2017

Semiparametric Approaches To Developing Models For Predicting Binary Outcomes Through Data And Information Integration

Xinglei Chai
University of Pennsylvania, asenna1333@gmail.com

Semiparametric Approaches To Developing Models For Predicting Binary Outcomes Through Data And Information Integration

Abstract

We developed statistical methods for evaluating the added value of biomarkers for predicting binary outcomes when biomarker data has limited availability. In the first project, we considered a cost effective study design called “two-phase study”, where data on the outcome and established risk predictors was collected for all study subjects in Phase I while biomarkers were measured only for a judiciously selected subset in Phase II. Using a logistic regression model to describe the relationship between the binary outcome and risk predictors, we developed three approaches to estimating the risk distribution and summary measures of predictive accuracy. We showed that all three estimators were consistent and asymptotically normally distributed, and compared the efficiency and robustness of the three methods through extensive simulation studies and application to an ongoing biomarker study of Gestational Diabetes. We also developed a novel sampling strategy for selecting Phase II subjects towards improved efficiency for estimating measures of predictive accuracy. In the second project, we developed a statistical method for alleviating the challenge of lack of independent data to validate biomarkers for prediction, focusing on model calibration. When a well-calibrated model with only standard predictors exists, we proposed to calibrate the new model to the existing model at the stage of model development. With data collected under a case-control study design, we developed a novel constrained maximum likelihood approach to fitting logistic regression models that brought this idea to fruition. We developed large sample theory for this method, and performed extensive simulation studies to assess the impact of constraints on the odds ratio parameter estimates. We applied our method to analyze a case-control study of breast cancer nested within the Breast Cancer Detection and Demonstration Project to evaluate the added value of mammographic density for predicting the 5-year risk of breast cancer. In the third project, we extended the statistical method developed in the second project to accommodate the cross-sectional study design. By simulation studies and the analysis of Gestational Diabetes, we demonstrated that our method ensured that the model was well calibrated.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Jinbo Chen

Subject Categories

Biostatistics

SEMIPARAMETRIC APPROACHES TO DEVELOPING MODELS FOR PREDICTING BINARY OUTCOMES THROUGH DATA AND INFORMATION INTEGRATION

Xinglei Chai

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Jinbo Chen, Associate Professor of Biostatistics

Graduate Group Chairperson

_____

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Russell T. Shinohara, Assistant Professor of Biostatistics

Andrea B. Troxel, Professor of Population Health

Emily F. Conant, Professor of Radiology

Hongzhe Li, Professor of Biostatistics

SEMIPARAMETRIC APPROACHES TO DEVELOPING MODELS FOR PREDICTING BINARY OUTCOMES THROUGH DATA AND INFORMATION INTEGRATION

ACKNOWLEDGEMENT

I would like to first thank my advisor and mentor, Dr. Jinbo Chen, for her guidance thoughout my dissertation. We have been working together for more than 4 years and she has always been patient, encouraging, willing to educate, listen and answer questions. As a fresh to research, Dr. Chen is the person who guided and led me into this field. Without Dr. Chen, it would be impossible for me to complete this work. I am also grateful to my committee members, Dr. Hongzhe Li, Dr. Russell T. Shinohara, Dr. Andrea B. Troxel, Dr. Emily Conant and Dr. Rui Xiao (my candidacy examination committee member) who provided numerous helpful suggestions for my research. I further would like to thank all the faculty members and staff in the Department of Biostatistics and Epidemiology. Their instruction, kindness and considerateness have led me to learn and grow up. I also need to thank the fellow students, Yu, Emily and Amy. We have spent much time together to get though the course work and qualifying examination. Their help made me quickly accustomed to the academic environment and life in University. Great thanks to my parents for their love and support of my PhD study. I would also like to thank my partner for her love and support during the dissertation. She is the one who cared and encouraged me when I was frustrated.

ABSTRACT

SEMIPARAMETRIC APPROACHES TO DEVELOPING MODELS FOR PREDICTING BINARY OUTCOMES THROUGH DATA AND INFORMATION INTEGRATION

Xinglei Chai

Jinbo Chen

We developed statistical methods for evaluating the added value of biomarkers for predicting binary outcomes when biomarker data has limited availability. In the first project, we considered a cost effective study design called two-phase study, where data on the outcome and established risk predictors was collected for all study subjects in Phase I while biomarkers were measured only for a judiciously selected subset in Phase II. Using a logistic regression model to describe the relationship between the binary outcome and risk predictors, we developed three approaches to estimating the risk distribution and summary measures of predictive accuracy. We showed that all three estimators were consistent and asymptotically normally distributed, and compared the efficiency and robustness of the three methods through extensive simulation studies and application to an ongoing biomarker study of Gestational Diabetes. We also developed a novel sampling strategy for selecting Phase II subjects towards improved efficiency for estimating measures of predictive accuracy. In the second project, we developed a statistical method for alleviating the challenge of lack of independent data to validate biomarkers for prediction, focusing on model calibration. When a well-calibrated model with only standard predictors exists, we proposed to calibrate the new model to the existing model at the stage of model development. With data collected under a case-control study design, we developed a novel constrained maximum likelihood approach to fitting logistic regression models that brought this idea to fruition. We developed large sample theory for this method, and performed extensive simulation studies to assess the impact of constraints on the odds ratio parameter estimates. We applied our method to analyze a case-control study of breast cancer nested within the Breast Cancer Detection and Demonstration Project to evaluate the added value of mammographic density for predicting the 5-year risk of breast cancer. In the third project, we extended the statistical method developed in the second project to accommodate the cross-sectional study design. By simulation studies and the analysis of Gestational Diabetes, we demonstrated that our method ensured that the model was well calibrated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1.

Accurate risk prediction is central to precision medicine and precision disease prevention. Modern technology has enabled widespead efforts of biomarker discovery, promising great possibilities of improving risk prediction for human diseases. For putative biomarkers, it is of ultmost interest to evaluate their added values for prediction. But this is a challenging task. The model development requires data for disease status, biomarkers, as well as established risk predictors from a sample that is of sufficiently large size to allow stable model fitting. In practice, initial investigation efforts most often afford collection of data only from a small number of subjects.

The two-phase study design, which was firstly introduced by Neyman (1938), is commonly conducted in medical research due to their economy and efficiency. In this design, data is collected in two phases (Neyman, 1938; White, 1982). In Phase I, data on the outcome and conventional risk predictors is collected for all study subjects. In Phase II, biomarkers are measured for a judiciously selected subset. In the study of gestational diabetes mellitus (GDM), a case-control study nested in the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Fetal Growth StudySingletons cohort (Zhu et al., 2016), data can be seen as being collected in two phases. In this prospective study cohort (Phase I) , data on conventional predictors such as age, race, body mass index, and family history was fully collected for 2,799 women. Blood samples were collected for the full cohort prior to the GDM screening. To study biomarkers in relation to the risk of GDM, a nested case-control subsample (Phase II) was assembled, where two controls were matched to each of the 107 cases on ethnicity, age within 2 years, and gestational weeks. Biomarkers were measured only for this case-control subset. It was of interest to develop a model for predicting the risk of GDM using both conventional risk predictors and new biomarkers. However, such two-phase studies induce incomplete data and pose significant statistical challenges both in developing a comprehensive risk model and its validation. It is well known that naive analysis using only Phase II data with the sampling procedure ignored leads to biased estimation and decreased power.

In Chapter 2, we develop statistical methods for estimating the absolute risk for a binary outcome and several summary measures of prediction accuracy with two-phase data. Specifically, we consider estimation of the receiver operating characteristic (ROC) curve and the area under the ROC

curve (AUC), which are routinely reported to gauge the discriminatory capacity of a prediction model. We also consider the proportion of cases followed (PCF; Pfeiffer and Gail, 2011) which is the probability that a case is included in a high-risk group, and the proportion needed to follow (PNF; Pfeiffer and Gail, 2011), which is the proportion of high-risk subjects that contain a specified proportion of cases. PCF and PNF measure the concentration of risk in the population and can be used to quantify the usefulness of a model in a risk-based screening program. PCF and PNF are of particular interest to our GDM study example, because the ultimate goal is to develop a model to help screen high-risk women upon validation. The current literature for analyzing two-phase data has been mainly focused on the estimation of association parameters that describe the relationship between an outcome variable and predictors. The available methods that account for biased sampling of a two-phase study largely fell into four class: semiparametric maximum likelihood methods (ML; Breslow and Holubkov, 1997; Scott and Wild, 1997; Wang and Zhou, 2010), pseudo-likelihood (PL; Breslow and Cain, 1988; Breslow and Zhao, 1988) or estimated likelihood (Carroll and Wand, 1991; Pepe and Fleming, 1991) methods, weighted likelihood methods (WL; Flanders and Greenland, 1991; Ibrahim, 1990; Lipsitz, Ibrahim, and Zhao, 1999; Robins, Rotnitzky, and Zhao, 1994), and methods based on modifying likelihood score functions for the complete data (Chatterjee, Chen, and Breslow, 2003; Reilly and Pepe, 1995). Most of these methods have been synthesized and further extended in the seminal paper (Lawless, Kalbfleisch, and Wild, 1999). Currently, there is a paucity of procedures that can be applied to the two-phase setting for estimating measures of accuracy for predicting binary outcomes. Existing methods were not able to fully accommodate stratified sampling of Phase II subjects (Huang, 2016; Huang and Pepe, 2010), limited to the estimation of AUC (Huang, 2016; Pepe, Fan, and Seymour, 2013), or focused on the failure time data (Cai and Zheng, 2012; Liu, Cai, and Zheng, 2012). To our best knowledge, no methods are yet available for estimating the key accuracy measures PCF and PNF for our GDM study example.

We consider the general two-phase sampling scheme where the probability that a subject is selected into Phase II depends only on his/her own characteristics, and adopt logistic regression models for prediction. The estimation of predictive accuracy measure requires estimation of the risk distribution, which is a function of the odds ratio (OR) parameters and the predictor distribution. For estimating the OR parameters with two-phase data, the semiparametric ML method (Lawless, Kalbfleisch, and Wild, 1999; Scott and Wild, 1997), the PL method (Breslow and Cain, 1988; Bres-

low and Zhao, 1988), and the WL method (Flanders and Greenland, 1991; Robins, Rotnitzky, and Zhao, 1994) have been widely used in practice. Each of them has its own pros and cons, and they have been implemented in widely used R packages. Therefore, we propose three corresponding methods for estimating risk distribution and predictive accuracy measures.

In this chapter, we also explore efficient sampling strategies for selecting Phase II subjects in order to improve statistical efficiency for estimating predictive accuracy measures. It has been shown that selection stratified by case-control status alone or together with a small number of discrete Phase I variables can lead to a more informative Phase II sample than that obtained via simple random sampling (Breslow and Cain, 1988; Breslow and Chatterjee, 1999). The latter selects similar numbers of Phase II subjects across strata, and is particularly attractive for increasing the efficiency for estimating OR parameters when the outcome is rare and some values of Phase I variables occur infrequently. This "balanced" design requires categorization of Phase I predictors, but there is no guidance on how to do so when multiple phase I predictors are available. One may choose to stratify by only a single Phase I predictor, or by coarse categories based on multiple predictors defined in such a way that a sufficient number of Phase I cases and controls are available for selection within each stratum. We propose to sample Phase II subjects based on a preliminary prediction model that includes only Phase I variables as predictors.

1.2.

To assess the value of biomarkers that adds to the established risk predictors to achieve improved prediction, it is important that the model is well calibrated since good calibration is essential in order to inform patients about their risks and make risk-based decision. However, the subsequent model validation is challenging because it requires data to be collected from sources independent of those for model development and these independent datasets most ofter are not readily available. Even if data for the outcome and standard risk predictors may well be available from existing studies or can be easily obtained, it is usually much more challenging to obtain biomarker data. Lack of independent validation data has been an obvious factor that hinders translation of biomarkers to clinics. For example, the breast cancer risk assessment tool (BCRAT), which was developed using data from a case-control study of breast cancer that was nested in the Breast Cancer Detection and Demonstration Project (BCDDP), was calibrated to the composite breast cancer rates reported in

the Surveillance, Epidemiology, End Results program, making it a useful tool for projecting individualized risk for the U.S. Caucasian women. Percent mammographic density was incorporated as a strong risk predictor into the BCRAT (Chen et al., 2006), and led to an increase in the area under the ROC curve that is comparable to that by incorporation of breast cancer risk SNPs identified to date. However, validation studies has yet to be conducted after more than 10 years since this updated model was published. Even if the data for validation is fully available, the sample size may limit the power for detecting lack of calibration. Recent work shows that greater than 10,000 subjects were required in order to detect meaningful differences between the predicted and observed risks in the upper tail of the risk distribution (Chatterjee et al., 2016).

Fortunately, data from multiple sources becomes more and more available that can be exploited to enhance model development and validation, which may compensate the scarce of biomarker data. The incidence rates for common diseases in the U.S. are publically available. Population level data for standard risk predictors is frequently available from national and international efforts or existing cohort studies. For example, the BCRAT risk predictors are fully represented in the National Health Interview Survey (NHIS III). The relationship between the outcome and standard risk predictors may have been assessed in multiple studies, or a risk prediction model that uses only standard predictors may have been developed and validated extensively. Data may be available for characterizing the relationship between biomarkers and standard risk predictors. Aiming to exploit these existing resources for model development and validation, we develop novel statistical methods for predicting the risk of a binary outcome using the logistic regression model.

In Chapter 3, we consider a scenario where a well calibrated model based on standard predictors exists, and focus on a central task of incorporating new risk predictors into the existing model, which henceforth is referred to as the “base model”. We consider a case-control study design, which is also a cost-effective option for recruiting subjects and collecting data on disease status, biomarkers and established risk predictors. The standard method for analysis would be to fit a prospective logistic regression model that includes an offset term to adjust for case-control sampling. Here, we develop a novel constrained maximum likelihood method, which has four features that distinguish it from the standard method. First, our method ensures that the new model calibrates similarly as the base model, in the sense that the predicted risk by the new model in the population strata defined by standard predictors is comparable to that by the base model. Because the base model

is well calibrated, such agreement between the two models lends support to the good calibration of the new model in the absence of independent validation data. A similar idea of indirect calibration was successfully applied for updating the BCRAT by incorporating percent mammographic density (Chen et al., 2006). Second, our method explicitly recognizes that the underlying population from which the case-control sample was assembled may not have the same distribution of standard risk predictors as the target population for prediction. In BCDDP, the participants were recruited from women who volunteered to undergo mammographic screening and they turned out to have higher average risk of breast cancer than the general U.S. Caucasian women. The distribution of BCRAT risk predictors in the BCDDP also differs from that estimated from the National Health Interview Survey (NHIS), where the latter better represents the general Caucasian woman population. Third, our method accommodates the known distribution of standard risk predictors in the target population, while relying on the case-control data for information on biomarkers. Fouth, our method more readily accommodates smaller sample sizes because of its high statistical efficiency.

Our method is based on maximizing the likelihood function under the constraints translated from the base model and external information on standard risk predictors. Using a parametric regression model to describe the relationship between the biomarker and standard risk predictors, we apply the Lagrange multiplier approach to deriving the profile likelihood for the Euclidean parameters. Constrained maximum likelihood methods have recently been developed to increase statistical efficiency for estimating odds ratio association parameters by exploiting external information through constraints (Chatterjee et al., 2016; Qin et al., 2015). Putting into the current context, Qin et al. (2015) exploited known outcome prevalence in strata defined by standard predictors to increase statistical efficiency, where the underlying population for the case-control sample shares the same risk and predictor distributions that define the stratum-specific prevalences. Chatterjee et al. (2016) exploited a known regression relationship between the outcome and standard predictors to increase statistical efficiency, when the joint distribution for both standard predictors and biomarkers is known externally. Our method differs in important ways: it exploits stratum-specific prevalence similarly as Qin et al. (2015), but accommodates known distribution of standard predictors for the target population of prediction while requiring information on biomarkers only from the case-control data. These differences lead to important practical implications: the model developed using our method calibrates to the target population, where the calibration is defined by the agreement between the predicted and estimated risks in discrete population strata as commonly done

for assessing goodness-of-fit of regression models. These differences also call for new theoretical development for statistical inference.

1.3.

In Chapter 4, we extend the statistical method developed in Chapter 3 to accommodate the cross-sectional study design, which often can not be treated as a random sample selected from the target population. We consider the same scenario where a well calibrated model based on standard predictors exists, and the goal is to incorporate the new risk predictors into the existing model. We assume that data on the outcome and all predictors is available from a cross-sectional sample. After re-deriving the likelihood function, we propose a similar constrained maximum likelihood method, which guarantees that the new model calibrates similarly as the existing model, allows the distribution of standard risk predictors in the sample to be different from that in the target population of prediction, and relies on the data to infer the relationship between biomarkers and standard predictors. This work is motivated by the study of Gestational Diabetes Mellitus as described in Chapter 2 (Zhu et al., 2016). The Phase I sample in the study can be seen as a prospective cohort study, where data on the outcome and predictors including age, race, BMI and family history was measured for all study subjects. The distribution of age, race and BMI is externally available in the National Vital Statistics Report (NVSR) from Centers for Disease Control and Prevention (CDC), which turns out to be quite different from that in the data. Given that the information of the relationship between family history and these predictors is limited or unknown, we formulate the problem into our constrained MLE approach framework by treating family history as the "new" predictor. The goal is to develop a logistic regression model with all four predictors included and the model is calibrated to an existing logistic regression model based on age, race and BMI with ORs reported in the previous literature (Berkowitz et al., 1992; Solomon et al., 1997). We provide simulation results to assess the performance of our method.

CHAPTER 2

DESIGN AND ANALYSES OF TWO-PHASE STUDIES FOR PREDICTING BINARY OUTCOMES

2.1. Introduction

This work was motivated by statistical challenges arising from the development of a model for predicting the risk of gestational diabetes mellitus (GDM) using a case-control study nested in the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Fetal Growth StudySingletons cohort. In this prospective study cohort (Zhu et al., 2016), data on conventional predictors such as age, race, body mass index, and family history was fully collected for 2,799 women. Blood samples were collected for the full cohort prior to the GDM screening. To study biomarkers in relation to the risk of GDM, a nested case-control subsample was assembled, where two controls were matched to each of the 107 cases on ethnicity, age within 2 years, and gestational weeks. Biomarkers were measured only for this case-control subset. It was of interest to develop a model for predicting the risk of GDM using both conventional risk predictors and new biomarkers. The data can be seen as being collected in two phases (Neyman, 1938; White, 1982). In Phase I, data on the outcome and conventional risk predictors was collected for all study subjects. In Phase II, biomarkers were measured from the blood only for a judiciously selected subset. Such two-phase studies, while commonly conducted in medical research, induce incomplete data and pose significant statistical challenges both in developing a comprehensive risk model and its validation. It is well known that naive analysis using only Phase II data with the sampling procedure ignored leads to biased estimation and decreased power.

In this work, we develop statistical methods for estimating the absolute risk for a binary outcome and several summary measures of prediction accuracy with two-phase data. Specifically, we consider estimation of the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC), which are routinely reported to gauge the discriminatory capacity of a prediction model. We also consider the proportion of cases followed (PCF; Pfeiffer and Gail, 2011) which is the probability that a case is included in a high-risk group, and the proportion needed to follow (PNF; Pfeiffer and Gail, 2011), which is the proportion of high-risk subjects that contain a specified proportion of cases. PCF and PNF measure the concentration of risk in the population and can be used to quantify the usefulness of a model in a risk-based screening program. PCF and PNF are of particular interest to our GDM study example, because the ultimate goal is to develop a model to help screen high-risk women upon validation. The current literature for analyzing two-phase data has been mainly focused on the estimation of association parameters that describe the relationship between an

outcome variable and predictors. The available methods that account for biased sampling of a two-phase study largely fell into four class: semiparametric maximum likelihood methods (ML; Breslow and Holubkov, 1997; Scott and Wild, 1997; Wang and Zhou, 2010), pseudo-likelihood (PL; Breslow and Cain, 1988; Breslow and Zhao, 1988) or estimated likelihood (Carroll and Wand, 1991; Pepe and Fleming, 1991) methods, weighted likelihood methods (WL; Flanders and Greenland, 1991; Ibrahim, 1990; Lipsitz, Ibrahim, and Zhao, 1999; Robins, Rotnitzky, and Zhao, 1994), and methods based on modifying likelihood score functions for the complete data (Chatterjee, Chen, and Breslow, 2003; Reilly and Pepe, 1995). Most of these methods have been synthesized and further extended in the seminal paper (Lawless, Kalbfleisch, and Wild, 1999). Currently, there is a paucity of procedures that can be applied to the two-phase setting for estimating measures of accuracy for predicting binary outcomes. Existing methods were not able to fully accommodate stratified sampling of Phase II subjects (Huang, 2016; Huang and Pepe, 2010), limited to the estimation of AUC (Huang, 2016; Pepe, Fan, and Seymour, 2013), or focused on the failure time data (Cai and Zheng, 2012; Liu, Cai, and Zheng, 2012). To our best knowledge, no methods are yet available for estimating the key accuracy measures PCF and PNF for our GDM study example.

We consider the general two-phase sampling scheme where the probability that a subject is selected into Phase II depends only on his/her own characteristics, and adopt logistic regression models for prediction. The estimation of predictive accuracy measure requires estimation of the risk distribution, which is a function of the odds ratio (OR) parameters and the predictor distribution. For estimating the OR parameters with two-phase data, the semiparametric ML method (Lawless, Kalbfleisch, and Wild, 1999; Scott and Wild, 1997), the PL method (Breslow and Cain, 1988; Breslow and Zhao, 1988), and the WL method (Flanders and Greenland, 1991; Robins, Rotnitzky, and Zhao, 1994) have been widely used in practice. Each of them has its own pros and cons, and they have been implemented in widely used R packages. Therefore, we propose three corresponding methods for estimating risk distribution and predictive accuracy measures.

In this work, we also explore efficient sampling strategies for selecting Phase II subjects in order to improve statistical efficiency for estimating predictive accuracy measures. It has been shown that selection stratified by case-control status alone or together with a small number of discrete Phase I variables can lead to a more informative Phase II sample than that obtained via simple random sampling (Breslow and Cain, 1988; Breslow and Chatterjee, 1999). The latter selects

similar numbers of Phase II subjects across strata, and is particularly attractive for increasing the efficiency for estimating OR parameters when the outcome is rare and some values of Phase I variables occur infrequently. This "balanced" design requires categorization of Phase I predictors, but there is no guidance on how to do so when multiple phase I predictors are available. One may choose to stratify by only a single Phase I predictor, or by coarse categories based on multiple predictors defined in such a way that a sufficient number of Phase I cases and controls are available for selection within each stratum. We propose to sample Phase II subjects based on a preliminary prediction model that includes only Phase I variables as predictors.

The rest of the chapter is organized as follows. In Section 2.2, we develop three methods for estimating risk distribution and predictive accuracy measures under the two-phase study design, and provide an inference procedure for each method. In Section 2.3, we conduct extensive simulation studies to evaluate the finite sample performance of the proposed methods under different sampling strategies. In Section 2.4, we apply our method to analyze the example GDM study to develop a preliminary prediction model. We conclude with some remarks in Section 2.5.

2.2. The Model, Estimation, and Inference Procedure

2.2.1. Model

We consider a prospective two-phase study design where Phase I is a representative cohort of size N drawn from a population of interest. Data on the binary outcome Y, with $Y = 1$ indicating cases and $Y = 0$ indicating controls, and sampling stratum S, is available for all N subjects. Stratified by Y and S, a subset of subjects is randomly selected into Phase II for measuring predictors $(\mathbf{X}, \mathbf{Z})$. Let N_{ys} and n_{ys}, $y = 0, 1$ and $s = 1, 2, ..., S$, denote the number of Phase I and Phase II subjects with $Y = y$ in the stratum $S = s$, respectively. We adopt a logistic regression model to describe the relationship between Y and $(\mathbf{X}, \mathbf{Z})$:

$$p(Y = 1|S = s, \mathbf{x}, \mathbf{z}) \equiv p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T\mathbf{x} + \boldsymbol{\gamma}^T\mathbf{z})}{1 + \exp(\alpha + \boldsymbol{\beta}^T\mathbf{x} + \boldsymbol{\gamma}^T\mathbf{z})}, \tag{2.1}$$

where $\boldsymbol{\theta}$ denotes $(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma})$. Here, we assume that the probability of Y depends on stratum S only through predictors $\mathbf{X}$. The requirement that S is not associated with Y given $(\mathbf{X}, \mathbf{Z})$, $p(Y = 1|S = s, \mathbf{x}, \mathbf{z}) = p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$, is naturally satisfied when S is the same as, a discretized version of, or a

surrogate of a subset of predictors $\mathbf{X}$. Let F denote the cumulative distribution function (CDF) for $(\mathbf{X}, \mathbf{Z})$. We are interested in estimating the distribution of risk $p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ and the three predictive accuracy measures, AUC, the PCF at risk quantile q ($0 < q < 1$), and the PNF for capturing percentage p cases ($0 < p < 1$), which can all be expressed as the risk distribution and CDF F as follows:

$$\text{AUC}(\boldsymbol{\theta}, F) = \int \text{TPR}_\nu(\boldsymbol{\theta}, F) d\{\text{FPR}_\nu(\boldsymbol{\theta}, F)\},$$

where $\text{TPR}_\nu(\boldsymbol{\theta}, F)$ and $\text{FPR}_\nu(\boldsymbol{\theta}, F)$ are the true and false positive rates at risk cutoff ν, respectively:

$$\text{TPR}_\nu = \text{Pr}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \nu | Y = 1\} = \frac{\int \text{I}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \nu\} p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) dF(\mathbf{x}, \mathbf{z})}{\int p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) dF(\mathbf{x}, \mathbf{z})},$$
$$\text{FPR}_\nu = \text{Pr}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \nu | Y = 0\} = \frac{\int \text{I}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \nu\} \{1 - p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})\} dF(\mathbf{x}, \mathbf{z})}{\int \{1 - p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})\} dF(\mathbf{x}, \mathbf{z})}.$$

Define the q^{th} upper quantile of the risk distribution ξ_q by equation $q = \text{Pr}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \xi_q\} = \int \text{I}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \xi_q\} dF(\mathbf{x}, \mathbf{z})$, which is also a function of parameters $(\boldsymbol{\theta}, F)$. Then $\text{PCF}_q(\boldsymbol{\theta}, F)$ has the same expression as $\text{TPR}(\boldsymbol{\theta}, F)$, except that the risk cutoff equals ξ_q:

$$\text{PCF}_q(\boldsymbol{\theta}, F) = \text{Pr}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \xi_q | Y = 1\} = \frac{\int \text{I}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \xi_q\} p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) dF(\mathbf{x}, \mathbf{z})}{\int p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) dF(\mathbf{x}, \mathbf{z})}.$$

$\text{PNF}_p(\boldsymbol{\theta}, F)$ is formally defined as the probability that the predicted risk is higher than the p^{th} quantile of the risk distribution for cases:

$$\text{PNF}_p(\boldsymbol{\theta}, F) = \text{Pr}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \xi_p\} = \int \text{I}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) > \xi_p\} dF(\mathbf{x}, \mathbf{z}),$$

where ξ_p is defined by equation $p = \text{PCF}_p(\boldsymbol{\theta}, F)$.

The estimates of these three measures can be obtained upon plugging in the estimates for $\boldsymbol{\theta}$ and F. If data for all predictors were collected for every subject in the cohort, the maximum likelihood estimate for $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, can be obtained by standard logistic regression, and the CDF of the predictors F can be estimated empirically as

$$\hat{F}(\mathbf{x}, \mathbf{z}) = \frac{1}{N} \sum_{k=1}^{N} \text{I}\{\mathbf{X}_k \leq \mathbf{x}, \mathbf{Z}_k \leq \mathbf{z}\}.$$

Plug-in estimators $\text{AUC}(\hat{\boldsymbol{\theta}}, \hat{F})$, $\text{PCF}_q(\hat{\boldsymbol{\theta}}, \hat{F})$, and $\text{PNF}_p(\hat{\boldsymbol{\theta}}, \hat{F})$ can be obtained accordingly.

2.2.2. Estimation and Sampling Strategies under the Two-Phase Design

Under the two-phase study design, $(\mathbf{X}, \mathbf{Z})$ is selectively measured based on Y and stratum information S collected in Phase I. Therefore, fitting model (2.1) to Phase II data alone will yield a biased estimate of OR parameters $\boldsymbol{\theta}$ and subsequent predictive accuracy measures. Below, we describe three estimators for $\boldsymbol{\theta}$ as mentioned in introduction, propose three corresponding estimators for the CDF $F(\mathbf{x}, \mathbf{z})$, and subsequently propose three corresponding methods for estimating the predictive accuracy measures. We derive the large sample distributions of the three sets of estimators. These methods are expected to have different statistical efficiencies and involve different levels of computational complexity.

