Fall 8-2020

SEMIPARAMETRIC METHODS TO IMPROVE RISK ASSESSMENT AND DYNAMIC PREDICTION

WEN LI

SEMIPARAMETRIC METHODS TO IMPROVE RISK ASSESSMENT AND DYNAMIC PREDICTION

by

WEN LI, BA

APPROVED:

Ruosha Li Digitally signed by Ruosha Li Date: 2020.07.08 14:38:29 -05'00'

RUOSHA LI, PHD

Jing Ning Digitally signed by Jing Ning Date: 2020.07.08 11:41:58 -05'00'

JING NING, PHD

Han Chen Digitally signed by Han Chen Date: 2020.07.08 15:00:22 -05'00'

HAN CHEN, PHD

Hongyu Miao Digitally signed by Hongyu Miao Date: 2020.07.08 16:47:21 -05'00'

HONGYU MIAO, PHD

DEAN, THE UNIVERSITY OF TEXAS SCHOOL OF PUBLIC HEALTH

DEDICATION

To Meiqi Li, Yahong Li

SEMIPARAMETRIC METHODS TO IMPROVE RISK ASSESSMENT AND DYNAMIC PREDICTION

by

WEN LI

BA, Shanghai Jiao Tong University, 2011

Presented to the Faculty of The University of Texas
School of Public Health
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
August, 2020

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my dissertation supervisor Dr. Jing Ning for her mentoring and support. Her dedication and passion for statistical research inspired me to become a researcher. I want to express my sincere gratitude to my committee chair and dissertation co-supervisor Dr. Ruosha Li for her guidance and persistent help. It's no doubt that Dr. Jing Ning and Dr. Ruosha Li have been the two most important people for me in my PhD study. Without their help, I could not have finished these exciting projects. The weekly meeting with them have always been my favorite time of each week. Their mentoring, enlightenment, and encouragement will continue to benefit my future career.

I would like to thank my minor advisor Dr. Han Chen, breadth advisor Dr. Hongyu Miao, and external reviewer Dr. Liang Zhu for their consistent help and support. I would also like to thank Dr. Hongjian Zhu for being my academic advisor during my master's study and teaching me advanced statistical theories. Special thanks to Dr. Wenyaw Chan and Dr. Peng Wei for their generous recommendation when I transferred to the Biostatistics programs. I would also like to thank Zhiling Wu, Jessica Swann, and LeeAnn Chastain for their professionally help of polishing the writing.

This dissertation would not have been possible without the support by my beloved family and friends. I would like to thank them for being great listeners and helpers. I would especially like to thank my husband Zhuojun for taking up most of the housework in the last two years of my PhD study and letting me pursue this goal. I would also like to thank my daughter Elizabeth for being my dearest sweetheart. I would like to thank my parents Meiqi Li and Yahong Li for their love, trust, respect, and being my safe harbor.

I am lucky to be surrounded with wonderful advisors, family, and friends. None of this would have been accomplished without their contribution.

SEMIPARAMETRIC METHODS TO IMPROVE RISK ASSESSMENT AND DYNAMIC PREDICTION

Wen Li, BA, PhD
The University of Texas
School of Public Health, 2020

Dissertation Chair: Ruosha Li, PhD

Incorporating promising biomarkers to improve risk assessment and prediction is the central goal in many biomedical studies. Cost-effective designs and longitudinal designs are often utilized for measuring biomarker information, but they pose challenges to the data analyses. Statistical analyses for these kinds of data are routinely performed using parametric models. When the model assumptions are violated, parametric models may lead to substantial bias in parameter estimation, risk evaluation and prediction. In this dissertation, we will develop robust, flexible statistical methods for risk assessment for matched case-control, nested case-control, and case-cohort designs, as well as a dynamic prediction tool for longitudinal data. In the first aim, we will develop a distribution-free method for identifying an optimal combination of biomarkers to differentiate cases and controls in matched case-control data. In the second aim, we will develop a semiparametric regression model with minimal assumptions on the link function for data from two-phase sampling designs with *binary* outcomes. In the third aim, we will develop a model-free dynamic prediction method for a *survival* outcome that provides dynamically updated risk scores using longitudinal biomarker(s).

Contents

List of Figures

List of Tables

Chapter 1

Rationale and Objectives

Numerous novel biomarkers have emerged with the advent of biotechnologies, and they have the potential to improve disease screening, diagnosis, and prognosis. The immediate questions are whether the novel biomarkers are useful (e.g., whether they can substantially improve current or standard performance) and how to incorporate biomarkers with routine clinical risk factors. To identify useful biomarkers for early detection of cancer, Pepe et al. (2001) defined a set of comprehensive guidelines that included five phases of biomarker development. The statistical methods in this dissertation address the statistical challenges in different phases of biomarker development. The first method in this dissertation can be employed in the phase 2 of the biomarker development where the cases (those with cancer) and controls (those without cancer) are selected through a cross-sectional study, usually a matched case-control study. The focus of the first method is to assess how well a combination of biomarkers can discriminate the cases from the controls in an individually matched study. The second method in this dissertation can be viewed as a tool for the phase 3 of the biomarker development where cases and controls are sampled from a prospective cohort study. While routine clinical variables are available for the entire cohort, only a portion of the cohort are sampled for biomarker measurement. This kind of study designs include the nested case-control (NCC) and the case-cohort study designs (Liddell et al., 1977; Prentice and Breslow, 1978; Prentice, 1986). The

goal of the second method is to combine novel biomarkers and routine variables for risk assessment as well as dealing with the outcome-dependent data missingness. The third method in this dissertation can aid the phase 3 of biomarker development where the prediction power over time is of interest. In this phase, biomarkers are measured repeatedly during the follow-up so that risk prediction can be refined utilizing the most recent information. Thus, the goal of the third method is to fully use the longitudinal information and to facilitate dynamically updated risk prediction through the information captured in repeated measurement data.

Statistical analyses for the aforementioned study designs are routinely performed using parametric or semiparametric models with strong model assumptions. However, when the assumptions are violated, these models may lead to biased parameter estimates and invalid risk prediction. Thus, it is more desirable to relax the assumptions by posing minimal constraints on the link function or using distribution-free methods. Therefore, in this dissertation, we aim to develop robust semiparametric methods to improve risk assessment and dynamic prediction while incorporating cost-effective study designs and longitudinal study designs.

1.1 Assessing discrimination capacity of a combination of biomarkers under matched case-control studies

1.1.1 Introduction

The performance of the current cancer screening programs is still far from satisfactory for many types of cancers. For example, the sensitivity of the current surveillance for hepatocellular carcinoma (ultrasound every 6 months in cirrhosis patients) only ranges from 32% to 65% (Singal et al., 2009, 2012). Under such circumstances, biomarkers and their combinations with patients' clinical characteristics can serve as promising tools and

will likely be the best option for future research to complement the current population screenings (Schütte et al., 2015). Pepe et al. (2001) provided a comprehensive set of guidelines and recommended five phases for biomarker development studies. In phase 2 studies, case-control studies are commonly used to assess the ability to distinguish cases and controls. Particularly, a matched case-control study is a popular option to reduce the confounding issue, in which each of the cases is matched to one or more controls based on variables believed to be confounders. There are several advantages of matching. First, it allows to assess the classification accuracy of the biomarkers beyond the contribution of the matching variables (Janes and Pepe, 2008). Second, it has been reported that a balanced number of cases and controls across the levels of the matching variables can reduce the variance for estimating parameters of interest compared to an unmatched study with the same sample size (Breslow et al., 2005; Rose and Van der Laan, 2009).

1.1.2 Literature review

Matched case-control data have been routinely analyzed by conditional logistic regression in literature. Then the combination of biomarkers, termed as composite score, is derived by maximizing the conditional likelihood, a global fit criterion. To quantify the discrimination ability of the derived composite score, sensitivity and specificity are two commonly used measures. The sensitivity and specificity associated with a cut-off can be calculated respectively using the percentage of positive results (e.g., composite score $>$ the cut-off) under cases, and the percentage of positive results (e.g., composite score $\leq$ the cut-off) under controls. The cut-off can be determined by a certain criterion such as Youden's index (Unal, 2017).

Maintaining a high specificity has been noted as the top priority for population screening since it can prevent a large number of disease-free subjects from going through unnecessary costly medical procedures and psychological stress (Pepe et al., 2001, 2008). Taking the ovarian cancer screening as an example, a clinically acceptable specificity should exceed 98% (Skates et al., 2004; Pepe et al., 2008). Although the aforementioned

composite score is derived by maximizing the likelihood function, it is not clear it is still an optimal score within this clinically meaningful region. Another limitation of the conditional logistic regression is the parametric link function that connects the composite score and disease risk. In practice, investigators often have little prior knowledge about the mathematical form of the underlying true link, although the logit link is routinely used. Misspecified link functions may lead to non-optimal composite scores (Shen et al., 2018). Thus, it is more desirable to make the link function unspecified and enjoy the robustness of semiparametric models.

Some recent works considered the unique features of population screening and constructed composite scores by maximizing a local criterion. For the data from case-control studies, Meisner et al. (2017) and Zhang et al. (2019) proposed to directly maximize the sensitivity under the constraint that the specificity is greater than a pre-specified threshold. Nevertheless, this method only included information from the cases and ignored the information from controls in the objective function. Consequently, the derived composite scores may not be able to maintain the pre-specified specificity in external validation studies, the top priority in population screening. Yan et al. (2018) alternatively derived the composite score by maximizing the partial area under the receiver operating characteristic (ROC) curve, which is a trade-off between the local and global criteria.

1.1.3 A motivating example

The first project in this dissertation is motivated by a prostate cancer data set in the Carotene and Retinol Efficacy Trial (CARET). It is a randomized trial that was originally designed to evaluate the efficacy of the combination of beta-carotene and retinol on reducing lung cancer risk. It enrolled 18,314 subjects at high risk for lung cancer. During the intervention phase of CARET, blood samples were collected and stored, and thus provided invaluable resources for future research.

Within the CARET, a matched case-control study was conducted. For each of the 71 prostate cancer cases diagnosed between 1998 and 1995, one control who was free of

prostate cancer by the study time was matched by age and number of blood samples. Two biomarkers for prostate cancer, the total prostate specific antigen (tPSA) and the free prostate-specific antigen (fPSA), were measured from the stored blood samples of the subjects in the data set. The details of this study were provided in Etzioni et al. (1999). It is of interest here to evaluate the discrimination ability of the biomarkers with a tool that can simultaneously address the matched design and offer robustness in terms of model mis-specification.

1.2 Risk assessment under two-phase sampling designs

1.2.1 Introduction

In disease risk assessment in a large prospective cohort, two-phase sampling designs are commonly adopted as a cost-effective alternative (Liddell et al., 1977; Prentice and Breslow, 1978; Prentice, 1986). Furthermore, to overcome the inherent problem of low incidence encountered with rare diseases, it is often necessary to employ two-phase designs for early detection in a cohort of disease-free subjects. In the first phase of a two-phase sampling design, a large cohort is sampled from the target population. The outcome variable is prospectively collected and some easy-to-obtain covariates such as routine clinical risk factors and demographic characteristics are recorded. In the second phase, a subcohort of all the cases and a fraction of the controls in the full cohort are selected for biomarker measurements. Two commonly used two-phase sampling designs are the NCC design and the case-cohort design, which differ in their approach for selecting controls. In the NCC design, controls are chosen without replacement from the risk set at each event time (Liddell et al., 1977; Prentice and Breslow, 1978). In the case-cohort design, controls are randomly selected at baseline (Prentice, 1986). However, these cost-effective sampling strategies create challenges for statistical analysis because of data missingness for biomarker measurements.

1.2.2 Literature review

The analysis of two-phase design data with binary outcomes has routinely been carried out using parametric models. Two popular methods are the conditional logistic regression model for the NCC design (Schwartz et al., 2017; Keizman et al., 2017) and the logistic regression model with inverse probability weighting (IPW), hereafter termed IPW-based logistic regression, for the case-cohort design (Noma and Tanaka, 2017; Landry et al., 2017). In application, researchers often have limited information regarding the mathematical specification of the true regression function, although a logit link between the disease probability and risk score is a convenient choice. However, the underlying relationship may differ from the logit link in many situations, leading to biased estimation of the regression coefficients and/or disease probabilities. It is more desirable to assume a semiparametric model with minimal assumptions on the link function. Isotonic regression is a least squares problem in which only monotonicity is assumed on the shape of the regression models. Pioneering work was done by Ayer et al. (1955) and comprehensive reviews were provided by Barlow et al. (1972) and Robertson et al. (1988). A unique solution to standard isotonic regression exists and can be obtained using the pool-adjacent violators algorithm (PAVA) (Barlow et al., 1972; Best and Chakravarti, 1990; Qin, 2017). The computational aspects and fast implementation of PAVA in R are discussed by Mair et al. (2009). However, standard isotonic regression with PAVA cannot be directly applied to the data from two-phase studies due to the outcome-dependent data missingness. Therefore, the goal of this aim is to handle such data under a semiparametric isotonic regression model and to develop a computationally appealing algorithm by integrating PAVA, the IPW method and the profiling method.

1.2.3 A motivating example

The Rotterdam breast cancer data include 2,982 primary breast cancer patients underwent primary surgery between 1978 and 1993. The details of the data can be found in Sauerbrei et al. (2007). Biomarkers such as progesterone receptor and estrogen recep-

tor were available for the full cohort. Other prognostic factors and treatment variables included age, menopausal status, tumor size, tumor grade, number of positive lymph nodes, hormonal therapy and chemotherapy. Using this full data set, we can create an NCC data set nested within this cohort and evaluate the risk of developing an important clinical event (e.g., death in two years after primary surgery).

1.3 Dynamic scoring system of a survival outcome using longitudinally collected biomarkers

1.3.1 Introduction

Longitudinal designs for biomarker traits are very appealing. The repeated collection of biomarker information of the same patient over time can update the prognosis and improve the time-varying classification of patients with different predicted risk levels. Several dynamic scoring systems emerged for assorted diseases such as the Dynamic International Prognostic Scoring System (DIPSS) and its refined version (DIPSS-plus) for primary myelofibrosis (Passamonti et al., 2010; Gangat et al., 2011); the Dynamic Stage, Size, Grade, and Necrosis (D-SSIGN) score for clear-cell renal cell carcinoma (Thompson et al., 2007); the dynamic thrombolysis in myocardial infarction (dynamic TIMI) risk score for ST-elevation myocardial infarction (Amin et al., 2013); and the dynamic prognostic score for head and neck squamous cell carcinoma (van der Schroeff et al., 2012). Unlike the static scores which are used to stratify patients only at study enrollment or baseline, these dynamic scores are designed to help guide treatment decisions at any time during the follow-up.

1.3.2 Literature review

Although dynamic scoring system attracts increasingly more attention in the medical field, there is a lack of methodology development. To our knowledge, current dynamic

scores were constructed by either repeating analysis at multiple follow-up times or using a time-dependent covariate Cox model. There are several drawbacks regarding these methods. First, repeated analyses utilize the information at one time point in each analysis; they do not make full use of the longitudinally collected information, and a set of follow-up times need to be pre-specified. Second, although the time-dependent Cox model is a convenient option to obtain biomarker effects or hazard ratios, it assumes biomarkers are available continuously over time, which is rarely true in biomarker measurement.

Motivated by the need to incorporate longitudinal biomarkers for a dynamic scoring system, the statistical challenge is how to efficiently use a tool capable of updating risk prediction as more longitudinal information is collected during follow-up. Designed for the dynamic prediction task, the partly conditional model is a system of prediction models that change with the follow-up time (Zheng and Heagerty, 2005). Its regression term that combines the biomarkers can be naturally treated as a dynamic score. A similar approach is called the landmark model (van Houwelingen and Putter, 2011). Within the partly conditional model framework, Maziarz et al. (2017) proposed a two-stage procedure that improves the prediction performance when large variation exists due to measurement errors in biomarkers. Nevertheless, even though the partly conditional model can be more easily implemented in practice, the validity of model inference requires the proportional hazards assumption for the sequence of Cox models. Another approach for dynamic prediction is the joint modeling, which models the longitudinal trend of the time-dependent covariates, usually through individual-specific random effects and parametric models (Rizopoulos, 2011). However, the joint modeling does not provide a direct combination of the longitudinal biomarkers that physicians can use for risk stratification, and thus will not be discussed in the rest of this study.

1.3.3 A motivating example

The third project in this dissertation is motivated by the data from the Terry Beirn Community Programs for Clinical Research on AIDS didanosine/zalcitabine trial, which

randomized 467 human immunodeficiency virus (HIV) infected patients to receive one of the two antiretroviral drugs: didanosine or zalcitabine. Absolute CD4 cell count in the peripheral blood was measured at baseline, 2nd, 6th, 12th, and 18th months during the follow-up. The primary outcome is time to death and about 40% of patients died at the end of the study. Details of the study design can be found in Abrams et al. (1994). To discriminate between patients with high-risk and low-risk of death by using all available information including the longitudinal CD4 count measurements, a dynamic prediction model must be constructed.

1.4 Public Health Significance

1.4.1 Assessing discrimination capacity of a combination of biomarkers for prostate cancer

The proposed method in the first project in this dissertation can identify the optimal combination of biomarkers for the data in matched case-control studies. The method is especially useful in early phase biomarker development for population screening. It can accurately detect true positive subjects, and then reduce morbidity and mortality. By constraining specificity to be higher than a cutoff, the proposed method can also help avoid unnecessary public health burdens caused by false positive results in population screening.

1.4.2 Risk assessment for breast cancer patients

The proposed method in the second project in this dissertation can generate accurate parameter estimates to facilitate optimal scoring systems for data from two-phase sampling designs. Successful implementation of the proposed study could ensure more accurate risk stratification for patients with breast cancer. Consequently, it will help health providers identify high-risk subjects and make good use of the limited healthcare resources. In particular, if high-risk subjects are identified and treated earlier, they can achieve better

health and quality of life.

1.4.3 Timely disease prognosis of AIDS

AIDS is a syndrome caused by HIV infection which destroys the immune system. The prognosis of HIV infected patients can be improved with 80% reduction of death rate and 20-50 years increase of life span if the patients are treated properly (Collaboration et al., 2008). So, it is critical to provide timely disease prognosis and adjusted medical treatments for HIV patients.

Our proposed method can provide updated risk stratification by taking into account longitudinally measured CD4 count. It can provide personalized information for patients and facilitate guided treatment decision making, too. In fact, our proposed method can be applied to typical longitudinal studies where longitudinal measurements are collected during the follow-up.

1.5 Specific Aims

Specific Aim 1: To develop a robust method to identify optimal combination of biomarkers given data in the matched case-control studies.

We will develop an objective function to maximize the discrimination ability of the composite score. This method is more robust than the commonly used conditional logistic regression model by leaving the link function unspecified. Moreover, it is also more tailored to clinical needs by imposing a constraint on specificity.

Specific Aim 2: To develop estimation procedures and computation algorithm for conducting semiparametric isotonic regression in two-phase studies.

We will develop estimation procedures under a semiparametric isotonic regression by integrating PAVA, the IPW method and the profiling method. This proposed method can combine multiple biomarkers, construct risk scores, assess absolute risks, and handle data from two-phase sampling designs with binary outcomes.

Specific Aim 3: To develop an optimal scoring system for dynamic prediction using longitudinal biomarkers.

We will develop a model-free dynamic prediction method that can facilitate timely disease prognosis at each biomarker measurement time. The estimates in this model will evolve with the ever-changing risk sets and handle both regularly and irregularly measured longitudinal data.

Chapter 2

Methods and Results for Aim 1

2.1 Method

2.1.1 Notations

Consider a matched case-control study that allows multiple cases or controls in each stratum. Denote Y_{ki} as the disease status for the ith subject in the kth stratum, $k = 1, \ldots, K$. $Y_{ki} = 1$ means diseased (e.g., case) and $Y_{ki} = 0$ means non-diseased (e.g., control). Let n_{kD} and $n_{k\bar{D}}$ be the number of cases and matched controls in stratum k, respectively, and denote $n_k = n_{kD} + n_{k\bar{D}}$ as the stratum total. Then $n_D = \sum_{k=1}^{K} n_{kD}$ and $n_{\bar{D}} = \sum_{k=1}^{K} n_{k\bar{D}}$ are the total numbers of cases and controls, respectively. For notation simplicity, we arrange the subjects in each stratum such that the first n_{kD} subjects are cases. Let $\boldsymbol{X}_{ki}$ be the p-dimensional vector of biomarkers for the ith subject in the kth stratum. We define the composite score as a linear combination $\boldsymbol{\beta}^T \boldsymbol{X}_{ki}$, where $\boldsymbol{\beta}$ is a vector of coefficients with the same dimension of $\boldsymbol{X}_{ki}$.

Given the composite score and a cut-off c, the sensitivity can be estimated as

$$\widehat{Se}(\boldsymbol{\beta}, c) = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_{kD}} I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} > c)}{\sum_{k=1}^{K} n_{kD}}. \tag{2.1}$$

Similarly, the study-specific specificity can be estimated as

$$\widehat{Sp}_s(\boldsymbol{\beta}, c) = \frac{\sum_{k=1}^{K} \sum_{i=n_{kD}+1}^{n_k} I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} \leq c)}{\sum_{k=1}^{K} n_{k\bar{D}}}. \tag{2.2}$$

Note that the controls are sampled based on the matching variables of their matched cases instead of random sampling, and thus they cannot represent the general control population. Denote the sampling probability as $p_{ki}, i \in \{n_{kD}+1, \ldots, n_k\}$. Then we can estimate the population-level specificity as follows:

$$\widehat{Sp}(\boldsymbol{\beta}, c) = \frac{\sum_{k=1}^{K} \sum_{i=n_{kD}+1}^{n_k} \widehat{w}_{ki} I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} \leq c)}{\sum_{k=1}^{K} \sum_{i=n_{kD}+1}^{n_k} \widehat{w}_{ki}}, \tag{2.3}$$

where $\widehat{w}_{ki} = 1/\widehat{p}_{ki}$ and the estimated sampling probability, $\widehat{p}_{ki}$, can be estimated empirically or via a logistic regression model.

2.1.2 Review of Existing Methods

Data from matched case-control studies are routinely analyzed using conditional logistic regression. The associated conditional likelihood is conditional on the total number of cases and the total number of subjects within each stratum, which avoids the estimation of stratum-specific nuisance parameters,

$$\mathcal{L}_{CL}(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\prod_{i=1}^{n_{kD}} \exp(\boldsymbol{\beta}^T \boldsymbol{X}_{ki})}{\sum_{J \in \mathcal{C}_k^D} \prod_{j \in J} \exp(\boldsymbol{\beta}^T \boldsymbol{X}_{kj})}, \tag{2.4}$$

where $\mathcal{C}_k^D$ are all subsets of size n_{kD} from $\mathcal{C}_k = \{1, \ldots, n_k\}$. Denote $\widehat{\boldsymbol{\beta}}_{CL}$ as the estimator of $\boldsymbol{\beta}$, which maximizes the conditional likelihood in (2.4). In cancer population screening, a high specificity is its top priority, so the cut-off value is usually determined by $\widehat{c}_{CL} = \inf\{c : \widehat{Sp}(\widehat{\boldsymbol{\beta}}_{CL}, c) \geq \tau\}$, where τ is a pre-specified specificity such as 0.98 for the ovarian cancer screening. Then the corresponding sensitivity is $\widehat{Se}_{CL} = \widehat{Se}(\widehat{\boldsymbol{\beta}}_{CL}, \widehat{c}_{CL})$.

