^{(m)} \{\alpha^* + \beta^T \mathbf{q}(\mathbf{Y}_j)\} \\ &\quad - \sum_{k=1}^n \log[1 + \exp\{\alpha^* + \beta^T \mathbf{q}(\mathbf{T}_k)\}]. \end{aligned} \quad (4.22)$$

Plugging (4.20) and (4.21) to (4.18) and (4.19), we can verify that $\{\alpha^{*(m)}, \beta^{(m)}\}$ solve the equations:

$$\frac{\partial h^{(m)}(\alpha^*, \beta)}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial h^{(m)}(\alpha^*, \beta)}{\partial \beta} = \mathbf{0}.$$

That is, $\{\alpha^{*(m)}, \beta^{(m)}\}$ is the stationary point of $h^{(m)}(\alpha^*, \beta)$. Due to the concavity of $h^{(m)}(\alpha^*, \beta)$, we have

$$\{\alpha^{*(m)}, \beta^{(m)}\} = \arg \max_{\alpha^*, \beta} h^{(m)}(\alpha^*, \beta).$$

4.6.3 Proof of Theorem 4.2

Note that

$$\begin{aligned}
& \log L(\boldsymbol{\theta}^{(m)}) - \log L(\boldsymbol{\theta}^{(m-1)}) \\
&= \sum_{i=1}^{n_0} \log \left(\frac{\lambda_0 + (1 - \lambda_0) \exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(X_i) \right]}{\lambda_0 + (1 - \lambda_0) \exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(X_i) \right]} \right) \\
&+ \sum_{j=1}^{n_1} \log \left(\frac{\lambda_1 \exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(Y_j) \right] + (1 - \lambda_1)}{\lambda_1 \exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(Y_j) \right] + (1 - \lambda_1)} \right) \\
&+ \sum_{k=1}^n \log \left\{ \frac{p_k^{(m)}}{p_k^{(m-1)}} \right\}. \tag{4.23}
\end{aligned}$$

We first observe that

$$\begin{aligned}
& \sum_{i=1}^{n_0} \log \left(\frac{\lambda_0 + (1 - \lambda_0) \exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(X_i) \right]}{\lambda_0 + (1 - \lambda_0) \exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(X_i) \right]} \right) \\
&= \sum_{i=1}^{n_0} \log \left(w_{i,0}^{(m)} + \frac{\exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(X_i) \right]}{\exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(X_i) \right]} \left\{ 1 - w_{i,0}^{(m)} \right\} \right) \\
&\geq \sum_{i=1}^{n_0} w_{i,0}^{(m)} \log(1) \\
&+ \sum_{i=1}^{n_0} \left\{ 1 - w_{i,0}^{(m)} \right\} \log \left(\frac{\exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(X_i) \right]}{\exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(X_i) \right]} \right) \\
&= \sum_{i=1}^{n_0} \left\{ 1 - w_{i,0}^{(m)} \right\} \log \left(\frac{\exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(X_i) \right]}{\exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(X_i) \right]} \right), \tag{4.24}
\end{aligned}$$

where the inequality follows from the concavity of the logarithm function.

Similarly, we have

$$\begin{aligned}
& \sum_{j=1}^{n_1} \log \left(\frac{\lambda_1 \exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(Y_j) \right] + (1 - \lambda_1)}{\lambda_1 \exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(Y_j) \right] + (1 - \lambda_1)} \right) \\
&= \sum_{j=1}^{n_1} w_{j,1}^{(m)} \log \left(\frac{\exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(Y_j) \right]}{\exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(Y_j) \right]} \right). \tag{4.25}
\end{aligned}$$

Combining equations (4.23)–(4.25), we obtain

$$\begin{aligned}
& \log L(\boldsymbol{\theta}^{(m)}) - \log L(\boldsymbol{\theta}^{(m-1)}) \\
&= \sum_{i=1}^{n_0} \left\{ 1 - w_{i,0}^{(m)} \right\} \log \left(\frac{\exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(X_i) \right]}{\exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(X_i) \right]} \right) \\
&+ \sum_{j=1}^{n_1} w_{j,1}^{(m)} \log \left(\frac{\exp \left[\alpha^{(m)} + \{\boldsymbol{\beta}^{(m)}\}^T \mathbf{q}(Y_j) \right]}{\exp \left[\alpha^{(m-1)} + \{\boldsymbol{\beta}^{(m-1)}\}^T \mathbf{q}(Y_j) \right]} \right) \\
&+ \sum_{k=1}^n \log \left\{ \frac{p_k^{(m)}}{p_k^{(m-1)}} \right\} \\
&= Q^{(m)}(\boldsymbol{\theta}^{(m)} | \boldsymbol{\theta}^{(m-1)}) - Q^{(m-1)}(\boldsymbol{\theta}^{(m)} | \boldsymbol{\theta}^{(m-1)}) \\
&\geq 0.
\end{aligned}$$

This completes the proof.