Semiparametric Maximum Likelihood Method

Let P_I and P_{II} denote subjects in Phase I and Phase II, respectively, and R_{ysk} be the indicator of whether or not the k^{th} subject with outcome $Y = y$ from stratum s in Phase I sample is included in Phase II ($R_{ysk} = 1$: yes; $R_{ysk} = 0$: no). Let $\eta_{\mathbf{xz}s} = p(\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}|S = s)$ denote the stratum-specific empirical distribution of predictors $(\mathbf{X}, \mathbf{Z})$ and $p_y(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ denote $p(Y = y|\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$. The MLE of $\boldsymbol{\theta}$, proposed by Scott and Wild (1997), is obtained by maximizing the empirical log-likelihood function of both Phase I and Phase II data, which can be written as

$$
\begin{aligned}
l_{ML} &= \log \prod_{P_I} p(Y, S) \prod_{P_{II}} p(\mathbf{X}, \mathbf{Z}|Y, S) \\
&= \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} \left[(1 - R_{ysk}) \log \int \int p_y(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \eta_{\mathbf{xz}s} d\mathbf{x} d\mathbf{z} \right] \\
&+ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} \left[R_{ysk} \left\{ \log p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \boldsymbol{\theta}) + \log \eta_{\mathbf{x}_{ysk}\mathbf{z}_{ysk}s} \right\} \right].
\end{aligned}
$$

Let $\sum_{(\mathbf{x},\mathbf{z}) \subset (y,s)}$ denote summation over predictors of all Phase II individuals with $Y = y$ in stratum s. Define new variables μ_{ys} which has a dimension equal to twice the number of Phase I sampling

strata:

$$\mu_{ys} = \frac{n_{ys} - \gamma_{ys}}{N_{ys} - \gamma_{ys}}$$
$$\text{with} \quad \gamma_{ys} = n_{ys} - \sum_{(\mathbf{x},\mathbf{z})\subset(1,s)} p_y^\star(\mathbf{x},\mathbf{z};\boldsymbol{\theta},s) - \sum_{(\mathbf{x},\mathbf{z})\subset(0,s)} p_y^\star(\mathbf{x},\mathbf{z};\boldsymbol{\theta},s), \tag{2.2}$$

where

$$p_y^\star(\mathbf{x},\mathbf{z};\boldsymbol{\theta},s) = \frac{\mu_{ys} p_y(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{\sum_y \mu_{ys} p_y(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}. \tag{2.3}$$

By maximizing the likelihood over η with OR parameters $\boldsymbol{\theta}$ fixed, an expression of $\eta_{\mathbf{xz}s}$ as a function of $\boldsymbol{\theta}$, $\eta_{\mathbf{xz}s}(\boldsymbol{\theta})$, was obtained as

$$\hat{\eta}_{\mathbf{xz}s}(\boldsymbol{\theta}) = \frac{n_{+\mathbf{xz}s}}{N_{+s}\sum_y \mu_{ys} p_y(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}, \tag{2.4}$$

Plugging expression (2.4) into the log likelihood function l_{ML} leads to the profile likelihood function for $\boldsymbol{\theta}$, which, upon maximization, yields the MLE of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{ML}$. It turns out that $\hat{\boldsymbol{\theta}}_{ML}$ can be obtained by iteratively fitting the "pseudo-model" $p^\star(\mathbf{x},\mathbf{z};\boldsymbol{\theta},s)$, using software for standard logistic regression analysis that includes an offset term $\log\mu_{1s} - \log\mu_{0s}$. This is because the pseudo-model (2.3) can be equivalently written as

$$\log\frac{p_1^\star(\mathbf{x},\mathbf{z};\boldsymbol{\theta},s)}{p_0^\star(\mathbf{x},\mathbf{z};\boldsymbol{\theta},s)} = \log\mu_{1s} - \log\mu_{0s} + \log\frac{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{p_0(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}.$$

The μ_{ys} values are updated during each iteration. Beginning with $\gamma_{ys}^{(0)} = 0$, i.e., $\mu_{ys}^{(0)} = \frac{n_{ys}}{N_{ys}}$, an initial estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}^{(1)}$, is obtained via standard logistic regression analysis with an offset term $\log(n_{1s}/n_{0s}) - \log(N_{1s}/N_{0s})$ using only Phase II subjects. γ_{ys} is then updated by plugging $\hat{\boldsymbol{\theta}}^{(1)}$ into formula (3). The process continues until convergence to obtain $\hat{\boldsymbol{\theta}}_{ML}$.

To estimate the CDF $F(\mathbf{x},\mathbf{z})$ that is needed for the estimation of AUC, PCF, and PNF, the MLE of the empirical distribution $\eta_{\mathbf{xz}s}(\boldsymbol{\theta})$ is obtained by plugging in equation (2.4) $\hat{\boldsymbol{\theta}}_{ML}$ and the corresponding $\hat{\mu}_{ys}$. The MLE of the stratum membership, $p(S=s)$, is simply obtained as $(N_{0s}+N_{1s})/N$. Therefore, we estimate $F(\mathbf{x},\mathbf{z})$ as

$$\hat{F}_{ML}(\hat{\boldsymbol{\theta}}) = \hat{p}(\mathbf{X}\le\mathbf{x},\mathbf{Z}\le\mathbf{z}) = \sum_s \hat{p}(S=s) \times \left\{ \sum_{(\mathbf{x},\mathbf{z}):\mathbf{X}\le\mathbf{x},\mathbf{Z}\le\mathbf{z}} \hat{\eta}_{\mathbf{xz}s}(\hat{\boldsymbol{\theta}}_{ML}) \right\}.$$

Here we make it explicit that $\hat{F}_{ML}$ depends on $\hat{\theta}_{ML}$.

With $\hat{\boldsymbol{\theta}}$ and $\hat{F}_{ML}$ obtained, the semiparametric maximum likelihood estimate of the risk distribution $R(r) = \text{Pr}\{p_1(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \leq r\}$ are now obtained as

$$\hat{R}_r(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}) = \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML}) \leq r\}}{N \sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})}.$$

Below, we refer to the ML estimates of PCF$_q$, PNF$_p$, and AUC correspondingly as PCF$_q(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML})$, PNF$_p(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML})$, and AUC$(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML})$. The expression of PCF$_q(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML})$ is given below, and those of the other two are provided in the Appendices:

$$\begin{aligned}
\widehat{\text{PCF}}_q(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}) &= \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})}{\sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})} \right\}^{-1} \times \\
&= \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML}) > \xi_q\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})}{\sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})} \right\},
\end{aligned}$$

where ξ_q is estimated from equation

$$q = \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML}) > \xi_q\}}{N \sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})}.$$

We derive the large sample distributions of PCF$_q(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML})$, PNF$_p(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML})$, and AUC$(\hat{\boldsymbol{\theta}}_{ML}$, $\hat{F}_{ML})$ using the influence function approach. Let $\hat{T}_{ML} \equiv T\{\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})\}$ be a generic term for any of these three estimators and $T \equiv T(\boldsymbol{\theta}, F)$ be the corresponding true value. We derive a large sample approximation in the form of

$$\sqrt{N}(\hat{T}_{ML} - T) = \sqrt{N} \left\{ \frac{1}{N} \sum_{k=1}^{N} H_{Tk}^{ML} \right\} + o_p(1),$$

where $H_{Tk}, k = 1, ..., N$, the influence function contributed by the k^{th} subject, are independent and identically distributed with mean 0 and variance σ^2. Then, $\sqrt{N}(\hat{T}_{ML} - T)$ is asymptotically

normally distributed with mean 0 and variance σ^2 by the central limit theorem. Note that $\hat{\boldsymbol{\theta}}_{ML}$ can be approximated by an asymptotically linear form, a $\sqrt{N}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) = \sqrt{N}\left\{\frac{1}{N}\sum_{k=1}^{N} H_{\boldsymbol{\theta}k}^{ML}\right\} + o_p(1)$ (Scott and Wild, 1997), where the influence function $H_{\boldsymbol{\theta}k}^{ML}$ is provided in Appendices. By applying the standard Taylor series expansion and Delta method for statistical functionals, we obtain the asymptotic linear approximation of $T\{\boldsymbol{\theta}, \hat{F}_{ML}(\boldsymbol{\theta})\}$,

$$\sqrt{N}\left[T\{\boldsymbol{\theta}, \hat{F}_{ML}(\boldsymbol{\theta})\} - T(\boldsymbol{\theta}, F)\right] = \sqrt{N}\left\{\frac{1}{N}\sum_{k=1}^{N} \varphi_F^{ML}(\mathbf{X}_k, \mathbf{Z}_k)\right\} + o_p(1).$$

Then H_{Tk}^{ML} can be accordingly calculated as

$$H_{Tk}^{ML} = \left\{\frac{\partial T}{\partial \boldsymbol{\theta}}|_{F=\hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})} + \frac{\partial T}{\partial \hat{F}_{ML}(\boldsymbol{\theta})}\frac{\partial \hat{F}_{ML}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\} H_{\boldsymbol{\theta}k}^{ML} + \varphi_F^{ML}(\mathbf{X}, \mathbf{Z}).$$

The detailed derivations are provided in Appendices. The variance of H_{Tk}^{ML}, σ^2, can be estimated as the empirical variance of H_{Tk}^{ML}, and all partial derivatives involved can be estimated numerically.

Pseudo-Likelihood (PL) Method

The PL approach to estimating $\boldsymbol{\theta}$ (Breslow and Cain, 1988; Breslow and Zhao, 1988) is actually the result from the first step of the iterative process in the ML approach above, $\hat{\boldsymbol{\theta}}^{(1)}$. In other words, the PL estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{PL}$, is obtained by maximizing a pseudo-likelihood based on the pseudo-model defined in model (2.3) over Phase II subjects, but with μ_{ys} replaced by n_{ys}/N_{ys}:

$$S_{PL}^{\star} = \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk}\frac{\partial \log p_y^{\star}(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \boldsymbol{\theta}, s)}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Note that n_{ys}/N_{ys} in the large sample converges to $\pi_{ys} \equiv p(R_{ysk} = 1|Y = y, S = s)$, the stratum-specific probability of being selected into Phase II sample under variable probability sampling (VPS) (Lawless, Kalbfleisch, and Wild, 1999), where units are considered independent and sequentially inspected until a total number of n_{ys} subjects are selected from each stratum $(Y = y, S = s)$. Let $\hat{\boldsymbol{\pi}}_s$ denote the vector of sampling probabilities in stratum s for both cases and controls, $(\hat{\pi}_{ys},\ y = 0, 1)$, and $\hat{\boldsymbol{\pi}} = \{\hat{\boldsymbol{\pi}}_s,\ s = 1, 2, \ldots, S\}$. For the estimation of $F(\mathbf{x}, \mathbf{z})$, motivated by the expression (2.4) and similarity between the PL and ML methods for estimating $\boldsymbol{\theta}$, we propose a novel pseudo-likelihood

type estimator of $p(\mathbf{x}, \mathbf{z}|s)$ as

$$\hat{\delta}_{\mathbf{xz}s}(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}}_s) = \frac{n_{+\mathbf{xz}s}}{N_{+s} \sum_y n_{ys} N_{ys}^{-1} p_y(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}.$$

With the stratum proportion still estimated as $\hat{p}(S = s) = (N_{0s} + N_{1s})/N$ as in the ML method, we obtain a PL estimate of $F(\mathbf{x}, \mathbf{z})$, which depends on the estimated parameters $\hat{\boldsymbol{\theta}}_{PL}$ and $\hat{\boldsymbol{\pi}}_s$:

$$\hat{F}_{PL}(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}}) = \sum_s \hat{p}(S = s) \times \left\{ \sum_{(\mathbf{x},\mathbf{z}):\mathbf{X}\le\mathbf{x},\mathbf{Z}\le\mathbf{z}} \hat{\delta}_{\mathbf{xz}s}(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}}_s) \right\}.$$

The corresponding PL estimates of the risk distribution $R(r)$, PCF$_q$, PNF$_p$, and AUC are respectively denoted as $R_r(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL})$, PCF$_q(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL})$, PNF$_p(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL})$, and AUC$(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL})$. The expressions are similar to those under the ML approach with details provided in Appendices. The inference procedures are similar to those of the ML method, except that the variability from estimating the sampling probabilities, $\hat{\boldsymbol{\pi}}$, needs to be taken into account in addition to that due to $\hat{\boldsymbol{\theta}}_{PL}$ and $\hat{F}_{PL}$. The details are provided in Appendices.

Weighted Likelihood (WL) Method

The WL approach analyzes only Phase II subjects but corrects the biased sampling by weighting the contribution of each subject to the standard likelihood function by the inverse of the subject's sampling probability. The corresponding estimate for $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{WL}$, is the solution to the weighted likelihood score function, defined as

$$S_{WL} = \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} \frac{R_{ysk}}{\hat{\pi}_{ys}} \frac{\partial \log p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

To estimate the CDF $F(\mathbf{x}, \mathbf{z})$, we similarly propose a weighted empirical likelihood estimator as

$$\hat{F}_{WL}(\hat{\boldsymbol{\pi}}) = \frac{1}{N} \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} \frac{R_{ysk} \mathrm{I}\{\mathbf{X}_{ysk} \le \boldsymbol{x}, \mathbf{Z}_{ysk} \le \boldsymbol{z}\}}{\hat{\pi}_{ys}}.$$

The weighted likelihood estimate of risk distribution $R(r)$ can then be obtained as

$$\hat{R}_r(\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}) = \frac{1}{N} \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} \frac{R_{ysk}}{\hat{\pi}_{ys}} I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) \le r\}.$$

We refer to the WL estimates of PCF$_q$, PNF$_p$, and AUC correspondingly as PCF$_q$($\hat{\boldsymbol{\theta}}_{WL}$,$\hat{F}_{WL}$), PNF$_p$($\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}$), and AUC($\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}$). The expression for PCF$_q$($\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}$) is given as

$$\widehat{\text{PCF}}_q(\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}) = \frac{\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk}\hat{\pi}_{ys}^{-1} I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) > \xi_q\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL})}{\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk}\hat{\pi}_{ys}^{-1} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL})},$$

where ξ_q is estimated from equation

$$q = \frac{1}{N}\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk}\hat{\pi}_{ys}^{-1} I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) > \xi_q\}.$$

Those of the estimates for PNF$_p$($\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}$) and AUC($\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}$), together with the large sample distributions, are provided in Appendices.

An Extended Balanced (E-balanced) Design

An important consideration in designing two-phase studies is to select more informative Phase II subjects. In a two-phase study that collects additional confounding variables for studying the effect of a rare exposure, a "balanced" design (Breslow and Cain, 1988), which samples roughly equal numbers of subjects from strata defined by case-control status and the rare exposure, was shown to have higher efficiency than sampling equal numbers of cases and controls. The efficiency of this balanced design is often comparable to that of the optimal design, where the latter is largely infeasible as it depends on unknown parameters. In addition, it is not clear how to form sampling strata in the presence of multiple Phase I variables. To increase the efficiency for estimating predictive accuracy measures, a desirable strategy would select more informative Phase II subjects to increase the efficiency of estimating not only the OR parameters, but also the risk predictor distribution. For the latter, we expect that a strategy that oversamples extreme values of risk predictors would be useful. Therefore, considering that the relationship between standard risk predictors and the outcome variable has been studied in external studies, we propose to extend the balanced design (Breslow and Cain, 1988) by incorporating this existing knowledge. Specifically, we propose to stratify Phase I subjects by the "preliminary" risk predicted by the standard predictors, e.g., based on a logistic regression model available from an external source. Subjects with extreme preliminary risks are over-sampled, and within each stratum the number of cases and controls are similar. We call this design "the E-balanced design". It naturally accommodates multiple Phase I variables through the

risk distribution jointly determined by them. Intuitively, this design may have efficiency advantage for estimating predictive accuracy measures because the "extreme" value of the multiple predictors is more appropriately defined in terms of classifying subjects based on their risk of developing the disease. Results from Simulation Studies described below indicated the high efficiency of this design, and that the efficiency improvement owed to a large extent to improved estimates of the risk predictor distribution and to a smaller extent to improved estimates of the OR parameters.

2.3. Simulation Studies

We conducted extensive simulation studies to evaluate the finite sample performance of our proposed methods for estimating risk distribution and predictive accuracy measures and to compare the efficiency of our proposed E-balanced sampling design with the case-control and balanced sampling designs. For the latter, we chose to stratify by one Phase I variable that was most strongly associated with the outcome based on prior evidence. Three Phase I risk predictors, X_1, X_2, and X_3, were generated from the standard normal distribution, uniform distribution between 0 and 1, and binomial distribution with success probability 0.2, respectively. Phase II variables consisted of one single predictor Z that followed the standard normal distribution. The binary outcome Y was generated from the logistic regression model (2.1), with $\mathbf{X} = (X_1, X_2, X_3)$. We set the log OR parameters for $(\mathbf{X}, Z)$ as $\{\log(0.6), \log(1.6), \log(0.6), \log(1.5)\}$, so that X_1 and X_3 were negatively associated with Y and X_2 was positively associated. We chose $\alpha = \log(0.03)$ so that the prevalence of Y, $p(Y = 1)$, was around 4%. We generated a cohort of size $N = 3000$ (Phase I) and selected all cases and stratum-matched controls at a ratio of $1 : 2$ into Phase II using the three sampling strategies. Under the case-control design, we randomly selected all cases and twice as many controls from phase I. Under the balanced design, we created four sampling strata based on quartiles of X_2 and then selected all cases and twice as many controls in each stratum. Under the E-balanced design, similar stratified sampling was performed, but the stratum S was defined based on a linear combination of X_1, X_2, and X_3. We assumed that information on the predictiveness of (X_1, X_2, X_3) is available externally and summarized by the following logistic regression model:

$$p^e(x_1, x_2, x_3) = \frac{\exp(\alpha' + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3)}{1 + \exp(\alpha' + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3)}. \tag{2.5}$$

Phase I subjects are classified into four strata based on the estimated risk $p^e(x_1, x_2, x_3)$ such that each stratum contained roughly equal numbers of cases. We considered two sets of OR parameters for this model. In one ("E-balanced I"), the values were chosen so that the risk distribution based on external information was reasonably close to the true risk distribution based on model (2.1). In the other ("E-balanced II"), we let the values deviate further from those in model (2.1), and also reversed the association between (X_1, X_3) and Y. The risk distribution shifted by a reasonably large amount compared to the truth. In both cases, the value of α' was set so that the prevalence of Y was comparable to that under model (2.1). The two corresponding risk distributions are displayed in Figure 2.1. Results from E-balanced II will inform the efficiency of our proposed design

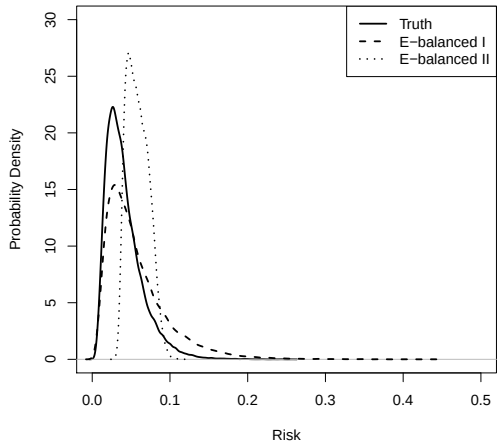

Figure 2.1: Risk distributions under three different prior risk models in the simulation study. "Truth" denotes the true model (1) with $(\beta_1, \beta_2, \beta_3)$=($\log(0.6)$, $\log(1.6)$, $\log(0.6)$). "E-balanced I" denotes the same model but with log OR parameters equal to $(\beta_1, \beta_2, \beta_3)$=$(\log(0.5), \log(1.7), \log(0.7))$. "E-balanced II" denotes the same model with log OR parameters equal to $(\beta_1, \beta_2, \beta_3)$ = ($\log(1.1)$, $\log(2.2)$, $\log(1.1)$).

when external information on the association between the outcome and Phase I variables agrees poorly with that in the data. All three proposed methods, ML, PL, and WL, were then applied to analyze each of 1000 simulated datasets. We also obtained estimates using full cohort data as the benchmark. For PCF and PNF, we considered threshold values $q = 0.2$ for PCF and $p = 0.9$ for PNF.

The results are summarized in Table 2.1. The averaged estimates of the three accuracy measures by all the three methods were close to the benchmark values. The averaged asymptotic standard

error ("ASE") estimates were all close to the empirical standard errors. The ML estimates always had the smallest asymptotic standard error estimates regardless of the sampling strategies, while the WL estimates had the largest as expected. The PL estimates mostly had similar standard errors as the ML estimates under E-balanced I when information used for stratification was relatively precise, although they became less efficient under the other two sampling designs. For example, under E-balanced I, estimates for $PCF_{0.2}$ by ML, PL, and WL were 0.433 (ASE: 0.038; 95% CI: 0.359, 0.507), 0.434 (ASE: 0.038; 95% CI: 0.360, 0.508), and 0.435 (ASE: 0.040; 95% CI: 0.357, 0.513), and the relative efficiency improvement of ML over WL was 10.8%. For $PNF_{0.9}$, the improvement was 17.9%. Under the balanced and case-control sampling, the efficiency improvement became greater: 25% and 36% for $PCF_{0.2}$ and 28% and 33% for $PNF_{0.9}$, respectively. These larger improvements most likely resulted from the higher efficiency for estimating the predictor distribution in the ML method, since the efficiency gain of ML compared to WL for estimating the ORs were only around 8% and 10%, respectively (data not shown). Across all designs, the efficiency of estimating AUC was similar for all three methods, although WL was slightly less efficient.

The efficiency of the E-balanced design appeared to depend on how precise the external information was for the relationship between Phase I predictors (X_1, X_2, X_3) and outcome Y. Under E-balanced I where the sampling model (2.5) was close to the true relationship induced by model (2.1), it yielded the most efficient estimates for each estimation method. The balanced and case-control designs had similar efficiencies although the former was at times slightly better. For estimating $PCF_{0.2}$, the efficiency of the E-balanced I design relative to the balanced and case-control design was around 22.0% by ML, 34.0% by PL, and 27.0% by WL. Compared to WL under the least efficient case-control design, which is the commonly used approach under the most widely used sampling design, ML under the E-balanced I gained around 40% efficiency for estimating PCF and PNF and 25% for AUC. The efficiency advantage over the other two sampling strategies largely vanished when the external information deviated substantially from the truth under E-balanced II, although there was some improvement (less than 10%) for WL.

2.4. The Analysis of a Study of Gestational Diabetes Mellitus

Using data from the GDM study described in Introduction, we applied our proposed methods to develop a preliminary model for the risk of GDM. We considered conventional risk factors and

Table 2.1: Estimates of the three predictive accuracy measures and their asymptotic standard errors under four Phase II sampling designs. Results are presented as the mean estimate (the empirical standard error estimate, the mean asymptotic standard error estimate).

	ML	PL	WL
E-Balanced Design I			
$\widehat{\text{PCF}}_{0.2}$	0.433 (0.038, 0.038)	0.434 (0.038, 0.038)	0.435 (0.040, 0.041)
$\widehat{\text{PNF}}_{0.9}$	0.722 (0.036, 0.035)	0.723 (0.036, 0.036)	0.723 (0.038, 0.038)
$\widehat{\text{AUC}}$	0.689 (0.025, 0.025)	0.689 (0.026, 0.025)	0.689 (0.027, 0.027)
Balanced Design			
$\widehat{\text{PCF}}_{0.2}$	0.435 (0.042, 0.042)	0.436 (0.044, 0.043)	0.437 (0.045, 0.047)
$\widehat{\text{PNF}}_{0.9}$	0.720 (0.038, 0.038)	0.721 (0.039, 0.039)	0.719 (0.040, 0.043)
$\widehat{\text{AUC}}$	0.690 (0.029, 0.028)	0.690 (0.029, 0.029)	0.691 (0.030, 0.029)
Case-control Design			
$\widehat{\text{PCF}}_{0.2}$	0.439 (0.043, 0.042)	0.441 (0.043, 0.043)	0.443 (0.047, 0.049)
$\widehat{\text{PNF}}_{0.9}$	0.715 (0.039, 0.039)	0.716 (0.039, 0.039)	0.713 (0.043, 0.045)
$\widehat{\text{AUC}}$	0.689 (0.028, 0.028)	0.689 (0.028, 0.028)	0.690 (0.029, 0.029)
E-Balanced Design II			
$\widehat{\text{PCF}}_{0.2}$	0.436 (0.042, 0.042)	0.436 (0.045, 0.044)	0.437 (0.045, 0.047)
$\widehat{\text{PNF}}_{0.9}$	0.719 (0.038, 0.038)	0.720 (0.040, 0.040)	0.718 (0.041, 0.043)
$\widehat{\text{AUC}}$	0.688 (0.029, 0.029)	0.688 (0.029, 0.030)	0.689 (0.030, 0.031)
Benchmark Using the Full Cohort			
$\widehat{\text{PCF}}_{0.2}$	0.432 (0.034, 0.034)		
$\widehat{\text{PNF}}_{0.9}$	0.720 (0.031, 0.031)		
$\widehat{\text{AUC}}$	0.686 (0.023, 0.023)		

glucose level, where the latter was measured only on the case-control subset. Complete data for both conventional risk predictors and glucose level was available for 104 cases and 208 controls, who were included in our analysis. We additionally assessed the efficiency of the E-balanced design by generating two-phase data involving only conventional risk factors, using each of the three sampling approaches as described in the simulation study. We discarded data on BMI for subjects not selected into Phase II, thereby treating BMI as the Phase II variable.

We used a logistic regression model to describe the relationship between gestational diabetes and all risk predictors. The two matching variables, age and race, are predictive of the risk of gestational diabetes (Berkowitz et al., 1992). We ignored gestational age in all analyses because it turned out not to be significantly associated with the risk of GDM in the full cohort. Because cases and controls in this study were more finely matched compared to the frequency-matching, conditional logistic regression analysis should have been the most appropriate approach to fitting the logistic regression model to estimate the OR parameters. But the effects of the matching variables, age and race, could not be estimated in such an analysis. We conducted an unconditional logistic regression analysis with matching variables adjusted for, and found that the estimates were very close to those from the conditional analysis (Table 2.2). Consequently, we adopted the unconditional regression analysis, treating the whole cohort as Phase I sample and the nested case-control subset as Phase II sample. We post-stratified the whole cohort by age (above or below 50 years) and race in all the three methods in the model development. The resultant data structure aligns with the two-phase design, allowing us to apply the proposed methods for analysis (Breslow and Chatterjee, 1999; Lawless, Kalbfleisch, and Wild, 1999). For WL, the sampling probabilities for cases were all equal to one, but for controls, they were calculated as the number of Phase II controls divided by the number of Phase I controls within each post-stratum. For the purpose of comparison, we also developed a model that included only conventional risk predictors using data from the full cohort and estimated the corresponding predictive accuracy measures. Age and BMI were fitted as continuous variables after exploring their functional forms by local polynomial regression, and race and family history were fitted as categorical variables. We dichotomized glucose level at median 94 in this analysis.

The estimates of the OR parameters are provided in Table 2.2, and those for the three predictive accuracy measures are provided in Table 2.3. Higher BMI, positive family history of diabetes, and higher glucose level appeared to be positively associated with the risk of gestational diabetes.

Table 2.2: Estimated odds ratio parameters (95% CI). Unconditional logistic regression model with conventional predictors only was fit to the full cohort ("Conventional predictors"). Conditional logistic regression analysis ("CL") with both conventional predictors and glucose level was performed using the nested case-control sample only. Unconditional logistic regression model with both conventional predictors and glucose level was fit to the full cohort using the three methods for fitting two-phase case-control data ("ML", "PL", "WL"). "*" represents that the p-value is less than 0.05 for testing the significance of the corresponding variable.

	Conventional predictors	Conventional predictors & Glucose			
		CL	ML	PL	WL
Age	1.073 (1.030, 1.117)$*$	NA	1.045 (0.999, 1.093)	1.040 (0.996, 1.087)	1.015 (0.951, 1.083)
Black	0.539 (0.265, 1.097)	NA	0.578 (0.236, 1.417)	0.576 (0.248, 1.342)	0.606 (0.241, 1.529)
Hispanics	1.532 (0.891, 2.632)	NA	1.414 (0.754, 2.650)	1.429 (0.762, 2.679)	1.258 (0.631, 2.509)
Asian	1.919 (1.022, 3.606)$*$	NA	1.814 (0.882, 3.732)	1.840 (0.888, 3.810)	1.812 (0.842, 3.902)
BMI	1.100 (1.062, 1.139)$*$	1.074 (1.021, 1.130)$*$	1.071 (1.020, 1.126)$*$	1.072 (1.020, 1.127)$*$	1.067 (1.015, 1.121)$*$
Family history	1.763 (1.136, 2.736)$*$	2.015 (1.098, 3.699)$*$	2.116 (1.201, 3.729)$*$	2.110 (1.182, 3.768)$*$	2.150 (1.157, 3.996)$*$
Glucose≥ 94	NA	3.068 (1.561, 6.029)$*$	2.950 (1.616, 5.385)$*$	2.938 (1.629, 5.298)$*$	3.791 (1.561, 6.029)$*$

The OR for the glucose level was estimated to be 2.95 (95% CI: 1.62, 5.39) by ML. To facilitate comparison with results when only conventional risk factors are included, the predictive accuracy measures were calculated with age and race included in the model, even if they were not significant. Results were largely similar to those when age and race were excluded (data not shown). If 20% of the women in the population at the highest risk by this model are screened for gestational diabetes, 51.5% (95% CI: 42.6%, 60.4%) of the cases can be identified when ML and PL are used. The corresponding estimates based on WL were 52.6% (95% CI: 42.1%, 63.1%). To be able to identify 90% of the cases, 69.8% (95% CI: 61.1%, 78.5%) of the population at the highest risk need to be followed according to the ML method. The AUC was estimated to be 0.673 (95% CI: 0.614, 0.732) by ML. Estimates by the other two methods were very close, and PL had similar while WL had larger estimated standard errors. Glucose level appeared to have predictive values independent of conventional risk predictors. With the glucose level included in the model, $\text{PCF}_{0.2}$, $\text{PNF}_{0.9}$, and AUC were estimated to increase by 2.9%, 1.4%, and 1.6%, respectively. We note that a more predictive model should have lower PNF values, but here the positive difference 1.4% was not expected to be significant.