Alternatively, for case-control studies, Meisner et al. (2017) and Zhang et al. (2019) proposed a direct method to maximize the sensitivity under the constraint that the speci-

ficity $\geq \tau$. For matched case-control studies, the population specificity in (2.3) instead of the study-specific specificity in (2.2) should be used. Denote the maximizers as ($\widehat{\boldsymbol{\beta}}_D$ and $\widehat{c}_D$). The corresponding sensitivity can be subsequently calculated by $\widehat{Se}_D = \widehat{Se}(\widehat{\boldsymbol{\beta}}_D, \widehat{c}_D)$. As expected, the direct method may derive a score with substantially higher sensitivities than that by the conditional logistic regression, since it maximizes the sensitivity directly. . However, the objective function of the direct method only includes information from the cases and ignores information from the controls. Given the external validation data sets, as shown in Meisner et al. (2017), its composite score cannot maintain the pre-specified specificity.

Yan et al. (2018) recently developed an optimal score by maximizing the partial area under the ROC curve, termed as pAUC method. This method was originally designed for data from case-control studies, so we will generalized this method to accommodate data from matched case-control studies and evaluate its performance in this setting in Section 2.2.

2.1.3 Proposed Method

Motivated by the limitations of the existing methods in Section 2.2 and the robustness of semiparametric models, we leave the link function unspecified and propose the following pseudo-conditional likelihood function:

$$\mathcal{L}(\boldsymbol{\beta}, c) = \prod_{k=1}^{K} \frac{\prod_{i=1}^{n_{kD}} I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} > c) \prod_{i=n_{kD}+1}^{n_k} \{1 - I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} > c)\}}{\sum_{J \in \mathcal{C}_k^D} \left[\prod_{j \in J} I(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} > c) \prod_{j \in \mathcal{C}_k \setminus J} \{1 - I(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} > c)\} \right]}. \tag{2.5}$$

The derivation of (2.5) is provided in the Appendix. To ensure identifiability, we set the Euclidean norm $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{i=1}^{p} \beta_i^2}$ to be 1. When maximizing this likelihood, we add a tiny number ϵ to the numerator and denominator to avoid a zero in the denominator. The simulation studies confirm that the estimation is not sensitive to the value of ϵ. Similar to the conditional likelihood, the pseudo-conditional likelihood characterizes the discrimination ability of the composite score within each case-control stratum, and eliminates the

need to estimate stratum-specific parameters. The pseudo-conditional likelihood makes a close connection with the final rules to calculate the sensitivity and specificity and avoids a parametric specification on the link function between the composite score and the probability of being diseased. The denominator describes all possible classifications while the numerator is the correct classification. Different from the objective function of the direct method, our pseudo-conditional likelihood unitizes the information from both cases and controls; and ensures a better control for specificity on independent validation data sets, which is confirmed in Section 2.2.

To ensure the clinically acceptable specificity as our priority, we maximize the pseudo-conditional likelihood subject to the constraint of $\widehat{Sp}(\boldsymbol{\beta}, c) \geq \tau$. The threshold τ is pre-specified and should be tailored to the study of interest. For example, a threshold of 80% might be reasonable in a study of high-risk subjects, and a much higher threshold (e.g., 98%) is usually required for general population screening. Maximizing (2.5) under the constraint is not computationally straightforward, so we propose a stable and computationally efficient algorithm based on the profiling approach. For any given $\boldsymbol{\beta}$, we can obtain an estimate of c, denoted as $\widehat{c}(\boldsymbol{\beta})$, by finding the τth quantile of $\boldsymbol{\beta}^T\boldsymbol{X}$ among controls after incorporating the sampling weights. We then plug $\widehat{c}(\boldsymbol{\beta})$ in equation (2.5) and maximize the profiled pseudo-conditional likelihood $\mathcal{L}\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\}$ with respect to $\boldsymbol{\beta}$. Given these estimates, the sensitivity and specificities can be calculated by equations (2.1), (2.2) and (2.3).

Note that the pseudo-likelihood is not a continuous function of the unknown parameters. With a small number of biomarkers, we can adopt the Nelder-Mead method and multiple starting values to identify the global maximizers. However, with a large number of biomarkers, this method is impractical due to the intensive computation. An alternative solution is to use a continuous kernel function to approximate the indicator function, $\int_{-\infty}^{\boldsymbol{\beta}^T\boldsymbol{X}_{ki}-c} K(u; h_n)du$, where $K(\cdot, h_n)$ is a symmetric kernel function and h_n is the bandwidth (Jones, 1990; Zeng and Lin, 2007; Shen et al., 2018). Accordingly, we

have the following kernel-smoothed pseudo-conditional likelihood:

$$\mathcal{L}_K(\boldsymbol{\beta}, c) = \prod_{k=1}^{K} \frac{\prod_{i=1}^{n_{kD}} \int_{-\infty}^{\boldsymbol{\beta}^T \boldsymbol{X}_{ki}-c} K(u; h_n)du \prod_{i=n_{kD}+1}^{n_k} \{1 - \int_{-\infty}^{\boldsymbol{\beta}^T \boldsymbol{X}_{ki}-c} K(u; h_n)du\}}{\sum_{J \in \mathcal{C}_k^D} \left[\prod_{j \in J} \int_{-\infty}^{\boldsymbol{\beta}^T \boldsymbol{X}_{kj}-c} K(u; h_n)du \prod_{j \in \mathcal{C}_k \setminus J} \{1 - \int_{-\infty}^{\boldsymbol{\beta}^T \boldsymbol{X}_{kj}-c} K(u; h_n)du\}\right]}. \tag{2.6}$$

Although any smooth and symmetric probability density functions can be used as the kernel function, the standard normal distribution is a popular choice in practice. Details about the Gaussian kernel for equation (2.6) are provided in the Appendix.

2.2 Asymptotic Properties

In this section, we establish the asymptotic properties of $(\widehat{\boldsymbol{\beta}}, \widehat{c})$ and $\widehat{Se}(\widehat{\boldsymbol{\beta}}, \widehat{c})$. Denote the true values of these parameters by $(\widetilde{\boldsymbol{\beta}}, \widetilde{c})$ and $\widetilde{Se} = \widetilde{Se}(\widetilde{\boldsymbol{\beta}}, \widetilde{c})$. The main technical challenge is the discontinuity of $\mathcal{L}(\boldsymbol{\beta}, c)$ due to the indicator function, since standard methods require the smoothness and differentiability of the likelihood function. Under the mild regularity conditions given in the Appendix, we apply the empirical processes techniques to prove that $\widehat{Se}$ is a consistent estimator of $\widetilde{Se}$.

2.3 Simulation Studies

We conducted extensive simulation studies to evaluate the finite sample performance of the proposed method and compared it to that of three existing methods: the conditional logistic regression, the direct method by Meisner et al. (2017) and Zhang et al. (2019), and pAUC method by Yan et al. (2018).

To enable fair comparisons, we first extended the pAUC method to accommodate data from matched case-control studies. We estimated the density function for the control

group by incorporating the sampling weights,

$$\widehat{f}_{\bar{D}}(s) = \frac{1}{\sum_{k=1}^{K}\sum_{i=n_{kD}+1}^{n_k}\{\widehat{w}_{ki}h_{\bar{D}}\}}\sum_{k=1}^{K}\sum_{i=n_{kD}+1}^{n_k}\widehat{w}_{ki}K\left(\frac{s-\boldsymbol{\beta}^T\boldsymbol{X}_{ki}}{h_{\bar{D}}}\right), \tag{2.7}$$

where $K(.)$ is the kernel function, and $h_{\bar{D}}$ is the bandwidth. The density function for the case group can be estimated in a similar fashion but without the weight, named as $\widehat{f}_D$. Then the two estimated survival functions were $\widehat{S}_D(s) = \int_s^\infty \widehat{f}_D(t)dt$ and $\widehat{S}_{\bar{D}}(s) = \int_s^\infty \widehat{f}_{\bar{D}}(t)dt)$, respectively. The kernel smoothed ROC was then given by $\widehat{\text{ROC}}_K(t) = \widehat{S}_D\{\widehat{S}_{\bar{D}}^{-1}(t)\}$. Integrating over the range of specificities of interest (t_0, 1), or equivalently the range of false positive rates $(0, 1-t_0)$, we derived the corresponding kernel smoothed pAUC, $\widehat{\text{pAUC}}_K = \int_0^{1-t_0}\widehat{\text{ROC}}_K(t)dt$. Given the coefficient estimates $\widehat{\boldsymbol{\beta}}_{pauc}$ that maximized $\widehat{\text{pAUC}}_K$, we can identify the cutoff value $\widehat{c}_{pauc}$ to make $\widehat{Sp}(\widehat{\boldsymbol{\beta}}_{pauc}, c) \geq \tau$. The subsequent sensitivity and specificity can be obtained by equations (2.1)-(2.3).

2.3.1 Data Generation

We considered different scenarios for the performance evaluation:

Scenario 1. We generated two independent biomarkers, X_1 and X_2, from the standard normal distribution. We generated two matching variables Z_1 and Z_2 from $Bernoulli(0.3)$ and $Bernoulli(0.1)$ independently. We then defined the matching group $\mathcal{S}$ based on the values of Z_1 and Z_2: $\mathcal{S} = 1$ if $Z_1 = 0$ and $Z_2 = 0$; $\mathcal{S} = 2$ if $Z_1 = 1$ and $Z_2 = 0$; $\mathcal{S} = 3$ if $Z_1 = 0$ and $Z_2 = 1$; and $\mathcal{S} = 4$ otherwise. We last generated the disease status from a *Bernoulli* distribution with a diseased probability of $\text{logit}^{-1}\{(X_1 + 3X_2 + 0.5Z_1 + 4Z_2)/1.5 - 7\}$, where $\text{logit}(t) = \log\{t/(1-t)\}$.

Scenario 2. We generated two biomarkers X_1 and X_2 as well as the matching variable Z_1 from a multivariate normal distribution conditional on the disease status. Among controls, X_1 followed $N(0, 3)$, and both X_2 and Z_1 followed $N(0, 1)$. They were pairwise correlated with a correlation coefficient of 0.3. Among cases, X_1, X_2, and Z_1 independently followed $N(3, 3)$, $N(3, 5)$, and $N(3, 5)$, respectively. Hence, both means and covariance

matrices of the biomarkers and the matching variable were dependent on the disease status, and the covariance matrices were disproportional for cases and controls. We then generated the matching group as $\mathcal{S} = I\{Z_1 \geq \Phi^{-1}(1/4)\} + I\{Z_1 \geq \Phi^{-1}(1/2)\} + I\{Z_1 \geq \Phi^{-1}(3/4)\} + 1$, where Φ is the standard normal cumulative distribution function.

Scenario 3. We used the same sampling mechanism used in Scenario 2, except that the correlation between biomarkers among cases was increased to 0.9 to further the extent of disproportion in the covariance matrices.

Scenario 4. We considered the same means but different correlation directions between controls and cases in this scenario. Specifically, X_1, X_2 and Z_1 were negatively correlated with a correlation coefficient of -0.3 among controls, whereas they were positively correlated with a correlation coefficient of 0.3 among cases. The marginal distribution of X_1 was $N(0, 3)$, and the marginal distributions of X_2 and Z_1 were $N(0, 1)$ for controls and $N(0, 5)$ for cases. In this scenario, because the cases and controls had the same means, it was not easy to separate them.

In all four scenarios, we used 1:1 matching to construct the matched case-control data; that is, for each case, we sampled one control among those nondiseased subjects in the same matching group as that in the case. We added a tiny number of .0001 to both the numerator and the denominator of the pseudo-conditional likelihood function to avoid the occurrence of zero in the denominator or the product. To ensure that we identify the global maxima of the proposed likelihood, we used 20 sets of starting values around the coefficient estimates by the conditional logistic regression. We then obtained the proposed estimates that achieve the largest value of the objective function over the 20 maximizations with different initial values. Note that the maximization converged quickly for our method even though multiple starting values were used. We adopted the bootstrap method for the variance estimation. In particular, we resampled the strata with replacement 200 times and calculated the sample standard deviation.

To generate true sensitivities under Scenario 1, we used the known true optimal combination of the biomarkers and 1,000 independent huge datasets to mimic the population data. The final truth was the average over the resulting 1000 sensitivities. To gener-

ate truth for reference under Scenario 2-4, we used the known true distributions of the biomarkers. Since there were only two biomarkers in simulation, a full grid search based on the equation of sensitivity under multivariate normal assumption for biomarkers, $\Phi(\frac{\mu_D - \mu_{\bar{D}} + \sigma_{\bar{D}}\Phi^{-1}(1-\tau)}{\sigma_D})$, was adopted to generate performance reference. Here, μ_D and $\mu_{\bar{D}}$ are the means of the composite scores for cases and controls, respectively; and σ_D and $\sigma_{\bar{D}}$ are the standard deviations of the composite scores for cases and controls, respectively.

When implementing the kernel-smoothing method, we chose bandwidth $h_n = C_h(n_D)^{-1/3}$, where $(n_D)^{-1/3}$ is the optimal bandwidth recommended by Jones (1990) and $C_h = 0.2$, 1, or 5. $n_D = n_{\bar{D}}$ varied from 50 to 400, and the pre-specified threshold of specificity τ varied from 0.70 to 0.98. For each setting, we used 1,000 simulation replicates to summarize the simulation results. We calculated the sensitivities and specificities of the composite score by the aforementioned four methods using independent external validation data sets with a large sample size of 20,000, such that the variability due to the external data sets was ignorable (Payne et al., 2016; Yan et al., 2018). The specificity range of interest for pAUC method was set to be (0.7, 1) or $t_0 = 0.7$. For fair comparison, the same 20 sets of starting values were used for the proposed, the direct, and the pAUC methods.

2.3.2 Simulation Results

Figures 2.1 & 2.2 show the average values and empirical standard errors (ESE) of estimated sensitivities (± ESE) and specificities (± ESE) on the validation data at various prespecified specificities τ (0.70, 0.75, 0.85, 0.90, 0.95, and 0.98). Here the composite scores and the cutoffs were estimated using the training data sets with a sample size of 200. To better differentiate the results of the four different methods, the error bars corresponding to different τs were shifted slightly along the x-axis. The corresponding summary tables are presented in Tables 6.1-6.4 in the Appendix.

Under Scenario 1, the logistic regression model is the underlying true model. When the sample size was small ($n_D = n_{\bar{D}} = 50$), all methods could not maintain specificity, as shown in Figure 2.1(B). But the proposed, the pAUC, and the conditional logistic meth-

ods had higher specificities than the direct method. When the sample size was increased, all methods except the direct method can maintain the prespecified specificities well, as seen in Figure 2.2(B). The direct method had slightly higher sensitivities compared with the other three methods, but had lower specificities than the prespecified levels. This finding made the direct method suboptimal, since a small drop in specificity will translate to a lot of subjects having false positive results in general population screening.

Under Scenarios 2-4, there is not a simple parametric model such as the logistic model to present the probability of having the disease. As expected, the proposed method clearly outperformed the other three methods. First, the proposed method produced the highest specificities. When the sample size was moderate or large (e.g., $n_D = n_{\bar{D}} \geq 100$), the specificities from the proposed method were close to the prespecified level of τ, and even higher than τ in some settings. Specifically, the difference between the average of estimated specificities and the prespecified level was between -0.02 and 0.01. This superiority of the proposed method can be explained by the full utilization of the control information in the proposed pseudo-conditional likelihood. On the other hand, the direct method again failed to preserve the specificity. For example, under Scenario 3 with n_D of 100 and τ of 0.80, the difference between the average of the estimated specificities and the prespecified level was as large as 0.06 (see Figure 2.2(F)). Similar to the direct method, the pAUC method could not maintain the specificity, especially when $\tau \leq 0.95$, even though it had a better control than the direct method. Second, the estimated sensitivities by the proposed method were consistently higher than those by the conditional logistic regression, due to the model flexibility of the semi-parametric property of the proposed method. For example, when $\tau = 0.98$, the *relative percentage difference*, defined as (mean $\widehat{Se}$−mean $\widehat{Se}_{CL}$)/mean $\widehat{Se}_{CL}\times 100\%$, ranged from 32% to 124%. Third, the proposed method, the direct method , and the pAUC method had much smaller empirical standard errors of the sensitivities than the conditional likelihood. For example, under Scenario 4, the empirical standard errors from the proposed method, the direct method, and the pAUC method were only 9% to 77% of those from the conditional logistic regression. This statistical efficiency gain was achieved mainly because all the three

methods focused on local or sub-global performances by focusing on clinically-relevant levels of the specificity; whereas the conditional logistic regression maximized the global performance including those clinically-irrelevant specificities, e.g, $\tau = 0.3$. Moreover, we evaluated the Youden's Index of the four methods and summarized the results in Tables 6.6-6.9 in the Appendix. Again, the proposed method showed better discrimination capacity than the conditional logistic regression when the evaluation metric placed equal importance on sensitivity and specificity.

Simulation results on the training data are summarized in Table 2.1 and Table 6.5 in the Appendix. The ESEs and the average of the estimated standard errors (ASEs) by the bootstrap method agreed well, indicating the bootstrap method can accurately capture the variability of the proposed method. Coverage probabilities based on the Fisher transformation were close to the nominal level except when both τ was close to 1 and sample size was moderate or large. We also implemented the method by maximizing the kernel-smoothed pseudo-conditional likelihood, and summarized the results in Table 6.10 in the Appendix. Overall, it results were very similar to those by the pseudo-conditional likelihood, suggesting the kernel-smoothed method is a reasonable alternative in our setting. In addition, we compared the results by using three different values of C_h and found that the kernel-smoothed method is quite robust to the choice of the bandwidth in our setting.

2.4 Application

We return to the aforementioned prostate cancer data set (Section 1.1.3) and illustrate the proposed method for disease status discrimination. For the illustrative purpose, we identified 68 matched pairs of cases and controls from that existing matched case-control study. Our goal was to compose a risk score using the biomarkers (tPSA and fPSA) to distinguish cases from controls under a matched study design. We performed bootstrap validation with a bootstrap sample size of 10,000 (Steyerberg et al., 2001). Due to the small sample size, we only focused on τ from 0.70 to 0.95. Overall, the proposed method

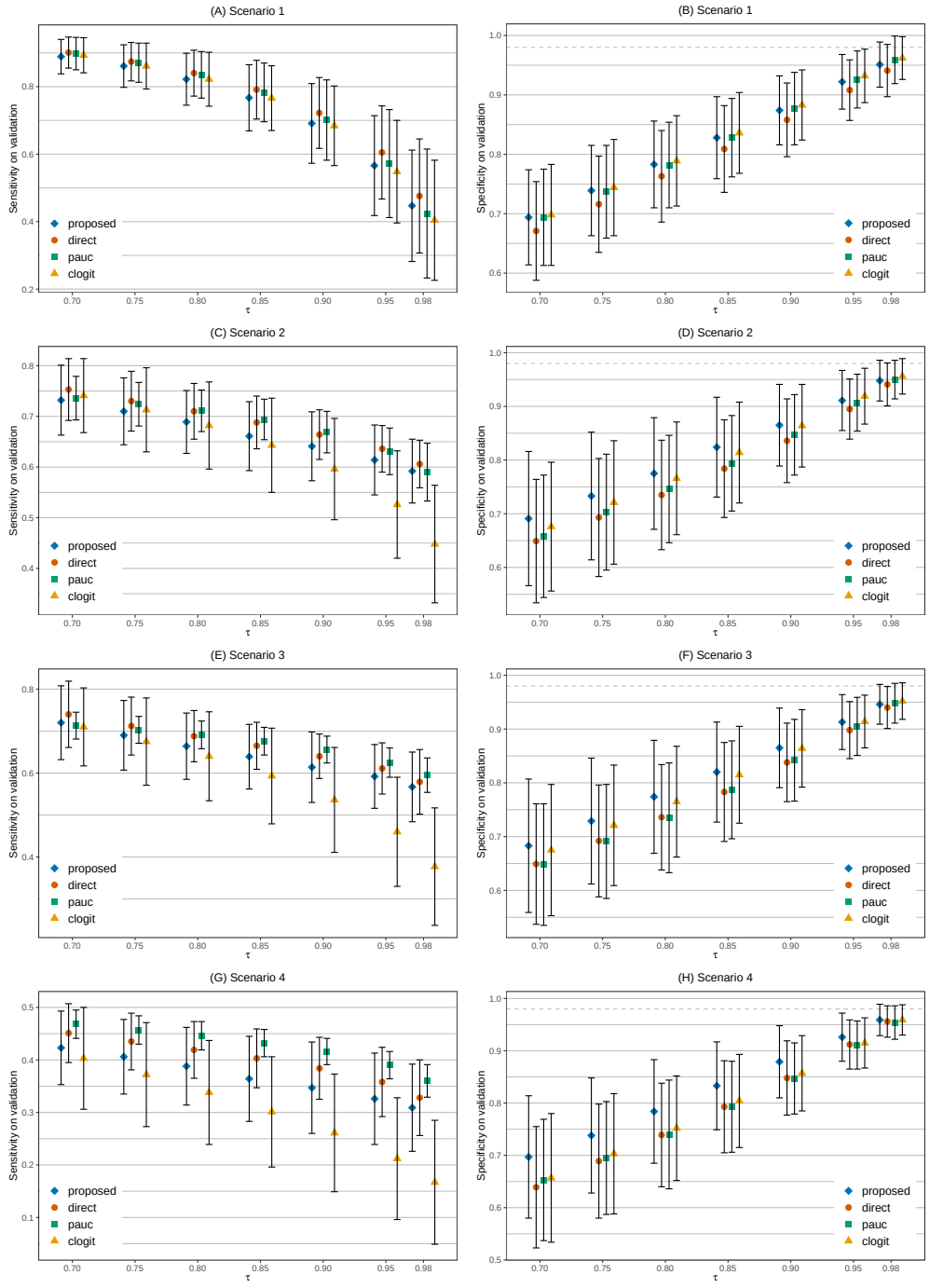

Figure 2.1: Visualization of simulation results on validation data when the sample size of the training data is $n_D = n_{\bar{D}} = 50$. Error bars are shifted slightly along the x-axis. τ: prespecified threshold of specificity. Gray dashed line: y-axis at 0.98.