Chapter 5

Discussion and Future Work

In this chapter, we provide a brief summary of the research problems and methodologies discussed in Chapters 2–4, and outline several potential directions for future research.

5.1 Summary of the thesis

In Chapter 2, we proposed a Bernstein polynomial method to estimate the cdfs of biomarkers in a two-sample setting under the likelihood ratio ordering assumption. Building on this, we addressed the estimation of the ROC curve and its summary indices. Using empirical process theory, we established the consistency of our estimators. Through simulation studies, we compared our method with existing approaches and found that it performs well in estimating both the ROC curve and its summary indices. We further applied our method to the DMD dataset to evaluate the diagnostic performance of the CK biomarker. To facilitate practical application, we implemented our numerical algorithm in an R package named BPLR, which is publicly available for use. The work presented in this chapter has

been developed into a research paper (Hu et al. 2023), which was published in *Statistics in Medicine*.

In Chapter 3, we proposed a likelihood-based approach for estimating the ROC curve and AUC in the presence of non-ignorable missingness in disease status. We established the identifiability of the model parameters under mild conditions and demonstrated the asymptotic normality of the proposed estimators for both the ROC curve and the AUC. Simulation studies highlighted the advantages of our method over existing approaches. Finally, we applied the proposed method to the Alzheimer’s disease dataset from the NACC to evaluate the diagnostic accuracy of the MMSE test. The work presented in this chapter has been submitted for publication in the *Canadian Journal of Statistics*, and we have been invited to revise the manuscript.

In Chapter 4, we considered the optimal combination of multiple biomarkers under the DRM model (4.1) and Assumptions 1–2 in Section 4.2.1, in the presence of an imperfect reference standard. We established the identifiability of the model parameters in (4.1) under mild conditions and proposed a maximum empirical likelihood method to estimate the optimal ROC curve and the corresponding AUC. Furthermore, we developed an EM algorithm for numerical implementation and demonstrated its monotonicity. A bootstrap method was introduced to construct confidence interval for the AUC and confidence band for the ROC curve. Simulation studies and real data applications demonstrated that the proposed method could consistently estimate the optimal ROC and the corresponding AUC, even when mislabeling was present due to the imperfect reference standard.

5.2 Future work

We observe that within the framework of Chapter 2, there are many interesting topics that can be explored in the future. First, we established the consistency of our proposed estimators, but not the convergence rates and the asymptotic distributions; if we are able to establish the asymptotic distributions of them in the future, we can construct the confidence bands/intervals for $F_0(\cdot)$, $F_1(\cdot)$, $ROC(\cdot)$, and other summary indices. Second, we used the BIC criterion to select the optimal order of the Bernstein polynomial basis functions. However, BIC may over-penalize the empirical likelihood function under certain circumstances. For example, in panel (a) of Figure 2.1, the linear trend suggests that the penalty term dominates $BIC(N)$. Alternative selection criteria, such as Akaike's information criterion (AIC) (Bozdogan 1987) or generalized cross-validation (Golub et al. 1979), may be more suitable and warrant further investigation in future work. Third, we proposed to employ the Youden index as the criterion, and maximize it to determine the optimal cutoff point; however, there may exist other criteria; for example, the closest to (0,1) criterion and equal sensitivity and specificity criterion (Sande et al. 2021). It could be interesting to explore how to extend our method to accommodate these criteria. Fourth, in Remark 2.3, we suggested to include the log transformed biomarker in our model and have applied this strategy in both the simulation and real data analysis. We observe that there may exist other reasonable transformations that can be incorporated in our method, for example, Yang et al. (2021) proposed a transformation method based on the likelihood ratio and discussed the scenarios under which that transformation is applicable. Also, in Section 2.3, our proposed BP method demonstrates advantages over existing approaches in estimating the ROC curve, AUC, Youden index, and optimal cutoff point, owing to the incorporation of likelihood ratio ordering in the statistical model. It would be interesting

to verify whether this advantage is indeed due to the constraint by comparing the performance of our proposed method with and without the likelihood ratio ordering. Other potential areas for future research include adapting our method to accommodate discrete biomarkers.