Table 2.3: Estimated predictive accuracy measures and standard errors for the prediction models in Table 2.2, one using conventional predictors only ("Conventional predictors"), and the other using both conventional predictors and glucose level fitted using each of the three methods ("ML", "PL", "WL").

	Conventional predictors	Conventional predictors & Glucose level		
		ML	PL	WL
$\widehat{\text{PCF}}_{0.2}$	0.486 (0.037)	0.515 (0.045)	0.515 (0.045)	0.526 (0.053)
$\widehat{\text{PNF}}_{0.9}$	0.684 (0.036)	0.698 (0.044)	0.692 (0.045)	0.718 (0.047)
$\widehat{\text{AUC}}$	0.657 (0.024)	0.673 (0.030)	0.674 (0.030)	0.677 (0.037)

To compare the efficiency of the three estimation methods and Phase II sampling strategies, we considered the conventional risk predictors only under the same model as that in Table 2.3 (the second column). We applied each of the three sampling methods, the E-balanced, case-control, and balanced sampling, for selecting Phase II subjects, with age, race, and family history considered as Phase I variables ($\mathbf{X}$) and BMI as Phase II variable (Z). Data for BMI was retained only for Phase II subjects and deleted in the rest in the analysis. Under the E-balanced sampling, we used the OR parameters $(1.63, 2.61, 1.43)$ for every 5-year increase in age, Asian race, and family history of diabetes reported by Berkowitz et al. (1992). These ORs turned out to be reasonably close to

the ones in our analysis (Table 2.2). Phase I subjects were then classified into three strata based on the predicted risk calculated with these ORs. Under the balanced sampling, the Phase I strata were simply defined as three age groups ($< 30, 30{\sim}35, > 35$). From each stratum, 20 cases and 20 controls were selected into Phase II. Under the case-control sampling, 60 cases and 60 controls were randomly selected from those who developed or did not develop GDM. We generated 500 two-phase samples with each sampling strategy, and the results that summarized the 500 corresponding sets of estimates are presented in Table 2.4. ML appeared to be the most efficient and WL the least efficient, which is best shown by comparing the ratio between the averaged estimate and the standard error. The E-balanced sampling always led to smaller standard errors compared with the balanced sampling, but the improvement was marginal in the current analysis. We conjecture that the improvement would be greater if the number of cases were larger to allow for finer stratification of the Phase I data. The averaged estimates appeared to be slightly biased upwards compared with the full cohort analysis ("Benchmark estimates"). Our exploration showed that was due to the small number of cases and controls.

Table 2.4: Estimated predictive accuracy measures and standard errors of conventional risk predictors for predicting the risk of gestational diabetes mellitus under different sampling designs. Results are presented as the mean estimate (the mean asymptotic standard error estimate)

	ML	PL	WL
E-Balanced Design			
$\widehat{\text{PCF}}_{0.2}$	0.531 (0.096)	0.550 (0.104)	0.551 (0.119)
$\widehat{\text{PNF}}_{0.9}$	0.669 (0.082)	0.660 (0.103)	0.650 (0.118)
$\widehat{\text{AUC}}$	0.684 (0.063)	0.693 (0.078)	0.692 (0.096)
Balanced Design			
$\widehat{\text{PCF}}_{0.2}$	0.544 (0.098)	0.558 (0.120)	0.551 (0.123)
$\widehat{\text{PNF}}_{0.9}$	0.657 (0.085)	0.654 (0.109)	0.652 (0.122)
$\widehat{\text{AUC}}$	0.691 (0.064)	0.698 (0.084)	0.692 (0.104)
Benchmark			
$\widehat{\text{PCF}}_{0.2}$	0.495 (0.038)		
$\widehat{\text{PNF}}_{0.9}$	0.679 (0.035)		
$\widehat{\text{AUC}}$	0.665 (0.024)		

2.5. Conclusion

Two-phase study design is a cost-effective option to collect expensive predictors for the development of risk models. To predict the risk of a binary outcome, we developed an arsenal of statistical approaches to estimate risk distribution and several popular statistical measures for quantifying

predictive accuracy of the model. These methods differ in both statistical efficiency and ease of implementation. The ML estimator had the highest efficiency in the simulation study and real data analysis, and the PL estimator was a close competitor. Computation of the latter did not require iteration and therefore was faster and stable. One may balance the computation burden and desire for efficiency when deciding which method to use. The weighted likelihood estimator was the least efficient, but its efficiency can be improved through augmentation (Robins, Rotnitzky, and Zhao, 1994). More interestingly, it is expected to converge to the same limit as that when the data is available for the full cohort, even when the prediction model deviates from the true relationship between the outcome and predictors (Scott and Wild, 2002), an appealing robustness property that ML and PL do not have. A comparison of the three methods under model misspecification will be pursued in future work.

In order to analyze the GDM study example, we treated Phase II data as if it were collected using a stratified case-control design instead of the matched case-control design that was actually adopted. We justified our analysis by very similar estimates from the conditional and unconditional logistic regression analyses for BMI and family history which were measured on the full cohort. In general, we do not claim that our methods are universally applicable for analyzing matched case-control data supplemented with cohort information. To our best knowledge, two-phase methods are yet to be developed to rigorously accommodate individual matching, although such methods are available for studying time-to-event outcomes (Samuelsen, 1997). In particular, the WL method cannot be applied because subjects who are not eligible to match as controls for any case (e.g., in the GDM example, those who did not fall into the 2-year range of any case's age) have sampling probability equal to zero. The PL and ML methods would not apply either because they required stratified sampling as reflected in their respective likelihood functions. On the other hand, performing unconditional logistic regression is not uncommon for analyzing matched case-control data when the matching is not very fine.

We proposed selecting Phase II subjects based on existing information on the estimated risk with only Phase I variables for improved estimation efficiency. This strategy is an extension of the balanced design where Phase II subjects are selected based on combined values of individual Phase I predictors. Post-stratification based on prior predicted risks to increase statistical efficiency in fitting a risk model enriched with new predictors has been considered in the literature (Cai and Zheng,

2012). But the efficiency of sampling Phase II subjects utilizing prior risk estimates compared to the case-control and balanced sampling schemes has not been explored. Through over-sampling of subjects with more extreme risks, this design was shown to lead to improved efficiency for estimating risk distribution and predictive accuracy measures in our numerical studies. The efficiency gain increases when the prior risk estimates are closer to the risks estimated from Phase I data. Although it is difficult to perform theoretical studies on the efficiency of this design, the fact that it is an extension of the balanced design along with our numerical results supports its practical usefulness. Instead of relying on external data to assess prior risk, one may fit a working model to Phase I variables and perform Phase II sampling accordingly. While efficiency gain is expected, this incurs serious methodological challenges partly because the sampling across individuals now becomes dependent. We will study this sampling strategy in the context of risk prediction in future work.

CHAPTER 3

A Semiparametric Approach to Developing Well-Calibrated Models for Predicting the Risk of Binary Outcomes Using Case-Control Data

3.1. Introduction

Accurate risk prediction is central to precision medicine and precision disease prevention. Modern technology has enabled widespead efforts of biomarker discovery, promising great possibilities of improving risk prediction for human diseases. For putative biomarkers, it is of ultmost interest to evaluate their added values for prediction. But this is a challenging task. The model development requires data for disease status, biomarkers, as well as established risk predictors from a sample that is of sufficiently large size to allow stable model fitting. In practice, initial investigation efforts most often afford collection of data only from a small number of subjects, who are often recruited using a case-control study design. Cases or controls are not necessarily representative of the diseased or non-diseased in the target population for prediction. For example, the breast cancer risk assessment tool (BCRAT) was developed using data from a case-control study of breast cancer that was nested in the Breast Cancer Detection and Demonstration Project (BCDDP). The participants were recruited from women who volunteered to undergo mammographic screening. The BCDDP women turned out to have higher average risk of breast cancer than the general U.S. Caucasian women. The distribution of BCRAT risk predictors in the BCDDP also differs from that estimated from the National Health Interview Survey (NHIS), where the latter better represents the general Caucasian woman population. In an ongoing study on evaluating the predictiveness of volumetric breast density, cases and controls were recruited from the University of Pennsylvania health system. The odds ratio estimates for the BCRAT risk predictors were quite different from those used in the BCRAT. The study sample used for model development may also have been assembled cross-sectionally, which often can not be treated as a random sample selected from the target population.

Once a model is developed, the subsequent validation, which is necessary for establishing practical usefulness of the model, requires data that is collected from sources independent of those for model development. In particular, good calibration is essential in order to inform patients about their risks and make risk-based decision. The BCRAT was calibrated to the composite breast cancer rates reported in the Surveillance, Epidemiology, End Results program, making it a useful tool for projecting individualized risk for the U.S. Caucasian women. While data for the outcome and standard risk predictors may well be available from existing studies or can be easily obtained, it is usually much more challenging to obtain biomarker data. For example, it is not possible to genotype

BCDDP cases and controls because blood samples were not collected. The resources required to validate the predictiveness of metabolomic biomarkers are non-trivial partly due to the cost of metabolomics technology. Lack of independent validation data has been an obvious factor that hinders translation of biomarkers to clinics. Percent mammographic density was incorporated as a strong risk predictor into the BCRAT (Chen et al., 2006), and led to an increase in the area under the ROC curve that is comparable to that by incorporation of breast cancer risk SNPs identified to date. However, validation studies has yet to be conducted after more than 10 years since this updated model was published. Even if the data for validation is fully available, the sample size may limit the power for detecting lack of calibration. Recent work shows that greater than 10,000 subjects were required in order to detect meaningful differences between the predicted and observed risks in the upper tail of the risk distribution (Chatterjee et al., 2016).

Fortunately, data from multiple sources becomes more and more available that can be exploited to enhance model development and validation, which may compensate the scarce of biomarker data. The incidence rates for common diseases in the U.S. are publically available. Population level data for standard risk predictors is frequently available from national and international efforts or existing cohort studies. For example, the BCRAT risk predictors are fully represented in the National Health Interview Survey (NHIS III). The relationship between the outcome and standard risk predictors may have been assessed in multiple studies, or a risk prediction model that uses only standard predictors may have been developed and validated extensively. Data may be available for characterizing the relationship between biomarkers and standard risk predictors. Aiming to exploit these existing resources for model development and validation, we develop novel statistical methods for predicting the risk of a binary outcome using the logistic regression model.

Considering a scenario where a well calibrated model based on standard predictors exists, we focus on a central task of incorporating new risk predictors into the existing model, which henceforth is referred to as the “base model”. We assume that data on the outcome and all predictors is available from a case-control sample. The standard method for analysis would be to fit a prospective logistic regression model that includes an offset term to adjust for case-control sampling. Here, we develop a novel constrained maximum likelihood method, which has four features that distinguish it from the standard method. First, our method ensures that the new model calibrates similarly as the base model, in the sense that the predicted risk by the new model in the population strata defined

by standard predictors is comparable to that by the base model. Because the base model is well calibrated, such agreement between the two models lends support to the good calibration of the new model in the absence of independent validation data. A similar idea of indirect calibration was successfully applied for updating the BCRAT by incorporating percent mammographic density (Chen et al., 2006). Second, our method explicitly recognizes that the underlying population from which the case-control sample was assembled may not have the same distribution of standard risk predictors as the target population for prediction. Third, our method accommodates the known distribution of standard risk predictors in the target population, while relying on the case-control data for information on biomarkers. Fouth, our method more readily accommodates smaller sample sizes because of its high statistical efficiency.

Our method is based on maximizing the likelihood function under the constraints translated from the base model and external information on standard risk predictors. Using a parametric regression model to describe the relationship between the biomarker and standard risk predictors, we apply the Lagrange multiplier approach to deriving the profile likelihood for the Euclidean parameters. Constrained maximum likelihood methods have recently been developed to increase statistical efficiency for estimating odds ratio association parameters by exploiting external information through constraints (Chatterjee et al., 2016; Qin et al., 2015). Putting into the current context, Qin et al. (2015) exploited known outcome prevalence in strata defined by standard predictors to increase statistical efficiency, where the underlying population for the case-control sample shares the same risk and predictor distributions that define the stratum-specific prevalences. Chatterjee et al. (2016) exploited a known regression relationship between the outcome and standard predictors to increase statistical efficiency, when the joint distribution for both standard predictors and biomarkers is known externally. Our method differs in important ways: it exploits stratum-specific prevalence similarly as Qin et al. (2015), but accommodates known distribution of standard predictors for the target population of prediction while requiring information on biomarkers only from the case-control data. These differences lead to important practical implications: the model developed using our method calibrates to the target population, where the calibration is defined by the agreement between the predicted and estimated risks in discrete population strata as commonly done for assessing goodness-of-fit of regression models. These differences also call for new theoretical development for statistical inference.

The rest of this Chapter is organized as follows. We describe our proposed method and inference procedures in Section 3.2. In Section 3.3, we assess the finite sample performance of our method using simulated data, considering small, moderate, or large differences between the source population of the case-control sample and the target population and statistical efficiency. In Section 3.4, we will apply our method to analyze the BCDDP data to develop a logistic regression model for predicting the 5-year risk of breast cancer, using both the BCRAT risk predictors and percent mammographic density. We estimate the distribution of the BCRAT risk predictors from the National Health Interview Survey (NIHS III), and use a Beta-regression model for the distribution of percent mammographic density given the BCRAT risk predictors. Therefore, implicitly, we assume the latter to be the same in the BCDDP and general U.S. Caucasian woman population. We report estimates for both odds ratio parameters and parameters in the Beta-regression model by the proposed constrained maximum likelihood method.

3.2. The Method

3.2.1. Notation and likelihood function

Let Y denote the binary outcome status with $Y = 1$ indicating cases and $Y = 0$ indicating controls. Let $\mathbf{X}$ and $\mathbf{Z}$ denote the standard risk predictors and biomarkers, respectively. We consider that data for $(\mathbf{X}, \mathbf{Z})$ is collected from a frequency-matched case-control sample, where cases and controls are matched on a random variable S with M levels. We consider that S is an established predictor, such as age categories and ethnicity. Let n_{1s} and n_{0s} denote the respective number of cases and controls in stratum s, $s = 1, 2, \ldots, M$. Suppose that $\mathbf{X}$ and $\mathbf{Z}$ have K and L unique values observed in the data, $\{\mathbf{x}_k, k = 1, ..., K\}$ and $\{\mathbf{z}_l, l = 1, ..., L\}$, respectively. Let n_{1skl} and n_{0skl} represent the number of cases and controls in stratum s with $\mathbf{X} = \mathbf{x}_k$ and $\mathbf{Z} = \mathbf{z}_l$, $\sum_k \sum_l n_{iskl} = n_{is}, i = 0, 1, s = 1, ..., M$. Our goal is to fit a logistic regression model for predicting Y with $(S, \mathbf{X}, \mathbf{Z})$:

$$P(Y = 1|S = s, \mathbf{X} = \mathbf{x}_k, \mathbf{Z} = \mathbf{z}_l) = \frac{\exp\{\alpha_s + \boldsymbol{\beta}_{\mathbf{x}}^T \mathbf{x}_k + \boldsymbol{\beta}_{\mathbf{z}}^T \mathbf{z}_l\}}{1 + \exp\{\alpha_s + \boldsymbol{\beta}_{\mathbf{x}}^T \mathbf{x}_k + \beta_{\mathbf{z}}^T \mathbf{z}_l\}}, \tag{3.1}$$

where $\alpha_s, s = 1, ..., M$ is the stratum-specific intercept, $(\boldsymbol{\beta}_{\mathbf{x}}, \boldsymbol{\beta}_{\mathbf{z}})$ are the log odds ratio parameters for $(\mathbf{X}, \mathbf{Z})$. Let $\pi_{sk} = P(\mathbf{X} = \mathbf{x}_k|S = s), k = 1, ..., K, s = 1, ..., M$ denote the empirical distribution of $\mathbf{X}$ in stratum $S = s$, and we use a parametric model $f_{\boldsymbol{\tau}}(\mathbf{z}|s, \mathbf{x}_k)$ to describe the conditional distribution

of $\mathbf{Z}$ given $(\mathbf{X}, S)$, where $\boldsymbol{\tau}$ is a vector of Euclidean parameters. Let $P_{iskl} \equiv P(Y = i|S = s, \mathbf{X} = \mathbf{x}_k, \mathbf{Z} = \mathbf{z}_l)$. The empirical log-likelihood function of the case-control data can be derived as

$$\begin{aligned} l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}) &= \log \prod_{i=0}^{1} \prod_{s=1}^{M} \prod_{k=1}^{K} \prod_{l=1}^{L} P(\mathbf{X} = \mathbf{x}_k, \mathbf{Z} = \mathbf{z}_l | Y = i, S = s)^{n_{iskl}} \\ &= \log \prod_{i=0}^{1} \prod_{s=1}^{M} \prod_{k=1}^{K} \prod_{l=1}^{L} \left\{ \frac{P_{iskl} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{P(Y = i|S = s)} \right\}^{n_{iskl}} \\ &= \sum_{i=0}^{1} \sum_{s=1}^{M} \sum_{k=1}^{K} \sum_{l=1}^{L} n_{iakl} \log P_{iakl} + \sum_{s=1}^{M} \sum_{k=1}^{K} n_{+sk+} \log \pi_{sk} \\ &+ \sum_{s=1}^{M} \sum_{k=1}^{K} \sum_{l=1}^{L} n_{+skl} \log f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k) - \sum_{i=0}^{1} \sum_{s=1}^{M} n_{is++} \log \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} P_{iskl} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k) \right\}, \end{aligned}$$

where $\boldsymbol{\alpha} = \{\alpha_s : s = 1, ..., M\}, \boldsymbol{\pi} = \{\pi_{sk} : s = 1, ..., M; k = 1, ..., K\}, \boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathbf{x}}, \boldsymbol{\beta}_{\mathbf{z}})$. Since π_{sk} is the empirical probability mass function, it should satisfy the following constraints:

$$\sum_{k=1}^{K} \pi_{sk} = 1, \ s = 1, ..., M. \tag{3.2}$$

We assume that a risk prediction model based on risk predictors $(S, \mathbf{X})$ is available and well-calibrated for strata defined by $(S, \mathbf{X})$, and denote the predicted risk as $\varphi(S, \mathbf{X})$. The stratum-specific distribution for $\mathbf{X}$ in the target population of prediction, $P^e(\mathbf{X} = \mathbf{x}_k|S = s)$, denoted by δ_{sk}, is known from external sources, where superscript "e" here and after indicates "external". We explicitly allow δ_{sk} and π_{sk} to be different, that is, $\delta_{sk} \neq \pi_{sk}$. Within each stratum s, we categorize predicted risk φ into I_s intervals, and use a_{sr} and b_{sr} to denote the beginning and end of each interval. We assume that the calibration of the model $\varphi(S, \mathbf{X})$ was assessed in external studies by comparing the averaged predicted risk within each risk interval, defined as

$$P^e(Y = 1|a_{sr} \leq \varphi(s, \mathbf{x}) \leq b_{sr}), s = 1, ..., M, r = 1, ..., I_s,$$

with the "observed" average risk. We impose the equality of these averaged risks between the new model (3.1) and model $\varphi(S, \mathbf{X})$, thereby ensuring a good calibration performance of the new model in stratum defined by $(S, \mathbf{X})$. The resultant constraints are expressed as below:

$$P^e(Y = 1|a_{sr} \leq \varphi(s, \mathbf{x}) \leq b_{sr}) = \frac{\sum_{k:a_{sr} \leq \varphi(s, \mathbf{x}_k) \leq b_{sr}} \sum_{l=1}^{L} P_{1skl} \delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{\sum_{k:a_{sr} \leq \varphi(s, \mathbf{x}_k) \leq b_{sr}} \delta_{sk}}, \tag{3.3}$$

where $s = 1, ..., M, r = 1, ..., I_s$.

3.2.2. Constrained maximum likelihood for estimating $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$

We propose to estimate $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$ by maximizing the log-likelihood function $l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi})$ subject to constraints (3.2) and (3.3). Because the number of nuisance parameter π_{sk} can potentially be large, we derive the profile likelihood for parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$ using the method of Lagrange multipliers. The objective function is written as

$$g^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^*) = l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}) + \sum_{s=1}^{M} \lambda_s^* \left\{ \sum_{k=1}^{K} \pi_{sk} - 1 \right\}$$
$$+ \sum_{s=1}^{M} \sum_{r=1}^{I_s} \lambda_{sr} \left\{ \frac{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \sum_{l=1}^{L} P_{1skl} \delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \delta_{sk}} - P^e(Y = 1 | a_{sr} \le \varphi(s, \mathbf{x}) \le b_{sr}) \right\}.$$

Let $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^*$ denote the two sets of Lagrange multipliers $(\lambda_{sr} : s = 1, ..., M, r = 1, ..., I_s)$ and $(\lambda_s^* : s = 1, ..., M)$, respectively. By maximizing function $g^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$ over $\boldsymbol{\pi}$ with parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$ fixed, we can show that $\boldsymbol{\lambda}^* = \mathbf{0}$ and are able to express $\boldsymbol{\pi}$ as a function of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu})$,

$$\hat{\pi}_{sk}(\alpha_s, \boldsymbol{\beta}, \boldsymbol{\tau}, \mu_{is}) = \frac{n_{+sk+}}{n_{+s++} \sum_{i=0}^{1} \mu_{is} \sum_{l=1}^{L} P_{iskl} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)},$$
$$\mu_{is} = \frac{n_{is++}}{n_{+s++} \sum_{k=1}^{K} \sum_{l=1}^{L} P_{iskl} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)},$$

where $\boldsymbol{\mu} = (\mu_{is} : i = 0, 1, s = 1, ...M)$ has a dimension that equals twice the number of strata and therefore is much smaller than K. $\boldsymbol{\mu}$ can be treated as a vector of independent parameters in subsequet estimation. The detailed calculations are provided in the Appendices. Upon plugging $\widehat{\boldsymbol{\pi}}$ back into $g^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$, we obtain the constrained profile likelihood function for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$ as

$$g(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_{i=0}^{1} \sum_{s=1}^{M} \sum_{k=1}^{K} \sum_{l=1}^{L} n_{iskl} \log \left\{ \frac{P_{iskl} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k) \mu_{is}}{\sum_{i=0}^{1} \mu_{is} \sum_{l=1}^{L} P_{iskl} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)} \right\}$$
$$+ \sum_{s=1}^{M} \sum_{r=1}^{I_s} \lambda_{sr} \left\{ \frac{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \sum_{l=1}^{L} P_{1skl} \delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \delta_{sk}} - P^e(Y = 1 | a_{sr} \le \varphi(s, \mathbf{x}) \le b_{sr}) \right\}. \tag{3.4}$$

In the above development, we note that the second constraint did not contribute to the derivation because of the absence of π_{sk} due to $\delta_{sk} \neq \pi_{sk}$. This, together with the needed parametric relationship between $\mathbf{Z}$ and $(S, \mathbf{X})$, defines the novelty of our profile likelihood function.

We maximize profile likelihood function (3.4) jointly with respect to all the unknown parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ to obtain estimates of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$. We derive the score functions by taking the first derivative of the function (3.4) with respect to each component of the unknown parameters, ($S_{\boldsymbol{\alpha}}$, $S_{\boldsymbol{\beta}}$, $S_{\boldsymbol{\tau}}$, $S_{\boldsymbol{\mu}}$, $S_{\boldsymbol{\lambda}}$). The detailed derivations are provided in the Appendices. The maximum likelihood estimates $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\lambda}})$ are obtained via solving $\left\{S_{\boldsymbol{\alpha}}^T, S_{\boldsymbol{\beta}}^T, S_{\boldsymbol{\tau}}^T, S_{\boldsymbol{\mu}}^T, S_{\boldsymbol{\lambda}}^T\right\}^T = \mathbf{0}$. The identifiability, consistency, and asymptotic normality of the estimators are given in the Theorems below.

3.2.3. Identifiability and consistency

When the model in (3.1) is correctly specified, the identifiability of the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}$ has been shown in the literature. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \boldsymbol{\pi}^T)^T$. For simplicity, we assume the parameter space $\boldsymbol{\Theta}$ is bounded. We show that the constrained MLE $\widehat{\boldsymbol{\theta}}$ converges to the unique true parameter $\boldsymbol{\theta}_0$ in probability. The details of the proof are provided in the Appendices. We also consider a more realistic and common case in the setting of risk prediction, where the model in (3.1) is arbitrary and not necessarily the true model. In this case, we need to establish the identifiability first, i.e., there are no two sets of parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}$ and $\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\tau}}, \widetilde{\boldsymbol{\pi}}$ so that they both satisfy the constraints in (3.2) and (3.3), and $P(\mathbf{x}, \mathbf{z}|y, s, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}) = P(\mathbf{x}, \mathbf{z}|y, s, \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\tau}}, \widetilde{\boldsymbol{\pi}})$ for all $(\mathbf{x}, \mathbf{z}, y, s)$ combinations. The detailed proof has been provided in the Appendices. Later, we show the consistency of the constrained maximum likelihood estimator and details are given in the Appendices again.