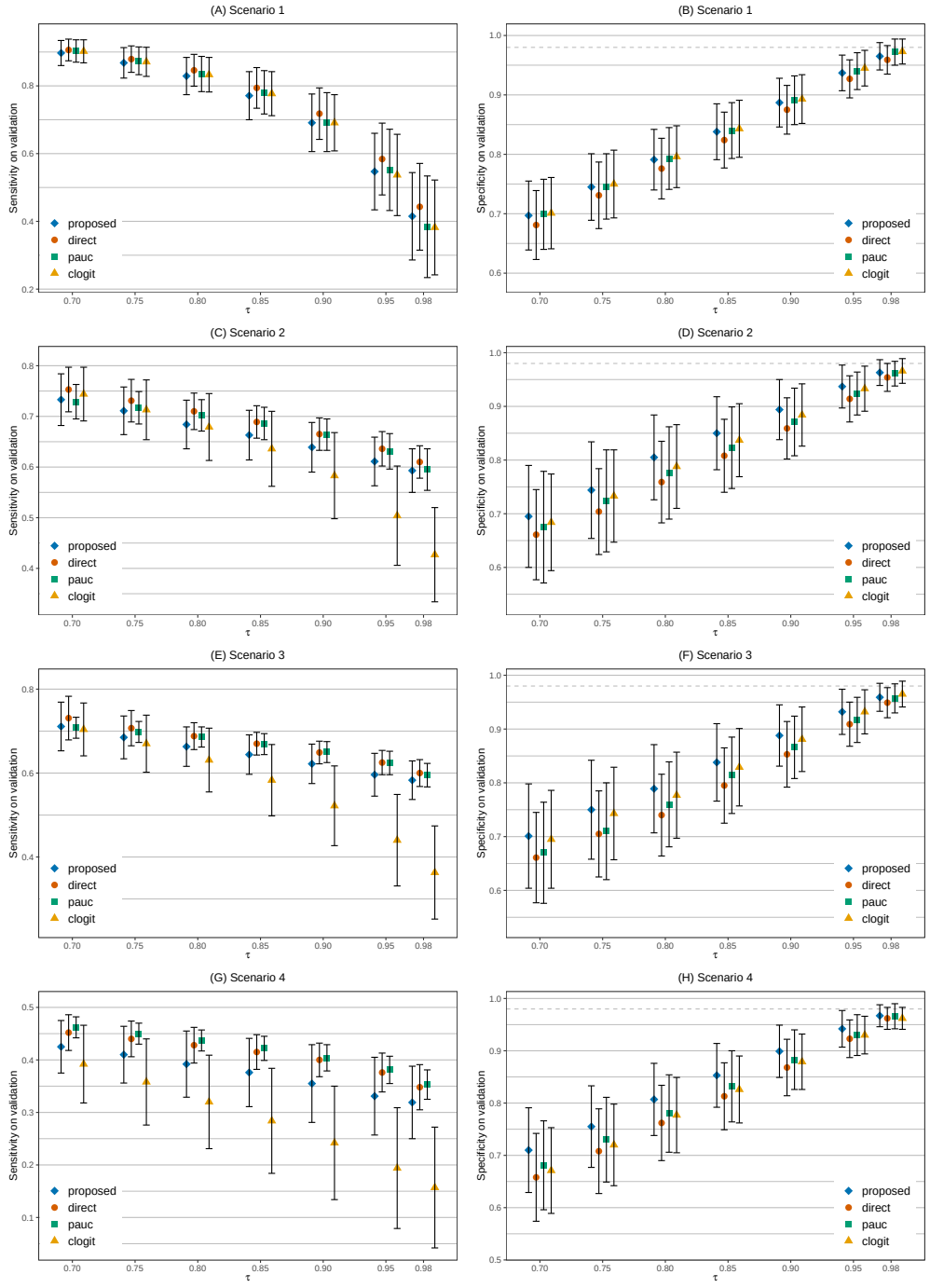

Figure 2.2: Visualization of simulation results on validation data when the sample size of the training data is $n_D = n_{\bar{D}} = 100$. Error bars are shifted slightly along the x-axis. τ: prespecified threshold of specificity. Gray dashed line: y-axis at 0.98.

Table 2.1: Summary statistics of estimated sensitivities on the training data. K: number of strata; τ: prespecified specificity; Clogit: conditional logistic regression; Mean: empirical mean sensitivity; ESE: empirical standard error; ASE: average of estimated standard errors; CP: 95% coverage probability.

Scenario	K	τ	Proposed				Direct		pAUC		Clogit	
			Mean	ESE	ASE	CP	Mean	ESE	Mean	ESE	Mean	ESE
1	50	.70	.911	.061	.067	.916	.925	.052	.901	.064	.898	.066
		.75	.885	.073	.079	.941	.900	.065	.873	.074	.867	.080
		.80	.848	.087	.094	.960	.867	.075	.838	.086	.829	.093
		.85	.794	.107	.113	.974	.820	.093	.789	.102	.773	.110
		.90	.717	.130	.132	.976	.752	.111	.709	.131	.692	.133
		.95	.597	.158	.140	.935	.636	.146	.582	.171	.555	.166
		.98	.476	.172	.130	.840	.506	.174	.435	.205	.413	.187
	100	.70	.914	.043	.046	.967	.924	.038	.908	.043	.906	.044
		.75	.885	.053	.056	.969	.898	.045	.880	.050	.876	.053
		.80	.846	.066	.068	.969	.865	.055	.841	.061	.837	.064
		.85	.791	.081	.083	.955	.815	.069	.788	.073	.783	.077
		.90	.711	.095	.103	.965	.739	.084	.701	.096	.696	.094
		.95	.565	.121	.123	.958	.603	.111	.558	.128	.540	.127
		.98	.432	.136	.120	.885	.461	.132	.389	.157	.384	.146
2	50	.70	.767	.091	.090	.977	.797	.077	.747	.078	.753	.095
		.75	.743	.092	.092	.981	.774	.078	.736	.079	.725	.105
		.80	.720	.086	.092	.974	.751	.077	.725	.078	.693	.108
		.85	.691	.091	.092	.980	.725	.075	.707	.078	.653	.113
		.90	.668	.090	.093	.968	.699	.075	.682	.078	.604	.122
		.95	.640	.093	.093	.977	.665	.076	.641	.081	.533	.130
		.98	.615	.092	.091	.957	.632	.080	.597	.087	.452	.137
	100	.70	.754	.067	.070	.986	.781	.057	.733	.057	.748	.067
		.75	.729	.065	.071	.978	.758	.055	.721	.057	.717	.072
		.80	.702	.066	.070	.976	.734	.056	.705	.058	.683	.082
		.85	.678	.067	.070	.968	.711	.054	.690	.058	.640	.092
		.90	.653	.067	.070	.970	.686	.054	.667	.057	.586	.100
		.95	.623	.068	.069	.957	.654	.054	.632	.059	.506	.112
		.98	.605	.063	.065	.957	.624	.055	.598	.064	.432	.107
3	50	.70	.746	.105	.108	.972	.770	.094	.709	.072	.714	.111
		.75	.714	.105	.109	.979	.741	.091	.701	.071	.680	.124
		.80	.685	.103	.109	.980	.713	.088	.690	.072	.643	.129
		.85	.659	.101	.108	.983	.688	.085	.675	.074	.596	.138
		.90	.633	.106	.107	.967	.660	.087	.657	.078	.540	.149
		.95	.609	.103	.107	.953	.629	.092	.630	.080	.468	.154
		.98	.588	.109	.108	.940	.598	.104	.601	.083	.385	.163
	100	.70	.726	.075	.081	.984	.749	.067	.706	.048	.705	.078
		.75	.697	.069	.078	.986	.723	.061	.697	.049	.671	.085
		.80	.674	.067	.077	.982	.702	.056	.685	.051	.634	.093
		.85	.655	.066	.077	.974	.682	.053	.671	.051	.587	.102
		.90	.632	.068	.078	.970	.661	.053	.653	.050	.528	.112
		.95	.607	.071	.078	.967	.637	.055	.627	.054	.448	.125
		.98	.594	.068	.077	.969	.612	.058	.598	.055	.373	.127
4	50	.70	.483	.097	.111	.984	.522	.082	.515	.054	.459	.117
		.75	.466	.097	.112	.983	.503	.080	.505	.055	.427	.121
		.80	.445	.099	.113	.985	.484	.080	.494	.054	.390	.123
		.85	.418	.105	.116	.987	.463	.081	.481	.053	.350	.128
		.90	.399	.109	.118	.975	.443	.084	.463	.054	.305	.138
		.95	.378	.112	.122	.976	.413	.092	.435	.055	.252	.143
		.98	.358	.109	.121	.981	.380	.099	.403	.059	.200	.146
	100	.70	.465	.065	.081	.985	.499	.052	.496	.036	.431	.087
		.75	.447	.071	.084	.979	.484	.052	.486	.036	.396	.095
		.80	.429	.076	.088	.974	.470	.050	.476	.037	.355	.104
		.85	.411	.079	.092	.974	.455	.051	.460	.040	.316	.113
		.90	.389	.087	.096	.957	.438	.049	.439	.042	.271	.123
		.95	.364	.088	.097	.967	.413	.053	.417	.042	.219	.130
		.98	.351	.082	.094	.966	.384	.059	.384	.045	.177	.130

Table 2.2: Study-specific results for the prostate cancer data. τ: prespecified threshold of specificity; Clogit: conditional logistic regression; Se: sensitivity; Sp: specificity.

τ	Proposed		Direct		pAUC		Clogit	
	Se	Sp	Se	Sp	Se	Sp	Se	Sp
.70	.912	.706	.912	.706	.897	.706	.897	.706
.75	.897	.750	.897	.750	.882	.750	.882	.750
.80	.867	.809	.868	.809	.838	.809	.838	.809
.85	.838	.853	.838	.853	.838	.853	.823	.853
.90	.809	.911	.809	.911	.794	.911	.809	.911
.95	.720	.956	.720	.956	.647	.956	.632	.956

outperformed the conditional logistic regression method. For example, when requiring 95% specificity, the proposed method could identify 72% of cases, while only 63% could be identified by the conditional logistic regression method (Table 2.2). The proposed method also showed advantages over the pAUC method in terms of optimizing sensitivity. The discrimination measures were almost identical for the proposed and the direct methods for this particular data example.

Since the sampling probabilities of the controls in the prostate cancer data are unavailable, the estimated cut-off, sensitivity, and specificity are study-specific, and as a result can not be generalized to the general population directly. To control the population-level specificity, one solution is to combine the current matched case-control data with the Census data. However, the population from the Census data differs systematically from the at-risk screening population, and thus is not an optimal source for this study. Instead, we can borrow information from the intervention arm of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial by comparing the age distribution of the controls in the prostate cancer data and the age distribution of the participants in the PLCO trial. The validation results by controlling the population-level specificity are shown in Table 2.3. Sampling probabilities were calculated using an approach similar to the propensity score method. We observed a consistent trend of performance between the different methods.

Table 2.3: Population-level results for the prostate cancer data. τ: prespecified threshold of specificity; Clogit: conditional logistic regression; Se: sensitivity; Sp: specificity.

τ	Proposed		Direct		pAUC		Clogit	
	Se	Sp	Se	Sp	Se	Sp	Se	Sp
.70	.897	.704	.911	.703	.911	.703	.897	.721
.75	.896	.765	.896	.759	.882	.757	.882	.757
.80	.882	.801	.882	.801	.882	.801	.838	.812
.85	.838	.851	.852	.855	.838	.866	.823	.874
.90	.809	.906	.809	.920	.809	.920	.809	.927
.95	.750	.956	.750	.956	.735	.956	.633	.966

Chapter 3

Methods and Results for Aim 2

3.1 Notation and Model

3.1.1 General Notations

Suppose that N subjects are followed prospectively in a study. Let Y_i be the binary outcome of interest, indicating whether subject i develops the disease of interest during the study. Cases and controls are respectively defined as subjects for whom $Y_i = 1$ and $Y_i = 0$. For notational simplicity, we use $\boldsymbol{X}_i$ to denote the p-dimensional vector of the covariates, including the routine variables and novel biomarkers. Note that under the two-phase sampling design, novel biomarkers are only ascertained at the second phase for the selected subcohort. Let n be the sample size of the subcohort. We consider two popular two-phase sampling designs, the case-cohort and the NCC designs, to introduce the selection probability.

In a case-cohort design, all cases are selected into the subcohort, and controls are randomly chosen at baseline from the full cohort. Accordingly, the probability of sampling the ith subject into the subcohort is

$$p_i = Y_i + (1 - Y_i)\alpha,$$

where α is a constant that represents the probability of being selected as a control. Under stratified case-cohort sampling, the full cohort is divided into L strata based on the baseline covariates. A subcohort is subsequently sampled from the full cohort using stratified sampling. Then $p_i = Y_i + (1 - Y_i)\alpha_l$ where α_l is the probability of being selected as a control for the lth stratum, and l denotes the stratum to which the ith subject belongs.

In the NCC design, cases that occur during the study are identified and for each case, a pre-specified number of controls are selected among those who have not developed the disease by the time the disease occurred for the case. Denote the risk set at time t as $R(t) = \{i : Z_i \geq t\}$, where $Z_i = \min(T_i, C_i)$, T_i is the event time and C_i is the follow-up time. Let the number of subjects in $R(t)$ be $n(t) = \sum_{i=1}^{N} I(Z_i \geq t)$. We define n_1 to be the number of cases and t_i, $i = 1, \ldots, n_1$ to be the failure times of the cases. At each failure time t_i, m controls are randomly selected without replacement from the risk set $R(t_i)$, excluding the case. Hence, the probability of sampling the ith subject into the subcohort is

$$p_i = Y_i + (1 - Y_i)\{1 - G(Z_i)\},$$

where $G(Z_i)$ denotes the probability that subject i has never been selected as a control up to the end of the study follow-up time Z_i. In stratified NCC sampling, at each case's failure time, controls are selected randomly without replacement among those who are in the risk set and matched to the case based on some covariates (Shiels et al., 2015). To accommodate stratified sampling, the $G(\cdot)$ in the sampling probability can be replaced by $G_K(Z_i, K_i)$, where K defines the covariate strata.

3.1.2 Regression Model

Our goal is to identify a scoring system $S(\boldsymbol{X})$, where a higher score is related to a higher risk of developing the given disease, and to estimate the absolute risk given the score. We assume that the probability of $Y_i = 1$ is related to the covariate vector through a

semiparametric regression model,

$$P(Y_i = 1|\boldsymbol{X}_i) = \pi\left\{S(\boldsymbol{X}_i;\boldsymbol{\beta})\right\}, \tag{3.1}$$

where $\pi(\cdot)$ denotes an unknown monotonic nondecreasing function, $S(\boldsymbol{X};\boldsymbol{\beta})$ is a pre-specified function of the subjects' characteristics, and $\boldsymbol{\beta}$ is an unknown vector of the same dimension as the covariate vector $\boldsymbol{X}$. A commonly used linear score summarizes the individual information as $S(\boldsymbol{X};\boldsymbol{\beta}) = \boldsymbol{\beta}^T\mathbf{X}$. In this case, the model is called a single index model (McCullagh and Nelder, 1989). Since π is left unspecified, we set the Euclidean norm of the coefficients $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{i=1}^{p}\beta_i^2}$ to be 1, to ensure identifiability. Such a semiparametric model covers a wide range of regression models, including the logistic and the probit regression models (Hristache et al., 2001; Ichimura, 1993). It offers substantial robustness and flexibility by relaxing the assumption regarding the form of the link function. Note that the monotonic assumption on $\pi(\cdot)$ is necessary to construct the scoring system; without this assumption, the scoring system cannot be used for risk stratification.

3.2 Likelihood and Estimation

We consider estimation procedures for both $\boldsymbol{\beta}$ and $\pi(\cdot)$ under model (3.1). After incorporating the unequal sampling probabilities, the weighted log-likelihood function of the data from the subcohort $(Y_i, \boldsymbol{X}_i, i = 1, \cdots, n)$ is

$$l(\boldsymbol{\beta}, \pi) = \sum_{i=1}^{n} \widehat{w}_i \Big(Y_i \log[\pi\left\{S(\boldsymbol{X}_i;\boldsymbol{\beta})\right\}] + (1 - Y_i)\log[1 - \pi\left\{S(\boldsymbol{X}_i;\boldsymbol{\beta})\right\}] \Big), \tag{3.2}$$

subject to the monotonic constraint for π, where $\widehat{w}_i = 1/\widehat{p}_i$ is the estimated version of $w_i = 1/p_i$. The weight w_i can be regarded as the contribution of the ith subject to the likelihood function (Støer and Samuelsen, 2013; Samuelsen, 1997). Extensive simulations confirmed that IPW estimators, which break the matching for two-phase designs, are

efficient (Kim, 2015; Delcoigne et al., 2017).

For the case-cohort design, the parameter α can be straightforwardly estimated by the empirical proportion n_0/N, where n_0 is the sample size of the random samples from the full cohort at baseline. For the NCC design, the unknown function of $G(\cdot)$ can be consistently estimated by a Kaplan-Meier-type estimator (Samuelsen, 1997), where

$$\widehat{G}(Z_i) = \prod_{j:Z_j<Z_i} \left\{1 - \frac{mY_j}{n(Z_j) - 1}\right\}.$$

Thus, the sampling probability for subject i can be estimated by

$$\widehat{p}_i = \begin{cases} Y_i + (1 - Y_i)\widehat{\alpha} & \text{in the case-cohort design,} \\ Y_i + (1 - Y_i)\{1 - \widehat{G}(Z_i)\} & \text{in the NCC design.} \end{cases}$$

In the presence of matching or stratification, we replace $\widehat{\alpha}$ with $\widehat{\alpha}_l = n_{0l}/N_l$ for the case-cohort studies, where n_{0l} and N_l denote the sample size of the random samples at baseline on the lth stratum and the full cohort on the lth stratum, respectively. Similarly, we replace $\widehat{G}(Z_i)$ with $\widehat{G}_K(Z_i, K_i)$ for the NCC studies, where

$$\widehat{G}_K(Z_i, K_i) = \prod_{j:Z_j<Z_i, K_j=K_i} \left\{1 - \frac{mY_j}{n_K(Z_j, K_j) - 1}\right\},$$

and $n_K(Z_j, K_j) = \sum_{i=1}^{N} I(Z_i \geq Z_j, K_i = K_j)$ is the size of the risk set at failure time Z_j after matching.

Note that directly maximizing the weighted likelihood in equation (3.2) with the monotonic constraint for π is computationally challenging. Considering that the likelihood in (3.2) belongs to the exponential family, we can apply PAVA to simplify the computational task (Best and Chakravarti, 1990; Qin et al., 2014). Following the theory of isotonic regression (Robertson et al., 1988), maximizing the likelihood in (3.2) under the monotonic constraint is equivalent to minimizing the following sum of squares,

denoted as $Q(\boldsymbol{\beta}, \pi)$, under the same constraint,

$$\underset{\pi\left\{S_{(1)}(\boldsymbol{\beta})\right\} \leq \cdots \leq \pi\left\{S_{(n)}(\boldsymbol{\beta})\right\}}{\arg\max} l(\boldsymbol{\beta}, \pi) = \underset{\pi\left\{S_{(1)}(\boldsymbol{\beta})\right\} \leq \cdots \leq \pi\left\{S_{(n)}(\boldsymbol{\beta})\right\}}{\arg\min} \sum_{i=1}^{n} \widehat{w}_i \left[Y_i - \pi\left\{S(\boldsymbol{X}_i; \boldsymbol{\beta})\right\}\right]^2, \quad (3.3)$$

where $S_{(1)}(\boldsymbol{\beta}), ..., S_{(n)}(\boldsymbol{\beta})$ denote the sorted $S(\boldsymbol{X}_i; \boldsymbol{\beta})$, $i = 1, 2, ..., n$ in ascending order. To minimize the right-hand side, we design a stable and efficient algorithm based on the method of profiling. For any given $\boldsymbol{\beta}$, we can apply PAVA to minimize the objective function $Q(\boldsymbol{\beta}, \pi)$, with respect to $\pi(\cdot)$ subject to the condition that if $S_{(1)}(\boldsymbol{\beta}) \leq S_{(2)}(\boldsymbol{\beta}) \leq \cdots \leq S_{(n)}(\boldsymbol{\beta})$, then $\pi\left\{S_{(1)}(\boldsymbol{\beta})\right\} \leq \pi\left\{S_{(2)}(\boldsymbol{\beta})\right\} \leq \cdots \leq \pi\left\{S_{(n)}(\boldsymbol{\beta})\right\}$. Denote the corresponding estimate as $\widehat{\pi}(\boldsymbol{\beta})$. We then minimize $Q\{\boldsymbol{\beta}, \widehat{\pi}(\boldsymbol{\beta})\}$ with respect to $\boldsymbol{\beta}$ and denote the minimizer as $\widehat{\boldsymbol{\xi}}_n = \{\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n(\cdot)\}$. Even though the estimation procedure involves the profiling idea, the computation is fast and can be easily implemented by existing programs. For example, the PAVA step can be accomplished using the R package *isotone* or *Iso*, and the minimization after profiling can be implemented using the R function *optim*.

3.3 Asymptotic Properties

We establish the asymptotic properties of $\widehat{\boldsymbol{\xi}}_n$, where true values of the parameters are denoted as $\boldsymbol{\xi}_0 = \{\boldsymbol{\beta}_0, \pi_0(\cdot)\}$. Technical challenges arise due to the infinite dimension of $\pi(\cdot)$, as well as the variability due to the estimated sampling probabilities. Under the mild regularity conditions given in the Appendix, we apply the empirical processes techniques (van der Vaart and Wellner, 1996; van der Vaart, 2002) to prove the consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$ and the uniform convergence of $\widehat{\pi}_n$. We further show that $\widehat{\pi}_n$ converges to π_0 at a convergence rate of $n^{-1/3}$ using the technique of bracketing entropy. Let $\|\cdot\|_2$ be the Euclidean distance and define the metric $\|\cdot\|$ by

$$\|\pi\| = \left\{\int \pi(u)^2 du\right\}^{1/2}.$$

The Hellinger distance h between the two density functions g_{ξ_1} and g_{ξ_2} is defined by

$$h^2(g_{\xi_1}, g_{\xi_2}) = \int \left\{ \sqrt{g_{\xi_1}(u)} - \sqrt{g_{\xi_2}(u)} \right\}^2 du.$$

We summarize the theoretical results in the following theorem and provide the detailed proof in the Appendix.

THEOREM 1. *Under the regularity conditions listed in the Appendix,* $\widehat{\boldsymbol{\beta}}_n$ *and* $\widehat{\pi}_n$ *are asymptotically consistent:*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \to 0, \textit{and } \|\widehat{\pi}_n - \pi_0\| \to 0,$$

in probability. Furthermore, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ *converges to a normal distribution, while* $\widehat{\pi}_n$ *has a convergence rate of* $n^{-1/3}$ *in the Hellinger distance.*

3.3.1 Variance Estimation

The explicit form of the asymptotic variance relies on many unknown quantities, preventing direct estimation of the variance. Alternatively, resampling techniques can be adopted for consistent variance estimation. Note that the standard bootstrap method cannot be applied to the full cohort, since the novel biomarkers are missing for subjects outside the subcohort. For the case-cohort study, we adopt the bootstrap method by Wacholder et al. (1989), in which the cases and controls are separately resampled from the subcohort with replacement such that the bootstrap data keep the same numbers of cases and controls.