In Chapter 3, we considered the problem setting in which disease statuses are subject to non-ignorable missingness. Several potential extensions of our method can be considered. First, we assume a logistic regression structure for both the disease and verification models. The effectiveness of our methodology relies on the correct specification of both the disease and verification models. To mitigate potential model misspecification, more robust semi-parametric or non-parametric models, such as the model proposed in [Tian et al. \(2025\)](#), could be explored to relax the parametric assumptions. Second, we focused on binary responses. However, in practice, responses with multiple categories often arise. For example, since Alzheimer’s disease can present with varying levels of severity, [Li et al. \(2012\)](#) analyzed an AD dataset by considering three distinct classes of disease status. Methodologies for handling multiple classes have been well studied in the literature; see [Li & Fine \(2008\)](#), [Li & Zhou \(2009\)](#), and [Wan & Zhang \(2013\)](#) for reviews and recent developments. We plan to extend the proposed method to accommodate non-ignorable missingness in disease status with multiple categories, and we will explore this direction in future research.

In Chapter 4, we addressed the problem of combining multiple biomarkers when the accuracy of the imperfect reference standard is known. However, several important questions remain open for future research. First, in practice, the accuracy of the imperfect reference standard may be unknown or biased. In such cases, sensitivity analyses can be conducted to assess the robustness of the results; see [Saltelli & Annoni \(2011\)](#) and [Christopher Frey & Patil \(2002\)](#) for reviews of existing methods. An interesting direction for future work would be to extend the estimation framework to account for unknown or biased misclassifi-

cation rates. Second, we have not established the asymptotic properties of our ROC curve and AUC estimators. A detailed investigation of their theoretical properties, including formal proofs and the necessary regularity conditions, is left for future research. Third, in Section 4.5, we combined four biomarkers from the DMD dataset to obtain an optimal combination of biomarkers. In practice, however, the number of available biomarkers can be large. Therefore, incorporating variable selection into our methodology is desirable to identify and combine only the most informative biomarkers. Finally, similar to Chapter 3, many diseases in practice are not limited to binary outcomes. Extending our method to accommodate categorical response variables is also an important direction for future research.

Bibliography

- Andrews, D. F. & Herzberg, A. M. (2012), *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Springer, New York.
- Bantis, L. E., Nakas, C. T. & Reiser, B. (2019), ‘Construction of confidence intervals for the maximum of the Youden index and the corresponding cutoff point of a continuous biomarker’, *Biometrical Journal* **61**, 138–156.
- Begg, C. B. & Greenes, R. A. (1983), ‘Assessment of diagnostic tests when disease verification is subject to selection bias’, *Biometrics* **39**, 207–215.
- Bozdogan, H. (1987), ‘Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions’, *Psychometrika* **52**, 345–370.
- Chen, B., Li, P., Qin, J. & Yu, T. (2016), ‘Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests’, *Journal of the American Statistical Association* **111**, 861–874.
- Chen, J., Li, P., Liu, Y. & Zidek, J. V. (2021), ‘Composite empirical likelihood for multi-sample clustered data’, *Journal of Nonparametric Statistics* **33**, 60–81.
- Chen, J. & Liu, Y. (2013), ‘Quantile and quantile-function estimations under density ratio model’, *Annals of Statistics* **41**, 1669–1692.

- Christopher Frey, H. & Patil, S. R. (2002), ‘Identification and review of sensitivity analysis methods’, *Risk analysis* **22**, 553–578.
- Du, Z., Du, P. & Liu, A. (2024), ‘Likelihood ratio combination of multiple biomarkers via smoothing spline estimated densities’, *Statistics in Medicine* **43**, 1372–1383.
- Dykstra, R., Kocher, S. & Robertson, T. (1995), ‘Inference for likelihood ratio ordering in the two-sample problem’, *Journal of the American Statistical Association* **90**, 1034–1040.
- Eguchi, S. & Copas, J. (2002), ‘A class of logistic-type discriminant functions’, *Biometrika* **89**, 1–22.
- Fluss, R., Reiser, B., Faraggi, D. & Rotnitzky, A. (2009), ‘Estimation of the ROC curve under verification bias’, *Biometrical Journal* **51**, 475–490.
- Folstein, M. F., Folstein, S. E. & McHugh, P. R. (1975), ‘“Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician’, *Journal of Psychiatric Research* **12**, 189–198.
- Friedman, J., Hastie, T. & Tibshirani, R. (2010), ‘Regularization paths for generalized linear models via coordinate descent’, *Journal of Statistical Software* **33**, 1–22.
- Gneiting, T. & Vogel, P. (2018), ‘Receiver operating characteristic (ROC) curves’, *arXiv preprint arXiv:1809.04808*.
- Golub, G. H., Heath, M. & Wahba, G. (1979), ‘Generalized cross-validation as a method for choosing a good ridge parameter’, *Technometrics* **21**, 215–223.
- Hosmer, D. W., Hosmer, T., Le Cessie, S. & Lemeshow, S. (1997), ‘A comparison of goodness-of-fit tests for the logistic regression model’, *Statistics in Medicine* **16**, 965–980.