3.2.4. Asymptotic properties

Regardless of the model in (3.1) is correct or misspecified, we have shown that the constrained MLE $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$ defined according to the specific situation. The constraint in (3.2) enables us to write $\pi_{sK} = 1 - \sum_{k=1}^{K-1} \pi_{sk}$ hence eliminate the parameter π_{sK}. Likewise, because P_{1skl} is a monotonically increasing funcion of α_s, the constraint in (3.3) enables us to solve for α_s uniquely as a function of other parameters, hence eliminate the parameter α_s, for $s = 1, \ldots, M$ when $I_s = 1$. When $I_s > 1$, we will be able to eliminate other parameters as well. Denote the remaining parameters $\boldsymbol{\psi}$ and write the parameters that are determined by the constraints $\boldsymbol{\gamma}(\boldsymbol{\psi})$. Note that the functional relation $\boldsymbol{\gamma}(\boldsymbol{\psi})$ does not depend on data. With a slight abuse of notation, we use $l\{\boldsymbol{\psi}, \boldsymbol{\gamma}(\boldsymbol{\psi})\}$ to denote the loglikelihood function of $\boldsymbol{\psi}$. The constrained maximization then can be

equivalently written as maximizing

$$M_n(\boldsymbol{\psi}) \equiv n^{-1} l\{\boldsymbol{\psi}, \boldsymbol{\gamma}(\boldsymbol{\psi})\}$$

with respect to ψ and we have shown that $\{\widehat{\boldsymbol{\psi}}_n, \boldsymbol{\gamma}(\widehat{\boldsymbol{\psi}}_n)\} \to \boldsymbol{\theta}_0$ in probability. This means we have

$$\begin{aligned} &= n^{1/2} \frac{\partial M_n(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\widehat{\boldsymbol{\psi}}_n} \\ &= n^{1/2} \frac{\partial M_n(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + \frac{\partial^2 M_n(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*} n^{1/2}(\widehat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0), \end{aligned}$$

where ψ^* is on the line connecting ψ_0 and $\widehat{\psi}_n$. Thus, we obtain the expansion

$$n^{1/2}(\widehat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0) = -\left[E\left\{\frac{\partial^2 M_n(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}_0^T}\right\}\right]^{-1} n^{1/2} \frac{\partial M_n(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}_0} + o_p(1).$$

This leads to that $\sqrt{n}(\widehat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0) \to N(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\psi}})$ in distribution, where

$$\boldsymbol{V}_{\boldsymbol{\psi}} = \left[E\left\{\frac{\partial^2 M_n(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}_0^T}\right\}\right]^{-1} \mathrm{var}\left\{\mathrm{n}^{1/2} \frac{\partial \mathrm{M_n}(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}_0}\right\} \left[\mathrm{E}\left\{\frac{\partial^2 \mathrm{M_n}(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}_0^{\mathrm{T}}}\right\}^{\mathrm{T}}\right]^{-1}.$$

Because $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\psi}}^T, \boldsymbol{\gamma}(\widehat{\boldsymbol{\psi}})^T\}^T$, we further have that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to N(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}})$ in distribution, where $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{A}\boldsymbol{V}_{\boldsymbol{\psi}}\boldsymbol{A}^T$ and

$$\boldsymbol{A} = \left\{ \begin{array}{c} \boldsymbol{I} \\ \frac{\partial \boldsymbol{\gamma}(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}_0^T} \end{array} \right\}.$$

3.3. The BCRAT and Percent Mammographic Density: Analysis of Data from the Breast Cancer Detection and Demonstration Project (BCDDP)

We applied our proposed method to analyze data from the BCDDP to develop a logistic regression model for predicting the risk of breast cancer for Caucasian women in the next five years following predictor meausurement. We intended to include both the BCRAT risk predictors and percent mammographic density (PD) as predictors. The BCRAT predictors include age at menarche (Agemen), age at first live birth (Ageflb), number of previous breast biopsies (Nbiops), and number of

first-degree relatives (mother/sisters; Numrel) who had breast cancer. PD, the percent of dense area on the mammogram image as a measure of breast density, has been established as one of the strongest risk predictors for breast cancer (Boyd et al., 1995; Byrne et al., 1995). The odds ratio function used in the BCRAT was developed from an age-stratified case-control study nested within the BCDDP, which included 2,808 cases and 3,119 controls. A subset of 1,217 cases and 1,616 controls had PD measurements. The details of this study were reported previously (Chen et al., 2006; 2008), where it was shown that the availability of PD only depended on the case-control status and age strata. Here, we analyzed data for the 2,833 (1217+1616) subjects who had complete predictors data, as inclusion of subjects who did not have PD data requires non-trivial extension of our method. Because weight confounded the relationship between breast cancer risk and PD, we also included weight as a predictor as in the previous work (Chen et al., 2006, 2008) The codings of BCRAT predictors and weight were the same as those in Chen et al. (2006), and they were all fitted as ordinal variables. In the model, we considered two age strata defined as $S = 1$ if age $\leq$ 50 and $S = 2$ otherwise, and finer categorization can be adopted in the same manner. We fitted model (3.1) with $\mathbf{X}$=(Ageflb, Agemen, Nbiops, Numrel, Weight) and Z=PD. In the data, around 10% of the subjects had zero values for PD, which we conjectured reflected both truly no dense tissues and PD values that were below the detection limit. We used a zero-inflated Beta regression model for the distribution of Z to accommodate the excess zero value (Chen and Chen, 2015). Let Z_{min} denote the minimal value that PD was detectable. The probability density function of Z can then be written as follows:

$$\begin{aligned} P(Z|S,\mathbf{X};\boldsymbol{\gamma},\boldsymbol{\omega},\rho) &= p(Z=0)^{I\{Z=0\}} p(Z>0)^{I\{Z>0\}} \\ &= [\rho + (1-\rho)Beta(Z<Z_{min}|\mathbf{X},S;\boldsymbol{\gamma},\boldsymbol{\omega})]^{I\{Z=0\}} \\ &\times [(1-\rho)Beta(Z|\mathbf{X},S;\boldsymbol{\gamma},\boldsymbol{\omega})]^{I\{Z>0\}}, \end{aligned}$$

where ρ denoted the probability that Z truly took zero value and $1-\rho$ the probability that Z was generated from a Beta regression model. The model was defined by its mean parameter κ and the precision parameter ϕ such that $Beta(Z|S,\mathbf{X};\boldsymbol{\gamma},\boldsymbol{\omega}) = Beta(Z|\kappa,\phi)$, where logit$(\kappa) = (1,S,\mathbf{X})\boldsymbol{\gamma}$, log$(\phi) = (1,S,\mathbf{X})\boldsymbol{\omega}$. In the logistic regression model (3.1), we categorized Z into 10 groups for taking values 0, 1-9 for $Z = 0$ or $Z \in (0, Z_{min})$, $[Z_{min}, 0.1), [0.1, 0.2), ..., [0.8, 1)$, respectively, denoted by Z^c.

To ensure a good calibration performance of our model, i.e., to accurately predict the expected number of breast cancer cases in defined population subgroups, we imposed the constraints based on 5-year absolute risks predicted from the BCRAT. Since the BCRAT has been shown to be generally well-calibrated in previous validation studies (Bondy et al., 1994; Costantino et al., 1999; Rockhill et al., 2001), by equating the prediction performance of our new model to that of the BCRAT, we would successfully maintain this "well-calibration" property. Let $\varphi^B(T_1, T_2, \mathbf{X}^B)$ denote the BCRAT absolute risk estimate starting from age T_1 till T_2 given $\mathbf{X}^B$=(Ageflb, Agemen, Nbiops, Numrel), which can be calculated online (http://www.cancer.gov/bcrisktool). The superscript "B" indicates "BCRAT". We calculated $\varphi^B(S, \mathbf{X}^B)$ as the weighted average of $\varphi^B(T_1, T_2, \mathbf{X}^B)$ with $T_1, T_2 \in S$, $T_2 - T_1 = 5$, and weight equal to $Pr(age \in [T_1, T_2])$ as estimated from NHIS. Further, we chose quartiles of $\varphi^B(S, \mathbf{X}^B)$ as the $a's$ and $b's$ in the constraints (3.3) and calculated $P^e(Y = 1|a_{sr} \leq \varphi^B(S, \mathbf{X}^B) \leq b_{sr}), s = 1, 2, r = 1, 2, 3, 4$. That is, we had a total of 8 constraints, 4 for each age stratum. Together with the distribution $P^e(\mathbf{X}|S)$ (Appendices) estimated from NHIS, whose variance was ignored in the current analysis, we calculated $P^e(\mathbf{X}^B|S) = \sum_{Weight} P^e(\mathbf{X}^B, Weight|S)$ and obtained the averaged 5-year risk:

$$P^e(Y = 1|\varphi^B(S = 1, \mathbf{X}^B) \in [\boldsymbol{a}_1, \boldsymbol{b}_1]) = (0.6\%, 1.3\%, 2.3\%, 3.4\%),$$
$$P^e(Y = 1|\varphi^B(S = 2, \mathbf{X}^B) \in [\boldsymbol{a}_2, \boldsymbol{b}_2]) = (1.5\%, 2.7\%, 4.2\%, 6.9\%).$$

We used these eight numbers as the benchmark for risk prediction calibration. The resultant constraints were slightly different form compared to (3.3), because weight was originally not included in $\mathbf{X}^B$ and had to be averaged over in the constraint equations:

$$P^e(Y = 1|a_{sr} \leq \varphi^B(S, \mathbf{X}^B) \leq b_{sr})$$
$$= \frac{\sum_{\mathbf{X}^B: a_{sr} \leq \varphi^B(S, \mathbf{X}^B) \leq b_{rs}} \sum_{Weight} \sum_{Z^c} P(Y = 1|S, \mathbf{X}, Z^c) P^e(\mathbf{X}|S) P(Z^c|S, \mathbf{X}; \boldsymbol{\gamma}, \boldsymbol{\omega}, \rho)}{\sum_{\mathbf{X}^B: a_{sr} \leq \varphi^B(S, \mathbf{X}^B) \leq b_{sr}} P^e(\mathbf{X}^B|S)}.$$

For the comparison purpose, we also applied a "standard" method to analyze the same data. Essentially the same method was applied to develop the BCRAT and other models for predicting the risk of various cancer types (ref?). First, a prospective logistic regression model was fitted to obtain the OR estimates for $(\mathbf{X}, Z)$. Then we fitted the Beta-regression model using data only from the controls to obtain an approximate estimate of the PD distribution. Lastly, we plugged these

estimates into the constraints above to solve for the stratum-specific intercept parameters corresponding to the effect of S. The estimated standard errors for the OR estimates and $(\boldsymbol{\gamma}, \boldsymbol{\omega}, \rho)$ were calculated using stantard MLE approach, while those for the intercept terms were estimated using the delta method.

The results were summarized in Table 3.1. The estimated log ORs obtained by our method for risk predictors could be quite different from those by the standard method. For Ageflb and Nbiops, the estimates were 1.6 times larger. The direction of association between weight and breast cancer was even reversed, changing from positive to negative. Because our method assumed the same model relating the risk of breast cancer with predictors in the general Caucasian woman population and the BCDDP population, these differences may indicate the effectiveness of our method for pursuing good calibration. We conjecture that the differences in the predictor distribution, $P(\mathbf{X}|S)$, between the BCDDP and NHIS largely drove the discrepancy in the two sets of OR estimates. As shown in Figure 3.1, the distribution of weight in the BCDDP controls differed substantially from that estimated from the NHIS in both age groups. On the other hand, the estimates for the log OR of PD^c and parameters involved in its distribution, except for the intercept term in the mean model that turned out to be not significant, were similar between the two approaches. Our assumption of the same distribution for PD in the general Caucasian woman population and BCDDP population, might explain these consistencies. In the mean model for the PD distribution, Ageflb and Nbiops were positively associated, and weight was negatively associated. Weight was the only significant variable in the variance model, where larger weights appeared to be associated with lower variance in PD. As indicated by the estimate of the mixture probability ρ which was significantly different from zero, only around 10% subjects truely had no dense breast tissue as reflected on the mammogram image, while the remaining 90% had non-zero but undetectable measurements. The estimates for the log OR of PD^c were similar by the two methods, and the estimated variance by our method was 10% smaller. The OR estimates for the age stratum, Agemen, and Numrel were also close, and the estimated variance by our method was around 30% smaller for the latter two variables. Estimates for parameters in the PD model were also similar, and our method yielded smaller standard errors because both cases and controls were used compared with the standard method where only controls were used.

Table 3.1: Analysis of the BCDDP data: estimates of stratum-specific intercept terms and log ORs for the BCRAT predictors, weight, and PD, together with estimates of parameters in the zero-inflated Beta regression model for the distribution of PD. In the parenthesis are the corresponding estimates of asymptotic standard errors. "cMLE" represents estimates from the proposed constrained maximum likelihood method, and "Standard" represents the estimates from the standard method.

		Predictors	cMLE	Standard
Logistic Regression Model for Breast Cancer Risk		Intercept	-6.105 (0.140)	-6.751 (0.182)
		Age$\geq$50	1.138 (0.030)	1.045 (0.035)
		Ageflb	0.286 (0.033)	0.105 (0.049)
		Agemen	0.241 (0.045)	0.213 (0.061)
		Nbiops	0.460 (0.048)	0.174 (0.070)
		Numrel	0.697 (0.019)	0.668 (0.090)
		Weight	-0.205 (0.033)	0.228 (0.044)
		PD^c	0.175 (0.017)	0.177 (0.018)
The PD distribution		Intercept	0.165 (0.073)	0.035 (0.088)
		Age$\geq$50	-0.538 (0.048)	-0.405 (0.057)
	The mean model	Ageflb	0.142 (0.027)	0.139 (0.033)
	γ	Nbiops	0.287 (0.035)	0.248 (0.045)
		Weight	-0.445 (0.025)	-0.421 (0.033)
	The variance model	Intercept	1.396 (0.097)	1.478 (0.101)
	ω	Age$\geq$50	0.198 (0.077)	0.202 (0.079)
		Weight	-0.093 (0.035)	-0.121 (0.045)
	Mixture probability	ρ	0.096 (0.008)	0.105 (0.010)

3.4. Simulation Studies

We conducted extensive simulation studies to evaluate the finite sample performance of our proposed methods. To make clear distinction between the target population of prediction and the population from which the cases and controls were sampled, we set up the simulation scheme in two steps. First, we defined a population where the true distribution of $\mathbf{X}$, $P^e(\mathbf{X}|S)$, was known, and a well-calibrated risk prediction model based on $(S, \mathbf{X})$ was available. Second, we generated data for the established risk predictors $(S, \mathbf{X})$ and biomarkers $\mathbf{Z}$ for a case-control sample, mimicking the BCDDP case-control study described above.

3.4.1. Step 1: Define $P^e(\mathbf{X}|S)$ and $\varphi(S, \mathbf{X})$

We considered that $\mathbf{X}$ consisted of four predictors, $\mathbf{X} = (X_1, X_2, X_3, X_4)$, where $X_1 \sim N(0.25, 1)$ was categorized into 4 groups by its quartiles, and X_2 was generated from a Multinomial distribution with probabilities (0.2, 0.6, 0.2). For X_3 and X_4, we generated two Poisson random variables with mean 0.75 and 1.25, respectively, and assigned X_3 values 0, 1, or 2 according to the first one and X_4 values 0, 1, $\geq$ 2 according to the second one. For the stratum variable S, we generated

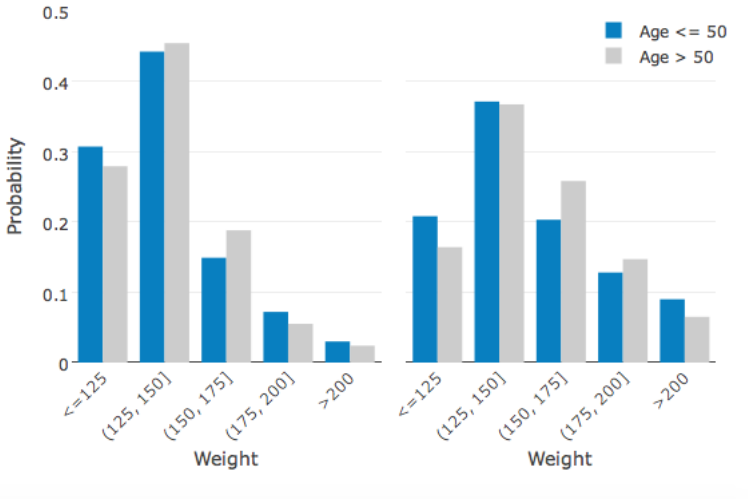

Figure 3.1: Distributions of weight for women with age $\leq$ 50 years or age $>$ 50 years. The left panel represents the distribution in the BCDDP controls. The right panel represents the distribution estimated from the National Health Interview Survey (NHIS).

a Uniform random variable in the range of (30, 80), and assign S value 1 if this variable is less than 50 and 2 otherwise. Therefore, $\mathbf{X}$ and S were mutually independent. The joint distribution of predictors, $P^e(\mathbf{X}|S)$, can then be expressed as $\prod_{i=1}^{4} P^e(X_i)$, which was treated as known. We assumed that the following logistic regression model for predicting the risk of Y based on $(S, \mathbf{X})$, where all variables in $\mathbf{X}$ were used as ordinal, was well calibrated:

$$\varphi(s, \mathbf{x}; \boldsymbol{\eta}) \equiv p^e(Y = 1|s, \mathbf{x}; \boldsymbol{\eta}) = \frac{\exp(\eta_0 + \eta_1 I\{s = 2\} + \eta_2 x_1 + \eta_3 x_2 + \eta_4 x_3 + \eta_5 x_4)}{1 + \exp(\eta_0 + \eta_1 I\{s = 2\} + \eta_2 x_1 + \eta_3 x_2 + \eta_4 x_3 + \eta_5 x_4)}.$$

We set parameters $(\eta_0, \eta_1, \eta_2, \eta_3, \eta_4, \eta_5)$ to be $(-4.5, 1.0, 0.2, 0.15, 0.15, 0.65)$ so that the prevalence of Y, $P(Y = 1)$, was around 8.5%. Consistent with the common goodness-of-fit test of calibration, we chose the quartiles of $\varphi(s, \mathbf{x}; \boldsymbol{\eta})$ as the cutoffs for each stratum used in the constraint (3.3). Together with $P^e(\mathbf{X}|S)$, we were able to obtain these cutoff values $[a_{sr}, b_{sr}]$, $s = 1, 2$, $r = 1, 2, 3, 4$ and the corresponding proportion of cases within each risk interval as

$$\begin{aligned} &P^e(Y = 1|\varphi(s = 1, \mathbf{x}; \boldsymbol{\eta}) \in [\boldsymbol{a}_1, \boldsymbol{b}_1]) = (1.9\%, 3.1\%, 5.0\%, 8.1\%), \\ &P^e(Y = 1|\varphi(s = 2, \mathbf{x}; \boldsymbol{\eta}) \in [\boldsymbol{a}_2, \boldsymbol{b}_2]) = (5.1\%, 8.2\%, 12.3\%, 19.0\%). \end{aligned} \tag{3.5}$$

These eight numbers were considered as the calibration benchmark, leading to eight corresponding constraints as described in (3.3).

3.4.2. Step 2: Generate the Case-Control Data

We considered three scenarios. In the first two scenarios, we generated $\mathbf{X}$ from distributions that were different from those in *Step 1* by using different parameter values. X_1 was generated from a standard normal distribution, X_2 from a Multinomial distribution with probabilities (0.3, 0.4, 0.3), and the Poisson distributions used for generating X_3 and X_4 had mean 1.25 and 1, respectively. All four variables were coded the same way as in *Step 1*. In the third scenario, $\mathbf{X}$ followed the same distribution as in *Step 1*. For all three scenarios, the stratum variable S was created the same way as in *Step 1*, and we considered a single biomarker Z that followed the Beta regression model,

$$
\begin{aligned}
&p(Z|s,\mathbf{x}) \sim Beta(\kappa\phi, \phi - \kappa\phi), \text{ where} \\
&\text{logit}(\kappa) = \gamma_0 + \gamma_1 I\{s=2\} + \gamma_2 x_1 + \gamma_3 x_3, \\
&\log(\phi) = \omega_0 + \omega_1 I\{s=2\} + \omega_2 x_2.
\end{aligned}
$$

We set $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ as $(1.2, -0.4, 0.2, -0.2)$ and $(\omega_0, \omega_1, \omega_2)$ as $(1.2, 0.2, 0.1)$. Then we categorized Z into 10 groups, denoted by Z^c, which takes integer values 1 to 10 for $Z \in (0, 0.1), [0.1, 0.2), \ldots,$ $[0.9, 1)$, respectively. The binary outcome Y was generated from the logistic regression model (3.1) with Z^c fitted instead of Z. In all three scenarios, the log OR parameters for $(X_1, X_2, X_3, X_4, Z^c)$ were set to be the same and equal to $(0.15, 0.2, 0.2, 0.6, 0.15)$. The log OR values for (X_1, X_2, X_3, X_4) were set to be reasonably close to those in the existing model $\varphi(s, \mathbf{x}; \boldsymbol{\eta})$, since we wanted to have similar association between standard predictors and the outcome in the case-control sample and in the population. We chose different α_s values so that the outcome prevalence in the three scenarios was 12.5%, 9.0%, and 9.8%, respectively. Above, we set up the covariate distribution and outcome prevalence to be different from those in the population described in *Step 1* to mimic the differences between the BCDDP study, studies on validating the BCRAT, and national data as represented in SEER and NHIS. That is, the source population for the case-control sample differs from the target population. The difference between the populations in the three scenarios and that in *Step 1* is depicted in Figure 3.2, where we plotted the "predicted" by model φ versus vs the "observed" in each scenario. To do this, we first generated a large sample of size 10^7 based

on the distribution of $\mathbf{X}$ and Z and $P(Y|S,\mathbf{X},Z)$ specified above. Using the same partition of the $\mathbf{X}$ space as that equivalent to $a_{sr} <= \varphi(s,\mathbf{x};\boldsymbol{\eta}) <= b_{sr}$ in *Step 1*, we obtained the proportion of cases in each of the subspaces ("observed"), denoted by $P^{cc}(Y = 1|\varphi(s,\mathbf{x};\boldsymbol{\eta}))$. Figure 3.2 plotted $P^{cc}(Y = 1|\varphi(s,\mathbf{x};\boldsymbol{\eta}) \in [a_{sr},b_{sr}]), s = 1,2, r = 1,2,3,4$ against the benchmark in (3.5), $P^{e}(Y = 1|\varphi(s,\mathbf{x};\boldsymbol{\eta}) \in [a_{sr},b_{sr}]), s = 1,2, r = 1,2,3,4$. The "observed" vs "predicted" appeared to differ appreciably in scenario 1 particularly for stratum 2 and be similar in scenarios 2 and 3. In all the three scenarios, we first generated a cross-sectional sample of 40,000 subjects, among which 500 cases and 1000 controls for each stratum were selected into the case-control sample, that is, a total of 1000 cases and 2000 controls were included in the analysis. We applied both the proposed and "standard" methods, as described in the analysis of the BCDDP data, to analyze each of the simulated datasets to estimate parameters $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\omega})$. We repeated the simulation 1000 times.

The results for the three simulation scenarios were summarized in Tables 3.2, 3.3, and 3.4, respectively. The "True" parameter values listed were for the source population of the case-control sample, and the "Diff" was calculated as the difference between the estimates and their true values. Note that "Diff" is the estimation bias that is routinely used for assessing the consistency of an estimator in finite samples. But we avoid the term "bias" because in our context, larger "Diff" actually indicates that the constraints served to pull the estimates away from those obtained by the standard method. Therefore, larger "Diff" implicates the effectiveness of our methods for pursusing good calibration. The standard approach yielded largely unbiased estimates for all log ORs and the distribution parameters for Z in all three scenarios as expected. In constrast, for the proposed method, the averaged estimates of log OR for Z^c and its distribution parameters were close to the true parameter values. But the log OR estimates for predictors $\mathbf{X}$ varied across different scenarios. For Scenarios 1 and 2, the Diff for (X_1,X_3,X_4) was quite noticeable, especially for X_3 where the estimate was almost twice the true value. Two reasons might explain why the OR estimates for $\mathbf{X}$ was affected by the contraints and for Z largely unchanged. First, as we enforced equality of the calibration performance between the models with and without Z, the effects of $\mathbf{X}$ on Y would be changed under Scenario 1 to accomodate the original large differences between the two as shown in Figure 3.2. Second, different predictor distributions between the target and source populations could also lead to changes in point estimates. Even though the "observed" and "expected" were similar under Scenario 2, the difference in the $\mathbf{X}$ distribution resulted in biases for estimating the correponding ORs. This similar reasoning can help explain results in Scenario 3, where our method

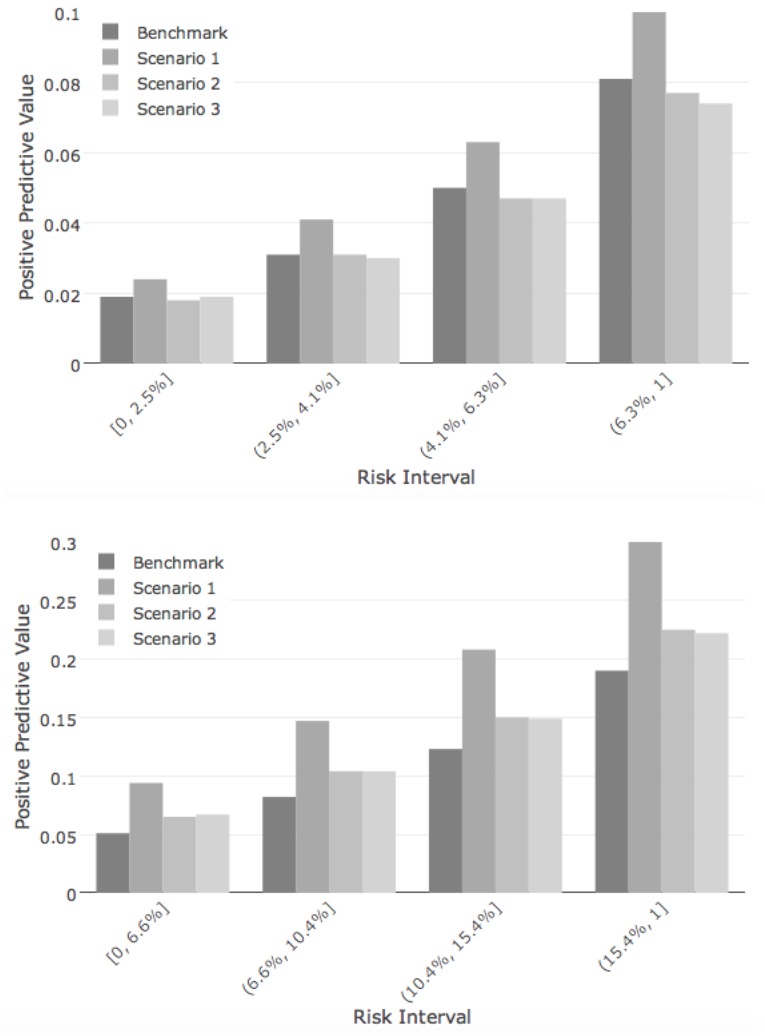

Figure 3.2: $P^e(Y = 1|\varphi(s, \mathbf{x}; \boldsymbol{\eta}))$ versus $P^{cc}(Y = 1|\varphi(s, \mathbf{x}; \boldsymbol{\eta}))$ under scenarios 1, 2, and 3. The upper panel represents the results for stratum 1 with the $1^{st}, 2^{nd}$, and 3^{rd} risk quartiles equal to 2.5%, 4.1%, and 6.3%, respectively. The lower panel represents the results for stratum 2 with the $1^{st}, 2^{nd}$, and 3^{rd} risk quartiles equal to 6.6%, 10.4%, and 15.4%, respectively.

also yielded unbiased estimates of ORs for predictor $\mathbf{X}$. Since neither the OR of Z^c nor its distribution were specified in the target population, heuristically, the inference of the related parameters should be largely dominated by the case-control sample. As observed in the results, estimates by the two methods were close in all three scenarios.

In all simulation scenarios, the averaged asymptotic standard error ("ASE) estimates were close to the empirical standard errors ("SE") for both methods under all three scenarios. For estimating parameters related to Z, our method was more efficient than the standard approach with 15%

reduction in the asymptotic variance for estimating the log OR and about 20% reduction on average for estimating parameters in the Beta-regression model, (γ, ω). In Scenario 3 in which the efficiency comparison of estimating $\beta_{\mathbf{x}}$ was sensible, our method yielded more efficienct estimates especially for X_1 and X_4, where the asymptotic variance reduced by more than 80%. The efficiency gains for estimating $(\beta_z, \gamma, \omega)$ were similar to those in the previous two scenarios.