For the NCC design, this modified bootstrap method cannot account for the complex dependence structure induced by repeatedly sampling without replacement from the risk sets. Thus, we apply the perturbation resampling method by Cai and Zheng (2013), where a more delicate resampling scheme was designed to account for the dependence. Specifically, let V_i be an indicator of whether subject i has ever been selected in the second phase, V_{0i} be a binary variable taking the value of 1 if subject i has ever been

sampled as a control, and V_{0ij} be a variable indicating if the jth subject has been chosen as a control for the ith subject. The resampling method perturbs these indicators with independent random numbers to mimic the Bernoulli sampling (with replacement). The sampling probabilities estimated from these perturbed indicators then recover the dependence structure in the finite population sampling (without replacement) and ensure that the corresponding perturbed IPW estimator has an appropriate limiting distribution. The formal justification of the resampling method can be found in Cai and Zheng (2013). We describe the perturbation procedure below.

(1) Generate non-negative random numbers $\{\mathcal{I}_{jk}, j = 1, \ldots, N; k = 1, \ldots, N\}$ independently from a known distribution with $E(\mathcal{I}_{jk}) = 1$ and $\text{var}(\mathcal{I}_{jk}) = 1$, such as the unit exponential distribution.

(2) Obtain perturbed weights $\widehat{w}_i^* = V_i^*/\widehat{p}_i^*$, where $V_i^* = Y_i\mathcal{I}_{ii} + (1 - Y_i)V_{0i}^*$, $\widehat{p}_i^* = Y_i + (1 - Y_i)\widehat{p}_{0i}^*$, $\widehat{p}_{0i}^* = 1 - \exp\{-\widehat{\Lambda}_{\text{marg}}^*(Z_i)\}$,

$$V_{0i}^* = 1 - \prod_{j:i\in R_j\setminus\{j\}} (1 - Y_j V_{0ji}\mathcal{I}_{ji}), \text{ and}$$

$$\widehat{\Lambda}_{\text{marg}}^*(t) = \sum_{j:Z_j\le t, Y_j=1} \frac{\sum_{k\in R_j\setminus\{j\}} V_{0jk}\mathcal{I}_{jk}}{n(Z_j) - 1}.$$

(3) Define $Q^*(\boldsymbol{\beta}, \pi)$ by replacing the $\widehat{w}_i$ in $Q(\boldsymbol{\beta}, \pi)$ with $\widehat{w}_i^*$ and apply the proposed algorithm. The resulting $(\widehat{\boldsymbol{\beta}}_n^*, \widehat{\pi}_n^*) = \arg\min Q^*(\boldsymbol{\beta}, \pi)$, under the monotone constraint, is a perturbed counterpart of $(\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)$.

Steps (1) - (3) can be repeated for B_0 times to obtain $\{(\widehat{\boldsymbol{\beta}}_n^*, \widehat{\pi}_n^*)_{(b)}, b = 1, ..., B_0\}$, where B_0 is the total number of perturbations. The variance of $\widehat{\boldsymbol{\beta}}_n$ can be estimated consistently by the empirical variance of its resampled counterparts, which would facilitate Wald-type confidence intervals and hypothesis testing.

3.4 Simulation Studies

We conducted simulation studies to examine the finite sample performance of the proposed method under two study designs: NCC study and case-cohort study.

3.4.1 Simulation Studies: NCC Study

Data Generation:

We generated X_1, X_2, and X_3 independently from $Beta(2, 2, 0, 2)$, $Bernoulli(0.5)$, and $Uniform(0, 2)$, respectively, such that the three covariates had similar variances. Here $Beta(2, 2, 0, 2)$ is a four-parameter beta distribution. The binary response Y_i was simulated following a Bernoulli distribution with a success probability of $\pi(\boldsymbol{\beta}^T\mathbf{X}) = \pi(\beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i})$, where the true regression coefficients were $(\beta_1, \beta_2, \beta_3)^T = (0.707, 0, 0.707)^T$. We generated T_i from a uniform distribution $Uniform(0, 5)$ for the cases and set the follow-up time to be 5.1 for all controls. Let $Z_i = \min(T_i, 5)$. For each case, three controls were sampled without replacement from its risk set, excluding the case.

We generated the event probabilities using different link functions. We considered four scenarios: Scenario 1 represented the case when the logistics regression model was the true model; Scenario 2 was used by other papers, such as Leitenstorfer and Tutz (2006); Scenario 3 was a combination of concave and convex curves, and Scenario 4 was for the sensitivity analysis when the monotonic assumption was violated. These curves were chosen to exemplify real-life relationships between the outcome probability and covariates, which may follow various shapes and curvatures.

Scenario 1: The true curve followed the logistic curve, $\pi(\boldsymbol{\beta}^T\mathbf{X}) = 1/[1+\exp\{-3.1(\boldsymbol{\beta}^T\mathbf{X} - 2.5)\}]$.

Scenario 2: The true curve had three curvatures, as shown in Figure 3.1, middle row, $\pi(\boldsymbol{\beta}^T\mathbf{X}) = 0.1/[1 + \exp\{-20(\boldsymbol{\beta}^T\mathbf{X} - 1.2)\}] + 0.2/[1 + \exp\{-16(\boldsymbol{\beta}^T\mathbf{X} - 2.3)\}]$.

Scenario 3: The true curve was concave at the left tail and convex at the right tail, as displayed in Figure 3.1, bottom row, $\pi(\boldsymbol{\beta}^T\mathbf{X}) = \tan\{1.65(\boldsymbol{\beta}^T\mathbf{X} - 0.56) - 1.4\}/95 + 0.06$.

Scenario 4: The true curve was not monotone, as shown in Figure 6.1 of the Appendix, $\pi(\boldsymbol{\beta}^T\mathbf{X}) = 0.12/[1+\exp\{-20(\boldsymbol{\beta}^T\mathbf{X}-1.2)\}]+0.24/[1+\exp\{-16(\boldsymbol{\beta}^T\mathbf{X}-2.3)\}]-0.05/[1+\exp\{-5(\boldsymbol{\beta}^T\mathbf{X}-2)\}]$.

We considered two sample sizes of 2500 and 5000 for the full cohort and two sample sizes of 700 and 1400 for the subcohort. To ensure that we would locate the global minima, we implemented 25 sets of random initial values around the estimates obtained from logistic regression and identified the estimators that achieved the lowest loss. The number of simulation replicates was 1000, and the resampling number B_0 was 499 (Dufour and Kiviet, 1998; Davidson and MacKinnon, 2000). For comparison, we implemented conditional logistic regression and IPW-based logistic regression.

Simulation Results:

The simulation results under the NCC design are summarized in Table 1 and Table 6.11 in the Appendix. The summary statistics are the empirical mean, empirical standard error (ESE) for $\boldsymbol{\beta}$ and the curve $\pi(\cdot)$ at the 25th, 50th, and 75th percentiles of the scores; average of the estimated standard error (ASE) based on resampling; and the empirical coverage probability of the 95% confidence interval. The estimate of β_1 was determined as a function of the estimates of β_2 and β_3 by the unit Euclidean norm constraint, and for completeness, all estimated coefficients were reported. The empirical means of the estimated π curves are plotted in Figure 3.1 and Figure 6.1 in the Appendix, in which the 5th and 95th percentiles of the scores $\boldsymbol{\beta}^T\mathbf{X}$ are chosen as the limits of the x-axis. In Scenario 1, all three methods performed well, since the true model followed logistic regression. As shown in the top row of Figure 3.1, both the proposed method and IPW-based logistic regression captured the underlying π function. The empirical biases of the estimates obtained by all three methods were negligible, and the coverage probabilities were reasonably close to the nominal level. As expected, the ESEs of the two logistic regression methods were smaller than those from the proposed method, since the two logistic regression methods utilized the information on the underlying link function (logistic curve) while the proposed method did not.

Table 3.1: Simulation results under the nested case-control design (ESE is the empirical standard error, ASE is the average of estimated standard error, and CP is the empirical coverage probability of the 95% confidence interval).

					Proposed method					IPW-based logistic regression					Conditional logistic regression				
Scenario	N	n	PARA	True	Mean	Bias	ESE	ASE	CP	Mean	Bias	ESE	ASE	CP	Mean	Bias	ESE	ASE	CP
Scenario 1	2500	700	β_1	.707	.683	-.024	.101	.104	.951	.705	-.002	.048	.048	.937	.704	-.003	.068	.068	.938
			β_2	.000	-.004	-.004	.111	.103	.928	.000	.000	.065	.065	.953	-.002	-.002	.096	.093	.936
			β_3	.707	.709	.001	.095	.098	.950	.703	-.004	.049	.047	.940	.702	-.005	.069	.067	.927
			$\pi(.99)$	.009	.010	.000	.006	-	-	.009	.000	.002	-	-	-	-	-	-	-
			$\pi(1.41)$	.034	.036	.002	.013	-	-	.034	.000	.006	-	-	-	-	-	-	-
			$\pi(1.84)$	.117	.124	.007	.037	-	-	.120	.003	.015	-	-	-	-	-	-	-
	5000	1400	β_1	.707	.700	-.007	.072	.076	.958	.703	-.004	.036	.034	.933	.704	-.003	.049	.048	.945
			β_2	.000	.000	.000	.085	.081	.942	.000	.000	.047	.046	.951	-.002	-.002	.069	.066	.933
			β_3	.707	.702	-.005	.069	.075	.956	.707	.000	.036	.034	.928	.706	-.001	.049	.047	.948
			$\pi(.99)$	.009	.009	.000	.004	-	-	.009	.000	.002	-	-	-	-	-	-	-
			$\pi(1.41)$	.034	.035	.001	.010	-	-	.034	.000	.004	-	-	-	-	-	-	-
			$\pi(1.84)$	.117	.119	.002	.027	-	-	.119	.001	.011	-	-	-	-	-	-	-
Scenario 2	2500	700	β_1	.707	.674	-.034	.126	.136	.932	.695	-.012	.078	.078	.943	.689	-.018	.088	.083	.922
			β_2	.000	.011	.011	.130	.126	.921	.000	.000	.106	.101	.943	.009	.009	.123	.118	.928
			β_3	.707	.708	.001	.111	.119	.947	.703	-.005	.076	.074	.933	.709	.002	.085	.080	.912
			$\pi(.99)$	.001	.011	.010	.018	-	-	.030	.028	.005	-	-	-	-	-	-	-
			$\pi(1.41)$	.107	.097	-.009	.017	-	-	.061	-.046	.008	-	-	-	-	-	-	-
			$\pi(1.84)$	.108	.116	.007	.017	-	-	.121	.013	.015	-	-	-	-	-	-	-
	5000	1400	β_1	.707	.694	-.013	.073	.089	.970	.695	-.012	.059	.056	.934	.697	-.010	.064	.058	.927
			β_2	.000	.000	.000	.094	.095	.942	.001	.001	.074	.073	.947	.000	.000	.087	.084	.933
			β_3	.707	.707	.000	.069	.082	.965	.710	.003	.057	.053	.921	.709	.002	.062	.057	.923
			$\pi(.99)$	.001	.005	.004	.011	-	-	.029	.028	.004	-	-	-	-	-	-	-
			$\pi(1.41)$	.107	.100	-.007	.012	-	-	.060	-.046	.005	-	-	-	-	-	-	-
			$\pi(1.84)$	.108	.112	.004	.010	-	-	.119	.011	.010	-	-	-	-	-	-	-
Scenario 3	2500	700	β_1	.707	.705	-.003	.072	.073	.921	.773	.066	.059	.058	.704	.764	.057	.098	.092	.803
			β_2	.000	.003	.003	.056	.058	.934	.003	.003	.092	.092	.946	-.002	-.002	.152	.141	.910
			β_3	.707	.700	-.007	.073	.074	.922	.621	-.086	.072	.070	.765	.624	-.083	.116	.108	.852
			$\pi(.99)$	.052	.049	-.003	.009	-	-	.028	-.024	.006	-	-	-	-	-	-	-
			$\pi(1.41)$	.061	.061	.000	.008	-	-	.061	.000	.008	-	-	-	-	-	-	-
			$\pi(1.84)$	.070	.074	.004	.012	-	-	.128	.058	.015	-	-	-	-	-	-	-
	5000	1400	β_1	.707	.706	-.001	.049	.051	.939	.775	.068	.042	.041	.588	.772	.065	.070	.065	.758
			β_2	.000	.000	.000	.033	.038	.955	.001	.001	.068	.066	.939	-.001	-.001	.108	.102	.919
			β_3	.707	.704	-.003	.049	.051	.940	.625	-.082	.052	.050	.636	.625	-.082	.084	.078	.801
			$\pi(.99)$	.052	.050	-.002	.007	-	-	.028	-.024	.004	-	-	-	-	-	-	-
			$\pi(1.41)$	.061	.061	.000	.006	-	-	.061	-.001	.006	-	-	-	-	-	-	-
			$\pi(1.84)$	.070	.072	.002	.009	-	-	.126	.056	.010	-	-	-	-	-	-	-

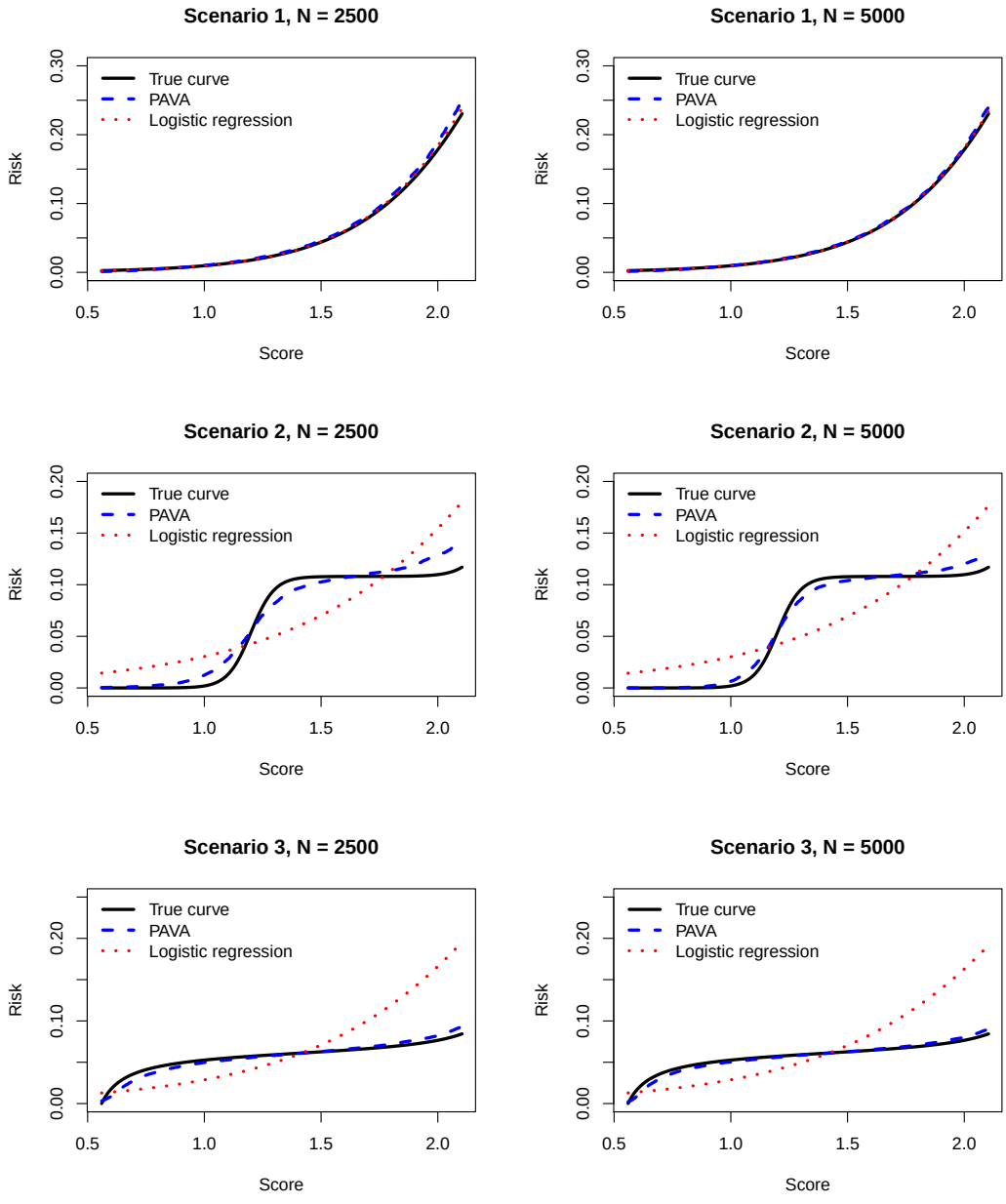

Figure 3.1: Estimated risk functions under the NCC design.

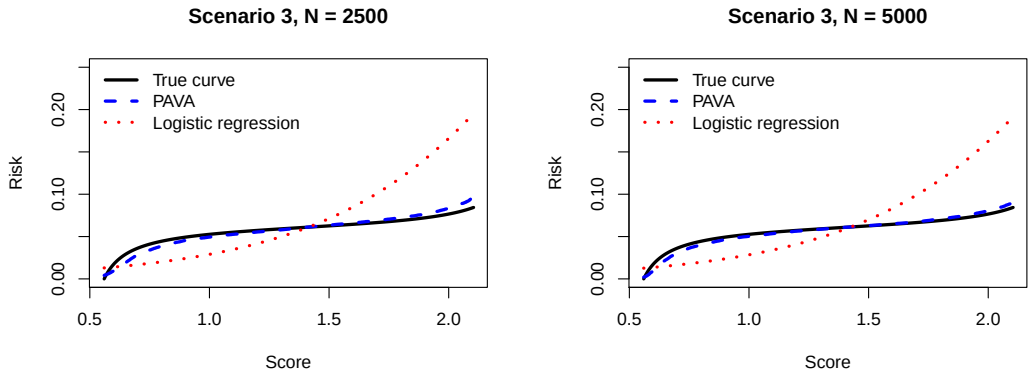

Figure 3.2: Estimated risk functions under the case-cohort design.

In Scenario 2, the underlying π curve exhibited more curvature than that assumed by the logistic regression model (Figure 3.1, middle row). The proposed method still performed well in terms of estimating the regression coefficients $\boldsymbol{\beta}$ and the link function π. Using IPW-based logistic regression, the estimates of $\boldsymbol{\beta}$ were close to the true values, but the estimated curve deviated from the true function. As a result, the risk probability could not be reliably estimated, due to the substantial bias in the estimated π. For example, for a subject with a median risk score, IPW-based logistic regression severely underestimated the risk probability by 50%, which would give misleading information to the subject. Under this setting, although the underlying true model was not the logistic regression model, conditional logistic regression was quite robust and performed similarly to the proposed method. However, the conditional regression model could not estimate the absolute risk, which was important in our setting.

In Scenario 3, the proposed method remained robust and accurate in terms of both the regression coefficients and the π function. In comparison, IPW-based logistic regression could not estimate the underlying link well. For example, the risk estimation around the 75th quantile of the risk score overestimated the true risk by 80% (Figure 3.1, bottom row). Both logistic regression methods had biased estimates for the regression coefficients due to the mis-specification of the underlying link function. For IPW-based logistic regression, the biases were larger than the corresponding ESEs for some regression

coefficients, leading to poor coverage probabilities as low as 59%.

In Scenario 4, the true curve was not monotonic; thus the monotonic assumption was not satisfied for the logistic methods, as well as the proposed method. Nevertheless, our proposed method still outperformed the commonly used logistic regression methods in estimating the link function (Figure 6.1 in the Appendix).

3.4.2 Simulation Studies: Case-cohort Study

Data generation:

We used the same sample sizes (2500 and 5000) and data generation scheme as specified in Scenario 3 of Section 3.4.1 to simulate the covariates and outcomes for the full cohort. For the subcohort, we selected 550 or 1100 controls from the full cohort at baseline. We compared the performance of our proposed design with that of IPW-based logistic regression.

Simulation Results:

Simulation results under the case-cohort design are summarized in Table 3.2. Similar to our previous findings under the NCC design, the estimates of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\pi}_n$ by the proposed method were close to the true values, and the empirical coverage probabilities of the confidence intervals of $\widehat{\boldsymbol{\beta}}_n$ were close to the nominal level. In contrast, IPW-based logistic regression overestimated β_1 by 10% and underestimated β_3 by 13%, which resulted in coverage probabilities as low as 53%. As shown in Figure 3.2, the proposed method fitted the true curve well with small biases, while logistic regression could not capture the true curve. The substantial differences between the estimated and true curves indicated that the use of logistic regression may result in misleading risk assessments when the model assumptions were not valid.

In summary, with the link function unspecified, the proposed method robustly estimated both the risk score using regression coefficients and the link function under various scenarios, given data from two-phase sampling designs. By comparison, logistic regres-

Table 3.2: Simulation results under the case-cohort design (ESE is the empirical standard error, ASE is the average of estimated standard error, and CP is the empirical coverage probability of the 95% confidence interval).

					Proposed method					IPW-based logistic regression				
Scenario	N	n	PARA	True	Mean	Bias	ESE	ASE	CP	Mean	Bias	ESE	ASE	CP
Scenario 3	2500	700	β_1	.707	.704	-.003	.076	.084	r.952	.770	.063	.058	.060	.749
			β_2	.000	.000	.000	.058	.066	.958	-.003	-.003	.094	.095	.951
			β_3	.707	.700	-.008	.075	.085	.957	.625	-.082	.071	.072	.792
			$\pi(.99)$	.052	.049	-.003	.010	-	-	.029	-.024	.006	-	-
			$\pi(1.41)$	.061	.061	.000	.008	-	-	.062	.001	.008	-	-
			$\pi(1.84)$	.070	.074	.004	.012	-	-	.129	.059	.015	-	-
	5000	1400	β_1	.707	.706	-.002	.049	.057	.966	.781	.074	.040	.041	.532
			β_2	.000	-.001	-.001	.033	.041	.973	.002	.002	.066	.067	.947
			β_3	.707	.705	-.003	.048	.057	.972	.618	-.089	.050	.051	.597
			$\pi(.99)$	.052	.050	-.002	.007	-	-	.028	-.024	.004	-	-
			$\pi(1.41)$	.061	.061	.000	.006	-	-	.061	.000	.006	-	-
			$\pi(1.84)$	.070	.073	.003	.009	-	-	.127	.057	.010	-	-

sion approaches generated severely biased estimates for both the risk score and the link function, even under the settings with large sample sizes. These results signified the advantage of relaxing the model assumptions by using the proposed method.

3.5 Application

We return to the Rotterdam breast cancer data set introduced in Section 1.2.3. To create an NCC data set, we first defined subjects who died in two years' of follow up as cases. Then each case was matched to two controls who were alive at the case's event time, and cases and controls are matched based on variables including age group ($\leq$40 years, 40-60 years, >60 years), tumor size ($\leq$20 mm, 21–50 mm, >50 mm), tumor grade ($\leq$ 2, 3), hormonal therapy, and chemotherapy. A total of 1340 subjects were included in the analysis. We then constructed risk scores by combining number of positive lymph nodes (NODES), progesterone receptor (PGR), and estrogen receptor (ER). The estimated regression coefficients, standard errors, and p-values using the proposed method and the conditional logistic regression method are reported in Table 3.3. The two methods resulted in different conclusions. The effect of ER was significant with a relatively high impact on the risk score in the model fitted by the proposed method, while this effect was insignificant with a relatively small impact in the model by the conditional logistic regression method. Differences in the coefficients and p-values of NODES were also seen between the results from the two methods. In Figure 3.3, the risk curve estimated by the proposed method was above the curve by the logistic regression method. This indicated the true risk curve may differ from the logistic curve, and as a result, the proposed method, which imposed less constraint on the shape of the link function, may be preferred for this data set.