- Hsieh, F. & Turnbull, B. (1996), ‘Nonparametric methods for evaluating diagnostic tests’, *Statistica Sinica* **6**, 47–62.
- Hu, D., Yuan, M., Yu, T. & Li, P. (2023), ‘Statistical inference for the two-sample problem under likelihood ratio ordering, with application to the ROC curve estimation’, *Statistics in Medicine* **42**, 3649–3664.
- Huang, Y. & Sanda, M. G. (2022), ‘Linear biomarker combination for constrained classification’, *Annals of Statistics* **50**, 2793–2815.
- Jokiel-Rokita, A. & Pulit, M. (2013), ‘Nonparametric estimation of the ROC curve based on smoothed empirical distribution functions’, *Statistics and Computing* **23**, 703–712.
- Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Springer, New York.
- Li, J. & Fine, J. P. (2008), ‘ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies’, *Biostatistics* **9**, 566–576.
- Li, J. & Zhou, X. (2009), ‘Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface’, *Journal of Statistical Planning and Inference* **139**, 4133–4142.
- Li, J., Zhou, X. & Fine, J. P. (2012), ‘A regression approach to ROC surface, with applications to Alzheimer’s disease’, *Science China Mathematics* **55**, 1583–1595.
- Li, P., Qin, J. & Liu, Y. (2023), ‘Instability of inverse probability weighting methods and a remedy for nonignorable missing data’, *Biometrics* **79**, 3215–3226.

- Lin, H., Zhou, L., Peng, H. & Zhou, X. (2011), ‘Selection and combination of biomarkers using ROC method for disease classification and prediction’, *Canadian Journal of Statistics* **39**, 324–343.
- Lin, H., Zhou, X. & Li, G. (2012), ‘A direct semiparametric receiver operating characteristic curve regression with unknown link and baseline functions’, *Statistica Sinica* **22**, 1427–1456.
- Liu, D. & Zhou, X. (2010), ‘A model for adjusting for nonignorable verification bias in estimation of ROC curve and its area with likelihood-based approach’, *Biometrics* **66**, 1119–1128.
- Liu, Y., Li, P. & Qin, J. (2022), ‘Full-semiparametric-likelihood-based inference for non-ignorable missing data’, *Statistica Sinica* **32**, 271–292.
- Lorentz, G. G. (1986), *Bernstein Polynomials*, second edn, Chelsea Publishing Company, New York.
- Ma, S. & Huang, J. (2005), ‘Regularized ROC method for disease classification and biomarker selection with microarray data’, *Bioinformatics* **21**, 4356–4362.
- McIntosh, M. W. & Pepe, M. S. (2002), ‘Combining several screening tests: optimality of the risk score’, *Biometrics* **58**, 657–664.
- Meisner, A., Carone, M., Pepe, M. S. & Kerr, K. F. (2021), ‘Combining biomarkers by maximizing the true positive rate for a fixed false positive rate’, *Biometrical Journal* **63**, 1223–1240.
- Okoye, G. A., Kamara, H. I., Strobeck, M., Mellman, T. A., Kwagyan, J., Sullivan, A.,

- Byrd, A. S., Shokrani, B. & Mighty, H. E. (2022), ‘Diagnostic accuracy of a rapid diagnostic test for the early detection of COVID-19’, *Journal of Clinical Virology* **147**, 105023.
- Owen, A. B. (2001), *Empirical Likelihood*, Chapman and Hall/CRC, New York.
- Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York.
- Percy, M. E., Andrews, D. F. & Thompson, M. W. (1982), ‘Duchenne muscular dystrophy carrier detection using logistic discrimination: Serum creatine kinase, hemopexin, pyruvate kinase, and lactate dehydrogenase in combination’, *American Journal of Medical Genetics* **13**, 27–38.
- Qin, J. (2017), *Biased Sampling, Over-identified Parameter Problems and Beyond*, Springer, New York.
- Qin, J. & Zhang, B. (2003), ‘Using logistic regression procedures for estimating receiver operating characteristic curves’, *Biometrika* **90**, 585–596.
- Qin, J. & Zhang, B. (2010), ‘Best combination of multiple diagnostic tests for screening purposes’, *Statistics in Medicine* **29**, 2905–2919.
- Saltelli, A. & Annoni, P. (2011), Sensitivity analysis, in ‘International encyclopedia of statistical science’, Springer, Berlin, Heidelberg, pp. 1298–1301.
- Sande, S. Z., Seng, L., Li, J. & D’Agostino, R. (2021), ‘Statistical learning in medical research with decision threshold and accuracy evaluation’, *Journal of Data Science* **19**, 634–657.
- Sun, J., Tang, C., Xie, W. & Zhou, X. (2024), ‘Nonparametric receiver operating characteristic curve analysis with an imperfect gold standard’, *Biometrics* **80**, ujae063.