Table 3.2: Estimation results under scenario 1. True: true parameter values; Est: mean estimates; Diff (%): the differences between mean estimates and true values divided by true values; SE: empirical standard error estimates; ASE: mean asymptotic standard error estimates;

		Proposed				Standard			
Parameter	True	Est	Diff (%)	SE	ASE	Est	Diff (%)	SE	ASE
α_1	-5.5	-5.887	7.04	0.193	0.193	-5.727	4.13	0.229	0.230
α_2	1.5	1.026	-31.60	0.021	0.022	1.072	-28.50	0.041	0.040
β_{x1}	0.15	0.179	19.33	0.008	0.008	0.148	-1.33	0.035	0.035
β_{x2}	0.2	0.207	3.50	0.055	0.053	0.198	-1.00	0.060	0.060
β_{x3}	0.2	0.481	140.50	0.039	0.041	0.200	0.00	0.052	0.052
β_{x4}	0.6	0.535	-10.83	0.018	0.017	0.601	0.17	0.051	0.051
β_{z^c}	0.15	0.146	-2.67	0.021	0.021	0.151	0.67	0.023	0.023
γ_0	1.2	1.200	0.00	0.047	0.047	1.198	-0.17	0.052	0.053
γ_1	-0.4	-0.418	4.50	0.038	0.037	-0.430	7.50	0.046	0.049
γ_2	0.2	0.198	-1.00	0.017	0.017	0.197	-1.50	0.018	0.018
γ_3	-0.2	-0.196	-2.00	0.023	0.023	-0.204	2.00	0.026	0.026
ω_0	1.2	1.205	0.42	0.055	0.053	1.199	-0.08	0.060	0.058
ω_1	0.2	0.193	-3.50	0.059	0.061	0.190	-5.00	0.061	0.062
ω_2	0.1	0.100	0.00	0.035	0.034	0.100	0.00	0.039	0.040

Table 3.3: Estimation results under scenario 2. True: true parameter values; Est: mean estimates; Diff (%): the differences between mean estimates and true values divided by true values; SE: empirical standard error estimates; ASE: mean asymptotic standard error estimates;

		Proposed				Standard			
Parameter	True	Est	Diff (%)	SE	ASE	Est	Diff (%)	SE	ASE
α_1	-5.8	-5.873	1.26	0.199	0.200	-5.721	-1.36	0.236	0.236
α_2	1.4	1.031	-26.36	0.022	0.023	1.067	-23.79	0.042	0.041
β_{x1}	0.15	0.175	16.67	0.008	0.008	0.152	1.33	0.036	0.036
β_{x2}	0.2	0.218	9.00	0.054	0.052	0.200	0.00	0.061	0.062
β_{x3}	0.2	0.442	121.00	0.044	0.045	0.196	-2.00	0.052	0.052
β_{x4}	0.6	0.548	-8.67	0.019	0.019	0.602	0.33	0.052	0.052
β_{z^c}	0.15	0.146	-2.67	0.022	0.021	0.149	-0.67	0.023	0.023
γ_0	1.2	1.200	0.00	0.047	0.045	1.196	-0.33	0.052	0.052
γ_1	-0.4	-0.406	1.50	0.040	0.039	-0.410	4.75	0.045	0.046
γ_2	0.2	0.199	-0.50	0.017	0.017	0.198	-1.00	0.019	0.019
γ_3	-0.2	-0.196	-2.00	0.023	0.023	-0.202	1.00	0.026	0.026
ω_0	1.2	1.203	0.25	0.054	0.055	1.196	-0.33	0.059	0.060
ω_1	0.2	0.196	-2.00	0.058	0.055	0.196	-2.00	0.060	0.058
ω_2	0.1	0.101	1.00	0.036	0.038	0.101	2.00	0.040	0.041

Table 3.4: Estimation results under scenario 3. True: true parameter values; Est: mean estimates; Diff (%): the differences between mean estimates and true values divided by true values; SE: empirical standard error estimates; ASE: mean asymptotic standard error estimates;

		Proposed				Standard			
Parameter	True	Est	Diff (%)	SE	ASE	Est	Diff (%)	SE	ASE
α_1	-5.8	-5.792	-0.14	0.208	0.210	-5.729	-1.22	0.248	0.247
α_2	1.4	1.150	-17.86	0.022	0.022	1.226	-12.43	0.039	0.040
β_{x1}	0.15	0.156	4.00	0.007	0.007	0.149	-0.67	0.038	0.038
β_{x2}	0.2	0.192	-4.00	0.060	0.058	0.202	1.00	0.065	0.064
β_{x3}	0.2	0.207	3.50	0.052	0.054	0.202	1.00	0.056	0.056
β_{x4}	0.6	0.638	6.33	0.019	0.019	0.599	-0.17	0.054	0.054
β_{z^c}	0.15	0.151	0.67	0.023	0.023	0.151	0.67	0.025	0.025
γ_0	1.2	1.204	0.33	0.047	0.047	1.197	-0.25	0.051	0.052
γ_1	-0.4	-0.411	2.75	0.041	0.038	-0.422	5.50	0.047	0.050
γ_2	0.2	0.201	0.50	0.017	0.017	0.198	-1.00	0.019	0.019
γ_3	-0.2	-0.201	0.50	0.025	0.025	-0.204	2.00	0.027	0.028
ω_0	1.2	1.200	0.00	0.063	0.062	1.197	-0.25	0.071	0.071
ω_1	0.2	0.202	1.00	0.065	0.064	0.196	-2.00	0.067	0.067
ω_2	0.1	0.101	1.00	0.043	0.044	0.100	0.00	0.048	0.049

3.5. Conclusion

Our method offers a way to exploit existing models for the development of new models that incoporate new predictors. Through suitable constraints for maximizing the likelihood function, our proposed method can yield a model that is calibrated similarly as the existing model, but it does not enforce that the model expanded with the biomarkers has to be the "true model" in the source population of the data. For predicting human diseases, the interest is often on identification of population subgroups who have high, or sometimes low, risk. Therefore, it is important that the model has good calibration in the tails of the risk distribution. Our method achieves this goal by setting up suitable constraints in our method. On the other hand, we put as loose constraints as possible on the predicted risk in the subgroup with moderate risk, which to a large extent allows the data to inform the relationship between the outcome, standard predictors, and biomarkers. When the data is collected under the case-control study design, such constraints can just be the outcome prevalence in that subgroup, which is necessary for estimating the intercept parameter in the logistic prediction model. To compare the impact of different constraints on the model developed, we also analyzed BCDDP data with eight constraints similarly as in Section 3, but with the BCRAT risk cutoff points placed at the (25%, 75%) percentiles for the stratum of age $\leq$ 50 and (15%, 25%, 75%, 85%) percentiles for the stratum of age $>$ 50. The OR estimates for the BCRAT predictors

can be quite different (Appendices). It is also computationally advantageous to place constraints in the risk groups that are of primary interest, since we found that too many constraints may lead to numerical problems. Of course, our method ensures good calibration only in population strata defined by standard predictors. It is ideal that the new model with biomarkers incorporated can be validated in a suitable cohort where data on standard predictors and biomarkers is available.

One limitation of our approach is that a parametric model is needed to relate biomarkers to standard predictors. Mis-specification of this model may negatively affect the calibration of the new model. For many human diseases which are rare, careful model selection can be performed using data from controls. It may also be helpful to explore more flexible model forms. When multiple biomarkers are involved, it is challenging to specify a parametric model, and it is largely infeasible to leave the biomarker distribution nonparametrically due to the curse of dimensionality. Extension of our work by relaxing the parametric distribution requirement is warranteed.

Our method particularly allows that the source and target populations can differ. As seen in the BCDDP, the BCDDP women and the general Caucasian women differed in the composite breast cancer rates and also breast cancer risk factor distributions. The BCRAT assumed that the odds ratio function for breast cancer was the same in the BCDDP and general Caucasian woman population. When the target population is well characterized and a prediction model exists, our method offers a way to develop a practical model using a data source that may deviate from the target population while ensuring desirable calibration. In the current work, we assumed that no information was available on the relationship between biomarkers and standard predictors. For biomarker discovery, it is frequent that the relationship between biomarkers and standard predictors is studied first, particularly when the outcome data is not readily available. It might be sensible to incorporate such established relationship into the model development by modifying the constraints that we exploited. Alternatively, it may make better sense to recognize that such relationship in the target population may better be approximated as an average of the external relationship and that relationship reflected in the case-control sample. We will investigate the feasibility of extending our method along these lines.

We analyzed a subset of the BCDDP cases and controls to demonstrate our methods, where the remaining subjects were excluded due to the lack of mammographic density data. The analysis was valid because the availability of mammograms only depended on the case-control status (Chen et

al., 2006). However, the efficiency can be enhanced if data on the BCRAT predictors for those excluded can be incorporated into the analysis. We will extend our method to accommodate the incomplete data in the future work.

CHAPTER 4

A Semiparametric Approach to Developing Well-Calibrated Models for Predicting the Risk of Binary Outcomes Using Cross-Sectional Data

4.1. Introduction

In this chapter, we extend the statistical methods developed in Chapter 3 to accommodate the cross-sectional studies, which often can not be treated as a random sample selected from the target population. We consider the same scenario where a well-calibrated model based on standard predictors exists, and the goal is to incorporate the new risk predictors into the existing model. We assume that data on the outcome and all predictors is available from a cross-sectional sample. We adopt the constrained maximum likelihood method proposed in Chapter 3, which guarantees that the new model calibrates similarly as the existing model, allows the distribution of standard risk predictors in the sample to be different from that in the target population of prediction, and relies on the data to infer the relationship between biomarkers and standard predictors. This work is actually motivated by the study of Gestational Diabetes Mellitus as described in Chapter 2 (Zhu et al., 2016). The Phase I sample in the study can be seen as a prospective cohort study, where data on the outcome and predictors including age, race, BMI and family history was measured for all study subjects. The distribution of age, race and BMI was externally available in the National Vital Statistics Report from Centers for Disease Control and Prevention (CDC), which turned out to be quite different from that in the data. Given that the information of the relationship between family history and these predictors was limited or unknown, we formulated the problem into our constrained MLE approach framework by treating family history as the "new" predictor. The goal was to develop a model with all four predictors included and the model was calibrated to an existing model based on age, race and BMI, where the latter was available in previous literature (Berkowitz et al., 1992; Solomon et al., 1997).

The differences of this work from what has been done in the previous chapter are listed here. First, there is no profile likelihood when deriving the objective function for our method because the empirical distribution of standard risk predictors will be totally factored out and irrelavent in the likelihood function in terms of estimating the association parameters. Second, the intercept term can be directly estimated via fitting a logistic regression model to the data under the standard approach, which is used for the comparison purpose. Lastly, the standard approach can use both cases and controls to estimate the distribution parameters for biomarkers because of the cross-sectional study designs.

The rest of this Chapter is organized as follows. We describe our proposed method and inference procedures in Section 4.2. In Section 4.3, we assess the finite sample performance of our method using simulated data, considering small or large differences between the data and the target population and statistical efficiency. In Section 4.4, we will apply our method to analyze the Phase I data of National Institute of Child Health and Human Development (NICHD) Fetal Growth Study Singletons to develop a logistic regression model for predicting the risk of gestational diabetes.

4.2. The Method

4.2.1. Notation and likelihood function

We consider a cross-sectional sample of size N, where the standard risk predictors and biomarkers are measured, denoted by $\mathbf{X}$ and $\mathbf{Z}$, respectively. Let Y represent the disease status with $Y = 1$ indicating cases and $Y = 0$ indicating controls. We use a logistic regression model to predict Y based on $(\mathbf{X}, \mathbf{Z})$:

$$P(Y = y|\mathbf{x}, \mathbf{z}) = \frac{\exp\{y(\alpha + \boldsymbol{\beta}_{\mathbf{x}}^T\mathbf{x} + \boldsymbol{\beta}_{\mathbf{z}}^T\mathbf{z})\}}{1 + \exp\{\alpha + \boldsymbol{\beta}_{\mathbf{x}}^T\mathbf{x} + \beta_{\mathbf{z}}^T\mathbf{z}\}}, y = 0, 1 \tag{4.1}$$

where $(\boldsymbol{\beta}_{\mathbf{x}}, \boldsymbol{\beta}_{\mathbf{z}})$ are the log odds ratios for $(\mathbf{X}, \mathbf{Z})$. We use a parametric model $f_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x})$ with Euclidean parameters $\boldsymbol{\tau}$ to describe the conditional distribution of $\mathbf{Z}|\mathbf{X}$. Let $P_{y\mathbf{xz}} \equiv P(Y = y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}), y = 0, 1$. The empirical log-likelihood function of the sample can be derived as

$$\begin{aligned} l(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}) &= \log \prod_{i=1}^{N} P(Y_i = y_i, \mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i) \\ &= \log \prod_{i=1}^{N} P(Y_i = y_i|\mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i) P(\mathbf{X}_i = \mathbf{x}_i) P(\mathbf{Z}_i = \mathbf{z}_i|\mathbf{X}_i = \mathbf{x}_i) \\ &\propto \sum_{i=1}^{N} \log P(y_i|\mathbf{x}_i, \mathbf{z}_i) + \sum_{i=1}^{N} \log f_{\boldsymbol{\tau}}(\mathbf{z}_i|\mathbf{x}_i) \end{aligned}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathbf{x}}, \boldsymbol{\beta}_{\mathbf{z}})$. The log-likelihood function is much simpler than that for the case-control data, since we can factor out the empirical distribution of $\mathbf{X}$ because it doesn't involve $(\alpha, \boldsymbol{\beta})$. We assume that a well-calibrated risk prediction model based on $\mathbf{X}$ is available and well-calibrated for strata defined by $\mathbf{X}$, and we use $\varphi(\mathbf{X})$ to represent the corresponding predicted risk. The distribution of $\mathbf{X}$ in the target population of prediction, $P^e(\mathbf{X})$, is known from external sources, where the superscript "e" here and after indicates "external". We explicitly allow $P^e(\mathbf{X})$ to be different from that in the

data. We categorize predicted risk φ into I intervals and use a_r and b_r to denote the beginning and end of each interval. We assume that the calibration of the model $\varphi(\mathbf{X})$ was evaluated in external studies by comparing the averaged predicted risk within each interval, defined as

$$P^e(Y=1|a_r \leq \varphi(\mathbf{x}) \leq b_r), r=1,...,I,$$

with the "observed" averaged risk. We impose the equality of these averaged risks between the newly developed model (4.1) and $\varphi(\mathbf{X})$ to ensure a good calibration performance of the new model. The resultant constraints are expressed as below:

$$P^e(Y=1|a_r \leq \varphi(\mathbf{x}) \leq b_r) = \frac{\int_{\mathbf{x}:a_r \leq \varphi(\mathbf{x}) \leq b_r} \int_{\mathbf{z}} P_{1\mathbf{xz}} P^e(\mathbf{x}) f_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} d\mathbf{x}}{\int_{x:a_r \leq \varphi(\mathbf{x}) \leq b_r} P^e(\mathbf{x}) d\mathbf{x}},\ r=1,...,I. \tag{4.2}$$

4.2.2. Constrained maximum likelihood for estimating $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$

We apply the similar approach as developed in the previous chapter, i.e., maximize the log-likelihood function $l(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau})$ subject to the constraint (4.2). The objective function using Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_r : r = 1,...,I)$ is written as

$$\begin{aligned} &g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda}) = l(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}) \\ &+ \sum_{r=1}^{I} \lambda_r \left\{ \frac{\int_{\mathbf{x}:a_r \leq \varphi(\mathbf{x}) \leq b_r} \int_{\mathbf{z}} P_{1\mathbf{xz}} P^e(\mathbf{x}) f_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}:a_r \leq \varphi(\mathbf{x}) \leq b_r} P^e(\mathbf{x}) d\mathbf{x}} - P^e(Y=1|a_r \leq \varphi(\mathbf{x}) \leq b_r) \right\}. \end{aligned} \tag{4.3}$$

To maximize the above function with respect to all unknown parameters, we derive the corresponding score functions by taking the first derivative of $g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ with respect to $(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$, respectively. The score functions are denoted by $\{S_\alpha, S_{\boldsymbol{\beta}}, S_{\boldsymbol{\tau}}, S_{\boldsymbol{\lambda}}\}$ and the maximum likelihood estimates $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\lambda}})$ are obtained via solving $\left\{S_\alpha^T, S_{\boldsymbol{\beta}}^T, S_{\boldsymbol{\tau}}^T, S_{\boldsymbol{\lambda}}^T\right\}^T = \mathbf{0}$. The details are provided in the Appendices.

4.2.3. Identifiability, consistency and asymptotic normality

The theorems developed in Chapter 3 for case-control data can be applied in this cross-sectional data setting.

4.3. Simulation Studies

We conducted extensive simulation studies to evaluate the finite sample performance of our proposed methods. To make clear distinction between the target population of prediction and the cross-sectional sample, we set up the simulation scheme in two steps. First, we defined a population where the true distribution of $\mathbf{X}$, $P^e(\mathbf{X})$, was known, and a well-calibrated risk prediction model based on $\mathbf{X}$ was available. Second, we generated data for the established risk predictors $\mathbf{X}$ and biomarker Z for a cross-sectional sample.

4.3.1. Step 1: Define $P^e(\mathbf{X})$ and $\varphi(\mathbf{X})$

We considered three standard predictors $\mathbf{X} = (X_1, X_2, X_3)$, where $X_1 \sim N(30, 5)$ was categorized into 5 groups by $\leq$ 25, (25, 30], (30, 35], (35, 40] or $>$ 40, X_2 was generated from a Multinomial distribution with probabilities (0.3, 0.2, 0.2, 0.3), and $X_3 \sim Poisson(25)$ was categorized into 6 groups by $\leq$ 20, (20, 25], (25, 30], (30, 35], (35, 40] or $>$ 40. The three predictors were mutually independent, thus their joint distribution, $P^e(\mathbf{X})$, can be expressed as $\prod_{i=1}^{3} P^e(X_i)$, which was treated as known. We assumed that the following logistic regression model for predicting the risk of Y based on $\mathbf{X}$, where X_1 and X_3 were fitted as ordinal while X_2 was a categorical variable, was well calibrated:

$$\varphi(\mathbf{x};\boldsymbol{\eta}) \equiv p^e(Y=1|\mathbf{x};\boldsymbol{\eta}) = \frac{\exp(\eta_0 + \eta_1 x_1 + \boldsymbol{\eta_2}^T \boldsymbol{x_2^*} + \eta_3 x_3)}{1 + \exp(\eta_0 + \eta_1 x_1 + \boldsymbol{\eta_2}^T \boldsymbol{x_2^*} + \eta_3 x_3)},$$

where $\mathbf{x}_2^*$ was a vector of indicator functions with length 3 that represented x_2. We set the parameters $(\eta_0, \eta_1, \boldsymbol{\eta_2}, \eta_3)$ to be (-5, 0.25, 0.5, 0.5, 0.5, 0.5) so that the disease prevalence of Y, $P(Y=1)$, was around 6.5%. Consistent with the common goodness-of-fit test of calibration, we chose the quartiles of $\varphi(\mathbf{x};\boldsymbol{\eta})$ as the risk cutoff values used in the constraints (4.2). Together with $P^e(\mathbf{X})$, we were able to calculate the averaged risk within each interval $[a_r, b_r], r = 1, 2, 3, 4$ as

$$P^e(Y=1|\varphi(\mathbf{x};\boldsymbol{\eta}) \in [\boldsymbol{a}, \boldsymbol{b}]) = (2.5\%, 4.2\%, 6.7\%, 12.2\%). \tag{4.4}$$

These four numbers were considered as the calibration benchmark, leading to four corresponding constraints as described in (4.2).

4.3.2. Step 2: Generate the Cross-Sectional Sample Data

We considered two scenarios. In the first scenario, we generated $\mathbf{X}$ from distributions that were different from those in *Step 1* by using different parameter values. $X_1 \sim N(33, 5)$, X_2 was generated from a Multinomial distribution with probabilities (0.25, 0.25, 0.25, 0.25), and X_3 from a Poisson distribution with mean 28. All three variables were coded the same way as in *Step 1*. In the second scenario, $\mathbf{X}$ followed the same distribution as that in *Step 1*. We considered a single biomarker Z with two values 1 or 0 that followed the logistic regression model,

$$\text{logit}P(Z = 1|x_1, x_3) = \tau_0 + \tau_1 x_1 + \tau_2 x_3.$$

We set (τ_0, τ_1, τ_2) as (-2, 0.2, 0.2). The binary outcome Y was generated from the logistic regression model (4.1), and the log OR parameters for $(X_1, \mathbf{X}_2^*, X_3, Z)$ were set to be the same under the two scenarios and equal to (0.2, 0.45, 0.45, 0.55, 0.55, 0.5). The log OR values for $(X_1, \mathbf{X}_2^*, X_3)$ were set to be reasonably close to those in the existing model $\varphi(\mathbf{x}; \boldsymbol{\eta})$, since we wanted to maintain the association between standard predictors and the disease in the sample and in the population. We chose different α values so that the disease prevalence in the two scenarios was 11.5% and 7.0%, respectively. Under scenario 1, we set up the covariate distribution and outcome prevalence to be different from those in the population described in *Step 1* to mimic the differences between the GDM study and the national data as represented in CDC. That is, the prospective cohort study differs from the target population. The difference between the populations in the two scenarios and that in *Step 1* is depicted in Figure 4.1, where we plotted the "predicted" by model φ versus vs the "observed" in each scenario. To do this, we first generated a large sample of size 10^7 based on the distribution of $\mathbf{X}$ and Z and $P(Y|\mathbf{X}, Z)$ specified above. Using the same partition of the $\mathbf{X}$ space as that equivalent to $a_r <= \varphi(\mathbf{x}; \boldsymbol{\eta}) <= b_r$ in *Step 1*, we obtained the proportion of cases in each of the subspaces ("observed"), denoted by $P^{cc}(Y = 1|\varphi(\mathbf{x}; \boldsymbol{\eta}))$. Figure 4.1 plotted $P^{cc}(Y = 1|\varphi(\mathbf{x}; \boldsymbol{\eta}) \in [a_r, b_r]), r = 1, 2, 3, 4$ against the benchmark in (4.4), $P^e(Y = 1|\varphi(\mathbf{x}; \boldsymbol{\eta}) \in [a_r, b_r]), r = 1, 2, 3, 4$. The "observed" vs "predicted" appeared to differ appreciably in scenario 1 while be similar in scenarios 2. In both scenarios, we generated 5,000 subjects for the analysis. We applied both the proposed and "standard" methods, as described in the analysis of the GDM data, to analyze each of the simulated datasets to estimate parameters $(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau})$. We repeated the simulation 1000 times.

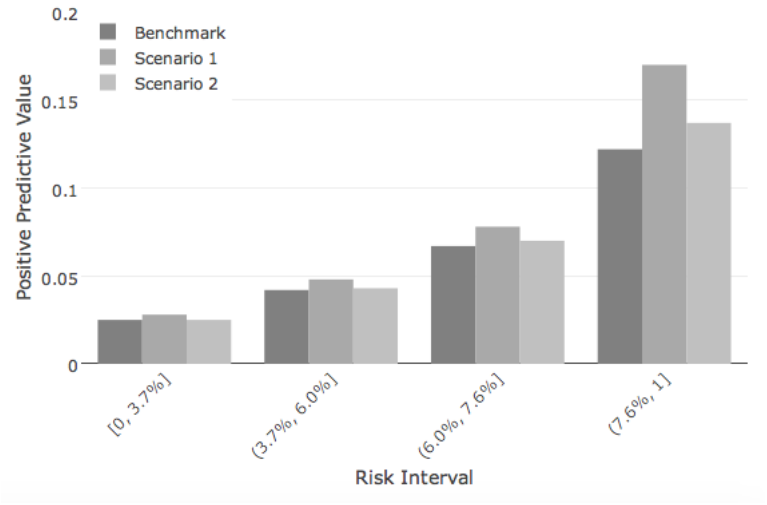

Figure 4.1: $P^e(Y = 1|\varphi(\mathbf{x};\boldsymbol{\eta}))$ versus $P^{cc}(Y = 1|\varphi(\mathbf{x};\boldsymbol{\eta}))$ under scenarios 1 and 2 with the $1^{st}, 2^{nd}$, and 3^{rd} risk quartiles equal to 3.7%, 6.0%, and 7.6%, respectively.

The results for the two simulation scenarios were summarized in Tables 4.1 and Table 4.2, respectively. The "True" parameter values listed were for the cross-sectional sample, and the "Diff" was calculated as the difference between the estimates and their true values. Note that "Diff" is the estimation bias that is routinely used for assessing the consistency of an estimator in finite samples. But we avoided the term "bias" because in our context, larger "Diff" actually indicated that the constraints served to pull the estimates away from those obtained by the standard method. Therefore, larger "Diff" implicated the effectiveness of our methods for pursuing good calibration. The standard approach yielded largely unbiased estimates for all log ORs and the distribution parameters for Z in both scenarios as expected. In constrast, for the proposed method, the averaged estimates of log OR for Z and its distribution parameters were close to the true parameter values. But the log OR estimates for predictors $\mathbf{X}$ varied across different scenarios. For Scenarios 1, the Diff for $(X_1, \mathbf{X}_2^*, X_3)$ was quite noticeable. Two reasons might explain why the OR estimates for $\mathbf{X}$ was affected by the contraints and for Z largely unchanged. First, as we enforced equality of the calibration performance between the models with and without Z, the effects of $\mathbf{X}$ on Y would be changed under Scenario 1 to accomodate the original large differences between the two as shown in Figure 4.1. Second, different predictor distributions between the target population and the data could also lead to changes in point estimates. On the other hand, under Scenario 2 where the "observed" and "expected" risks were similar and $P(\mathbf{X})$ was the same between the data and the

Table 4.1: Estimation results under scenario 1. True: true parameter values; Est: mean estimates; Diff (%): the differences between mean estimates and true values divided by true values; SE: empirical standard error estimates; ASE: mean asymptotic standard error estimates;

		Proposed				Standard			
Parameter	True	Est	Diff (%)	SE	ASE	Est	Diff (%)	SE	ASE
α	-5	-4.982	-0.36	0.080	0.079	-5.170	3.40	0.192	0.193
β_{x1}	0.2	0.222	11.00	0.040	0.040	0.199	-0.50	0.047	0.046
β_{x21}	0.45	0.361	-19.78	0.134	0.132	0.448	-0.44	0.139	0.139
β_{x22}	0.45	0.363	-19.33	0.133	0.132	0.450	0.00	0.138	0.139
β_{x23}	0.55	0.381	-30.73	0.128	0.128	0.551	0.18	0.134	0.135
β_{x3}	0.55	0.495	-10.00	0.021	0.022	0.551	0.18	0.042	0.042
β_z	0.5	0.496	-0.80	0.091	0.092	0.497	-0.60	0.091	0.092
τ_0	-2	-1.998	-0.10	0.136	0.136	-2.007	0.35	0.137	0.136
τ_1	0.2	0.202	1.00	0.031	0.031	0.202	1.00	0.031	0.031
τ_2	0.2	0.194	-3.00	0.029	0.029	0.200	0.00	0.029	0.030

population, our method yielded almost unbiased estimates of ORs for predictor $\mathbf{X}$. Since neither the OR of Z nor its distribution were specified in the target population, heuristically, the inference of the related parameters should be largely dominated by the data itself. As observed in the results, estimates by the two methods were close in both scenarios.

In the two simulation scenarios, the averaged asymptotic standard error ("ASE) estimates were close to the empirical standard errors ("SE") for both methods. For estimating parameters related to Z, the standard approach was as efficient as our method for estimating $(\beta_z, \boldsymbol{\tau})$. Unlike the standard approach in Chapter 3 where only controls could be used to estimate the distribution parameters for biomarkers, both cases and controls were used for the estimation in current work because of the cross-sectional study design. In Scenario 2 where the efficiency comparison of estimating $\boldsymbol{\beta}_{\mathbf{x}}$ was sensible, our method yielded more efficienct estimates especially for X_3, where the asymptotic variance reduced by more than 70%.

4.4. The Analysis of a Study of Gestational Diabetes Mellitus

Using Phase I data (as described in Chapter 2) from the National Institute of Child Health and Human Development (NICHD) Fetal Growth Study Singletons, we applied our proposed methods to build a logistic regression model for predicting the risk of gestational diabetes mellitus (GDM). In this prospective study cohort (Zhu et al., 2016), data on age, race, body mass index, and family history of diabetes was fully collected for 2,799 women. In the model, we considered 5-year age

Table 4.2: Estimation results under scenario 2. True: true parameter values; Est: mean estimates; Diff (%): the differences between mean estimates and true values divided by true values; SE: empirical standard error estimates; ASE: mean asymptotic standard error estimates;

		Proposed				Standard			
Parameter	True	Est	Diff (%)	SE	ASE	Est	Diff (%)	SE	ASE
α	-5.1	-4.964	-2.67	0.102	0.098	-5.177	1.51	0.238	0.239
β_{x1}	0.2	0.187	-6.50	0.052	0.052	0.200	0.00	0.057	0.057
β_{x21}	0.45	0.423	-6.00	0.169	0.167	0.457	1.56	0.173	0.174
β_{x22}	0.45	0.425	-5.56	0.164	0.165	0.459	2.00	0.172	0.172
β_{x23}	0.55	0.526	-4.36	0.150	0.150	0.560	1.82	0.156	0.157
β_{x3}	0.55	0.515	-6.36	0.027	0.026	0.550	0.00	0.055	0.055
β_z	0.5	0.501	0.20	0.120	0.121	0.501	0.20	0.120	0.122
τ_0	-2	-1.995	-0.25	0.123	0.123	-2.002	0.10	0.123	0.123
τ_1	0.2	0.200	0.00	0.033	0.034	0.201	0.50	0.033	0.033
τ_2	0.2	0.197	-1.50	0.033	0.032	0.199	-0.50	0.033	0.032

intervals and fitted them as ordinal variable after exploring their functional forms by local polynomial regression. BMI (kg/m^2) was categorized into three groups as normal if the value was from 18.5 to 25, overweight if 25 to 30, and obese if over 30, and fitted as categorical variable. Both race (White, Black, Hispanic, Asian) and family history (Yes or No) were fitted as categorical variables as well.

To ensure a good calibration performance of our model, i.e., to accurately predict the expected number of GDM cases in defined population subgroups, we imposed the constraints based on the results published in Berkowitz et al. (1992) and Solomon et al. (1997). In their work, the association between (Age, Race, BMI) and gestational diabetes has been well studied with large cohort data using logistic regression models. We assumed that the model based on $\mathbf{X} =$ (Age, Race, BMI) with log ORs reported in these paper to be well-calibrated. Let $\varphi(\mathbf{X})$ denote the corresponding predicted risk and the expression was provided as below

$$\begin{aligned}\varphi(\mathbf{x}) &= P^e(Y=1|\mathbf{x}) \\ &= \frac{\exp\{\eta_0+\eta_1 Age^*+\eta_2 Black+\eta_3 Hispanic+\eta_4 Asian+\eta_5 Overweight+\eta_6 Obese\}}{1+\exp\{\eta_0+\eta_1 Age^*+\eta_2 Black+\eta_3 Hispanic+\eta_4 Asian+\eta_5 Overweight+\eta_6 Obese\}},\end{aligned}$$

where Age^* represented the recoded 5-year age intervals and $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ reported by Berkowitz et al. (1992) and Solomon et al. (1997) was equal to $\log(1.63)$, $\log(1.75)$, $\log(1.45)$, $\log(2.32)$, $\log(2.13)$, $\log(2.90)$, respectively. Given the disease prevalence of 4% reported in literature, we were able to recover the intercept term η_0. By equating the prediction performance of our new model which also included $Z =$ family history, which was treated as biomarker in this analysis,

to that of $\varphi(\mathbf{X})$, we would successfully maintain the "well-calibration" property. Moreover, the distribution of $\mathbf{X}$, $P^e(\mathbf{X})$, was externally available from the National Vital Statistics Reports (NVSR) of Centers for Disease Control and Prevention (CDC). Due to the limited knowledge of the population-level distribution of family history at different ages and BMI values in different race groups, we proposed to model $f(Z|\mathbf{X})$ via a logistic regression model. Implicitly, we assumed $f(Z|\mathbf{X})$ to be the same in the data and in the general population. Based on $\varphi(\mathbf{X})$ and $P^e(\mathbf{X})$, we were able to calculate $P^e(Y = 1|a_r \leq \varphi(\mathbf{x}) \leq b_r)$. We considered two sets of $[a_r, b_r]$ values, where the first set was equal to the quartiles of $\varphi(\mathbf{X})$ while the second set was equal to the (15%, 25%, 75%, 85%) percentiles. Note that the latter imposed much finer constraints on the tails of the risk distribution, i.e., subjects classified as low or high risk were given more weights. The calculated averaged risks within each set were summarized as below

$$P^e(Y = 1|\varphi(\mathbf{x}) \in [\boldsymbol{a}^1, \boldsymbol{b}^1]) = (1.3\%, 3.2\%, 5.6\%, 11.0\%),$$
$$P^e(Y = 1|\varphi(\mathbf{x}) \in [\boldsymbol{a}^2, \boldsymbol{b}^2]) = (1.1\%, 1.9\%, 4.2\%, 9.2\%, 12.8\%).$$

We used these numbers as the benchmark for risk prediction calibration and the resultant constraints were expressed as in (4.2). Again, we compared our proposed constrained maximum likelihood method to a "standard" approach, where two logistic regression models were fitted to the data. The first one was to estimate the OR parameters in the risk prediction model and the second one was to estimate the distribution parameters for family history. The stanard error estimates for the standard approach were calculated using stantard MLE approach. Unlike the "standard" approach for case-control data, we were able to directly estimate the intercept parameter and use both cases and controls to estimate the distribution parameters for family history because of the non-retrospective nature of the data.