As one feature of the proposed method, the derived risk scores can be used for risk stratification. Using the estimated median risk score as the cutoff, we divided the subjects equally into two groups (high-risk and low-risk groups). For the proposed method, 64.6% of the subjects in the high-risk group died in two years, and 35.4% of the subjects in the

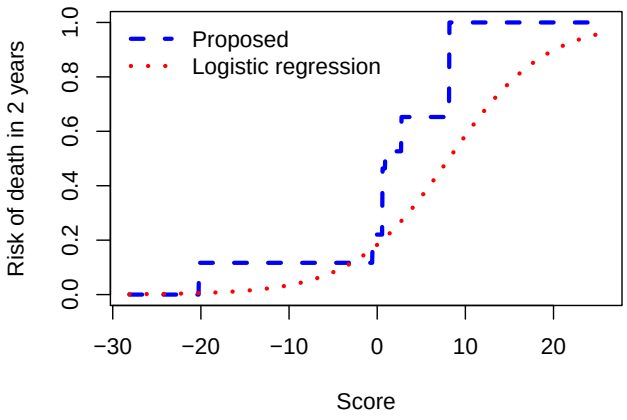

Figure 3.3: Estimated risk of death in two years in the Rotterdam breast cancer population.

low-risk group died in two years. In comparison, for the conditional logistic regression method, 63.3% of the high-risk and 36.7 % of the low-risk died.

Table 3.3: Estimated regression coefficients, standard errors, and p-values using the proposed method and the conditional logistic regression model in the Rotterdam breast cancer data.

	Proposed			Conditional logistic regression		
	Coefficient	SE	P-value	Coefficient	SE	P-value
Nodes	0.317	0.200	0.11	0.709	0.137	< 0.001
PGR/100	-0.511	0.255	0.045	-0.586	0.223	0.04
ER/100	-0.799	0.266	0.003	-0.393	0.271	0.17

Chapter 4

Methods and Results for Aim 3

4.1 Method

4.1.1 Notations

Suppose there are n patients followed prospectively in a study. For the ith subject, let T_i and C_i be the event time and censoring time, respectively. We observe $Y_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = I(T_i \leq C_i)$. At a given time point t, define the risk set as $\mathcal{R}(t) = \{j : Y_j > t\}$. Let $t_{i1} < t_{i2} < \ldots < t_{in_i} < Y_i$ be the measurement times, where n_i is the total number of measurement times of the ith subject, and the measurement times may be irregular and are not the same for different subjects. Denote by $X_i(t_{ij})$, a $p \times 1$ vector of risk factors collected on the ith subject at time $t_{ij}, j = 1, \ldots, n_i$. The notation $X_i(t_{ij})$ includes both baseline and time-dependent variables. Let $\widetilde{X}_i(t_{ij})$ be the summary information up to time t_{ij}, such as average values, changes or rates of changes of risk factors. We assume that the measurement times are independent of the longitudinal risk factors and the event time.

4.1.2 Estimation

At any time t, we aim to develop a dynamic prediction score, denoted $S\{\widetilde{\boldsymbol{X}}(t); \boldsymbol{\beta}(t)\}$, to characterize the risk of the event of interest using all collected information. Ideally, given

the collected information from subject i at time t_{ij}, we can update the risk prediction by calculating $S\{\widetilde{\boldsymbol{X}}_i(t_{ij}); \boldsymbol{\beta}(t_{ij})\}$. A commonly used linear model summarizes the patient information by a linear form, $S\{\widetilde{\boldsymbol{X}}_i(t); \boldsymbol{\beta}(t)\} = \boldsymbol{\beta}(t)'\widetilde{\boldsymbol{X}}_i(t)$, which will be used for the illustration. The unknown function $\boldsymbol{\beta}(t)$ describes the time-varying effects of risk factors. We can impose smoothing constraints, such as fractional polynomials and splines, for the unknown parameter function $\boldsymbol{\beta}(t)$. As an illustration, for each scalar risk factor $\widetilde{X}_i^k(t)(k = 1, \ldots, p)$, we assume

$$\boldsymbol{\beta}(t)\widetilde{X}_i^k(t) = \{\beta_0^k + \beta_1^k ln(t+1) + \beta_2^k\sqrt{t} + \beta_3^k/\sqrt{t+1} + \beta_4^k t + \beta_5^k/(t+1)\}\widetilde{X}_i^k(t).$$

Assume that we can observe $\widetilde{\boldsymbol{X}}(t)$ at $\{t_{ij}, j = 1, \ldots, n_i; i = 1, \ldots, n\}$, we then construct the composite-likelihood function as

$$\ell_n(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left(\frac{\sum_{k=1}^{n} I(Y_k \geq Y_i) I[S\{\widetilde{\boldsymbol{X}}_i(t_{ij}); \boldsymbol{\beta}(t_{ij})\} - S\{\widetilde{\boldsymbol{X}}_k(t_{ij}); \boldsymbol{\beta}(t_{ij})\} \geq 0]}{\sum_{k=1}^{n} I(Y_k \geq Y_i)}\right)^{\delta_i}. \tag{4.1}$$

This composite-likelihood reflects the concept of concordance, which are also used in Payne et al. (2016) and Shen et al. (2018) to form objective functions. For identifiability purposes, we set $\beta_0 = 1$ in model (4.1). Directly maximizing the above composite-likelihood is computationally challenging due to the indicator function, so we propose an approximation to $\ell_n(\boldsymbol{\beta})$ by applying the smoothing kernel method:

$$\ell_n^s(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left(\frac{\sum_{k=1}^{n} I(Y_k \geq Y_i) \int_{-\infty}^{S\{\widetilde{\boldsymbol{X}}_i(t_{ij}); \boldsymbol{\beta}(t_{ij})\} - S\{\widetilde{\boldsymbol{X}}_k(t_{ij}); \boldsymbol{\beta}(t_{ij})\}} K_{h_1}(u)du}{\sum_{k=1}^{n} I(Y_k \geq Y_i)}\right)^{\delta_i}, \tag{4.2}$$

where $K_{h_1}(u) = \frac{1}{h_1}K(\frac{u}{h_1})$, and $K(\cdot)$ is a symmetric kernel function with a pre-specified bandwidth h_1. Note that the measurement times could be irregular and different for different subjects. We are faced with the challenge that the time-varying risk factors are

not observed at all measurement times. So, if the risk factor $\widetilde{X}_i^k$ is not available at time s, we borrow the observed information around s to approximate the missing value by using a kernel weight: $\widetilde{X}_i^k(s)^* = \sum_{j=1}^{n_i} K_{h_2}(t_{ij} - s)\widetilde{X}_{ij}^k / \sum_{j=1}^{n_i} K_{h_2}(t_{ij} - s)$, where $K_{h_2}(.)$ is similarly defined as K_{h_1}. Accordingly, we can revise the likelihood function in (4.2) by plugging in $\widetilde{X}_i^k(s)^*$:

$$\ell_n^s(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left(\frac{\sum_{k=1}^{n} I(Y_k \geq Y_i)\int_{-\infty}^{S\{\widetilde{\boldsymbol{X}}_i(t_{ij});\boldsymbol{\beta}(t_{ij})\}-S\{\widetilde{\boldsymbol{X}}_k^*(t_{ij});\boldsymbol{\beta}(t_{ij})\}} K_{h_1}(u)du}{\sum_{k=1}^{n} I(Y_k \geq Y_i)}\right)^{\delta_i}. \quad (4.3)$$

Theoretically, any smooth and symmetric probability density function can be adopted as the kernel function $K(\cdot)$. The Gaussian kernel is a popular choice in practice and thus is implemented for illustration. The bandwidth can be selected either via cross validation or the recommendation made by Jones (1990). Please see Section 4.2 for more details. The regression coefficients in the model can be obtained by maximizing $\ell_n^s(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, and we denote the maximizer as $\widehat{\boldsymbol{\beta}}$.

4.1.3 Prediction Discrimination

After the model fitting, we then need to find out to what extent can the derived dynamic score discriminate between patients who will experience the event in the next w time period and those who will not given these patients have survived up to time s. To this end, one important pair of metrics for summarizing the discrimination capacity of the prediction models is sensitivity (Se) and specificity (Sp). To adapt the metrics to the longitudinal setting and utilize the risk scores that may include multiple time-independent and time-varying covariates, we define

$$\mathrm{Se}_{s,w}(c) = P(S\{\widetilde{\boldsymbol{X}}_i(s);\widehat{\boldsymbol{\beta}}(s)\} > c \mid \ Y_i \geq s, T_i \leq s + w) \quad \text{and} \quad (4.4)$$

$$\mathrm{Sp}_{s,w}(c) = P(S\{\widetilde{\boldsymbol{X}}_i(s);\widehat{\boldsymbol{\beta}}(s)\} \leq c \mid \ Y_i \geq s, T_i > s + w), \quad (4.5)$$

where c is a pre-specified threshold.

We may estimate $\mathrm{Se}_{s,w}(c)$ and $\mathrm{Sp}_{s,w}(c)$ empirically by

$$\widehat{\mathrm{Se}}_{s,w}(c) = \frac{\sum_i I(S\{\widetilde{\boldsymbol{X}}_i(s); \widehat{\boldsymbol{\beta}}(s)\} > c) I(Y_i \geq s) I(T_i \leq s+w)}{\sum_i I(Y_i \geq s) I(T_i \leq s+w)} \quad \text{and} \tag{4.6}$$

$$\widehat{\mathrm{Sp}}_{s,w}(c) = \frac{\sum_i I(S\{\widetilde{\boldsymbol{X}}_i(s); \widehat{\boldsymbol{\beta}}(s)\} \leq c) I(Y_i \geq s) I(T_i > s+w)}{\sum_i I(Y_i \geq s) I(T_i > s+w)}, \tag{4.7}$$

However, in most real cases, some subjects are censored, and the event times are unobserved for those subjects. To adjust for the censoring and obtain unbiased estimates for evaluation, here we propose the following estimators which are adapted from those in Uno et al. (2007):

$$\widehat{\mathrm{Se}}_{s,w}(c) = \frac{\sum_i I(S\{\widetilde{\boldsymbol{X}}_i(s); \widehat{\boldsymbol{\beta}}(s)\} > c) I(s \leq Y_i \leq s+w) \delta_i / \widehat{G}(Y_i)}{\sum_i I(s \leq Y_i \leq s+w) \delta_i / \widehat{G}(Y_i)} \quad \text{and} \tag{4.8}$$

$$\widehat{\mathrm{Sp}}_{s,w}(c) = \frac{\sum_i I(S\{\widetilde{\boldsymbol{X}}_i(s); \widehat{\boldsymbol{\beta}}(s)\} \leq c) I(Y_i > s+w)}{\sum_i I(Y_i > s+w)}, \tag{4.9}$$

where $\widehat{G}(t) = Pr(C_i > t)$ is the Kaplan-Meier-type estimator for the distribution of the censoring time. Subsequently, the area under the Receiver Operating Characteristic (AUC) can be calculated: $\widehat{\mathrm{AUC}}_{s,w} = \int \widehat{\mathrm{Se}}_{s,w}(c) d\widehat{\mathrm{Sp}}_{s,w}(c)$.

4.2 Simulation

We conducted extensive simulation studies and compared the performance of the proposed method to that of the four existing methods that are capable of handling longitudinal measurements as well as derive dynamic risk scores, namely, the Cox model with time-dependent covariates (COX), the Cox model with both time-dependent covariates and time-dependent coefficients (VCOX) (Therneau and Grambsch, 2000), the partly conditional model without time-dependent coefficients (PC), and the partly conditional model with time-dependent coefficients (VPC) (Zheng and Heagerty, 2005).

4.2.1 Data generation

The data were generated according to the following three scenarios:

Scenario 1. We generated two longitudinal biomarkers $\widetilde{X}_i^1(t) = \gamma_1 + \gamma_2 t$, and $\widetilde{X}_i^2(t) = \gamma_3 + \gamma_4 t$, where $\gamma_1 \sim Unif(0,2), \gamma_2 \sim 4 \times Beta(2,5), \gamma_3 \sim Unif(0,2)$, and $\gamma_4 \sim 3 \times Beta(1,3)$. Then we generated the failure time T_i from the hazard function of the form: $\lambda_i(t) = 0.02I(t \leq 1)\exp\{2\widetilde{X}_i^1(t) + 0.5\widetilde{X}_i^2(t)\} + 0.02I(t > 1)\exp\{0.5\widetilde{X}_i^2(t)\}$, so that the coefficient of $\widetilde{X}_i^1$ would change with t.

Scenario 2. We generated the biomarkers in the same way as that in Scenario 1. Then we generated the failure time T_i from the hazard function $\lambda_i(t) = 2(1-Z_i)\text{logit}^{-1}\{30\widetilde{X}_i^1(t) - 5\widetilde{X}_i^2(t) - 1\} + 2Z_i\text{logit}^{-1}\{0.1\widetilde{X}_i^1(t) + 10\widetilde{X}_i^2(t) - 15\}$, which was a mixture of two inverse logit functions. Here, $Z_i = I\{\widetilde{X}_i^1(t=0) > 1\}$, and $\text{logit}(p) = \log\{p/(1-p)\}$.

Scenario 3. The data generation for the biomarkers remained the same as that in the previous scenarios, but the hazard function was changed to $\lambda_i(t) = 2\text{logit}^{-1}\{0.15\widetilde{X}_i^1(t) + 15\widetilde{X}_i^2(t) - 15\}$.

For each scenario, $C_i \sim Unif(4,6)$. We considered both regular visits and irregular visits in all three scenarios. For regular visits, longitudinal biomarkers were recorded at pre-determined scheduled times, such as $t = 0, 0.6, 1.2, \ldots, 6$, as long as the subject was at risk. For irregular visits, each observation time was randomly generated from a uniform distribution with the support to be the scheduled time ± 0.3, and not earlier than the last observation time. This mimicked the situation in which the study subjects may visit the clinics slightly before or after the scheduled time. A summary of the number of measurements in each scenario is presented in Table S3.

We set the bandwidth $h_1 = n^{-1/3}$, which was the optimal bandwidth for density estimation problems recommended by Jones (1990). In fact, we conducted a sensitivity analysis using several different bandwidths for h_1 and found that the choice of bandwidth did not affect estimation as long as the bandwidth was on the same scale of the optimal value. To choose bandwidth h_2 for borrowing information in the presence of missing data,

we employed a grid search with an independent data set under each scenario, and the bandwidth resulting the highest value of averaged Uno's C-statistic across s was selected (Uno et al., 2011).

Moreover, for fair comparison, the same bases in fractional polynomials were used in the proposed method, the VCOX method, and the VPC method. The implementation of the existing methods were accomplished by existing programs and sample codes. Specifically, we employed the *survival* package in R for COX, and specified the *tt* option for VCOX. We used the *partlyconditional* package in R for PC (Therneau, 2015; Maziarz et al., 2018). Since the *partlyconditional* package did not allow for varying coefficients, we implemented the method in Zheng and Heagerty (2005) for VPC by adapting the sample codes in Maziarz et al. (2017).

In each simulation replicate, we first obtained the coefficient estimates for the biomarkers on the training data set, and derived the dynamic score term for each method. Then we applied the dynamic score term to an independent data set that had the same sample size as the training data set. Next, we calculated $\text{AUC}_{s,w}$ discussed in Section 4.1.3 on multiple combinations of s and w, where $s = 0, 1.2, 2,$ and 4, and $w = 0.6$ and 1.2. We also evaluated the difference in the restricted mean survival time (RMST) (Tian et al., 2014). Specifically, subjects at-risk at s were divided into two groups (high-risk and low-risk) of equal size with respect to the median survival time. Then the difference in RMST between the high-risk and the low-risk groups were reported. The truncation time point for RMST was set to be the 95% quantile of the observed survival time. Sample size $n = 200$ and 400, and the number of simulation replicates were 1,000.

4.2.2 Results

Tables 4.1 and 6.12 in the Appendix present the simulation results for regular measurement. When the true underlying model was a Cox-like model and the coefficient changed with time (Scenario 1), the proposed method and the VPC had the best discrimination performance. On the other hand, although the COX and the PC correctly specified the

model form, they wrongly assumed the coefficients were fixed over time, and as a result their discrimination ability was compromised.

When the true model deviated from a Cox-like model (Scenarios 2&3), the proposed method outperformed the competing methods. In particular, the proposed method had the highest $\text{AUC}_{s,w}$ and difference in RMST, as well as the smallest standard deviation. Interestingly, the VPC was quite robust at the baseline and when the landmark time was close to the baseline, yet it did not perform well when s was large. For example, in Scenario 2 at $s = 2.4$ and $w = 0.6$, its value for $\text{AUC}_{s,w}$ was less than that of the proposed method by 0.077 .

The PC performed better than its varying-coefficient counterpart (VPC) when $s = 1.2$ and 2.4 and the true coefficients were fixed (Scenarios 2&3). This was expected because the PC used the correct trend of the coefficients. We also found that in some cases, the performance of VCOX was not as good as that of its fixed-coefficient counterpart (COX), even when the true biomarker effect changed with time. We checked the regression coefficients of VCOX, and we discovered that although empirical biases of VCOX were close to zero, the variations were unusually large. Moreover, the Cox model maximizes the partial likelihood, which is a global criterion and not necessarily translates into classification ability.

Tables 4.2 and 6.13 in the Appendix present the simulation results for irregular measurement. Here we observed similar patterns as those in the regular measurement situation, and even the variations were similar in both cases. These suggested that the data borrowing in our proposed method worked well, and the irregular measurements had minimum impact on our proposed method.

4.3 Application

We illustrate the proposed method on the AIDS data set discussed in Section 1.3.3. In this data application, we aim to see if the dynamic scoring system including longitudinal CD4 cell count information and gender can discriminate between patients who had high risk of

Table 4.1: Simulation results for Scenario 1-3 when measurements were regular: estimated area under the ROC curve and difference in restricted mean survival time between low-risk and high-risk groups obtained from the proposed method, the Cox model with time-varying covariates (COX), the Cox model with time-varying covariates and coefficients (VCOX), the partly conditional cox model (PC), and the partly conditional cox model with time-varying coefficients (VPC) on validation data. Number of replicates was $1,000$, $n = 200$.

Scenario		s	w	Proposed		COX		VCOX		PC		VPC	
				Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	$\text{AUC}_{s,w}$	0	0.6	.823	.031	.794	.039	.739	.145	.737	.054	.823	.030
		0	1.2	.819	.030	.790	.037	.736	.143	.733	.052	.819	.029
		1.2	0.6	.642	.201	.599	.201	.559	.200	.629	.200	.646	.200
		1.2	1.2	.671	.135	.625	.133	.582	.138	.657	.134	.675	.134
		2.4	0.6	.721	.153	.657	.162	.700	.167	.699	.158	.723	.154
		2.4	1.2	.751	.103	.680	.112	.726	.121	.726	.107	.754	.103
	RMST_{diff}	0		1.985	.276	1.796	.322	1.449	.922	1.388	.386	1.991	.276
		1.2		.735	.295	.534	.300	.339	.348	.679	.289	.755	.286
		2.4		.629	.221	.452	.233	.565	.264	.567	.221	.637	.218
2	$\text{AUC}_{s,w}$	0	0.6	.729	.037	.674	.047	.558	.134	.647	.058	.724	.038
		0	1.2	.791	.041	.737	.052	.571	.172	.707	.064	.787	.041
		1.2	0.6	.804	.056	.789	.062	.790	.061	.787	.067	.778	.070
		1.2	1.2	.856	.054	.836	.061	.838	.063	.836	.066	.827	.071
		2.4	0.6	.862	.086	.789	.124	.781	.136	.805	.120	.778	.148
		2.4	1.2	.885	.078	.802	.115	.798	.129	.821	.113	.793	.144
	RMST_{diff}	0		.947	.162	.824	.174	.265	.647	.747	.206	.940	.162
		1.2		1.333	.234	1.256	.251	1.270	.258	1.255	.270	1.226	.288
		2.4		.891	.226	.659	.298	.659	.326	.710	.301	.647	.363
3	$\text{AUC}_{s,w}$	0	0.6	.793	.032	.793	.032	.615	.182	.792	.032	.793	.032
		0	1.2	.815	.034	.814	.035	.622	.195	.814	.035	.814	.035
		1.2	0.6	.831	.052	.823	.055	.809	.061	.821	.056	.814	.059
		1.2	1.2	.890	.051	.879	.055	.861	.064	.877	.056	.867	.060
		2.4	0.6	.885	.077	.856	.096	.803	.138	.849	.103	.828	.122
		2.4	1.2	.922	.064	.885	.088	.826	.134	.879	.093	.855	.118
	RMST_{diff}	0		1.063	.143	1.062	.144	.397	.637	1.060	.142	1.062	.144
		1.2		1.269	.235	1.239	.235	1.186	.252	1.231	.240	1.203	.245
		2.4		.837	.199	.750	.225	.617	.300	.734	.239	.682	.278

Table 4.2: Simulation results for Scenario 1-3 when measurements were irregular: estimated area under the ROC curve and difference in restricted mean survival time between low-risk and high-risk groups obtained from the proposed method, the Cox model with time-varying covariates (COX), the Cox model with time-varying covariates and coefficients (VCOX), the partly conditional cox model (PC), and the partly conditional cox model with time-varying coefficients (VPC) on validation data. Number of replicates was $1,000$, $n = 200$.