- Sun, Y., Sang, P., Tian, Q. & Li, P. (2025), ‘Likelihood-based nonparametric receiver operating characteristic curve analysis in the presence of imperfect reference standard’, *arXiv preprint arXiv: 2502.08569* .
- Taghipour Zahir, S., Dehghani, F. et al. (2011), ‘Evaluation of diagnostic value of CT scan and MRI in brain tumors and comparison with biopsy’, *Iranian Journal of Pediatric Hematology and Oncology* **1**, 121–125.
- Tian, Q., Zeng, D. & Zhao, J. (2025), ‘Identification and efficient estimation in regression analysis with response missing not at random’, *Statistica Sinica* . In press.
- Tombaugh, T. N. & McIntyre, N. J. (1992), ‘The mini-mental state examination: a comprehensive review’, *Journal of the American Geriatrics Society* **40**, 922–935.
- van der Vaart, A. (2000), *Weak Convergence and Empirical Processes with Application to Statistics*, Cambridge University Press, New York.
- van der Vaart, A. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes with Application to Statistics*, Springer, New York.
- Wan, S. & Zhang, B. (2013), ‘Semiparametric ROC surface estimation for continuous diagnostic tests via polytomous logistic regression procedures’, *Journal of Statistical Computation and Simulation* **83**, 2195–2205.
- Wang, C. (2017), *Empirical Likelihood and Bootstrap Inference with Constraints*, PhD thesis, University of Waterloo.
- Wang, D., Tian, L. & Zhao, Y. (2017), ‘Smoothed empirical likelihood for the Youden index’, *Computational Statistics & Data Analysis* **115**, 1–10.

- Wang, J. & Ghosh, S. (2012), ‘Shape restricted nonparametric regression with Bernstein polynomials’, *Computational Statistics & Data Analysis* **56**, 2729–2741.
- Yamayoshi, S., Sakai-Tagawa, Y., Koga, M., Akasaka, O., Nakachi, I., Koh, H., Maeda, K., Adachi, E., Saito, M., Nagai, H. et al. (2020), ‘Comparison of rapid antigen tests for COVID-19’, *Viruses* **12**, 1420.
- Yang, J., Kuan, P.-F. & Li, J. (2021), ‘Transformation based on likelihood ratio’, *Statistical Methods in Medical Research* **30**, 354–356.
- Yin, J. & Tian, L. (2014), ‘Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index’, *Computational Statistics & Data Analysis* **77**, 1–13.
- Youden, W. J. (1950), ‘Index for rating diagnostic tests’, *Cancer* **3**, 32–35.
- Yu, T., Li, P. & Qin, J. (2017), ‘Density estimation in the two-sample problem with likelihood ratio ordering’, *Biometrika* **104**, 141–152.
- Yu, W., Kim, J. K. & Park, T. (2018), ‘Estimation of area under the ROC curve under nonignorable verification bias’, *Statistica Sinica* **28**, 2149–2166.
- Yuan, M. (2021), Semiparametric Empirical Likelihood Inference under Two-sample Density Ratio Models, PhD thesis, University of Waterloo.
- Yuan, M., Li, P. & Wu, C. (2021), ‘Semiparametric inference of the Youden index and the optimal cut-off point under density ratio models’, *Canadian Journal of Statistics* **49**, 965–986.
- Zhou, X. (1998), ‘Comparing correlated areas under the ROC curves of two diagnostic tests in the presence of verification bias’, *Biometrics* **54**, 453–470.

- Zhou, X. & Castelluccio, P. (2004), ‘Adjusting for non-ignorable verification bias in clinical studies for Alzheimer’s disease’, *Statistics in Medicine* **23**, 221–230.
- Zhou, X. & Lin, H. (2008), ‘Semi-parametric maximum likelihood estimates for ROC curves of continuous-scale tests’, *Statistics in Medicine* **27**, 5271–5290.
- Zhou, X., McClish, D. K. & Obuchowski, N. A. (2011), *Statistical Methods in Diagnostic Medicine*, second edn, Wiley, New York.