The results were summarized in Table 4.3. Older age, higher BMI and positive family history of diabetes appeared to be positively associated with the risk of gestational diabetes. Asian and Hispanic women had a higher chance of developing GDM than White women, while Black women were not statistically significant different from White women. However, the estimated log ORs obtained by our method for risk predictors could be quite different from those by the standard method. When using the quartiles of $\varphi(\mathbf{X})$ as the risk cutoff points in the constraints, the estimate for age was 0.5 times larger while the estimates for Asian, Overweight and Obese were smaller. Because our method

assumed the same model relating the risk of GDM with predictors in the general population and the data, these differences might indicate the effectiveness of our method for pursuing good calibration. We conjectured that the differences in the standard predictors distribution, $P(\mathbf{X})$, between the data and the NVSR largely drove the discrepancy in the OR estimates. As shown in Table 4.4, the distribution of age and race differed substantially from the reported estimates in NVSR while the distribution of BMI was reasonably different from that in NVSR. We noticed that different constraints impacted differently on the log OR estimates for $\mathbf{X}$. Using (15%, 25%, 75%, 85%) percentiles of $\varphi(\mathbf{X})$ as the cutoffs, cMLE^2 yielded smaller log OR estimate for age than cMLE^1 but still larger than the standard approach. For race, cMLE^2 yielded larger estimates for Asian women than the standard approach, while cMLE^1 was the opposite. Since the new contraints were more focused on women classified as low or high risk instead of women with moderate risk, the log OR estimates would be changed accordingly to accomodate the different association effects between $\mathbf{X}$ and Y in these "extreme" groups. With regard to family history, both the estimates of log OR and distribution parameters were similar across the three approaches. Our assumption of the same distribution of family history in the population and in the data might explain these consistencies. In the logistic regression model for family history, all predictors were significant and positively associated with the probability of having a positive family history. In terms of statistical efficiency, our approach with either type of constraints yielded more efficient log OR estimates for $\mathbf{X}$ compared to the standard method, with more than 80% efficient gains for intercept, age and obese and more than 40% for the remainings. On the other hand, the standard errors for the log OR estimate of family history and its distribution parameters estimates were similar between our method and the standard approach because the latter was able to use both cases and controls for the estimation due to the prospective nature of the data.

4.5. Conclusion

Our proposed method offers a way to exploit the existing models for the development of new models that incorporate new predictors. By calibrating the newly developed model to the existing model through suitable constraints, our approach achieves similar calibration performance, but it doesn't enforce that the expanded model has to be the "true" model in the cross-sectional study. We also explore the effect of different types of constraints on the odds ratio estimates in the logistic regression model developed. In the analysis of GDM, we first impose the constraints based on the

Table 4.3: Analysis of Gestational Diabetes Mellitus data: estimates of intercept term and log ORs for 5-year age intervals, race, BMI and family history, together with estimates of parameters in the logistic regression model for the distribution of family history. In the paranthesis are the corresponding estimates of asymptotic standard errors; "cMLE^1" represents estimates from the proposed constrained maximum likelihood method when using quartiles of $\varphi(\mathbf{x})$ as $a's$ and $b's$ in the constraints; "cMLE^2" represents estimates from the proposed constrained maximum likelihood method when using (15%, 25%, 75%, 85%) percentiles of $\varphi(\mathbf{x})$ as $a's$ and $b's$ in the constraints; "Standard" represents the estimates from the standard approach.

		cMLE^1	cMLE^2	Standard
Predictors	Intercept	-5.810 (0.064)	-5.529 (0.059)	-5.300 (0.442)
	Age*	0.522 (0.020)	0.419 (0.027)	0.339 (0.098)
	Black	-0.381 (0.290)	-0.103 (0.133)	-0.539 (0.352)
	Hispanic	0.484 (0.216)	0.640 (0.150)	0.447 (0.274)
	Asian	0.592 (0.237)	0.779 (0.238)	0.692 (0.319)
	Overweight	0.599 (0.190)	0.601 (0.183)	0.805 (0.261)
	Obese	1.175 (0.104)	1.202 (0.104)	1.435 (0.274)
	Family history	0.581 (0.218)	0.602 (0.210)	0.585 (0.220)
Family history Model	Intercept	-2.590 (0.193)	-2.618 (0.192)	-2.574 (0.193)
	Age*	0.191 (0.044)	0.197 (0.044)	0.187 (0.044)
	Black	0.542 (0.143)	0.557 (0.143)	0.536 (0.143)
	Hispanic	0.678 (0.136)	0.690 (0.136)	0.675 (0.137)
	Asian	0.692 (0.155)	0.704 (0.155)	0.692 (0.155)
	Overweight	0.348 (0.114)	0.352 (0.114)	0.351 (0.114)
	Obese	0.699 (0.128)	0.709 (0.128)	0.703 (0.128)

Table 4.4: Distribution of age, race and BMI for pregnant women estimated in the Gestational Diabetes Mellitus data and reported in the National Vital Statistics Report (NVSR).

		GDM data	NVSR
Age	$(15, 20)$	10.1%	19.6%
	$[20, 25)$	23.5%	20.8%
	$[25, 30)$	31.2%	20.5%
	$[30, 35)$	25.1%	20.1%
	$[35, 40]$	10.1%	19.0%
Race	White	26.8%	63.1%
	Black	27.8%	13.4%
	Hispanic	28.8%	17.2%
	Asian	16.6%	6.3%
BMI	Normal	56.5%	51.2%
	Overweight	26.5%	24.9%
	Obese	17.0%	23.9%

quartiles of the risk distribution from external model, i.e., we divide the group of interests evenly according to their predicted risks. Second, we place finer constraints on the tails of the risk distribution while cruder constraints in the middle, i.e., we focus more on the group with higher or lower risk and less on the group with moderate risk. Due to the numerical problems caused by too many constraints, we might prefer to put constraints in the risk groups that are of primary interest.

Our approach also explicitly allows the distribution of standard risk predictors to be different between the cross-sectional study and the target population of prediction. As seen in the GDM data, the distribution of race and 5-year age intervals differed substantially from that estimated in the National Vital Statistics Report. Our method provides a practical way of integrating data sources that may be different from each other while ensuring desirable calibration performance. On the other hand, we relies on the data itself to infer the relationship between biomarkers and the standard predictors. We use a parametric model to relate the two and one limitation of this approach is model misspecification, which might negatively affect the calibration performance of our model. Therefore, it might be attractive to relax the parametric assumption and adopt more flexible model forms. For biomarker discovery, the relationship between biomarkers and standard risk predictors is always studied first, which makes it sensible to incorporte this knowledge into our model development via constraints. We will investigate the feasibility of extending our method along these lines.

There are two main differences between applying our appoach to a cross-sectional study and applying it to a case-control study. First, the likelihood function of a cross-sectional study is straightforward and simpler than that of a case-control study, because the empirical distribution of standard predictors is irrelavent during the estimation process for the former, but has to be factored out via profile likelihood approach for the latter. Second, in term of the statistical efficiency of estimating the distribution parameters of biomarkers, our method is as efficient as the standard approach for analyzing a cross-sectional study while is more efficient for analyzing a case-control study. This is because the current sampling scheme allows the standard method to use both cases and controls to estimate the distribution of biomarkers. On the contrary, the standard method can only use controls for the estimation in a case-control study and thereby lose the efficiency.

CHAPTER 5

DISCUSSION

In this dissertation, we developed an arsenal of statistical methods to address the statistical challenges in developing and validating models for predicting binary outcomes when the biomarkers have limited availability. We first proposed three approaches, known as semiparametric maximum likelihood, pseudo-likelihood and weighted likelihood, to estimating the risk distribution and summary measures of predictive accuracy under two-phase studies. We also developed a novel sampling strategy for selecting Phase II subjects based on a preliminary model that included only Phase I predictors to improve the efficiency of estimating the predictive accuracy measures. We then applied the proposed methods and sampling strategy to develop and evaluate a risk prediction model for Gestational Diabetes Mellitus. Later, motivated by the lack of independent data to validate biomarkers for prediciotn, we incorporated the external knowledge of the distribution of standard risk predictors and a known well-calibrated model built on them in the target population of prediction, and proposed a novel constrained maximum likelihood approach to develop a model that guaranteed the good calibration performance. We applied the proposed statistical methods to both case-control studies and cross-sectional studies. With the application to a case-control study nested in BCDDP to evaluate the added value of percent mammographic density on breast cancer risk prediction and an ongoing prospective study of Gestational Diabetes, respectively, we demonstrated that our approach ensured good calibration performance and yielded more efficient odds ratio estimates.

These methods although have their own limitations as described in each chapter, they are readily applicable to binary outcome risk prediction problems when new predictors have missing data, or developing a well-calibrated model when there was lack of independent data to validate the biomarkers for prediction and external information of established predictors is available.

APPENDIX A

THEORETICAL DERIVATIONS

A.1. Chapter 2

A.1.1. Estimators of PCF under PL, and estimators of PNF and AUC under ML, PL and WL methods

$$\widehat{\mathsf{PCF}}_q(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}) = \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})}{\sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})} \right\}^{-1} \times \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL}) > \xi_q\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})}{\sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})} \right\},$$

where ξ_q is defined by equation

$$q = \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL}) > \xi_q\}}{N \sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})}.$$

$$\widehat{\mathsf{PNF}}_p(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}) = \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML}) > \xi_p\}}{N \sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})},$$

where ξ_p is defined by equation

$$p = \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})}{\sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})} \right\}^{-1} \times \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML}) > \xi_p\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})}{\sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})} \right\}.$$

$$\widehat{\mathsf{PNF}}_p(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}) = \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL}) > \xi_p\}}{N \sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})},$$

where ξ_p is defined by equation

$$p = \left\{ \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \frac{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})}{\sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})} \right\}^{-1} \times \left\{ \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL}) > \xi_p\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})}{\sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})} \right\}.$$

$$\widehat{\mathsf{PNF}}_p(\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}) = \frac{1}{N} \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \hat{\pi}_{ys}^{-1} I\{p(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) > \xi_p\},$$

where ξ_p is defined by equation

$$p = \frac{\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \hat{\pi}_{ys}^{-1} I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) > \xi_p\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL})}{\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \hat{\pi}_{ys}^{-1} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL})}.$$

$$\widehat{\mathsf{AUC}}(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}) = \int \widehat{\mathsf{TPR}}_\nu(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}) d\{\widehat{\mathsf{FPR}}_\nu(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML})\},$$

where

$$\widehat{\mathsf{TPR}}_\nu(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}) = \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})}{\sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})} \right\}^{-1} \times \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML}) > \nu\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})}{\sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})} \right\}$$

and

$$\widehat{\mathsf{FPR}}_\nu(\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}) = \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{\{1 - p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})\}}{\sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})} \right\}^{-1} \times \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML}) > \nu\} \left\{1 - p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})\right\}}{\sum_{y=0}^{1} \hat{\mu}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{ML})} \right\}.$$

$$\widehat{\mathsf{AUC}}(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}) = \int \widehat{\mathsf{TPR}}_\nu(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}) d\{\widehat{\mathsf{FPR}}_\nu(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL})\},$$

where

$$\widehat{\mathsf{TPR}}_\nu(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}) = \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{p(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})}{\sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})} \right\}^{-1} \times \left\{ \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL}) > \nu\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})}{\sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})} \right\}$$

and

$$\widehat{\mathsf{FPR}}_\nu(\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}) = \left\{ \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \frac{\{1 - p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})\}}{\sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})} \right\}^{-1} \times \left\{ \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \frac{I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL}) > \nu\} \left\{1 - p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})\right\}}{\sum_{y=0}^{1} \hat{\pi}_{ys} p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL})} \right\}.$$

$$\widehat{\mathsf{AUC}}(\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}) = \int \widehat{\mathsf{TPR}}_\nu(\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}) d\{\widehat{\mathsf{FPR}}_\nu(\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL})\},$$

where

$$\widehat{\mathsf{TPR}}_\nu(\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}) = \left\{ \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \hat{\pi}_{ys}^{-1} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) \right\}^{-1} \times \left\{ \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \hat{\pi}_{ys}^{-1} I\{p(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) > \nu\} p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) \right\}$$

and

$$\widehat{\mathsf{FPR}}_\nu(\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}) = \left\{ \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \hat{\pi}_{ys}^{-1} \left\{1 - p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL})\right\} \right\}^{-1} \times \left\{ \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk} \hat{\pi}_{ys}^{-1} I\{p_1(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL}) > \nu\} \left\{1 - p(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{WL})\right\} \right\}.$$

A.1.2. Large Sample Theories for the Proposed Estimators of Predictive Accuracy Measures

For ML, PL and WL methods, recall that we write the estimated predictive accuray statistics as $\hat{T}_{ML} = T\{\hat{\boldsymbol{\theta}}_{ML}, \hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})\}, \hat{T}_{PL} = T\{\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}})\}$ and $\hat{T}_{WL} = T\{\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}(\hat{\boldsymbol{\pi}})\}$, respec-

tively. For ML, we perform the following decomposition,

$$\begin{aligned}\sqrt{N}\left[T\{\hat{\boldsymbol{\theta}}_{ML},\hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})\}-T(\boldsymbol{\theta},F)\right] &= \sqrt{N}\left[T\{\hat{\boldsymbol{\theta}}_{ML},\hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})\}-T\{\boldsymbol{\theta},\hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})\}\right]\\ &+ \sqrt{N}\left[T\{\boldsymbol{\theta},\hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})\}-T\{\boldsymbol{\theta},\hat{F}_{ML}(\boldsymbol{\theta})\}\right]\\ &+ \sqrt{N}\left[T\{\boldsymbol{\theta},\hat{F}_{ML}(\boldsymbol{\theta})\}-T(\boldsymbol{\theta},F)\right].\end{aligned}$$

Apply the first-order Taylor series expansion, we obtain

$$\begin{aligned}&\sqrt{N}\left[T\{\hat{\boldsymbol{\theta}}_{ML},\hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})\}-T(\boldsymbol{\theta},F)\right]\\ &=\left\{\frac{\partial T}{\partial\boldsymbol{\theta}}|_{F=\hat{F}_{ML}(\hat{\boldsymbol{\theta}}_{ML})}+\frac{\partial T}{\partial\hat{F}_{ML}(\boldsymbol{\theta})}\frac{\partial\hat{F}_{ML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\}\sqrt{N}(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta})\\ &+\left[T\{\boldsymbol{\theta},\hat{F}_{ML}(\boldsymbol{\theta})\}-T(\boldsymbol{\theta},F)\right]+o_p(1).\end{aligned}$$

We obtain similar expressions for PL and WL as

$$\begin{aligned}&\sqrt{N}\left[T\{\hat{\boldsymbol{\theta}}_{PL},\hat{F}_{PL}(\hat{\boldsymbol{\theta}}_{PL},\hat{\boldsymbol{\pi}})\}-T(\boldsymbol{\theta},F)\right]\\ &=\left\{\frac{\partial T}{\partial\boldsymbol{\theta}}|_{F=\hat{F}_{PL}(\hat{\boldsymbol{\theta}}_{PL},\hat{\boldsymbol{\pi}})}+\frac{\partial T}{\partial\hat{F}_{PL}(\boldsymbol{\theta},\boldsymbol{\pi})}\frac{\partial\hat{F}_{PL}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial\boldsymbol{\theta}}\right\}\sqrt{N}(\hat{\boldsymbol{\theta}}_{PL}-\boldsymbol{\theta})\\ &+\frac{\partial T}{\partial\boldsymbol{\pi}}\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi})+\left[T\{\boldsymbol{\theta},\hat{F}_{PL}(\boldsymbol{\theta},\boldsymbol{\pi})\}-T(\boldsymbol{\theta},F)\right]+o_p(1)\end{aligned}$$

and

$$\begin{aligned}\sqrt{N}\left[T\{\hat{\boldsymbol{\theta}}_{WL},\hat{F}_{WL}(\hat{\boldsymbol{\pi}})\}-T(\boldsymbol{\theta},F)\right] &= \frac{\partial T}{\partial\boldsymbol{\theta}}\sqrt{N}(\hat{\boldsymbol{\theta}}_{WL}-\boldsymbol{\theta})+\frac{\partial T}{\partial\boldsymbol{\pi}}\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi})\\ &+\left[T\{\boldsymbol{\theta},\hat{F}_{WL}(\boldsymbol{\pi})\}-T(\boldsymbol{\theta},F)\right]+o_p(1).\end{aligned}$$

All three methods share the components of $\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}$ and $T(\hat{F})-T(F)$, while PL and WL have an additional term of $\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}$. The contribution to the influence function of $\hat{T}$ by each component is derived as below.

Asymptotic Properties of $\hat{\pi}$

We know that

$$\hat{\pi}_{ys} = \frac{n_{ys}}{N_{ys}} = \frac{\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} I\{Y_k = y, S_k = s\} \times R_{ysk}}{\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} I\{Y_k = y, S_k = s\}},$$

where $I\{Y_k = y, S_k = s\}$ is the indicator of whether the k^{th} subject with $Y = y$ is in the stratum $S = s$. Then we derive the influence function $H_{\pi_{ys}k}$ by expanding $\hat{\pi}_{ys} - \pi_{ys}$ as

$$\hat{\pi}_{ys} - \pi_{ys} = \frac{1}{N}\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} H_{\pi_{ys}k},$$

$$\text{where } H_{\pi_{ys}k} = \left\{\frac{1}{N}\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} I\{Y_k = y, S_k = s\}\right\}^{-1} \times I\{Y_k = y, S_k = s\}(R_{ysk} - \pi_{ys}).$$

Asymptotic Properties of $\hat{\theta}$

Maximum Likelihood Method

We refer readers to Scott and Wild (1997) for the detailed derivation of the profile likelihood function. Given that $\hat{\eta}_{\mathbf{x}\mathbf{z}s}$, μ_{ys}, and γ_{ys}, as respectively defined in the main paper, are all functions of $\boldsymbol{\theta}$, we are able to derive the profile likelihood score function by taking the first derivative of the profile log-likelihood function as

$$S^p(\boldsymbol{\theta}) = \frac{\partial l^p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} S^p_{ysk}(\boldsymbol{\theta}),$$

where

$$S^p_{ysk}(\boldsymbol{\theta}) = (1 - R_{ysk})\frac{\int_{\mathbf{x}}\int_{\mathbf{z}} \frac{\partial}{\partial \boldsymbol{\theta}}\{p_y(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})\eta_{\mathbf{x}\mathbf{z}s}\}d\mathbf{x}d\mathbf{z}}{\int_{\mathbf{x}}\int_{\mathbf{z}} p_y(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})\eta_{\mathbf{x}\mathbf{z}s}d\mathbf{x}d\mathbf{z}} + R_{ysk}\left\{\frac{\frac{\partial p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{p_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \boldsymbol{\theta})} + \frac{\frac{\partial \eta_{\mathbf{x}_{ysk}\mathbf{z}_{ysk}s}}{\partial \boldsymbol{\theta}}}{\eta_{\mathbf{x}_{ysk}\mathbf{z}_{ysk}s}}\right\},$$

and

$$\frac{\partial \eta_{\mathbf{xz}s}}{\partial \boldsymbol{\theta}} = \frac{n_{+\mathbf{xz}s}}{N_{+s}} \frac{-1}{\left\{\sum_y \mu_{ys} p_y(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\right\}^2} \sum_y \left\{ p_y(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) \frac{\partial \mu_{ys}}{\partial \boldsymbol{\theta}} + \mu_{ys} \frac{\partial p_y(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\},$$

$$\frac{\partial \mu_{ys}}{\partial \boldsymbol{\theta}} = \frac{\partial \mu_{ys}}{\partial \gamma_{ys}} \frac{\partial \gamma_{ys}}{\partial \boldsymbol{\theta}},$$

$$\frac{\partial \gamma_{1s}}{\partial \boldsymbol{\theta}} = -\frac{B_s}{1 - a_{+s} W_s}, \text{with } B_s = \sum_{(\mathbf{x},\mathbf{z}) \subset s} \frac{\partial p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, a_{ys} = \frac{1}{n_{ys} - \gamma_{ys}} - \frac{1}{N_{ys} - \gamma_{ys}},$$

$$W_s = \sum_{(\mathbf{x},\mathbf{z}) \subset s} \left[\text{diag}\{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\} - p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})^T \right],$$

$$\frac{\partial p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (1,\mathbf{x},\mathbf{z}) p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) p_0(\mathbf{x},\mathbf{z};\boldsymbol{\theta}).$$

Then, the influence function can be derived as

$$\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta} = \frac{1}{N} \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} H^{ML}_{\boldsymbol{\theta} ysk} + o_p(1),$$

$$\text{where } H^{ML}_{\boldsymbol{\theta} ysk} = \left\{ -\frac{1}{N} \frac{\partial S^p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} S^p_{ysk}(\boldsymbol{\theta}).$$

Pseudo-Likelihood Method

The PL estimator $\hat{\boldsymbol{\theta}}_{PL}$ is obtained by fitting the standard logistic regression model to Phase II subjects with an offset term $\log(\hat{\pi}_{1s}/\hat{\pi}_{0s})$. Then, we can rewrite the pseudo-model as

$$p^\star_y(\mathbf{x},\mathbf{z};\boldsymbol{\theta},\hat{\boldsymbol{\pi}}_s) = \frac{\exp\left[y\{log(\hat{\pi}_{1s}) - \log(\hat{\pi}_{0s}) + \alpha + \boldsymbol{\beta}^T\mathbf{x} + \boldsymbol{\gamma}^T\mathbf{z}\} \right]}{1 + \exp\{\log(\hat{\pi}_{1s}) - \log(\hat{\pi}_{0s}) + \alpha + \boldsymbol{\beta}^T\mathbf{x} + \boldsymbol{\gamma}^T\mathbf{z}\}},$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}), \hat{\boldsymbol{\pi}}_s = (\hat{\pi}_{0s}, \hat{\pi}_{1s})$. Then, $\hat{\boldsymbol{\theta}}_{PL}$ solves the following pseudo-score function

$$S^\star(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}}) = \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} S^\star_{ysk}(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}}_s) = \sum_{y=0}^{1} \sum_{s=1}^{S} \sum_{k=1}^{N_{ys}} R_{ysk} \frac{\partial \log p^\star_y(\mathbf{x}_{ysk}, \mathbf{z}_{ysk}; \hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}}_s)}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

After the first-order Taylor series expansion

$$\mathbf{0} = S^\star(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}}) = S^\star(\boldsymbol{\theta}, \boldsymbol{\pi}) + \frac{\partial S^\star(\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial \boldsymbol{\theta}} (\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}) + \frac{\partial S^\star(\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial \boldsymbol{\pi}} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + o_p(1),$$

we obtain

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta} &= \left\{-\frac{1}{N}\frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \boldsymbol{\theta}}\right\}^{-1} \frac{1}{N}\left\{S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi}) + \frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})\right\} + o_p(1) \\ &= \left\{-\frac{1}{N}\frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \boldsymbol{\theta}}\right\}^{-1} \frac{1}{N}\left\{\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} S^{\star}_{ysk}(\boldsymbol{\theta},\boldsymbol{\pi}_s) - \left\{-\frac{1}{N}\frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \boldsymbol{\pi}_s}\right\} H_{\pi_{ys}k}\right\} + o_p(1),\end{aligned}$$

where

$$\begin{aligned}\frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \boldsymbol{\theta}} &= -\sum_{y=1}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} R_{ysk}(1,\mathbf{x}_{ysk},\mathbf{z}_{ysk})(1,\mathbf{x}_{ysk},\mathbf{z}_{ysk})^T p_1^{\star}(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta},\boldsymbol{\pi}_s) p_0^{\star}(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta},\boldsymbol{\pi}_s) \\ \frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \pi_{0s}} &= \frac{\sum_{y=0}^{1}\sum_{k=1}^{N_{ys}} R_{ysk}(1,\mathbf{x}_{ysk},\mathbf{z}_{ysk}) p_1^{\star}(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta},\boldsymbol{\pi}_s) p_0^{\star}(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta},\boldsymbol{\pi}_s)}{\pi_{0s}}, \\ \frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \pi_{1s}} &= -\frac{\sum_{y=0}^{1}\sum_{k=1}^{N_{ys}} R_{ysk}(1,\mathbf{x}_{ysk},\mathbf{z}_{ysk}) p_1^{\star}(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta},\boldsymbol{\pi}_s) p_0^{\star}(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta},\boldsymbol{\pi}_s)}{\pi_{1s}}.\end{aligned}$$

The matrices $-N^{-1}\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})/\partial\boldsymbol{\theta}$ and $-N^{-1}\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})/\partial\boldsymbol{\pi}_s$ can be shown to converge to a constant matrix by the law of large numbers. Therefore, we are able to write

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta} &= \frac{1}{N}\sum_{y=0}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N} H^{PL}_{\boldsymbol{\theta} ysk} + o_p(1), \\ \text{where } H^{PL}_{\boldsymbol{\theta} ysk} &= \left\{-\frac{1}{N}\frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \boldsymbol{\theta}}\right\}^{-1}\left[S^{\star}_{ysk}(\boldsymbol{\theta},\boldsymbol{\pi}_s) - \left\{-\frac{1}{N}\frac{\partial S^{\star}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \boldsymbol{\pi}_s}\right\} H_{\pi_{ys}k}\right].\end{aligned}$$

Weighted Likelihood Method

The inference procedures are identical to those for PL. Using the weighted score function instead, the following quantities will be updated accordingly:

$$\begin{aligned}\frac{\partial S_{WL}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \boldsymbol{\theta}} &= -\sum_{y=1}^{1}\sum_{s=1}^{S}\sum_{k=1}^{N_{ys}} \frac{R_{ysk}}{\pi_{ys}}(1,\mathbf{x}_{ysk},\mathbf{z}_{ysk})(1,\mathbf{x}_{ysk},\mathbf{z}_{ysk})^T p_1(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta}) p_0(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta}), \\ \frac{\partial S_{WL}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \pi_{0s}} &= \frac{\sum_{y=0}^{1}\sum_{k=1}^{N_{ys}} R_{ysk}(1,\mathbf{x}_{ysk},\mathbf{z}_{ysk})\{0 - p_1(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta})\}}{-\pi_{0s}^2}, \\ \frac{\partial S_{WL}(\boldsymbol{\theta},\boldsymbol{\pi})}{\partial \pi_{1s}} &= -\frac{\sum_{y=0}^{1}\sum_{k=1}^{N_{ys}} R_{ysk}(1,\mathbf{x}_{ysk},\mathbf{z}_{ysk})\{1 - p_1(\mathbf{x}_{ysk},\mathbf{z}_{ysk};\boldsymbol{\theta})\}}{-\pi_{1s}^2}.\end{aligned}$$

Asymptotic Properties of $T(\hat{F})$

Let $\delta_{\mathbf{x},\mathbf{z}}$ be the point mass at $(\mathbf{x},\mathbf{z})$. Applying theory for deriving the influence function for the statistical functionals, we obtain

$$\begin{aligned}\sqrt{N}\{T(\hat{F})-T(F)\} &= \dot{T}\{F;\sqrt{N}(\hat{F}-F)\}+o_p(1)\\ &= \sqrt{N}\left\{\frac{1}{N}\sum_{k=1}^{N}\varphi_F(\mathbf{X}_k,\mathbf{Z}_k)\right\}+o_p(1),\end{aligned}$$

where $\varphi_F(\mathbf{X}_k,\mathbf{Z}_k)$ is the influence curve of T at F and is calculated as

$$\lim_{\epsilon\to 0}\frac{T\{(1-\epsilon)F+\epsilon\delta_{\mathbf{x},\mathbf{z}}\}-T(F)}{\epsilon}.$$

Let $A(\mathbf{x},\mathbf{z})$ and $B(\mathbf{x},\mathbf{z})$ denote $I\{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})>\nu\}$ and $p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$, respectively. The detailed derivations of $\varphi_F(\mathbf{X},\mathbf{Z})$ for TPR_ν and FPR_ν are given below as

$$\begin{aligned}\varphi_F^{\mathsf{TPR}}(\mathbf{x},\mathbf{z}) &= \lim_{\epsilon\to 0}\frac{\frac{\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})d\{(1-\epsilon)F(\mathbf{x},\mathbf{z})+\epsilon\delta_{\mathbf{x},\mathbf{z}}\}}{\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},z)d\{(1-\epsilon)F(\mathbf{x},\mathbf{z})+\epsilon\delta_{\mathbf{x},\mathbf{z}}\}}-\frac{\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}{\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}}{\epsilon}\\ &= \lim_{\epsilon\to 0}\frac{\frac{(1-\epsilon)\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})+\epsilon\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})d\delta_{\mathbf{x},z}}{(1-\epsilon)\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})+\epsilon\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})d\delta_{\mathbf{x},\mathbf{z}}}-\frac{\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}{\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}}{\epsilon}\\ &= \lim_{\epsilon\to 0}\frac{\frac{(1-\epsilon)\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})+\epsilon A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})}{(1-\epsilon)\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})+\epsilon B(\mathbf{x},\mathbf{z})}-\frac{\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}{\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}}{\epsilon}\\ &= \lim_{\epsilon\to 0}\frac{A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})-B(\mathbf{x},\mathbf{z})\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}{\left\{(1-\epsilon)\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})+\epsilon B(\mathbf{x},\mathbf{z})\right\}\times\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}\\ &= \frac{A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})-B(\mathbf{x},\mathbf{z})\int_{\mathbf{x},\mathbf{z}}A(\mathbf{x},\mathbf{z})B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}{\left\{\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})\right\}^2}\\ &= \frac{B(\mathbf{x},\mathbf{z})\{A(\mathbf{x},\mathbf{z})-\mathsf{TPR}\}}{\int_{\mathbf{x},\mathbf{z}}B(\mathbf{x},\mathbf{z})dF(\mathbf{x},\mathbf{z})}\\ &= \frac{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\left[I\{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})>\nu\}-\mathsf{TPR}\right]}{E_{\mathbf{x},\mathbf{z}}\{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\}}.\end{aligned}$$

Similarly, we can derive the influence function for FPR_ν by simply replacing $p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$ with $1-p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$:

$$\varphi_F^{\mathsf{FPR}}(\mathbf{x},\mathbf{z})=\frac{\{1-p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\}\left[I\{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})>\nu\}-\mathsf{FPR}\right]}{E_{\mathbf{x},\mathbf{z}}[\{1-p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\}]}.$$

The influence function for AUC is derived accordingly as

$$\varphi_F^{\mathsf{AUC}}(\mathbf{x},\mathbf{z}) = \int \mathsf{TPR}_\nu d\varphi_F^{\mathsf{FPR}}(\mathbf{x},\mathbf{z}) + \int \varphi_F^{\mathsf{TPR}}(\mathbf{x},\mathbf{z}) d\mathsf{FPR}_\nu.$$

The influence functions for PCF_q and PNF_p can be derived following the same line as Pfeiffer and Gail, 2011 (equation (10) and (12)).