Scenario		s	w	Proposed		COX		VCOX		PC		VPC	
				Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	$\text{AUC}_{s,w}$	0	0.6	.822	.031	.790	.040	.691	.187	.734	.055	.823	.031
		0	1.2	.819	.030	.787	.039	.689	.186	.731	.055	.819	.030
		1.2	0.6	.641	.207	.610	.217	.567	.219	.634	.214	.635	.212
		1.2	1.2	.671	.143	.635	.146	.587	.152	.664	.145	.667	.145
		2.4	0.6	.730	.160	.669	.163	.710	.174	.710	.161	.736	.157
		2.4	1.2	.749	.104	.684	.115	.730	.124	.727	.110	.756	.102
	RMST_{diff}	0		1.995	.266	1.778	.321	1.156	1.178	1.365	.402	1.997	.265
		1.2		.719	.298	.551	.304	.331	.380	.686	.305	.706	.304
		2.4		.625	.223	.455	.241	.577	.279	.567	.235	.643	.218
2	$\text{AUC}_{s,w}$	0	0.6	.729	.037	.680	.047	.561	.134	.650	.059	.724	.038
		0	1.2	.793	.038	.745	.049	.575	.171	.710	.063	.788	.039
		1.2	0.6	.799	.060	.787	.064	.788	.064	.789	.064	.781	.069
		1.2	1.2	.848	.057	.832	.065	.835	.063	.838	.066	.832	.068
		2.4	0.6	.856	.092	.780	.130	.791	.129	.805	.114	.779	.144
		2.4	1.2	.880	.085	.791	.123	.804	.125	.820	.109	.794	.141
	RMST_{diff}	0		.952	.156	.841	.169	.287	.639	.759	.207	.944	.156
		1.2		1.304	.260	1.245	.282	1.260	.273	1.267	.280	1.244	.290
		2.4		.881	.247	.638	.311	.676	.324	.709	.293	.648	.370
3	$\text{AUC}_{s,w}$	0	0.6	.794	.032	.793	.033	.616	.185	.793	.033	.794	.032
		0	1.2	.816	.036	.815	.036	.623	.197	.815	.036	.815	.036
		1.2	0.6	.833	.053	.827	.056	.809	.064	.825	.057	.817	.060
		1.2	1.2	.891	.049	.884	.052	.862	.065	.881	.053	.872	.058
		2.4	0.6	.882	.078	.854	.094	.801	.139	.848	.099	.824	.120
		2.4	1.2	.919	.067	.884	.090	.826	.138	.878	.096	.851	.119
	RMST_{diff}	0		1.075	.147	1.071	.148	.412	.657	1.069	.148	1.071	.146
		1.2		1.286	.235	1.265	.238	1.199	.262	1.255	.238	1.225	.245
		2.4		.827	.207	.740	.236	.616	.309	.726	.250	.664	.287

death and those had not. We hope that by timely identifying high-risk patients, we can inform physicians so that they adjust medical treatments accordingly. Here besides the proposed method, we also implemented the other methods mentioned in the simulation section.

To reduce bias, we adopted K-fold cross validation for the performance evaluation. Specifically, we randomly divided the study subjects into K folds of approximately equal sizes. For each $k, k = 1, \ldots, K$, we estimated the model coefficients using the subjects outside the kth fold, and calculated the $\text{AUC}_{s,w}$ and difference in RMST for the subjects inside the kth fold. This process was repeated 100 times, and then the average of the $K \times 100$ values for each evaluation metric was the final cross validation result. The bandwidth h_2 was tuned using a similar cross validation procedure and the decision was made based on the average of Uno's C-statistic across s (Uno et al., 2011). The truncation time point for RMST was 18, which was the 95% quantile of the time-to-events. Since the VPC and the COX failed to converge, we decreased the number of fractional polynomial bases from 6 to 4 for the two methods.

The resulting $\text{AUC}_{s,w}$ and the difference in RMST on different s and w were presented in Table 4.3. Overall, the proposed method performed similarly to the other methods, except that the VCOX method had lower values given baseline measurement ($s = 0$), suggesting that the true model underlying the AIDS data set may not deviate much from the Cox-like model. $\text{AUC}_{s,w}$ ranged from 0.659 to 0.756, indicating that the dynamic scoring system that incorporated the longitudinal CD4 cell count measurements had fair to moderate discrimination ability on advanced HIV patients, which is consistent to the previous findings in the literature (Goldman et al., 1996; Rizopoulos, 2011). The estimates of difference in RMST also revealed little differences among those methods on this data set. For example, for subjects who survived up to 2 months, the difference in RMST were around 3.5 for all methods, which meant the subjects in the high-risk group survived about 3.5 months shorter than those in the low-risk group on average, when following up the patients 18 months.

Table 4.3: Area under the ROC curve and the difference in restricted mean survival time (based on 5-fold cross validation repeated for 100 times) on $s = 0$, 2, and 6 months, and w=2, 4, and 6 months, applied in the AIDS data.

	s	w	Proposed		COX		VCOX		PC		VPC	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\text{AUC}_{s,w}$	0	2	.756	.218	.740	.214	.659	.226	.723	.215	.732	.214
		4	.746	.088	.742	.088	.693	.101	.733	.090	.738	.089
		6	.737	.072	.736	.071	.698	.079	.731	.072	.734	.072
	2	2	.718	.098	.729	.096	.727	.097	.728	.096	.724	.096
		4	.710	.077	.722	.074	.720	.074	.721	.074	.718	.074
		6	.721	.060	.726	.060	.725	.060	.726	.060	.725	.060
	6	2	.670	.124	.669	.122	.669	.122	.671	.123	.673	.124
		4	.669	.081	.664	.081	.666	.081	.669	.082	.672	.082
		6	.710	.058	.708	.059	.709	.058	.711	.058	.712	.057
RMST_{diff}	0		3.961	.955	3.957	.954	3.548	1.020	3.866	.973	3.911	.954
	2		3.498	.920	3.469	.916	3.510	.904	3.506	.904	3.541	.899
	6		2.056	.767	2.072	.746	2.087	.742	2.084	.755	2.077	.761

Chapter 5

Discussion

5.1 Assessing discrimination capacity of a combination of biomarkers under matched case-control studies

In Aim 1, we proposed an alternative semiparametric method to the conditional logistic regression given the data from matched case-control studies. We developed a pseudo-conditional likelihood function to avoid the need to estimate stratum-specific parameters. In the meanwhile, instead of using parametric link functions as in the conditional logistic regression, we directly used the decision rule on the construction of the pseudo-conditional likelihood. We maximized the proposed likelihood with a constraint of achieving a clinically acceptable specificity, based on the general guidance in cancer population screening practice. Different from the objective function of the direct method, the proposed likelihood utilized information from both cases and controls, and was shown advantageous to maintain the pre-specified specificity in the independent validation data.

Being able to maintain specificity is a pre-requisite for a good screening tool since even tiny loss in specificity has severe consequences. For instance, in liver cancer screening, per 1% drop in specificity would result in 1,000 more subjects getting false positive

results, experiencing psychological trauma, or even going through biopsy for diagnosis in a population screening program of 100,000 subjects considering the low incidence of liver cancer (U.S. Cancer Statistics Working Group, 2019). Thus, being able to keep specificity on external validation data makes the proposed method more appealing than other existing methods in population screening. Although the focus of this project is individually matched data, the proposed method can be straightforwardly extended to studies that use frequency matching (e.g, case and control groups have similar proportion of smokers, females in a lung cancer study) by post-hoc forming strata.

Of note, we maximized the proposed likelihood by using 20 different sets of initial values, to minimize the possibility that the algorithm converged to a local maxima of the likelihood function depending on the starting values. Even though we applied multiple starting values, the computation burden was not heavy. For example, in a 100-run simulation with a sample size of 400 under Scenario 1, the CPU time of a desktop with 3.30GHz CPU was 0.86 minutes for the point estimation and 2.37 hours for the variance estimation. If there is a large number of risk factors, we can then use the kernel-smoothed method since it has shown satisfactory performance as shown in our simulation studies.

5.2 Risk assessment under two-phase sampling designs

In Aim 2, we proposed a semiparametric isotonic regression model for constructing risk scores and assessing absolute risks given data from two-phase studies. This aim will help identify high-risk patients and improve the shared decision making between at-risk patients and their physicians by providing a quantifiable personalized risk assessment. We leave the link function unspecified, other than the monotonicity assumption, which is a necessary assumption to achieve a sensible risk score. Although our model involves a nonparametric component, a profiling method and PAVA are utilized to improve computational efficiency and can be easily implemented using existing software. Thus, the

proposed method offers robustness, easy implementation, and computational efficiency.

One challenge of the proposed method is model specification for the risk score: how to select the best set of risk factors and how to determine an appropriate form for combining multiple risk factors. A simple screening procedure (e.g., marginal correlation) or stepwise model selection procedures can be applied to select risk factors for the risk score, however, it may not be able to identify the optimal subset of risk factors. Next, to combine multiple factors, we focus on the linear form due to its simplicity and popularity. Standard model comparison tools, such as the likelihood ratio test, are not directly applicable and cannot handle the additional variation due to the estimated weights. Developing rigorous tools that simultaneously select the optimal set of risk factors and identify the best way to combine them is beyond the scope of this project, though worthy of future research.

The ranked set sampling (RSS) design, as an alternative design to the two-phase sampling designs, has been proposed by McIntyre (1952) and has received increasing attention. Recently, Zamanzade and Vock (2015) showed that the RSS design is more efficient than the two-phase sampling designs under certain cases, and Zamanzade and Mahdizadeh (2019) used a variation of RSS to efficiently estimate the prevalence of a rare disease in a given population. It is of interest to extend our work to RSS-based designs for constructing risk scores and assessing absolute risks simultaneously for future research.

5.3 Dynamic scoring system of a survival outcome using longitudinally collected biomarkers

In Aim 3, we proposed a dynamic scoring system that takes into account all available information and the ever-changing risk set. It is dynamic in the sense that the scoring system can provide updated risk stratification to physicians at any time during the follow-up. This method is model-free and hence not restricted to the proportional hazard assumption, which is a drawback in the partly conditional model. Our approach can be

widely applied to typical longitudinal studies with survival outcomes where longitudinal measurements are collected during follow-up, either regularly or irregularly.

Since we focused on developing a dynamic prediction rule, in evaluation of the methods, we used existing point-wise metrics, such as AUC given the score at a prediction time s and the binary outcome in an additional time interval w. A measure of prediction performance tailored for dynamic prediction rules for a survival outcome is beyond the scope of this project. Moreover, how to identify risk factors to be included in the dynamic risk score is an important topic, especially when biomarkers and other risk factors are abundant. Pre-selection procedures are often used in literature, but they may not lead to an optimal set of risk factors when the models on which the procedures rely is incorrect. Hence, a dynamic scoring system with an integrated variable selection function is attractive and is of future research interest.

Chapter 6

Appendices

6.1 Appendix for Aim 1

6.1.1 Derivation of the pseudolikelihood

In kth stratum, the conditional likelihood of the observed data given that one of the patients is the case and the remaining patients are controls may be written as:

$$\begin{aligned}
\mathcal{L}_k(\boldsymbol{\beta}, c) &= \frac{\prod_{i=1}^{n_{kD}} \Pr(\boldsymbol{X}_{ki}|Y_{ki}=1,\boldsymbol{\beta},c) \prod_{i=n_{kD}+1}^{n_k} \Pr(\boldsymbol{X}_{ki}|Y_{ki}=0,\boldsymbol{\beta},c)}{\sum_{J\in\mathcal{C}_k^D}\left\{\prod_{j\in J}\Pr(\boldsymbol{X}_{kj}|Y_{kj}=1,\boldsymbol{\beta},c)\prod_{j\in\mathcal{C}_k\setminus J}\Pr(\boldsymbol{X}_{kj}|Y_{kj}=0,\boldsymbol{\beta},c)\right\}} \\
&= \frac{\prod_{i=1}^{n_{kD}} \frac{\Pr(Y_{ki}=1|\boldsymbol{X}_{ki},\boldsymbol{\beta},c)\Pr(\boldsymbol{X}_{ki}|\boldsymbol{\beta},c)}{\Pr(Y_{ki}=1)} \prod_{i=n_{kD}+1}^{n_k} \frac{\Pr(Y_{ki}=0|\boldsymbol{X}_{ki},\boldsymbol{\beta},c)\Pr(\boldsymbol{X}_{ki}|\boldsymbol{\beta},c)}{\Pr(Y_{ki}=0)}}{\sum_{J\in\mathcal{C}_k^D}\left\{\prod_{j\in J}\frac{\Pr(Y_{kj}=1|\boldsymbol{X}_{kj},\boldsymbol{\beta},c)\Pr(\boldsymbol{X}_{kj}|\boldsymbol{\beta},c)}{\Pr(Y_{kj}=1)}\prod_{j\in\mathcal{C}_k\setminus J}\frac{\Pr(Y_{kj}=0|\boldsymbol{X}_{kj},\boldsymbol{\beta},c)\Pr(\boldsymbol{X}_{kj}|\boldsymbol{\beta},c)}{\Pr(Y_{kj}=0)}\right\}} \\
&= \frac{\prod_{i=1}^{n_{kD}} \Pr(Y_{ki}=1|\boldsymbol{X}_{ki},\boldsymbol{\beta},c) \prod_{i=n_{kD}+1}^{n_k} \Pr(Y_{ki}=0|\boldsymbol{X}_{ki},\boldsymbol{\beta},c)}{\sum_{J\in\mathcal{C}_k^D}\left\{\prod_{j\in J}\Pr(Y_{kj}=1|\boldsymbol{X}_{kj},\boldsymbol{\beta},c)\prod_{j\in\mathcal{C}_k\setminus J}\Pr(Y_{kj}=0|\boldsymbol{X}_{kj},\boldsymbol{\beta},c)\right\}} \\
&= \frac{\prod_{i=1}^{n_{kD}} \Pr(\boldsymbol{\beta}^T\boldsymbol{X}_{ki}>c) \prod_{i=n_{kD}+1}^{n_k}\{1-\Pr(\boldsymbol{\beta}^T\boldsymbol{X}_{ki}>c)\}}{\sum_{J\in\mathcal{C}_k^D}\left[\prod_{j\in J}\Pr(\boldsymbol{\beta}^T\boldsymbol{X}_{kj}>c)\prod_{j\in\mathcal{C}_k\setminus J}\{1-\Pr(\boldsymbol{\beta}^T\boldsymbol{X}_{kj}>c)\}\right]}.
\end{aligned} \tag{6.1}$$

Substitute the probabilities in (6.1) with indicator functions, we then have the pseudo-conditional likelihood function:

$$\mathcal{L}_k(\boldsymbol{\beta}, c) = \frac{\prod_{i=1}^{n_{kD}} I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} > c) \prod_{i=n_{kD}+1}^{n_k} \{1 - I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} > c)\}}{\sum_{J \in \mathcal{C}_k^D} \left[\prod_{j \in J} I(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} > c) \prod_{j \in \mathcal{C}_k \setminus J} \{1 - I(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} > c)\} \right]}. \tag{6.2}$$

6.1.2 Kernel smoother

In order to solve the optimization problem stated in (2.6), we choose to use the Gaussian kernel $K(u, h_n) = \frac{1}{h_n} K(\frac{u}{h_n})$, where $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$. Then we have

$$\begin{aligned}
I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} > c) &= \int_{-\infty}^{\boldsymbol{\beta}^T \boldsymbol{X}_{ki} - c} K(u; h_n) du \\
&= \int_{-\infty}^{\boldsymbol{\beta}^T \boldsymbol{X}_{ki} - c} \frac{1}{h_n} K(\frac{u}{h_n}) du \\
&= \int_{-\infty}^{\boldsymbol{\beta}^T \boldsymbol{X}_{ki} - c} \frac{1}{h_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(u/h_n)^2}{2}} du \quad (\text{ Let } x = \frac{u}{h_n}) \\
&= \frac{1}{\sqrt{2\pi} h_n} \int_{-\infty}^{\frac{\boldsymbol{\beta}^T \boldsymbol{X}_{ki} - c}{h_n}} e^{\frac{-x^2}{2}} h_n dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\boldsymbol{\beta}^T \boldsymbol{X}_{ki} - c}{h_n}} e^{\frac{-x^2}{2}} dx \\
&= \Phi\left(\frac{\boldsymbol{\beta}^T \boldsymbol{X}_{ki} - c}{h_n} \right)
\end{aligned}$$

6.1.3 Simulation Results on Validation Data

Table 6.1: Simulation results on validation data under Scenario 1. K: number of strata; τ: prespecified specificity; Clogit: conditional logistic regression; ESE: empirical standard error.

		Proposed				Direct				pAUC				Clogit			
K	τ	Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity	
		Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
50	.70	.889	.051	.694	.080	.901	.046	.671	.083	.898	.048	.694	.081	.893	.052	.698	.085
	.75	.861	.063	.739	.076	.874	.057	.716	.081	.871	.058	.737	.078	.861	.068	.744	.081
	.80	.822	.077	.783	.073	.840	.068	.763	.077	.835	.069	.782	.072	.822	.080	.789	.076
	.85	.767	.098	.828	.069	.791	.087	.809	.073	.783	.087	.828	.066	.766	.096	.836	.068
	.90	.691	.118	.874	.058	.722	.105	.858	.062	.701	.119	.877	.061	.684	.118	.883	.059
	.95	.566	.148	.922	.046	.605	.138	.908	.051	.572	.160	.926	.048	.548	.152	.932	.045
	.98	.447	.165	.951	.038	.476	.169	.941	.044	.424	.191	.959	.040	.404	.178	.962	.036
100	.70	.897	.037	.697	.058	.906	.032	.681	.058	.903	.033	.699	.059	.902	.034	.701	.060
	.75	.868	.045	.745	.056	.879	.039	.731	.056	.874	.041	.746	.055	.871	.043	.750	.057
	.80	.829	.055	.791	.051	.846	.047	.776	.051	.835	.052	.793	.052	.833	.051	.796	.052
	.85	.771	.071	.838	.047	.794	.060	.824	.047	.781	.064	.840	.047	.777	.065	.843	.048
	.90	.691	.085	.887	.041	.718	.076	.875	.041	.693	.087	.891	.041	.691	.083	.893	.041
	.95	.547	.113	.937	.030	.584	.106	.927	.032	.552	.120	.940	.031	.537	.120	.945	.030
	.98	.415	.129	.965	.023	.443	.128	.959	.024	.384	.150	.972	.022	.382	.140	.973	.021
200	.70	.903	.025	.698	.043	.910	.021	.688	.043	.906	.022	.702	.042	.906	.022	.703	.043
	.75	.873	.032	.747	.040	.882	.028	.738	.040	.877	.029	.751	.039	.876	.029	.752	.041
	.80	.835	.039	.793	.037	.846	.034	.784	.036	.836	.037	.799	.037	.838	.037	.797	.038
	.85	.778	.049	.842	.033	.793	.043	.833	.033	.780	.046	.847	.033	.780	.046	.847	.033
	.90	.694	.062	.892	.029	.714	.056	.883	.029	.692	.061	.898	.028	.694	.061	.897	.029
	.95	.553	.082	.942	.021	.579	.076	.935	.022	.545	.084	.947	.020	.546	.086	.946	.021
	.98	.401	.098	.971	.015	.422	.096	.967	.016	.377	.107	.976	.015	.375	.106	.976	.015
400	.70	.905	.016	.700	.030	.910	.015	.694	.030	.908	.015	.703	.031	.908	.015	.704	.030
	.75	.875	.021	.750	.027	.882	.019	.744	.027	.879	.020	.752	.028	.878	.020	.754	.028
	.80	.836	.025	.797	.025	.844	.023	.791	.025	.838	.025	.801	.026	.838	.024	.801	.025
	.85	.778	.033	.847	.023	.789	.030	.841	.022	.782	.032	.849	.023	.780	.032	.851	.022
	.90	.691	.044	.897	.020	.707	.040	.891	.020	.693	.043	.900	.020	.692	.043	.901	.020
	.95	.546	.060	.946	.014	.564	.055	.942	.014	.540	.060	.950	.013	.540	.061	.950	.014
	.98	.387	.073	.975	.010	.403	.072	.973	.011	.364	.074	.980	.010	.368	.077	.979	.010

Table 6.2: Simulation results on validation data under Scenario 2. K: number of strata; τ: prespecified specificity; Clogit: conditional logistic regression; ESE: empirical standard error.

		Proposed				Direct				pAUC				Clogit			
K	τ	Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity	
		Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
50	.70	.732	.069	.691	.125	.753	.061	.649	.115	.736	.043	.658	.114	.741	.073	.676	.120
	.75	.710	.066	.733	.119	.730	.059	.693	.110	.724	.043	.703	.108	.713	.083	.721	.115
	.80	.689	.062	.775	.104	.710	.055	.735	.102	.711	.041	.746	.100	.682	.086	.766	.105
	.85	.661	.068	.824	.093	.688	.052	.784	.091	.694	.040	.794	.089	.643	.093	.814	.094
	.90	.641	.068	.865	.076	.664	.049	.836	.078	.669	.041	.847	.075	.596	.100	.864	.077
	.95	.614	.069	.911	.056	.636	.046	.895	.056	.631	.046	.907	.053	.526	.106	.919	.052
	.98	.592	.063	.948	.038	.606	.047	.941	.040	.590	.057	.950	.036	.448	.116	.956	.033
100	.70	.733	.051	.695	.095	.753	.044	.661	.084	.729	.034	.675	.104	.744	.053	.684	.090
	.75	.711	.047	.744	.090	.731	.042	.704	.080	.717	.032	.724	.095	.713	.059	.733	.086
	.80	.684	.048	.805	.079	.710	.036	.759	.076	.702	.031	.776	.086	.679	.066	.788	.078
	.85	.663	.049	.850	.068	.689	.032	.808	.068	.686	.032	.823	.076	.636	.074	.837	.068
	.90	.639	.049	.894	.056	.665	.032	.859	.057	.664	.031	.871	.063	.583	.085	.884	.058
	.95	.611	.048	.937	.040	.636	.034	.914	.043	.631	.035	.924	.040	.504	.098	.933	.042
	.98	.593	.043	.963	.024	.610	.032	.954	.026	.595	.041	.961	.023	.427	.093	.966	.023
200	.70	.736	.034	.696	.064	.752	.030	.669	.056	.728	.019	.680	.064	.746	.034	.687	.061
	.75	.712	.031	.748	.060	.728	.027	.716	.054	.717	.020	.727	.067	.715	.038	.737	.059
	.80	.687	.030	.811	.053	.705	.024	.778	.051	.701	.020	.785	.061	.676	.041	.802	.052
	.85	.664	.033	.861	.046	.686	.022	.826	.048	.684	.021	.838	.054	.635	.047	.848	.048
	.90	.641	.034	.906	.037	.664	.021	.876	.040	.659	.023	.892	.043	.580	.056	.896	.040
	.95	.611	.032	.950	.026	.634	.023	.929	.031	.626	.026	.940	.032	.491	.074	.946	.028
	.98	.586	.035	.973	.017	.604	.026	.964	.018	.588	.036	.972	.018	.406	.089	.973	.018
400	.70	.737	.024	.697	.043	.749	.021	.679	.039	.725	.014	.692	.048	.745	.024	.691	.041
	.75	.713	.020	.748	.040	.726	.018	.725	.037	.713	.013	.744	.045	.715	.026	.739	.040
	.80	.691	.020	.810	.039	.705	.016	.785	.038	.699	.012	.799	.036	.680	.029	.798	.037
	.85	.669	.021	.860	.033	.685	.015	.835	.034	.682	.013	.850	.032	.635	.032	.849	.0333
	.90	.645	.020	.908	.027	.662	.015	.887	.028	.660	.013	.895	.024	.578	.040	.899	.028
	.95	.613	.023	.954	.019	.631	.016	.938	.021	.627	.015	.946	.019	.488	.054	.948	.021
	.98	.582	.025	.979	.011	.599	.020	.971	.013	.588	.025	.976	.013	.392	.069	.978	.014

Table 6.3: Simulation results on validation data under Scenario 3. K: number of strata; τ: prespecified specificity; Clogit: conditional logistic regression; ESE: empirical standard error.

		Proposed				Direct				pAUC				Clogit			
K	τ	Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity	
		Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
50	.70	.720	.088	.683	.124	.740	.079	.649	.112	.713	.032	.648	.113	.710	.093	.675	.122
	.75	.690	.083	.729	.117	.712	.069	.692	.104	.703	.032	.691	.106	.675	.104	.721	.112
	.80	.664	.079	.774	.105	.688	.061	.736	.098	.691	.033	.735	.102	.640	.106	.765	.103
	.85	.639	.077	.820	.093	.665	.056	.783	.092	.676	.033	.787	.091	.593	.114	.815	.090
	.90	.614	.084	.865	.074	.640	.053	.838	.073	.656	.032	.842	.076	.536	.125	.864	.072
	.95	.592	.076	.913	.051	.611	.061	.898	.053	.625	.035	.905	.054	.460	.130	.914	.049
	.98	.567	.083	.946	.037	.579	.077	.940	.039	.595	.041	.948	.037	.377	.140	.952	.034
100	.70	.711	.058	.701	.097	.731	.052	.661	.084	.708	.025	.670	.094	.704	.063	.695	.091
	.75	.685	.051	.750	.092	.707	.042	.705	.080	.698	.025	.710	.090	.670	.068	.743	.086
	.80	.663	.047	.789	.082	.688	.032	.740	.076	.686	.024	.760	.079	.631	.076	.777	.080
	.85	.644	.047	.838	.072	.670	.027	.795	.070	.669	.025	.814	.071	.583	.085	.829	.072
	.90	.622	.047	.888	.057	.649	.027	.853	.061	.650	.025	.866	.058	.522	.095	.881	.060
	.95	.596	.051	.932	.042	.625	.029	.909	.041	.624	.028	.917	.042	.440	.109	.932	.041
	.98	.583	.046	.959	.026	.600	.032	.949	.028	.595	.028	.957	.027	.363	.111	.965	.024
200	.70	.710	.038	.701	.065	.726	.034	.668	.054	.707	.016	.679	.066	.705	.040	.697	.061
	.75	.688	.030	.752	.059	.704	.025	.715	.054	.696	.017	.727	.062	.670	.043	.747	.059
	.80	.668	.025	.797	.057	.686	.020	.758	.057	.683	.018	.776	.060	.630	.048	.788	.058
	.85	.648	.029	.849	.049	.669	.018	.812	.051	.667	.019	.826	.052	.582	.055	.840	.052
	.90	.627	.030	.898	.038	.650	.018	.867	.042	.646	.018	.883	.039	.520	.068	.892	.042
	.95	.600	.032	.943	.029	.623	.020	.922	.031	.615	.021	.939	.029	.428	.087	.943	.029
	.98	.578	.037	.969	.020	.597	.024	.958	.020	.583	.031	.971	.019	.345	.097	.971	.018
400	.70	.708	.022	.702	.043	.720	.020	.677	.037	.704	.012	.692	.052	.704	.028	.699	.042
	.75	.689	.016	.752	.041	.701	.014	.723	.037	.693	.011	.744	.044	.668	.029	.750	.040
	.80	.672	.016	.796	.038	.686	.012	.767	.038	.680	.013	.790	.041	.630	.030	.786	.039
	.85	.654	.018	.849	.034	.669	.012	.822	.035	.664	.013	.841	.035	.581	.037	.842	.035
	.90	.632	.021	.898	.027	.649	.012	.876	.028	.645	.015	.888	.032	.518	.047	.895	.029
	.95	.603	.023	.947	.021	.621	.013	.930	.022	.617	.016	.940	.022	.422	.059	.947	.020
	.98	.577	.025	.977	.014	.593	.018	.967	.016	.586	.019	.975	.015	.321	.075	.978	.013

Table 6.4: Simulation results on validation data under Scenario 4. K: number of strata; τ: prespecified specificity; Clogit: conditional logistic regression; ESE: empirical standard error.

		Proposed				Direct				pAUC				Clogit			
K	τ	Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity	
		Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
50	.70	.423	.070	.697	.117	.451	.056	.639	.116	.468	.027	.653	.116	.403	.097	.657	.123
	.75	.406	.071	.738	.110	.435	.054	.689	.109	.457	.027	.695	.108	.372	.099	.703	.115
	.80	.388	.074	.784	.099	.419	.054	.739	.099	.446	.027	.740	.104	.338	.099	.752	.100
	.85	.364	.081	.833	.084	.403	.056	.793	.088	.432	.026	.793	.087	.301	.105	.804	.089
	.90	.347	.087	.879	.069	.384	.059	.848	.071	.416	.025	.847	.068	.261	.112	.857	.072
	.95	.326	.087	.926	.046	.358	.066	.912	.047	.390	.026	.911	.046	.212	.116	.915	.048
	.98	.309	.083	.959	.030	.328	.072	.956	.030	.360	.031	.954	.032	.167	.118	.959	.029
100	.70	.425	.050	.710	.081	.452	.034	.658	.084	.462	.020	.681	.085	.392	.074	.671	.082
	.75	.410	.054	.755	.078	.440	.034	.708	.081	.450	.020	.730	.081	.358	.082	.720	.078
	.80	.392	.063	.807	.069	.428	.034	.762	.072	.437	.020	.780	.074	.320	.089	.777	.072
	.85	.376	.065	.853	.061	.415	.033	.813	.064	.422	.023	.832	.068	.284	.100	.826	.064
	.90	.355	.074	.899	.050	.400	.032	.868	.054	.404	.025	.883	.057	.242	.108	.879	.053
	.95	.331	.074	.942	.035	.376	.037	.923	.036	.381	.026	.930	.039	.194	.115	.930	.036
	.98	.319	.069	.967	.021	.348	.043	.962	.021	.353	.028	.966	.024	.157	.115	.962	.021
200	.70	.435	.033	.709	.054	.454	.022	.669	.055	.461	.012	.688	.056	.386	.063	.681	.053
	.75	.420	.038	.760	.051	.445	.021	.721	.055	.450	.013	.736	.053	.353	.075	.730	.053
	.80	.407	.042	.810	.048	.434	.016	.776	.050	.438	.013	.785	.047	.320	.088	.784	.051
	.85	.389	.044	.860	.041	.421	.017	.827	.045	.424	.013	.834	.041	.284	.098	.835	.045
	.90	.365	.054	.907	.035	.403	.018	.880	.039	.406	.014	.887	.036	.242	.109	.888	.037
	.95	.337	.060	.953	.023	.378	.021	.935	.028	.379	.019	.940	.027	.193	.116	.940	.025
	.98	.318	.055	.976	.016	.351	.027	.967	.017	.353	.021	.971	.014	.154	.118	.970	.016
400	.70	.443	.022	.709	.038	.458	.012	.679	.039	.459	.010	.698	.043	.378	.059	.687	.036
	.75	.429	.025	.760	.036	.448	.011	.730	.037	.448	.010	.749	.040	.343	.072	.737	.036
	.80	.416	.026	.808	.033	.436	.011	.781	.033	.435	.010	.797	.036	.314	.085	.790	.036
	.85	.399	.030	.860	.029	.422	.011	.835	.031	.421	.009	.847	.029	.277	.097	.842	.031
	.90	.378	.035	.909	.023	.405	.011	.888	.025	.403	.010	.899	.023	.235	.109	.894	.026
	.95	.347	.047	.954	.016	.379	.015	.939	.019	.379	.011	.946	.017	.185	.117	.945	.018
	.98	.324	.037	.980	.010	.351	.018	.972	.012	.347	.017	.977	.012	.145	.117	.975	.012

6.1.4 Additional Simulation Results on Training Data

Table 6.5: Additional summary statistics of estimated sensitivities on the training data. K: number of strata; τ: prespecified specificity; Clogit: conditional logistic regression; Mean: empirical mean sensitivity; ESE: empirical standard error; ASE: average of estimated standard errors; CP: 95% coverage probability.

Scenario	K	τ	Proposed				Direct		pAUC		Clogit	
			Mean	ESE	ASE	CP	Mean	ESE	Mean	ESE	Mean	ESE
1	200	.70	.914	.030	.032	.979	.922	.026	.909	.031	.909	.030
		.75	.884	.037	.039	.972	.894	.032	.880	.038	.879	.036
		.80	.846	.046	.047	.968	.859	.040	.840	.046	.840	.045
		.85	.790	.056	.058	.948	.807	.050	.784	.055	.784	.056
		.90	.708	.071	.073	.956	.730	.064	.698	.069	.699	.071
		.95	.568	.090	.092	.938	.595	.083	.551	.091	.551	.094
		.98	.415	.103	.103	.906	.437	.100	.385	.111	.380	.112
	400	.70	.913	.021	.022	.982	.919	.019	.909	.021	.910	.021
		.75	.883	.025	.027	.967	.891	.023	.880	.025	.881	.025
		.80	.844	.029	.033	.975	.853	.027	.839	.030	.840	.029
		.85	.787	.037	.041	.967	.799	.033	.784	.037	.783	.036
		.90	.700	.049	.052	.959	.717	.045	.696	.049	.695	.048
		.95	.554	.064	.066	.950	.574	.059	.544	.064	.543	.066
		.98	.396	.075	.077	.917	.412	.074	.370	.078	.374	.080
2	200	.70	.749	.048	.050	.976	.770	.041	.728	.038	.748	.048
		.75	.723	.046	.050	.970	.746	.040	.717	.037	.717	.053
		.80	.696	.044	.050	.970	.720	.037	.701	.037	.676	.053
		.85	.671	.047	.051	.963	.698	.037	.683	.037	.635	.060
		.90	.647	.047	.052	.963	.674	.037	.659	.041	.579	.067
		.95	.616	.046	.051	.959	.642	.039	.627	.045	.491	.083
		.98	.591	.049	.047	.944	.612	.041	.586	.051	.407	.097
	400	.70	.746	.034	.036	.963	.762	.029	.725	.031	.747	.033
		.75	.721	.032	.034	.968	.738	.028	.713	.029	.717	.036
		.80	.696	.030	.034	.965	.714	.027	.699	.028	.681	.037
		.85	.673	.031	.034	.957	.692	.027	.682	.027	.637	.041
		.90	.649	.031	.036	.962	.668	.027	.660	.026	.577	.047
		.95	.617	.033	.037	.960	.637	.028	.627	.029	.487	.059
		.98	.586	.035	.035	.949	.604	.031	.584	.040	.392	.073
3	200	.70	.719	.050	.054	.980	.737	.046	.709	.036	.705	.052
		.75	.695	.044	.050	.974	.713	.040	.697	.035	.671	.057
		.80	.675	.041	.049	.971	.695	.037	.685	.036	.631	.061
		.85	.655	.044	.050	.963	.677	.037	.670	.037	.583	.068
		.90	.634	.047	.053	.958	.658	.038	.650	.037	.523	.080
		.95	.608	.048	.053	.962	.632	.039	.618	.042	.433	.097
		.98	.584	.053	.049	.946	.604	.043	.585	.048	.352	.107
	400	.70	.714	.033	.035	.980	.727	.031	.703	.025	.705	.037
		.75	.694	.029	.032	.968	.707	.027	.691	.024	.670	.040
		.80	.676	.029	.032	.966	.690	.025	.680	.025	.631	.042
		.85	.658	.030	.034	.962	.674	.025	.665	.027	.581	.047
		.90	.637	.031	.036	.962	.655	.026	.647	.028	.520	.055
		.95	.608	.034	.038	.968	.628	.027	.618	.027	.427	.066
		.98	.581	.035	.036	.948	.597	.031	.584	.030	.326	.081
4	200	.70	.463	.045	.057	.991	.488	.035	.487	.025	.416	.072
		.75	.448	.049	.060	.980	.477	.035	.477	.026	.382	.084
		.80	.433	.052	.065	.978	.465	.030	.466	.027	.346	.097
		.85	.414	.053	.070	.987	.451	.031	.452	.028	.308	.108
		.90	.391	.062	.075	.974	.432	.031	.434	.027	.264	.119
		.95	.362	.067	.076	.964	.405	.032	.404	.029	.212	.128
		.98	.343	.062	.069	.976	.377	.037	.377	.032	.170	.130
	400	.70	.460	.031	.040	.984	.481	.022	.480	.021	.397	.067
		.75	.447	.033	.043	.985	.470	.021	.469	.022	.362	.079
		.80	.432	.035	.046	.985	.457	.021	.456	.021	.332	.093
		.85	.415	.038	.051	.985	.443	.021	.443	.019	.293	.106
		.90	.394	.040	.056	.979	.424	.021	.423	.021	.250	.118
		.95	.362	.051	.061	.970	.397	.023	.396	.021	.198	.126
		.98	.340	.043	.053	.972	.368	.025	.366	.025	.155	.126

6.1.5 Simulation Results of Youden's Index

Table 6.6: Summary statistics of Youden's Index on the validation data under Scenario 1. K: number of strata in the training data; τ: prespecified specificity; Clogit: conditional logistic regression; ESE: empirical standard error.

K	τ	Proposed		Direct		pAUC		Clogit	
		Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
50	.70	.583	.047	.572	.053	.592	.041	.591	.042
	.75	.599	.040	.590	.047	.608	.033	.606	.036
	.80	.605	.040	.604	.040	.616	.026	.611	.033
	.85	.596	.054	.601	.046	.611	.035	.602	.043
	.90	.565	.074	.580	.062	.578	.067	.568	.069
	.95	.488	.111	.513	.100	.498	.118	.480	.113
	.98	.398	.133	.417	.134	.383	.156	.367	.146
100	.70	.593	.035	.587	.036	.602	.029	.603	.030
	.75	.612	.029	.610	.029	.620	.020	.621	.022
	.80	.619	.026	.621	.022	.628	.015	.629	.015
	.85	.609	.038	.619	.025	.621	.023	.620	.023
	.90	.578	.052	.592	.042	.585	.050	.584	.046
	.95	.485	.088	.511	.079	.492	.092	.481	.092
	.98	.380	.109	.402	.108	.356	.129	.354	.121
200	.70	.601	.024	.598	.025	.608	.021	.609	.022
	.75	.620	.020	.619	.018	.628	.013	.628	.014
	.80	.628	.016	.630	.012	.635	.008	.635	.008
	.85	.620	.023	.626	.015	.627	.016	.627	.015
	.90	.585	.038	.598	.030	.590	.034	.591	.033
	.95	.495	.063	.514	.057	.492	.065	.493	.066
	.98	.372	.084	.389	.082	.353	.093	.351	.092
400	.70	.605	.018	.604	.017	.611	.016	.612	.016
	.75	.625	.013	.626	.012	.631	.009	.632	.009
	.80	.634	.010	.635	.008	.639	.004	.639	.004
	.85	.625	.015	.630	.011	.631	.010	.631	.010
	.90	.589	.027	.598	.022	.593	.023	.593	.023
	.95	.492	.047	.506	.043	.491	.047	.491	.048
	.98	.362	.063	.375	.062	.344	.065	.347	.067

Table 6.7: Summary statistics of Youden's Index on the validation data under Scenario 2. K: number of strata in the training data; τ: prespecified specificity; Clogit: conditional logistic regression; ESE: empirical standard error.

		Proposed		Direct		pAUC		Clogit	
K	τ	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
50	.70	.423	.070	.402	.064	.394	.081	.417	.058
	.75	.443	.069	.423	.062	.427	.073	.434	.054
	.80	.464	.065	.445	.059	.457	.065	.449	.049
	.85	.485	.066	.472	.054	.488	.055	.458	.049
	.90	.506	.063	.499	.049	.516	.044	.459	.059
	.95	.525	.060	.531	.038	.538	.038	.445	.079
	.98	.540	.054	.547	.035	.540	.049	.404	.098
100	.70	.429	.053	.414	.046	.404	.076	.428	.041
	.75	.455	.052	.435	.046	.441	.068	.447	.036
	.80	.490	.051	.469	.046	.478	.058	.467	.032
	.85	.513	.048	.497	.042	.509	.047	.473	.035
	.90	.533	.045	.525	.036	.535	.038	.467	.048
	.95	.548	.037	.550	.026	.555	.023	.437	.071
	.98	.556	.033	.564	.022	.556	.033	.393	.079
200	.70	.432	.035	.421	.030	.408	.048	.433	.027
	.75	.459	.035	.445	.031	.444	.050	.452	.024
	.80	.498	.037	.484	.033	.486	.043	.478	.019
	.85	.524	.033	.512	.031	.522	.034	.483	.022
	.90	.547	.029	.540	.024	.551	.022	.476	.031
	.95	.561	.023	.563	.015	.566	.013	.437	.055
	.98	.559	.025	.568	.014	.560	.024	.379	.077
400	.70	.434	.024	.428	.020	.417	.037	.436	.018
	.75	.461	.025	.451	.022	.457	.034	.455	.015
	.80	.501	.027	.490	.025	.498	.025	.478	.013
	.85	.529	.024	.520	.022	.531	.020	.484	.014
	.90	.553	.018	.549	.016	.555	.011	.478	.021
	.95	.567	.015	.570	.009	.572	.007	.437	.039
	.98	.561	.018	.570	.011	.564	.016	.370	.059

Table 6.8: Summary statistics of Youden's Index on the validation data under Scenario 3. K: number of strata in the training data; τ: prespecified specificity; Clogit: conditional logistic regression; ESE: empirical standard error.

		Proposed		Direct		pAUC		Clogit	
K	τ	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
50	.70	.403	.063	.389	.055	.362	.084	.385	.055
	.75	.420	.072	.404	.055	.394	.077	.397	.056
	.80	.438	.074	.424	.061	.426	.071	.405	.061
	.85	.459	.077	.448	.063	.463	.060	.408	.073
	.90	.479	.081	.478	.053	.498	.047	.399	.093
	.95	.504	.070	.509	.054	.531	.032	.374	.113
	.98	.513	.076	.519	.069	.543	.032	.329	.131
100	.70	.412	.052	.392	.045	.377	.071	.399	.037
	.75	.435	.056	.412	.048	.408	.066	.413	.038
	.80	.451	.057	.429	.050	.445	.055	.408	.042
	.85	.482	.053	.465	.049	.483	.047	.412	.053
	.90	.510	.046	.502	.042	.516	.035	.404	.069
	.95	.528	.042	.534	.025	.541	.019	.372	.091
	.98	.541	.037	.550	.024	.552	.016	.328	.103
200	.70	.411	.036	.394	.028	.386	.050	.402	.024
	.75	.440	.039	.419	.034	.423	.046	.417	.025
	.80	.465	.041	.444	.040	.459	.042	.417	.031
	.85	.498	.038	.482	.036	.493	.033	.421	.038
	.90	.525	.030	.517	.026	.528	.022	.412	.052
	.95	.544	.025	.545	.016	.553	.011	.371	.073
	.98	.547	.027	.555	.014	.554	.018	.316	.089
400	.70	.410	.026	.397	.020	.397	.040	.403	.016
	.75	.441	.028	.424	.026	.436	.033	.418	.017
	.80	.468	.029	.453	.027	.471	.029	.416	.022
	.85	.503	.025	.491	.024	.504	.023	.423	.027
	.90	.530	.020	.525	.017	.532	.017	.413	.036
	.95	.550	.017	.551	.010	.557	.008	.369	.050
	.98	.553	.018	.559	.008	.560	.007	.299	.068

Table 6.9: Summary statistics of Youden's Index on the validation data under Scenario 4. K: number of strata in the training data; τ: prespecified specificity; Clogit: conditional logistic regression; ESE: empirical standard error.

		Proposed		Direct		pAUC		Clogit	
K	τ	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
50	.70	.120	.085	.090	.081	.120	.090	.059	.074
	.75	.145	.085	.123	.081	.153	.082	.075	.081
	.80	.172	.085	.158	.080	.186	.078	.091	.088
	.85	.198	.085	.196	.075	.225	.063	.105	.097
	.90	.227	.085	.233	.066	.263	.046	.117	.105
	.95	.252	.082	.271	.063	.301	.026	.127	.114
	.98	.268	.078	.285	.066	.315	.016	.126	.117
100	.70	.135	.070	.110	.066	.143	.066	.063	.065
	.75	.165	.071	.149	.066	.180	.062	.079	.076
	.80	.199	.069	.190	.060	.217	.055	.097	.086
	.85	.229	.067	.228	.053	.254	.047	.110	.096
	.90	.254	.070	.268	.044	.286	.035	.121	.105
	.95	.273	.066	.299	.032	.311	.019	.124	.112
	.98	.286	.061	.310	.036	.318	.012	.119	.114
200	.70	.144	.050	.123	.047	.149	.043	.067	.060
	.75	.180	.051	.166	.047	.186	.040	.083	.073
	.80	.217	.048	.210	.039	.223	.035	.103	.084
	.85	.249	.046	.248	.035	.258	.028	.119	.096
	.90	.272	.052	.284	.027	.293	.022	.130	.106
	.95	.290	.054	.313	.014	.319	.011	.132	.114
	.98	.294	.049	.318	.018	.324	.010	.124	.116
400	.70	.152	.036	.137	.033	.158	.034	.065	.058
	.75	.189	.035	.179	.030	.197	.030	.081	.072
	.80	.224	.033	.218	.025	.232	.026	.103	.086
	.85	.259	.032	.257	.023	.268	.020	.119	.099
	.90	.286	.032	.293	.016	.302	.013	.130	.109
	.95	.301	.043	.319	.009	.324	.008	.130	.117
	.98	.304	.033	.322	.010	.324	.008	.119	.117

6.1.6 Simulation Results of the Kernel Smoothing Method

Table 6.10: Simulation results of the kernel smoothing method on training data under Scenario 1. K: number of strata in the training data; τ: prespecified specificity; Mean: empirical mean sensitivity; ESE: empirical standard error.

K	τ	Proposed		Kernel $C_h = 1/5$		Kernel $C_h = 1$		Kernel $C_h = 5$	
		Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE
100	.80	.846	.066	.836	.067	.832	.067	.838	.064
	.85	.791	.081	.780	.080	.773	.081	.782	.076
	.90	.711	.095	.696	.095	.683	.100	.696	.094
	.95	.565	.121	.550	.121	.524	.123	.542	.125
	.98	.432	.136	.407	.139	.377	.139	.387	.145
200	.80	.846	.046	.838	.047	.837	.047	.839	.045
	.85	.790	.056	.781	.058	.778	.059	.785	.055
	.90	.708	.071	.695	.072	.688	.073	.702	.069
	.95	.568	.090	.552	.091	.536	.091	.559	.091
	.98	.415	.103	.396	.105	.376	.104	.389	.112
400	.80	.844	.029	.839	.030	.839	.030	.839	.028
	.85	.787	.037	.780	.038	.778	.039	.784	.035
	.90	.700	.049	.693	.048	.686	.051	.699	.048
	.95	.554	.064	.542	.064	.531	.066	.551	.065
	.98	.396	.075	.383	.076	.370	.073	.381	.083

6.1.7 Asymptotic Properties

In the following, we prove the consistency of $(\widehat{\boldsymbol{\beta}}, \widehat{c})$ and $\widehat{Se}(\widehat{\boldsymbol{\beta}}, \widehat{c})$. Denote the true values of these parameters as $(\widetilde{\boldsymbol{\beta}}, \widetilde{c})$ and $\widetilde{Se}(\widetilde{\boldsymbol{\beta}}, \widetilde{c})$. For simplicity, we focus on data with one case in each stratum.