Asymptotic Properties of $T\left\{\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}})\right\}$

Under the PL method, we write $\hat{T}_{PL} \equiv T\left\{\hat{\boldsymbol{\theta}}_{PL}, \hat{F}_{PL}(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}})\right\}$ as the generic notation for all three measures. By applying the standard Taylor series expansion and Delta method for statistical functionals, we obtain the following asymptotic linear approximations with details provided above:

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}) &= \sqrt{N}\left\{\frac{1}{N}\sum_{k=1}^{N} H_{\boldsymbol{\theta}k}^{PL}\right\} + o_p(1), \\
\sqrt{N}\{T(\boldsymbol{\theta}, \hat{F}_{PL}) - T(\boldsymbol{\theta}, F)\} &= \sqrt{N}\left\{\frac{1}{N}\sum_{k=1}^{N} \varphi_F^{PL}(\mathbf{X}_k, \mathbf{Z}_k)\right\} + o_p(1), \\
\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) &= \sqrt{N}\left\{\frac{1}{N}\sum_{k=1}^{N} H_{\boldsymbol{\pi}k}\right\} + o_p(1).
\end{aligned}$$

Then the asymptotic linear approximation of $\hat{T}_{PL}$ can be obtained as

$$\sqrt{N}(\hat{T}_{PL} - T) = \sqrt{N}\left\{\frac{1}{N}\sum_{k=1}^{N} H_{Tk}^{PL}\right\} + o_p(1),$$

where H_{Tk}^{PL} is calculated as

$$H_{Tk}^{PL} = \left\{\frac{\partial T}{\partial \boldsymbol{\theta}}|_{F=\hat{F}_{PL}(\hat{\boldsymbol{\theta}}_{PL}, \hat{\boldsymbol{\pi}})} + \frac{\partial T}{\partial \hat{F}_{PL}(\boldsymbol{\theta}, \boldsymbol{\pi})}\frac{\partial \hat{F}_{PL}(\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial \boldsymbol{\theta}}\right\} H_{\boldsymbol{\theta}k}^{PL} + \frac{\partial T}{\partial \boldsymbol{\pi}} H_{\boldsymbol{\pi}k} + \varphi_F^{PL}(\mathbf{X}_k, \mathbf{Z}_k).$$

Asymptotic Properties of $T\{\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}(\hat{\boldsymbol{\pi}})\}$

Now write $\hat{T}_{WL} \equiv T\{\hat{\boldsymbol{\theta}}_{WL}, \hat{F}_{WL}(\hat{\boldsymbol{\pi}})\}$ as the generic notation for the three predictive accuracy measures. We obtain its asymptotic linear approximation as $\sqrt{N}(\hat{T}_{WL} - T) = \sqrt{N}\left\{\frac{1}{N}\sum_{k=1}^{N} H_{Tk}^{WL}\right\} +$

$o_p(1)$, where the influence function H_{Tk}^{WL} takes the form

$$H_{Tk}^{WL} = \frac{\partial T}{\partial \boldsymbol{\theta}} H_{\boldsymbol{\theta}k}^{WL} + \frac{\partial T}{\partial \boldsymbol{\pi}} H_{\boldsymbol{\pi}k} + \varphi_F^{WL}(\mathbf{X}_k, \mathbf{Z}_k).$$

The terms $H_{\boldsymbol{\theta}k}^{WL}$, $H_{\boldsymbol{\pi}k}$, and $\varphi_F^{WL}(\mathbf{X}_k, \mathbf{Z}_k)$ are defined similar to those of the PL approach above (Supplementary Material). Then $\hat{T}_{WL}$ is asymptotically normally distributed, and the asymptotic variance-covariance matrix can be estimated as $N^{-2}\sum_{k=1}^{N} H_{Tk}^{WL}(H_{Tk}^{WL})^T$.

A.2. Chapter 3

A.2.1. Derivation of the profile likelihood

Given

$$g^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^*) = l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}) + \sum_{s=1}^{M} \lambda_s^* \left\{ \sum_{k=1}^{K} \pi_{sk} - 1 \right\}$$
$$+ \sum_{s=1}^{M} \sum_{r=1}^{I_s} \lambda_{sr} \left\{ \frac{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \sum_{l=1}^{L} P_{1skl} \delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \delta_{sk}} - P^e(Y = 1 | a_{sr} \le \varphi(s, \mathbf{x}) \le b_{sr}) \right\},$$

we notice that the second constraint doesn't involve $\boldsymbol{\pi}$, therefore we focus on the first two terms of $g^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$. By differentiating with respect to $\boldsymbol{\pi}$, we obtain the following

$$\frac{\partial g^*}{\partial \pi_{sk}} = \frac{n_{+sk+}}{\pi_{sk}} - \sum_{i=0}^{1} n_{is++} \frac{\sum_{l=1}^{L} P_{iskl} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{\sum_{k=1}^{K} \sum_{l=1}^{L} P_{iskl} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)} + \lambda_s^* = 0.$$

Multiplying π_{sk} by both sides of the equation and summing over k, we obtain

$$\begin{aligned}
\sum_{k=1}^{K} n_{+sk+} - \sum_{k=1}^{K} \pi_{sk} \sum_{i=0}^{1} n_{is++} \frac{\sum_{l=1}^{L} P_{iskl} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{\sum_{k=1}^{K} \sum_{l=1}^{L} P_{iskl} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)} + \lambda_s^* \sum_{k=1}^{K} \pi_{sk} &= 0, \\
n_{+s++} - \sum_{i=0}^{1} n_{is++} \sum_{k=1}^{K} \pi_{sk} \frac{\sum_{l=1}^{L} P_{iskl} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{\sum_{k=1}^{K} \sum_{l=1}^{L} P_{iskl} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)} + \lambda_s^* &= 0, \\
n_{+s++} - \sum_{i=0}^{1} n_{is++} \sum_{k=1}^{K} \frac{\sum_{l=1}^{L} P_{iskl} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)}{\sum_{k=1}^{K} \sum_{l=1}^{L} P_{iskl} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l | s, \mathbf{x}_k)} + \lambda_s^* &= 0, \\
n_{+s++} - \sum_{i=0}^{1} n_{is++} + \lambda_s^* &= 0, \\
\lambda_s^* &= 0.
\end{aligned}$$

Therefore, we are able to express π_{sk} as a function of $(\alpha_s, \boldsymbol{\beta}, \boldsymbol{\tau}, \mu_{is})$:

$$\hat{\pi}_{sk}(\alpha_s, \boldsymbol{\beta}, \boldsymbol{\tau}, \mu_{is}) = \frac{n_{+sk+}}{n_{+s++}\sum_{i=0}^{1}\mu_{is}\sum_{l=1}^{L}P_{iskl}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)},$$
$$\text{where } \mu_{is} = \frac{n_{is++}}{n_{+s++}\sum_{k=1}^{K}\sum_{l=1}^{L}P_{iskl}\pi_{sk}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}.$$

Plugging $\boldsymbol{\lambda}^* = \mathbf{0}$ and $\hat{\pi}_{sk}$ back into $g^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$ then gives the profile likelihood equivalent to (3.4).

A.2.2. The Score Function and Negative Information Matrix for Constrained Maximum Likelihood Method

Given the constrained profile likelihood function $g(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ (equation (4) in the main text), we can calculate the corresponding score functions for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\lambda})$, respectively. After we re-parametrize $\mu_s = \log\frac{\mu_{1s}}{\mu_{0s}}, s = 1, ..., M$, we obtain the score functions as

$$S_{\alpha_s} = \sum_{i=0}^{1}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{iskl}(i - P_{1skl}) - (e^{\mu_s}-1)\sum_{k=1}^{K} n_{+sk+}\frac{\sum_{l=1}^{L}P_{1skl}P_{0skl}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl})f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}$$
$$+\sum_{r=1}^{I_s}\lambda_{sr}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L}P_{1skl}P_{0skl}\delta_{sk}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}}\right\},$$
$$S_{\boldsymbol{\beta}} = \sum_{i=0}^{1}\sum_{s=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{iskl}(\mathbf{x}_k,\mathbf{z}_l)(i - P_{1skl})$$
$$-\sum_{s=1}^{M}\sum_{k=1}^{K}(e^{\mu_s}-1)n_{+sk+}\frac{\sum_{l=1}^{L}(\mathbf{x}_k,\mathbf{z}_l)P_{1skl}P_{0skl}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl})f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}$$
$$+\sum_{s=1}^{M}\sum_{r=1}^{I_s}\lambda_{sr}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L}(\mathbf{x}_k,\mathbf{z}_l)P_{1skl}P_{0skl}\delta_{sk}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}}\right\},$$
$$S_{\boldsymbol{\tau}} = \sum_{s=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{+skl}\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)/\partial\boldsymbol{\tau}}{f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)} - \sum_{s=1}^{M}\sum_{k=1}^{K} n_{+sk+}\frac{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl})\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)/\partial\boldsymbol{\tau}}{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl})f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}$$
$$+\sum_{s=1}^{M}\sum_{r=1}^{I_s}\lambda_{sr}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L}P_{1skl}\delta_{sk}\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)/\partial\boldsymbol{\tau}}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}}\right\},$$
$$S_{\mu_s} = \sum_{i=0}^{1}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{iskl}\left\{i - \frac{e^{\mu_s}\sum_{l=1}^{L}P_{1skl}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl})f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}\right\},$$
$$S_{\lambda_{sr}} = \frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L}P_{1skl}\delta_{sk}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}},$$

where $s = 1, ..., M, r = 1, ..., I_s$. The component matrices for

$$I = -\frac{\partial^2 g(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\lambda})\partial(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\lambda})^T}$$

are calculated as

$$\begin{aligned}
&I_{\alpha_s\alpha_s} = -\frac{\partial S_{\alpha_s}}{\partial \alpha_s} \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L} n_{+skl} P_{1skl} P_{0skl} - (e^{\mu_s} - 1)\sum_{k=1}^{K} n_{+sk+} \frac{\sum_{l=1}^{L} P_{1skl} P_{0skl}(P_{1skl} - P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)}{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)} \\
&- (e^{\mu_s} - 1)\sum_{k=1}^{K} n_{+sk+} \frac{\left\{\sum_{l=1}^{L} P_{1skl} P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)\right\}^2}{\left\{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)\right\}^2} \\
&+ \sum_{r=1}^{I_s} \lambda_{sr} \left\{ \frac{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \sum_{l=1}^{L} P_{1skl} P_{0skl}(P_{1skl} - P_{0skl}) \delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)}{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \delta_{sk}} \right\}, \\
&I_{\boldsymbol{\beta}\boldsymbol{\beta}} = -\frac{\partial S_{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} \\
&= \sum_{s=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{+skl} P_{1skl} P_{0skl} (\mathbf{x}_k, \mathbf{z}_l)^T (\mathbf{x}_k, \mathbf{z}_l) \\
&- \sum_{s=1}^{M} (e^{\mu_s} - 1)\sum_{k=1}^{K} n_{+sk+} \frac{\sum_{l=1}^{L} P_{1skl} P_{0skl}(P_{1skl} - P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)(\mathbf{x}_k, \mathbf{z}_l)^T (\mathbf{x}_k, \mathbf{z}_l)}{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)} - \sum_{s=1}^{M}\sum_{k=1}^{K} \\
&(e^{\mu_s} - 1) n_{+sk+} \frac{\left\{\sum_{l=1}^{L} P_{1skl} P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)(\mathbf{x}_k, \mathbf{z}_l)\right\}^T \left\{\sum_{l=1}^{L} P_{1skl} P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)(\mathbf{x}_k, \mathbf{z}_l)\right\}}{\left\{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)\right\}^2} \\
&+ \sum_{r=1}^{I_s} \lambda_{sr} \left\{ \frac{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \sum_{l=1}^{L} P_{1skl} P_{0skl}(P_{1skl} - P_{0skl}) \delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)(\mathbf{x}_k, \mathbf{z}_l)^T (\mathbf{x}_k, \mathbf{z}_l)}{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \delta_{sk}} \right\}, \\
&I_{\boldsymbol{\tau}\boldsymbol{\tau}} = -\frac{\partial S_{\boldsymbol{\tau}}}{\partial \boldsymbol{\tau}} \\
&= -\sum_{s=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{+skl} \frac{\partial^2 f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)/\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T}{f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)} + \sum_{s=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{+skl} \frac{\left\{\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)}{\partial \boldsymbol{\tau}}\right\}^T \left\{\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)}{\partial \boldsymbol{\tau}}\right\}}{f_{\boldsymbol{\tau}}^2(\mathbf{z}_l|s, \mathbf{x}_k)} \\
&+ \sum_{s=1}^{M}\sum_{k=1}^{K} n_{+sk+} \frac{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) \partial^2 f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)/\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T}{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)} \\
&- \sum_{s=1}^{M}\sum_{k=1}^{K} n_{+sk+} \frac{\left\{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) \frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)}{\partial \boldsymbol{\tau}}\right\}^T \left\{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) \frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)}{\partial \boldsymbol{\tau}}\right\}}{\left\{\sum_{l=1}^{L}(e^{\mu_s} P_{1skl} + P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)\right\}^2} \\
&- \sum_{s=1}^{M}\sum_{r=1}^{I_s} \lambda_{sr} \left\{ \frac{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \sum_{l=1}^{L} P_{1skl} \delta_{sk} \partial^2 f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)/\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T}{\sum_{k:a_{sr} \le \varphi(s,\mathbf{x}_k) \le b_{sr}} \delta_{sk}} \right\}
\end{aligned}$$

and

$$
\begin{aligned}
&I_{\mu_s\mu_s} = -\frac{\partial S_{\mu_s}}{\partial \mu_s} \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L} n_{+skl} \frac{\left\{e^{\mu_s}\sum_{l=1}^{L} P_{1skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}\left\{\sum_{l=1}^{L} P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}}{\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}^2} \\
&I_{\alpha_s\boldsymbol{\tau}} = I_{\boldsymbol{\tau}\alpha_s} = -\frac{S_{\alpha_s}}{\partial \boldsymbol{\tau}} \\
&= (e^{\mu_s}-1)\sum_{k=1}^{K} n_{+sk+} \frac{\sum_{l=1}^{L} P_{1skl}P_{0skl}\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\partial \boldsymbol{\tau}}}{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)} \\
&- (e^{\mu_s}-1)\sum_{k=1}^{K} n_{+sk+} \frac{\left\{\sum_{l=1}^{L} P_{1skl}P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl})\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\partial \boldsymbol{\tau}}\right\}}{\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}^2} \\
&- \sum_{r=1}^{I_s}\lambda_{sr}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L} P_{1skl}P_{0skl}\delta_{sk}\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\partial \boldsymbol{\tau}}}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}}\right\}, \\
&I_{\boldsymbol{\beta}\boldsymbol{\tau}} = I_{\boldsymbol{\tau}\boldsymbol{\beta}} = -\frac{S_{\boldsymbol{\beta}}}{\partial \boldsymbol{\tau}} \\
&= \sum_{s=1}^{M}(e^{\mu_s}-1)\sum_{k=1}^{K} n_{+sk+} \frac{\sum_{l=1}^{L} P_{1skl}P_{0skl}(\mathbf{x}_k,\mathbf{z}_l)\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\partial \boldsymbol{\tau}}}{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)} - \sum_{s=1}^{M}\sum_{k=1}^{K} \\
&(e^{\mu_s}-1) n_{+sk+} \frac{\left\{\sum_{l=1}^{L} P_{1skl}P_{0skl}(\mathbf{x}_k,\mathbf{z}_l) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl})\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\partial \boldsymbol{\tau}}\right\}}{\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}^2} \\
&- \sum_{r=1}^{I_s}\lambda_{sr}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L} P_{1skl}P_{0skl}\delta_{sk}(\mathbf{x}_k,\mathbf{z}_l)\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\partial \boldsymbol{\tau}}}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}}\right\}, \\
&I_{\alpha_s\mu_s} = I_{\mu_s\alpha_s} = -\frac{\partial S_{\alpha_s}}{\partial \mu_s} \\
&= \sum_{k=1}^{K} n_{+sk+} \frac{e^{\mu_s}\sum_{l=1}^{L} P_{1skl}P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}^2}, \\
&I_{\boldsymbol{\beta}\mu_s} = I_{\mu_s\boldsymbol{\beta}} = -\frac{\partial S_{\boldsymbol{\beta}}}{\partial \mu_s} \\
&= \sum_{s=1}^{M}\sum_{k=1}^{K} n_{+sk+} \frac{e^{\mu_s}\sum_{l=1}^{L} P_{1skl}P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)(\mathbf{x}_k,\mathbf{z}_l)}{\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}^2}, \\
&I_{(\alpha_s,\boldsymbol{\beta})\lambda_{sr}} = I_{\lambda_{sr}(\alpha_s,\boldsymbol{\beta})} = -\frac{\partial S_{\lambda_{sr}}}{\partial(\alpha_s,\boldsymbol{\beta})} \\
&= \frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L} P_{1skl}P_{0skl}\delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)(1,\mathbf{x}_k,\mathbf{z}_l)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}}
\end{aligned}
$$

and

$$
\begin{aligned}
I_{\boldsymbol{\tau}\mu_s} = I_{\mu_s\boldsymbol{\tau}} &= -\frac{\partial S_{\boldsymbol{\tau}}}{\partial \mu_s} \\
&= \sum_{k=1}^{K} n_{+sk+} \frac{e^{\mu_s}\left\{\sum_{l=1}^{L} P_{1skl}\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\partial \boldsymbol{\tau}}\right\}\left\{\sum_{l=1}^{L} P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}}{\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}^2} \\
&\quad - \sum_{k=1}^{K} n_{+sk+} \frac{e^{\mu_s}\left\{\sum_{l=1}^{L} P_{0skl}\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\partial \boldsymbol{\tau}}\right\}\left\{\sum_{l=1}^{L} P_{1skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}}{\left\{\sum_{l=1}^{L}(e^{\mu_s}P_{1skl}+P_{0skl}) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\}^2}, \\
I_{\boldsymbol{\tau}\lambda_{sr}} = I_{\lambda_{sr}\boldsymbol{\tau}} &= -\frac{\partial S_{\lambda_{sr}}}{\partial \boldsymbol{\tau}} \\
&= -\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L} P_{1skl}\delta_{sk}\partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)/\partial\boldsymbol{\tau}}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}}.
\end{aligned}
$$

The matrices of $I_{\boldsymbol{\lambda}}$ and $I_{\boldsymbol{\lambda\mu}} = I_{\boldsymbol{\mu\lambda}}$ are zero.

A.2.3. Proofs of the Identifiability and Consistency

Correct model

Assume the model in (3.1) is correct. The identifiability of the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}$ has been shown in the literature. We now show that using the constrained ML approach above, the maximizer converges to the true parameter values. Proof:

Consider maximizing

$$
\begin{aligned}
&M_{n,c}(\boldsymbol{\theta}) \equiv n^{-1} l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}) - c^2 \sum_{s=1}^{M}(\sum_{k=1}^{K}\pi_{sk} - 1)^2 \\
&- c^2 \sum_{s=1}^{M}\sum_{r=1}^{I_s}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L} P_{1skl}\delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}} - P^e(Y=1|a_{sr}\le\varphi(s,\mathbf{x})\le b_{sr})\right\}^2
\end{aligned}
$$

with respect to $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \boldsymbol{\pi}^T)^T$. For simplicity, we assume the parameter space $\boldsymbol{\Theta}$ is bounded. Define

$$
\begin{aligned}
&M_c(\boldsymbol{\theta}) = E\{n^{-1} l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi})\} - c^2 \sum_{s=1}^{M}(\sum_{k=1}^{K}\pi_{sk} - 1)^2 \\
&- c^2 \sum_{s=1}^{M}\sum_{r=1}^{I_s}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L} P_{1skl}\delta_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}} - P^e(Y=1|a_{sr}\le\varphi(s,\mathbf{x})\le b_{sr})\right\}^2.
\end{aligned}
$$

It is easy to see that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |M_{n,c}(\boldsymbol{\theta}) - M_c(\boldsymbol{\theta})| \to 0$ in probability. Let $\boldsymbol{\theta}_0$ be the true parameter. Furthermore, based on the maximum likelihood property and the uniqueness of the true parameter values following from the identifiability, it is clear that for any $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon > 0$,

$$
\begin{aligned}
&M_c(\boldsymbol{\theta}) = E\{n^{-1}l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\pi})\} - c^2\sum_{s=1}^{M}(\sum_{k=1}^{K}\pi_{sk} - 1)^2 \\
&- c^2\sum_{s=1}^{M}\sum_{r=1}^{I_s}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^{L}P_{1skl}\delta_{sk}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}} - P^e(Y=1|a_{sr}\le\varphi(s,\mathbf{x})\le b_{sr})\right\}^2 \\
&\le E\{n^{-1}l(\boldsymbol{\theta})\} \\
&= \int \log\{f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta})\}f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta}_0)d\mu(\mathbf{x},\mathbf{z},s,y) \\
&= \int \log\{f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta})/f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta}_0)\}f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta}_0)d\mu(\mathbf{x},\mathbf{z},s,y) \\
&+ \int \log f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta}_0)f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta}_0)d\mu(\mathbf{x},\mathbf{z},s,y) \\
&< \int \{f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta})/f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta}_0) - 1\}f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta}_0,)d\mu(\mathbf{x},\mathbf{z},s,y) \\
&+ E\{\log f(\mathbf{X},\mathbf{Z},S,Y,\boldsymbol{\theta}_0)\} \\
&= E\{n^{-1}l(\boldsymbol{\theta}_0)\} \\
&= E\{n^{-1}l(\boldsymbol{\theta}_0)\} - c^2\sum_{s=1}^{M}(\sum_{k=1}^{K}\pi_{sk0} - 1)^2 \\
&- c^2\sum_{s=1}^{M}\sum_{r=1}^{I_s}\left\{\frac{P^e(Y=1, a_{sr}\le\varphi(s,\mathbf{x})\le b_{sr})}{P^e(a_{sr}\le\varphi(s,\mathbf{x})\le b_{sr})} - P^e(Y=1|a_{sr}\le\varphi(s,\mathbf{x})\le b_{sr})\right\}^2 \\
&= M_c(\boldsymbol{\theta}_0),
\end{aligned}
$$

where we used $\log(t) < t-1$ for $t \neq 1$ and $t > 0$. Thus, following Theorem 5.7 of Van der Vaart (1998), the maximizer of $M_{n,c}(\boldsymbol{\theta})$ converges to the unique true parameter $\boldsymbol{\theta}_0$ in probability. When $c \to \infty$, the maximizer of $M_{n,c}(\boldsymbol{\theta})$ can be made arbitrarily close to the constrained MLE $\widehat{\boldsymbol{\theta}}$. Thus, the constrained MLE $\widehat{\boldsymbol{\theta}}$ converges to the unique true parameter $\boldsymbol{\theta}_0$ in probability.

Arbitrary model

Without assuming the model in (3.1) is correctly specified, the identifiability needs to be established first. Here, by identifiability, we mean that there are no two sets of parameters $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\pi}$ and $\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\tau}},\widetilde{\boldsymbol{\pi}}$ so that they both satisfy the constraints in (3.2) and (3.3), and $P(\mathbf{x},\mathbf{z}|y,s,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\pi}) =$

$P(\mathbf{x}, \mathbf{z}|y, s, \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\tau}}, \widetilde{\boldsymbol{\pi}})$ for all $(\mathbf{x}, \mathbf{z}, y, s)$ combinations.

Proof: We assume that there exist $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi}$ and $\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\tau}}, \widetilde{\boldsymbol{\pi}}$ so that

$$\begin{aligned}
& \frac{\exp\{i(\alpha_s + \boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)\}}{1 + \exp(\alpha_s + \boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k) \\
& \Big/ \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\exp\{i(\alpha_s + \boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)\}}{1 + \exp(\alpha_s + \boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k) \right\} \\
= \quad & \frac{\exp\{i(\widetilde{\alpha}_s + \widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)\}}{1 + \exp(\widetilde{\alpha}_s + \widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)} \widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s, \mathbf{x}_k) \\
& \Big/ \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\exp\{i(\widetilde{\alpha}_s + \widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)\}}{1 + \exp(\widetilde{\alpha}_s + \widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)} \widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s, \mathbf{x}_k) \right\}
\end{aligned} \tag{A.1}$$

for all $i = 0, 1$, $k = 1, \ldots, K$, $l = 1, \ldots, L$ and $s = 1, \ldots, S$. This leads to

$$\begin{aligned}
& \frac{\exp\{i(\boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)\}}{1 + \exp(\alpha_s + \boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k) \\
& \Big/ \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\exp\{i(\boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)\}}{1 + \exp(\alpha_s + \boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k) \right\} \\
= \quad & \frac{\exp\{i(\widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)\}}{1 + \exp(\widetilde{\alpha}_s + \widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)} \widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s, \mathbf{x}_k) \\
& \Big/ \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\exp\{i(\widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)\}}{1 + \exp(\widetilde{\alpha}_s + \widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)} \widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s, \mathbf{x}_k) \right\}
\end{aligned}$$

for all $i = 0, 1$, $k = 1, \ldots, K$, $l = 1, \ldots, L$ and $s = 1, \ldots, S$. Taking ratio of the value at $i = 1$ and $i = 0$ on each side, we get

$$\begin{aligned}
& \exp(\boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l) \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)}{1 + \exp(\alpha_s + \boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)} \right\} \\
& \Big/ \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\exp(\boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)}{1 + \exp(\alpha_s + \boldsymbol{\beta}_x^T \mathbf{x}_k + \boldsymbol{\beta}_z^T \mathbf{z}_l)} \pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k) \right\} \\
= \quad & \exp(\widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l) \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s, \mathbf{x}_k)}{1 + \exp(\widetilde{\alpha}_s + \widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)} \right\} \\
& \Big/ \left\{ \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\exp(\widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)}{1 + \exp(\widetilde{\alpha}_s + \widetilde{\boldsymbol{\beta}}_x^T \mathbf{x}_k + \widetilde{\boldsymbol{\beta}}_z^T \mathbf{z}_l)} \widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s, \mathbf{x}_k) \right\}
\end{aligned}$$

for all $k = 1, \ldots, K$, $l = 1, \ldots, L$ and $s = 1, \ldots, S$. Note that on both sides, other than the first

exponential term, the remaining quantities are not functions of $\mathbf{x}, \mathbf{z}$, hence we actually get $\exp(\boldsymbol{\beta}_x^T\mathbf{x}+\boldsymbol{\beta}_z^T\mathbf{z}) = c\exp(\widetilde{\boldsymbol{\beta}}_x^T\mathbf{x}+\widetilde{\boldsymbol{\beta}}_z^T\mathbf{z})$ for some c that does not depend on $\mathbf{x}, \mathbf{z}$. Thus, $\boldsymbol{\beta}_x = \widetilde{\boldsymbol{\beta}}_x, \boldsymbol{\beta}_z = \widetilde{\boldsymbol{\beta}}_z$ and $c = 1$. Thus, letting $i = 0$ and $i = 1$ respectively, (A.1) further leads to

$$\begin{aligned} & \frac{\pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \Big/ \left\{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{\pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)}\right\} \\ = & \frac{\widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \Big/ \left\{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{\widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)}\right\} \end{aligned} \tag{A.2}$$

and

$$\begin{aligned} & \frac{\pi_{sk}\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \\ & \Big/ \left\{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)}{1+\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)}\pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\} \\ = & \frac{\widetilde{\pi}_{sk}\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l) f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \\ & \Big/ \left\{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)}{1+\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)}\widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s,\mathbf{x}_k)\right\} \end{aligned} \tag{A.3}$$

for all $k = 1, \ldots, K$, $l = 1, \ldots, L$ and $s = 1, \ldots, S$.

Now assume that for any (s, k), we can select a suitable interval $[a_{sr}, b_{sr}]$ so that $\varphi(s, x_k)$ is the only φ value in the interval. I need the above assumption. This essentially says the different combinations of (s, k) are distinguishable in the external study. I guess this assumption is okay? Then (3.3) yields $P^e(Y = 1|s, \mathbf{x}_k) = \sum_{l=1}^{L} P_{1skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)$. This also implies $P^e(Y = 0|s, \mathbf{x}_k) = \sum_{l=1}^{L} P_{0skl} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)$. Plugging this into (A.2) and (A.3), we get

$$\begin{aligned} & \frac{\pi_{sk} f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \Big/ \{\sum_{k=1}^{K}\pi_{sk}P^e(Y=0|s,\mathbf{x}_k)\} \\ & = \frac{\widetilde{\pi}_{sk} f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \Big/ \{\sum_{k=1}^{K}\widetilde{\pi}_{sk}P^e(Y=0|s,\mathbf{x}_k)\} \\ & \text{and} \\ & \frac{\pi_{sk}\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l) f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \Big/ \{\sum_{k=1}^{K}\pi_{sk}P^e(Y=1|s,\mathbf{x}_k)\} \\ & = \frac{\widetilde{\pi}_{sk}\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l) f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \Big/ \{\sum_{k=1}^{K}\widetilde{\pi}_{sk}P^e(Y=1|s,\mathbf{x}_k)\} \end{aligned} \tag{A.4}$$

Summing the above for $l = 1$ to L, we further get

$$\begin{aligned}
\frac{\pi_{sk}P^e(Y=0|s,\mathbf{x}_k)}{\sum_{k=1}^K \pi_{sk}P^e(Y=0|s,\mathbf{x}_k)} &= \frac{\widetilde{\pi}_{sk}P^e(Y=0|s,\mathbf{x}_k)}{\sum_{k=1}^K \widetilde{\pi}_{sk}P^e(Y=0|s,\mathbf{x}_k)} \\
\frac{\pi_{sk}P^e(Y=1|s,\mathbf{x}_k)}{\sum_{k=1}^K \pi_{sk}P^e(Y=1|s,\mathbf{x}_k)} &= \frac{\widetilde{\pi}_{sk}P^e(Y=1|s,\mathbf{x}_k)}{\sum_{k=1}^K \widetilde{\pi}_{sk}P^e(Y=1|s,\mathbf{x}_k)}.
\end{aligned}$$

Adding the above two equalities, we get $\widetilde{\pi}_{sk} = \pi_{sk}c_s$ for some constant c_s at each s. Since $\sum_{k=1}^K \widetilde{\pi}_{sk} = \sum_{k=1}^K \pi_{sk} = 1$, this implies $c_s = 1$ and $\widetilde{\pi}_{sk} = \pi_{sk}$. Using this relation in (A.4) we get

$$\begin{aligned}
\frac{f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} &= \frac{f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} \\
\frac{\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\alpha_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)} &= \frac{\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l|s,\mathbf{x}_k)}{1+\exp(\widetilde{\alpha}_s+\boldsymbol{\beta}_x^T\mathbf{x}_k+\boldsymbol{\beta}_z^T\mathbf{z}_l)},
\end{aligned}$$

which leads to $f_{\boldsymbol{\tau}}(\mathbf{z}_l \mid s,\mathbf{x}_k) = f_{\widetilde{\boldsymbol{\tau}}}(\mathbf{z}_l \mid s,\mathbf{x}_k)$, hence $\boldsymbol{\tau} = \widetilde{\boldsymbol{\tau}}$. This further leads to $\alpha_s = \widetilde{\alpha}_s$. Thus, the problem is indeed identifiable.

Consistency of the constrained MLE.

Proof:

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \boldsymbol{\pi}^T)^T$. We define

$$\begin{aligned}
&M_{n,c}(\boldsymbol{\theta}) \equiv n^{-1}l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\pi}) - c^2\sum_{s=1}^M(\sum_{k=1}^K \pi_{sk}-1)^2 \\
&- c^2\sum_{s=1}^M\sum_{r=1}^{I_s}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^L P_{1skl}\delta_{sk}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}} - P^e(Y=1|a_{sr}\le\varphi(s,\mathbf{x})\le b_{sr})\right\}^2, \\
&M_c(\boldsymbol{\theta}) \equiv E\{n^{-1}l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\pi})\} - c^2\sum_{s=1}^M(\sum_{k=1}^K \pi_{sk}-1)^2 \\
&- c^2\sum_{s=1}^M\sum_{r=1}^{I_s}\left\{\frac{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\sum_{l=1}^L P_{1skl}\delta_{sk}f_{\boldsymbol{\tau}}(\mathbf{z}_l|s,\mathbf{x}_k)}{\sum_{k:a_{sr}\le\varphi(s,\mathbf{x}_k)\le b_{sr}}\delta_{sk}} - P^e(Y=1|a_{sr}\le\varphi(s,\mathbf{x})\le b_{sr})\right\}^2.
\end{aligned}$$

Let a maximizer of $M_{n,c}(\boldsymbol{\theta})$ be $\widehat{\boldsymbol{\theta}}_{n,c}$ and a maximizer of $M_c(\boldsymbol{\theta})$ be $\boldsymbol{\theta}_{0,c}$. We assume $\boldsymbol{\theta}_{0,c}$ is unique for sufficiently large c. Further, we define $\widehat{\boldsymbol{\theta}}_n$ to be a solution of the constrained maximization problem. It is easy to see that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \to 0$ in probability. Further, the definition of $\boldsymbol{\theta}_{0,c}$ yields

that for any $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{0,c}\| > \epsilon > 0$,

$$M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_{0,c}).$$

Thus, following Theorem 5.7 of Van der Vaart (1998), the maximizer of $M_{n,c}(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\theta}}_{n,c}$ converges to the unique maximizer $\boldsymbol{\theta}_{0,c}$ in probability. Now since $\boldsymbol{\Theta}$ is bounded, it is not stringent to assume that $n^{-1}l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi})$ and $E\{n^{-1}l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\pi})\}$ are bounded in $\boldsymbol{\Theta}$. Thus, when $c \to \infty$, we have $\widehat{\boldsymbol{\theta}}_{n,c} \to \widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_{0,c} \to \boldsymbol{\theta}_0$, where $\widehat{\boldsymbol{\theta}}_n$ is the solution of the constrainted maximization problem and $\boldsymbol{\theta}_0$ is the parameter that yields the minimum Kullback-Leibler distance to the true density among all the parameters that satisfy the constraints in both (3.2) and (3.3). These results combined together yield $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ in probability.

A.3. Chapter 4

A.3.1. The Score Function and Negative Information Matrix for Constrained Maximum Likelihood Method

Given the constrained likelihood function $g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ (equation (3) in the main text), we can calculate the corresponding score functions for $(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$, respectively. Let $P_{y\mathbf{xz}} \equiv P(Y = y|\mathbf{x}, \mathbf{z}), y = 0, 1$. The expressions are given as below:

$$S_{(\alpha,\boldsymbol{\beta})} = \sum_{i=1}^{N}(1, \mathbf{x}_i, \mathbf{z}_i)(y_i - P_{1\mathbf{x}_i\mathbf{z}_i}) + \sum_{r=1}^{I}\lambda_r \left\{ \frac{\int_{\mathbf{x}:a_r \le \varphi(\mathbf{x}) \le b_r} \int_{\mathbf{z}} (1, \mathbf{x}, \mathbf{z}) P_{1\mathbf{xz}} P_{0\mathbf{xz}} P^e(\mathbf{x}) f_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}:a_r \le \varphi(\mathbf{x}) \le b_r} P^e(\mathbf{x}) d\mathbf{x}} \right\},$$

$$S_{\boldsymbol{\tau}} = \sum_{i=1}^{N} \frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_i|\mathbf{x}_i)/\partial \boldsymbol{\tau}}{f_{\boldsymbol{\tau}}(\mathbf{z}_i|\mathbf{x}_i)} + \sum_{r=1}^{I}\lambda_r \left\{ \frac{\int_{\mathbf{x}:a_r \le \varphi(\mathbf{x}) \le b_r} \int_{\mathbf{z}} P_{1\mathbf{xz}} P^e(\mathbf{x}) \partial f_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)/\partial \boldsymbol{\tau} d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}:a_r \le \varphi(\mathbf{x}) \le b_r} P^e(\mathbf{x}) d\mathbf{x}} \right\},$$

$$S_{\lambda_r} = \frac{\int_{\mathbf{x}:a_r \le \varphi(\mathbf{x}) \le b_r} \int_{\mathbf{z}} P_{1\mathbf{xz}} P^e(\mathbf{x}) f_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}:a_r \le \varphi(\mathbf{x}) \le b_r} P^e(\mathbf{x}) d\mathbf{x}} - P^e(Y = 1|a_r \le \varphi(\mathbf{x}) \le b_r),\ r = 1, ..., I.$$

The component matrices for $I = -\partial^2 g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})/\partial(\alpha^T, \boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \boldsymbol{\lambda}^T)^T \partial(\alpha^T, \boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \boldsymbol{\lambda}^T)$ are calculated as

$$
\begin{aligned}
I_{(\alpha,\boldsymbol{\beta})(\alpha,\boldsymbol{\beta})} &= -\frac{\partial^2 g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})}{\partial(\alpha, \boldsymbol{\beta})\partial(\alpha, \boldsymbol{\beta})^T} \\
&= \sum_{i=1}^{N} P_{1\mathbf{x}_i\mathbf{z}_i} P_{0\mathbf{x}_i\mathbf{z}_i} (1, \mathbf{x}_i, \mathbf{z}_i)^T (1, \mathbf{x}_i, \mathbf{z}_i) \\
&\quad + \sum_{r=1}^{I} \lambda_r \left\{ \frac{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} \int_{\mathbf{z}} (1, \mathbf{x}, \mathbf{z})^T (1, \mathbf{x}, \mathbf{z}) P_{1\mathbf{xz}} P_{0\mathbf{xz}} (P_{1\mathbf{xz}} - P_{0\mathbf{xz}}) P^e(\mathbf{x}) f_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} P^e(\mathbf{x}) d\mathbf{x}} \right\}, \\
I_{\boldsymbol{\tau}\boldsymbol{\tau}} &= -\frac{\partial^2 g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T} \\
&= \sum_{i=1}^{N} \left\{ \frac{\frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_i|\mathbf{x}_i)}{\partial \boldsymbol{\tau}} \frac{\partial f_{\boldsymbol{\tau}}(\mathbf{z}_i|\mathbf{x}_i)}{\partial \boldsymbol{\tau}^T}}{f^2_{\boldsymbol{\tau}}(\mathbf{z}_i|\mathbf{x}_i)} - \frac{\partial^2 f_{\boldsymbol{\tau}}(\mathbf{z}_i|\mathbf{x}_i)/\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T}{f_{\boldsymbol{\tau}}(\mathbf{z}_i|\mathbf{x}_i)} \right\} \\
&\quad + \sum_{r=1}^{I} \lambda_r \left\{ \frac{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} \int_{\mathbf{z}} P_{1\mathbf{xz}} P^e(\mathbf{x}) \partial f^2_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x})/\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} P^e(\mathbf{x}) d\mathbf{x}} \right\}, \\
I_{\boldsymbol{\lambda}\boldsymbol{\lambda}} &= -\frac{\partial^2 g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^T} \\
&= \mathbf{0}
\end{aligned}
$$

and

$$
\begin{aligned}
I_{(\alpha,\boldsymbol{\beta})\boldsymbol{\tau}} = I^T_{\boldsymbol{\tau}(\alpha,\boldsymbol{\beta})} &= -\frac{\partial^2 g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})}{\partial(\alpha, \boldsymbol{\beta})\partial \boldsymbol{\tau}} \\
&= -\frac{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} \int_{\mathbf{z}} (1, \mathbf{x}, \mathbf{z}) P_{1\mathbf{xz}} P_{0\mathbf{xz}} P^e(\mathbf{x}) \partial f_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x})/\partial \boldsymbol{\tau} d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} P^e(\mathbf{x}) d\mathbf{x}}, \\
I_{(\alpha,\boldsymbol{\beta})\lambda_r} = I^T_{\lambda_r(\alpha,\boldsymbol{\beta})} &= -\frac{\partial^2 g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})}{\partial(\alpha, \boldsymbol{\beta})\partial \lambda_r} \\
&= \frac{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} \int_{\mathbf{z}} (1, \mathbf{x}, \mathbf{z}) P_{1\mathbf{xz}} P_{0\mathbf{xz}} P^e(\mathbf{x}) f_{\boldsymbol{\tau}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} P^e(\mathbf{x}) d\mathbf{x}}, \\
I_{\boldsymbol{\tau}\lambda_r} = I^T_{\lambda_r\boldsymbol{\tau}} &= -\frac{\partial^2 g(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda})}{\partial \boldsymbol{\tau} \partial \lambda_r} \\
&= \frac{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} \int_{\mathbf{z}} P_{1\mathbf{xz}} P^e(\mathbf{x}) \partial f^2_{\boldsymbol{\tau}}(\mathbf{z}_l|s, \mathbf{x}_k)/\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T d\mathbf{z} d\mathbf{x}}{\int_{\mathbf{x}: a_r \le \varphi(\mathbf{x}) \le b_r} P^e(\mathbf{x}) d\mathbf{x}}.
\end{aligned}
$$

APPENDIX B

SOFTWARE

B.1. Chapter 2: Software Package for Implementing the Proposed Methods

We developed an R package "TwoPhaseAccuracy" that implemented the three approaches described in Chapter 2, which is freely available and easy for users to install. In this package, function "evalTwoPhase" is for calculating AUC, PCF, and PNF using all the three approaches, and function "seTwoPhase" is for estimating the corresponding standard errors using bootstrap resampling. The input arguments of the two functions include the estimation method ("MLE", "PL" or "WL"), threshold values for estimating PCF and PNF, and data that contains case-control status, stratum membership, Phase I predictors, Phase II predictors, variable names, and a variable indicating whether a subject was selected into Phase II. Function "seTwoPhase" includes an additional input argument for the number of bootstrap samples. A third function, "summaryTwoPhase", summarizes the results from "evalTwoPhase" and "seTwoPhase" by outputing the estimates of AUC, PCF, and PNF together with their standard error estimates.

B.2. Chapter 3

B.2.1. Additional BCDDP Analysis

Table B.1 presents further analysis of BCDDP. In this analysis, we put different constraints on the prediction model with finer constraints on the tails of the risk distribution while cruder in the middle 50%. In contrast to the quartiles used in the main text, we chose (25%, 75%) percentiles of the BCRAT risk as cutoff points for stratum 1 (age $\leq$ 50) and (15%, 25%, 75%, 85%) percentiles for stratum 2 (age $>$ 50).

B.2.2. $P^e(\mathbf{X}|S)$ Estimated from the National Health Interview Surve (NHIS)

Table B.2 and Table B.3 respectively displays the joint probability distribution of (Ageflb, Agemen, Weight) and (Nbiops, Numrel) given age $\leq$ 50 or age $>$ 50. Then, $P^e(\mathbf{X}|S)$ can be calculated as the product of the two by assuming their independence of each other.

Table B.1: Analysis of the BCDDP data with BCRAT risk cutoffs placed at the (25%, 75%) for stratum 1 and (15%, 25%, 75%, 80%) for stratum 2: estimates of stratum-specific intercept terms and log ORs for the BCRAT predictors, weight, and PD, together with estimates of parameters in the zero-inflated Beta regression model for the distribution of PD. In the parenthesis are the corresponding estimates of asymptotic standard errors. "cMLE" represents estimates from the proposed constrained maximum likelihood method, and "Standard" represents the estimates from the standard method.

		Predictors	cMLE	Standard
Logistic Regression Model for Breast Cancer Risk		Intercept	-6.512 (0.135)	-6.751 (0.182)
		Age$\geq$50	1.198 (0.032)	1.045 (0.035)
		Ageflb	0.513 (0.025)	0.105 (0.049)
		Agemen	0.217 (0.045)	0.213 (0.061)
		Nbiops	0.543 (0.043)	0.174 (0.070)
		Numrel	0.485 (0.019)	0.668 (0.090)
		Weight	-0.199 (0.033)	0.228 (0.044)
		PD^c	0.174 (0.017)	0.177 (0.018)
The PD distribution		Intercept	0.163 (0.073)	0.035 (0.088)
		Age$\geq$50	-0.541 (0.049)	-0.405 (0.057)
	The mean model	Ageflb	0.146 (0.027)	0.139 (0.033)
	γ	Nbiops	0.289 (0.035)	0.248 (0.045)
		Weight	-0.446 (0.025)	-0.421 (0.033)
	The variance model	Intercept	1.386 (0.097)	1.478 (0.101)
	ω	Age$\geq$50	0.206 (0.078)	0.202 (0.079)
		Weight	-0.093 (0.035)	-0.121 (0.045)
	Mixture probability	ρ	0.096 (0.008)	0.105 (0.010)

Table B.2: Joint probability distribution of (Ageflb, Agemen Weight): "Before50" represents estimated probability given age ≤ 50; "After50" represents estimated probability given age > 50

Ageflb	Agemen	Weight	Before50	After50
<20	<12	(100, 125]	0.00789	0.00544
20-24	<12	(100, 125]	0.00887	0.00855
25-29 or never	<12	(100, 125]	0.00871	0.00796
30+	<12	(100, 125]	0.00207	0.00132
<20	12-13	(100, 125]	0.01471	0.01642
20-24	12-13	(100, 125]	0.02629	0.03377
25-29 or never	12-13	(100, 125]	0.05274	0.02836
30+	12-13	(100, 125]	0.02071	0.00663
<20	14+ or never	(100, 125]	0.01003	0.00739
20-24	14+ or never	(100, 125]	0.01574	0.02166
25-29 or never	14+ or never	(100, 125]	0.02821	0.02122
30+	14+ or never	(100, 125]	0.01215	0.00549
<20	<12	(125, 150]	0.00744	0.01290
20-24	<12	(125, 150]	0.01015	0.02310
25-29 or never	<12	(125, 150]	0.02186	0.01849
30+	<12	(125, 150]	0.00828	0.00266
<20	12-13	(125, 150]	0.03201	0.03899
20-24	12-13	(125, 150]	0.05249	0.08446
25-29 or never	12-13	(125, 150]	0.09420	0.05729
30+	12-13	(125, 150]	0.03438	0.01446
<20	14+ or never	(125, 150]	0.01842	0.02178
20-24	14+ or never	(125, 150]	0.02821	0.04803
25-29 or never	14+ or never	(125, 150]	0.03977	0.03581
30+	14+ or never	(125, 150]	0.02349	0.00868
<20	<12	(150, 175]	0.00679	0.01182
20-24	<12	(150, 175]	0.01080	0.01831
25-29 or never	<12	(150, 175]	0.01503	0.01233
30+	<12	(150, 175]	0.00619	0.00257
<20	12-13	(150, 175]	0.02009	0.03423
20-24	12-13	(150, 175]	0.03376	0.05725
25-29 or never	12-13	(150, 175]	0.04709	0.03544
30+	12-13	(150, 175]	0.01888	0.01078
<20	14+ or never	(150, 175]	0.00540	0.01270
20-24	14+ or never	(150, 175]	0.01542	0.03556
25-29 or never	14+ or never	(150, 175]	0.01633	0.02253
30+	14+ or never	(150, 175]	0.00780	0.00432
<20	<12	(175, 200]	0.00896	0.00859
20-24	<12	(175, 200]	0.00925	0.00818
25-29 or never	<12	(175, 200]	0.01131	0.00871
30+	<12	(175, 200]	0.00354	0.00225
<20	12-13	(175, 200]	0.01544	0.01703
20-24	12-13	(175, 200]	0.01997	0.03626
25-29 or never	12-13	(175, 200]	0.02822	0.01976
30+	12-13	(175, 200]	0.01113	0.00726
<20	14+ or never	(175, 200]	0.00500	0.00829
20-24	14+ or never	(175, 200]	0.00606	0.01862
25-29 or never	14+ or never	(175, 200]	0.00619	0.00841
30+	14+ or never	(175, 200]	0.00282	0.00327
<20	<12	>200	0.00829	0.00297
20-24	<12	>200	0.00377	0.00283
25-29 or never	<12	>200	0.00782	0.00532
30+	<12	>200	0.00139	0.00067
<20	12-13	>200	0.00954	0.00997
20-24	12-13	>200	0.01390	0.01246
25-29 or never	12-13	>200	0.02451	0.01349
30+	12-13	>200	0.00421	0.00236
<20	14+ or never	>200	0.00417	0.00476
20-24	14+ or never	>200	0.00765	0.00567
25-29 or never	14+ or never	>200	0.00395	0.00317
30+	14+ or never	>200	0.00048	0.00103

Table B.3: Joint probability distribution of (Nbiops, Numrel): “Before50” represents estimated probability given age ≤ 50; “After50” represents estimated probability given age > 50

Nbiops	Numrel	Before50	After50
0	0	0.81122	0.72053
1	0	0.07079	0.10636
2+	0	0.02231	0.04024
0	1	0.08102	0.09572
1	1	0.00880	0.01532
2+	1	0.00231	0.00779
0	2+	0.00354	0.00976
1	2+	0.00000	0.00207
2+	2+	0.00000	0.00219

BIBLIOGRAPHY

Berkowitz, GS, Lapinski, RH, Wein, R, and Lee, D (1992). Race/ethnicity and other risk factors for gestational diabetes. *American Journal of Epidemiology* 135, 965–973.

Bondy, ML, Lustbader, ED, Halabi, S, Ross, E, and Vogel, VG (1994). Validation of a breast cancer risk assessment model in women with a positive family history. *Journal of the National Cancer Institute* 86, 620–625.

Boyd, NF, Byng, JW, Jong, RA, Fishell, EK, Little, LE, Miller, AB, Lockwood, GA, Tritchler, DL, and Yaffe, MJ (1995). Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. *Journal of the National Cancer Institute* 87, 670–675.

Breslow, NE and Cain, KC (1988). Logistic regression for two-stage case-control data. *Biometrika* 75, 11–20.

Breslow, NE and Chatterjee, N (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics* 48, 457–468.

Breslow, NE and Holubkov, R (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B* 59, 447–461.

Breslow, NE and Zhao, LP (1988). Logistic Regression for Stratified Case-Control Studies. *Biometrics* 44, 891–899.

Byrne, C, Schairer, C, Wolfe, J, Parekh, N, Salane, M, Brinton, LA, Hoover, R, and Haile, R (1995). Mammographic features and breast cancer risk: effects with time, age, and menopause status. *Journal of the National Cancer Institute* 87, 1622–1629.

Cai, T and Zheng, Y (2012). Evaluating prognostic accuracy of biomarkers in nested case-control studies. *Biostatistics* 13, 89–100.

Carroll, RJ and Wand, MP (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B* 53, 573585.

Chatterjee, N, Chen, YH, and Breslow, NE (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association* 98, 158–168.

Chatterjee, N, Chen, YH, Maas, P, and Carroll, RJ (2016). Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-level Information from External Big Data Sources. *Journal of the American Statistical Association* 111, 107–117.

Chen, J, Pee, D, Ayyagari, R, Graubard, B, Schairer, C, Byrne, C, Benichou, J, and Gail, MH (2006). Projecting Absolute Invasive Breast Cancer Risk in White Women With a Model That Includes Mammographic Density. *Journal of the National Cancer Institute* 98, 1215–1226.

Chen, J, Ayyagari, R, Chatterjee, N, Pee, D, Schairer, C, Byrne, C, Benichou, J, and Gail, MH (2008). Breast Cancer Relative Hazard Estimates From CaseControl and Cohort Designs With

Missing Data on Mammographic Density. *Journal of the American Statistical Association* 103, 976–988.

Chen, L and Chen, J (2015). *Cancer Absolute Risk Projection with Incomplete Predictor Variables*. PhD thesis. University of Pennsylvania.

Costantino, JP, Gail, MH, Pee, D, Anderson, S, Redmond, CK, Benichou, J, and Wieand, HS (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute* 91, 1541–1548.

Flanders, WD and Greenland, S (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* 10, 739–747.

Huang, Y (2016). Evaluating and comparing biomarkers with respect to the area under the receiver operating characteristics curve in two-phase case-control studies. *Biostatistics* 17, 499–522.

Huang, Y and Pepe, MS (2010). Assessing risk prediction models in case-control studies using semiparametric and nonparametric methods. *Statistics in Medicine* 29, 1391–1410.

Ibrahim, JG (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* 85, 765–769.

Lawless, JF, Kalbfleisch, JD, and Wild, CJ (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B* 61, 413–438.

Lipsitz, SR, Ibrahim, JG, and Zhao, LP (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* 94, 11471160.

Liu, D, Cai, T, and Zheng, Y (2012). Evaluating the Predictive Value of Biomarkers with Stratified Case-Cohort Design. *Biometrics* 68, 1219–1227.

Neyman, J (1938). Contribution to the theory of sampling from human populations. *Journal of the American Statistical Association* 33, 101–116.

Pepe, MS, Fan, J, and Seymour, CW (2013). Estimating the ROC Curve in Studies that Match Controls to Cases on Covariates. *Academic Radiology* 20, 863–873.

Pepe, MS and Fleming, TR (1991). A non-parametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* 86, 108–113.

Pfeiffer, RM and Gail, MH (2011). Two criteria for evaluating risk prediction models. *Biometrics* 67, 1057–1065.

Qin, J, Zhang, H, Li, P, Albanes, D, and Yu, K (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* 102, 169–180.

Reilly, M and Pepe, MS (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 82, 299–314.

Robins, JM, Rotnitzky, A, and Zhao, LP (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846866.

Rockhill, B, Spiegelman, D, Byrne, C, Hunter, DJ, and Colditz, GA (2001). Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *Journal of the National Cancer Institute* 93, 358–366.

Samuelsen, SO (1997). A psudolikelihood approach to analysis of nested case-control studies. *Biometrika* 84, 379–394.

Scott, AJ and Wild, CJ (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* 84, 57–71.

Scott, AJ and Wild, CJ (2002). On the robustness of weighted methods for fitting models to case-control data. *Journal of the Royal Statistical Society, Series B* 64, 207–219.

Solomon, CG, Willett, WC, Carey, VJ, Rich-Edwards, J, Hunter, DJ, Colditz, GA, Stampfer, MJ, Speizer, FE, Spiegelman, D, and Manson, JE (1997). A prospective study of pregravid determinants of gestational diabetes mellitus. *Journal of the American Medical Association* 278, 1078–1083.

Wang, X and Zhou, H (2010). Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. *Biometrics* 66, 502–511.

White, JE (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* 115, 119–128.

Zhu, Y, Mendola, P, Albert, PS, Bao, W, Hinkle, SN, Tsai, M, and Zhang, C (2016). Longitudinal study of insulin-like growth factor 1 and binding proteins 2 and 3 and subsequent risk of gestational diabetes among women in a multiracial cohort. *Diabetes* 65, 3495–3504.