Regularity conditions

We summarize the regularity conditions as follows.

1. Observations are randomly sampled conditional on disease status Y.
2. $n_D + n_{\bar{D}} \to \infty$ and $n_D / n_{\bar{D}} \to \lambda \in (0, 1)$.
3. The covariate vector $\boldsymbol{X}$ is in a bounded compact set $\mathcal{X}$ in $\mathbb{R}^d$.
4. The parameters $(\boldsymbol{\beta}, c)$ are in the space $\mathcal{B} \times \mathcal{C}$, where $\mathcal{B} \times \mathcal{C}$ is a compact space in $\mathbb{R}^{d+1}$, and $\mathcal{B} = \{\boldsymbol{\beta} \mid \|\boldsymbol{\beta}\|_2 = 1, \boldsymbol{\beta} \in \mathbb{R}^d\}$.
5. At least one component of $\boldsymbol{X}$ is continuous.
6. There exists a constant $k_r > 0$ such that $\inf_{\boldsymbol{\beta} \in \mathcal{B}} \text{eigmin}[\mathbf{J}\{\widetilde{c}(\boldsymbol{\beta}); \boldsymbol{\beta}\}] > k_r$.
7. $\sup_{\boldsymbol{\beta}: d(\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}) \geq \epsilon} M(\boldsymbol{\beta}) < M(\widetilde{\boldsymbol{\beta}})$ for every $\epsilon > 0$.

Proof of consistency

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from a probability distribution P on a measurable space. Given a measurable function f, we define Pf for the expectation $\mathbb{E}f(\boldsymbol{X}) = \int f dP$ and $\mathbb{P}_n f$ for the average $n^{-1} \sum_{i=1}^n f(\boldsymbol{X}_i)$.

For simplicity, we scale the weights such that $\sum_{k=1}^{n_D} \sum_{j=2}^{n_k} \widehat{w}_{kj} = n_{\bar{D}}$. Since the estimation of c given $\boldsymbol{\beta}$ depends only on the rank of the control data, $\widehat{c}$ is invariant to the scaling. We define $W_n(\boldsymbol{\beta}, c) = \frac{1}{n_{\bar{D}}} \sum_{k=1}^{n_D} \sum_{j=2}^{n_k} \left\{ \widehat{w}_{kj} I(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} \leq c) - \tau \right\} = \mathbb{P}_n \left\{ \widehat{w}_{kj} I(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} \leq c) - \tau \right\} =: \mathbb{P}_n z_{\boldsymbol{\beta}, c}(\boldsymbol{X}_{kj})$, where τ is the prespecified threshold. Define $\widetilde{z}$ as the counterpart of z using true weight $\widetilde{w}_{kj}$, and $\widetilde{W}_n$ as the counterpart of W_n using $\widetilde{z}$. Let

$W(\boldsymbol{\beta}, c) = \mathbb{E}W_n(\boldsymbol{\beta}, c)$, and $\widehat{c}(\boldsymbol{\beta}) = \inf\{c : W_n(\boldsymbol{\beta}, c) \geq 0\}$. By regularity conditions (C5), for any given $\boldsymbol{\beta}$, $W_n\{\boldsymbol{\beta}, c(\boldsymbol{\beta})\}$ is monotone in c with a jump size that converges to 0. So,

$$W_n\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\} \xrightarrow{p} 0. \tag{6.3}$$

By regularity conditions, $|\boldsymbol{\beta}^T\mathbf{x}| < C_{\boldsymbol{\beta}}|\mathbf{x}|$ for $C_{\boldsymbol{\beta}} = \sup_{\boldsymbol{\beta}\in\mathcal{B}}|\boldsymbol{\beta}|$, the class of functions $\mathbf{x} \mapsto \boldsymbol{\beta}^T\mathbf{x}$ is therefore Glivenko-Cantelli. For any $c \in \mathcal{C}$, the class of functions $s \mapsto I(s \leq c)$ is monotone and is in the Glivenko-Cantelli class by Theorem 2.7.5 in Van Der Vaart and Wellner (1996). By the permanence properties of Glivenko-Cantelli classes (Van Der Vaart and Wellner, 1996, Section 2.6.5), the class of functions $\mathbf{x} \mapsto I(\boldsymbol{\beta}^T\mathbf{x} \leq c)$ is in the Glivenko-Cantelli class. It follows that $\widetilde{z}_{\boldsymbol{\beta},c}$ is in the Glivenko-Cantelli class. Applying the property of Glivenko-Cantelli class (Van der Vaart, 2000, page 269), we have the uniform convergence $\sup_{\boldsymbol{\beta}}|\widetilde{W}_n(\boldsymbol{\beta}, c) - W(\boldsymbol{\beta}, c)| = o_p(1)$, for any given c. Since $\max_{kj}|\widehat{w}_{kj} - \widetilde{w}_{kj}| = O_p(n^{1/2})$, we conclude that

$$\sup_{\boldsymbol{\beta}}|W_n(\boldsymbol{\beta}, c) - W(\boldsymbol{\beta}, c)| = o_p(1), \text{ for any given } c. \tag{6.4}$$

Combining (6.3) and (6.4), we have

$$\begin{aligned}
\sup_{\boldsymbol{\beta}}|W\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\} - W\{\boldsymbol{\beta}, \widetilde{c}(\boldsymbol{\beta})\}| &= \sup_{\boldsymbol{\beta}}|W\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\}| \\
&= \sup_{\boldsymbol{\beta}}|W\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\} - W_n\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\} + W_n\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\}| \\
&\leq \sup_{\boldsymbol{\beta}}|W\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\} - W_n\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\}| + \sup_{\boldsymbol{\beta}}|W_n\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\}| \\
&= o_p(1).
\end{aligned} \tag{6.5}$$

By Taylor expansion of $W\{\boldsymbol{\beta}, \widehat{c}(\boldsymbol{\beta})\}$ around $\widetilde{c}(\boldsymbol{\beta})$ and combining with the boundedness of $\mathbf{J}^{-1}(\widetilde{c}; \boldsymbol{\beta})$ in condition (C6), we immediately have

$$\sup_{\boldsymbol{\beta}}|\widehat{c}(\boldsymbol{\beta}) - \widetilde{c}(\boldsymbol{\beta})| \xrightarrow{p} 0. \tag{6.6}$$

Next, we define the random function

$$M_n(\boldsymbol{\beta}) = \frac{1}{n_D}\sum_{k=1}^{n_D}\log\frac{\mathbb{1}\{\boldsymbol{\beta}^T\boldsymbol{X}_{k1} > \widetilde{c}(\boldsymbol{\beta})\}\prod_{i=2}^{n_k}\mathbb{1}\{\boldsymbol{\beta}^T\boldsymbol{X}_{ki} \le \widetilde{c}(\boldsymbol{\beta})\}}{\sum_{l=1}^{n_k}\mathbb{1}\{\boldsymbol{\beta}^T\boldsymbol{X}_{kl} > \widetilde{c}(\boldsymbol{\beta})\}\prod_{j=1,j\neq l}^{n_k}\mathbb{1}\{\boldsymbol{\beta}^T\boldsymbol{X}_{kj} \le \widetilde{c}(\boldsymbol{\beta})\}} \tag{6.7}$$

$$=: \mathbb{P}_n m_{\boldsymbol{\beta},\widetilde{c}(\boldsymbol{\beta})}\{\boldsymbol{X}_{k1},\ldots,\boldsymbol{X}_{kn_k}\}. \tag{6.8}$$

Let $M = \mathbb{E}M_n = Pm$ denote the expectation of the random function. By the similar arguments to the proof for (6.4) , we may conclude

$$\sup_{\boldsymbol{\beta}}|M_n(\boldsymbol{\beta}) - M(\boldsymbol{\beta})| \xrightarrow{p} 0, \tag{6.9}$$

Define $\widehat{M}_n$ as the counterpart of M_n with $\widetilde{c}$ being replaced with $\widehat{c}$. Since $\widehat{\boldsymbol{\beta}}$ is the maximizer of $\widehat{M}_n(\boldsymbol{\beta})$, we get

$$\widehat{M}_n(\widehat{\boldsymbol{\beta}}) - \widehat{M}_n(\widetilde{\boldsymbol{\beta}}) \ge 0. \tag{6.10}$$

Following the consistency of $\widehat{c}$ to $\widetilde{c}$ in (6.6), we have

$$\widehat{M}_n(\widetilde{\boldsymbol{\beta}}) - M_n(\widetilde{\boldsymbol{\beta}}) \xrightarrow{p} 0 \text{ and } \widehat{M}_n(\widehat{\boldsymbol{\beta}}) - M_n(\widehat{\boldsymbol{\beta}}) \xrightarrow{p} 0. \tag{6.11}$$

The combination of (6.10) and (6.11) thus gives $M_n(\widehat{\boldsymbol{\beta}}) - M_n(\widetilde{\boldsymbol{\beta}}) = \{M_n(\widehat{\boldsymbol{\beta}}) - \widehat{M}_n(\widehat{\boldsymbol{\beta}})\} + \{\widehat{M}_n(\widehat{\boldsymbol{\beta}}) - \widehat{M}_n(\widetilde{\boldsymbol{\beta}})\} + \{\widehat{M}_n(\widetilde{\boldsymbol{\beta}}) - M_n(\widetilde{\boldsymbol{\beta}})\} \ge -o_p(1)$. That is,

$$M_n(\widehat{\boldsymbol{\beta}}) \ge M_n(\widetilde{\boldsymbol{\beta}}) - o_p(1). \tag{6.12}$$

It follows from (6.9), (C7), (6.12) and Theorem 5.7 by Van der Vaart (2000) that

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \widetilde{\boldsymbol{\beta}}. \tag{6.13}$$

Last, we write $\widehat{Se}(\widehat{\boldsymbol{\beta}},\widehat{c}) - \widetilde{Se}(\widetilde{\boldsymbol{\beta}},\widetilde{c}) = \{\widehat{Se}(\widehat{\boldsymbol{\beta}},\widehat{c}) - \widetilde{Se}(\widehat{\boldsymbol{\beta}},\widehat{c})\} + \{\widetilde{Se}(\widehat{\boldsymbol{\beta}},\widehat{c}) - \widetilde{Se}(\widetilde{\boldsymbol{\beta}},\widetilde{c})\}$, where $\widetilde{Se}$ denotes the true sensitivity, which is a probability. The law of large numbers

and the continuous mapping theorem then yield

$$\widehat{Se}(\widehat{\boldsymbol{\beta}}, \widehat{c}) \xrightarrow{p} \widetilde{Se}(\widetilde{\boldsymbol{\beta}}, \widetilde{c}). \tag{6.14}$$

6.2 Appendix for Aim 2

6.2.1 Simulation results under Scenario 4

Simulation results under Scenario 4 were summarized in Table 6.11 and Figure 6.1.

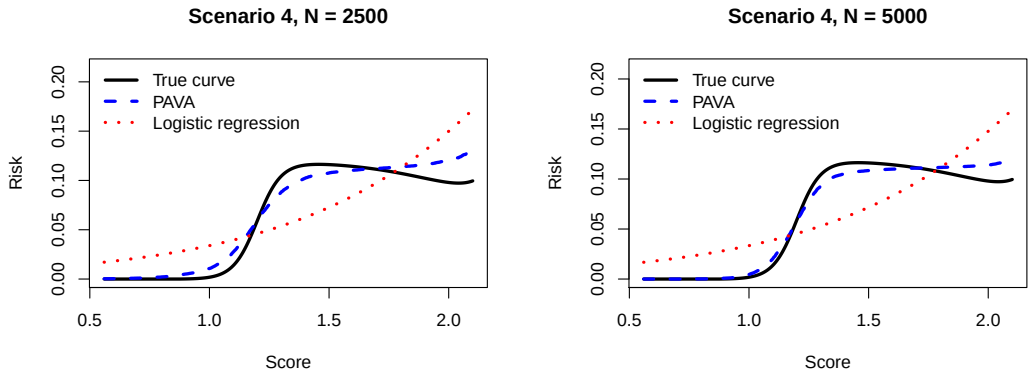

Figure 6.1: Estimated risk functions under the NCC design under Scenario 4.

6.2.2 Asymptotic Proofs

Regularity conditions

Denote the parameters $\boldsymbol{\xi} = (\boldsymbol{\beta}, \pi)$. We summarize the regularity conditions as follows.

1. The covariate vector $\boldsymbol{X}$ is in a bounded compact set $\mathcal{X}$ in $\mathbb{R}^p$.

2. The parameters $\boldsymbol{\xi}$ are in the space $\Theta = \mathcal{B} \times \mathcal{P}$, where $\mathcal{B}$ is a compact space in $\mathbb{R}^p$ and $\mathcal{P}$ is a space of monotonic nondecreasing functions of bounded variation.

3. The true value $\boldsymbol{\beta}_0$ is in the interior of the set $\mathcal{B}$. The true curve $\pi_0(\cdot)$ is continuously Fréchet differentiable on $\mathcal{K}$, where $\mathcal{K} = [\eta_1, \eta_2]$ is the support of $\boldsymbol{\beta}^T\mathbf{X}$, $\boldsymbol{\beta} \in \mathcal{B}$, and $\boldsymbol{X} \in \mathcal{X}$.

Due to identifiability constraint, we restrict that $\mathcal{B} = \{\boldsymbol{\beta} \mid \|\boldsymbol{\beta}\|_2 = 1, \boldsymbol{\beta} \in \mathbb{R}^p\}$; particularly, we treat $(\beta_2, \cdots, \beta_p)$ as unknown parameters and β_1 as a function of $(\beta_2, \cdots, \beta_p)$. Without loss of generality, we let $S(\boldsymbol{X}; \boldsymbol{\beta}) = \boldsymbol{\beta}^T \mathbf{X}$ throughout the proof, and the results can be easily extended to other forms of scoring. Let $U_1, \ldots, U_n$ be a random sample from a probability distribution P on a measurable space. Given a measurable function f, we define $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(U_i)$, $Pf = \int f dP$, and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)f = n^{-1} \sum_{i=1}^n \{f(\boldsymbol{U}_i) - Pf\}$.(van der Vaart and Wellner, 1996) The weighted log-likelihood function for a single subject with data $\boldsymbol{D}_i = (\boldsymbol{X}_i, Y_i)$ is

$$l_{\boldsymbol{\xi}}(\boldsymbol{D}_i) = \widehat{w}_i [Y_i \log\{\pi(\boldsymbol{\beta}^T \mathbf{X}_i)\} + (1 - Y_i) \log\{1 - \pi(\boldsymbol{\beta}^T \mathbf{X}_i)\}].$$

The estimator $\widehat{\boldsymbol{\xi}}_n = (\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)$ maximizes the weighted log-likelihood function $n^{-1} \sum_{i=1}^n l_{\boldsymbol{\xi}}(\boldsymbol{D_i})$.

Proof of consistency

Define a metric $\|\cdot\|$ on the parameter space Θ by $\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|^2 = \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \|\pi_1 - \pi_2\|^2$, where $\|\cdot\|_2$ is the Euclidean metric on a finite dimensional space, and $\|\pi\|^2 = \int_{\eta_1}^{\eta_2} \pi(u)^2 du$. The consistency of $\widehat{\boldsymbol{\xi}}_n$ can be proven by verifying the conditions in Theorem 5.7 for M-estimators.(Van der Vaart, 2000)

Given a known sampling weight, the likelihood function of the observed data from a single subject, denoted as $\boldsymbol{d} = (\boldsymbol{x}, y)$, can be written as

$$p_{\boldsymbol{\xi}}(\boldsymbol{d}) = \{\pi(\boldsymbol{\beta}^T \mathbf{x})\}^{wy} \{1 - \pi(\boldsymbol{\beta}^T \mathbf{x})\}^{w(1-y)}.$$

Also write $\widehat{p}_{\boldsymbol{\xi}}(\boldsymbol{d})$ as the estimated counterpart of $p_{\boldsymbol{\xi}}(\boldsymbol{d})$, obtained by replacing the true weight w with $\widehat{w}$. We consider the following class of functions that is related to Kullback-Leibler information,

$$m_{\boldsymbol{\xi}}(\boldsymbol{d}) = \log \left\{ \frac{p_{\boldsymbol{\xi}}(\boldsymbol{d}) + p_0(\boldsymbol{d})}{2p_0(\boldsymbol{d})} \right\}, \tag{6.15}$$

where $p_0(\boldsymbol{d}) = p_{\boldsymbol{\xi}_0}(\boldsymbol{d})$. Note that the function $m_{\boldsymbol{\xi}}(\boldsymbol{d}) - m_{\boldsymbol{\xi}_0}(\boldsymbol{d})$ is bounded from below, $m_{\boldsymbol{\xi}}(\boldsymbol{d}) - m_{\boldsymbol{\xi}_0}(\boldsymbol{d}) \geq -\log 2$, which is necessary for being the Glivenko-Cantelli class. Define

$M_n(\boldsymbol{\xi}) = \mathbb{P}_n \widehat{m}_{\boldsymbol{\xi}}$, $\widetilde{M}_n(\boldsymbol{\xi}) = \mathbb{P}_n m_{\boldsymbol{\xi}}$, and $M(\boldsymbol{\xi}) = E\widetilde{M}_n$, where $\widehat{m}_{\boldsymbol{\xi}} = \log\{(\widehat{p}_{\boldsymbol{\xi}} + \widehat{p}_0)/(2\widehat{p}_0)\}$.

We first show the uniform convergence of $\widetilde{M}_n(\boldsymbol{\xi})$ to $M(\boldsymbol{\xi})$ using the permanence properties of the Glivenko-Cantelli class (Section 2.6.5 of van der Vaart and Wellner (1996)) . We partition the likelihood into two parts according to the values of Y ($\{0, 1\}$). We then show that the following two classes of functions are P-Glivenko-Cantelli:

$$\mathcal{F}_1 = \left\{\{\pi(\boldsymbol{\beta}^T\mathbf{X})\}^w : \boldsymbol{\beta} \in \mathcal{B}, \pi \in \mathcal{P}\right\}; \tag{6.16}$$

$$\mathcal{F}_2 = \left\{\{1 - \pi(\boldsymbol{\beta}^T\mathbf{X})\}^w : \boldsymbol{\beta} \in \mathcal{B}, \pi \in \mathcal{P}\right\}. \tag{6.17}$$

Given these results, the density function $\{p_{\boldsymbol{\xi}}(\boldsymbol{d}) : \boldsymbol{\beta} \in \mathcal{B}, \pi \in \mathcal{P}\}$ is also Glivenko-Cantelli.

By regularity conditions, $|\boldsymbol{\beta}^T\boldsymbol{x}| < C_{\boldsymbol{\beta}}|\boldsymbol{x}|$ for $C_{\boldsymbol{\beta}} = \sup_{\boldsymbol{\beta}\in\mathcal{B}} |\boldsymbol{\beta}|$, the class of functions $\boldsymbol{x} \mapsto \boldsymbol{\beta}^T\boldsymbol{x}$ is therefore Glivenko-Cantelli. The class of functions $s \mapsto \pi(s)1(s \in \mathcal{K})$ is in the Glivenko-Cantelli class by Theorem 2.7.5 in van der Vaart and Wellner (1996). By the permanence properties of Glivenko-Cantelli classes, the class of functions $\boldsymbol{x} \mapsto \pi(\boldsymbol{\beta^T x})^w$ is Glivenko-Cantelli. It follows that the classes of functions $\mathcal{F}_1, \mathcal{F}_2$ are all Glivenko-Cantelli. By the regularity conditions, we have that $p_0(\boldsymbol{d})$ is bounded away from zero, and $P\{p_0^{-1}(\boldsymbol{D})\} < \infty$. Finally, the following class of functions is Glivenko-Cantelli as its envelope function is bounded, $\left\{\log\{(p_{\boldsymbol{\xi}} + p_0)/2p_0\} : \boldsymbol{\xi} \in \Theta\right\}$. Then we can apply the property of the Glivenko-Cantelli class to show that $\sup_{\boldsymbol{\xi}\in\Theta} |\widetilde{M}_n(\boldsymbol{\xi}) - M(\boldsymbol{\xi})| = \sup_{\boldsymbol{\xi}\in\Theta} |\mathbb{P}_n m_{\boldsymbol{\xi}} - Pm_{\boldsymbol{\xi}}| \xrightarrow{p} 0$.

Next, we verify the uniform convergence of $|M_n(\boldsymbol{\xi}) - \widetilde{M}_n(\boldsymbol{\xi})|$ to 0. Without loss of generality, we let $\widehat{w}_i$ be the weight under the NCC design. The results can be easily extended to the case-cohort design. Given the asymptotic behavior of the product limit estimator (Andersen and Gill, 1982), $\sup_t |\widehat{G}(t) - G(t)| = O_p(N^{-\frac{1}{2}})$, and we have $\max_i |\widehat{w}_i - w_i| = O_p(N^{-\frac{1}{2}})$. It follows that

$$\frac{\widehat{p}_0}{p_0} = \frac{\pi_0^{\widehat{w}y}(1-\pi_0)^{\widehat{w}(1-y)}}{\pi_0^{wy}(1-\pi_0)^{w(1-y)}} = \pi_0^{(\widehat{w}-w)y}(1-\pi_0)^{(\widehat{w}-w)(1-y)} \xrightarrow{p} 1. \tag{6.18}$$

Therefore, $\widehat{p}_0 - p_0 \xrightarrow{p} 0$ and $\log(\widehat{p}_0/p_0) \xrightarrow{p} 0$. Similarly, $\widehat{p}_{\boldsymbol{\xi}}/p_{\boldsymbol{\xi}} \xrightarrow{p} 1$, $\widehat{p}_{\boldsymbol{\xi}} - p_{\boldsymbol{\xi}} \xrightarrow{p} 0$, and $\log(\widehat{p}_{\boldsymbol{\xi}} - p_{\boldsymbol{\xi}} + \widehat{p}_0 - p_0 + p_{\boldsymbol{\xi}} + p_0) \xrightarrow{p} \log(p_{\boldsymbol{\xi}} + p_0)$. Combing these results, we have that

$$
\begin{aligned}
& M_n(\boldsymbol{\xi}) - \widetilde{M}_n(\boldsymbol{\xi}) = \mathbb{P}_n \widehat{m}_{\boldsymbol{\xi}} - \mathbb{P}_n m_{\boldsymbol{\xi}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{m}_{\boldsymbol{\xi}}(\boldsymbol{d}_i) - m_{\boldsymbol{\xi}}(\boldsymbol{d}_i)\right\} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\Big(\log\left\{\widehat{p}_{\boldsymbol{\xi}}(\boldsymbol{d}_i) + \widehat{p}_0(\boldsymbol{d}_i)\right\} - \log\left\{p_{\boldsymbol{\xi}}(\boldsymbol{d}_i) + p_0(\boldsymbol{d}_i)\right\} - \left[\log\{\widehat{p}_0(\boldsymbol{d}_i)\} - \log\{p_0(\boldsymbol{d}_i)\}\right]\Big) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left\{\log(\widehat{p}_{\boldsymbol{\xi}} - p_{\boldsymbol{\xi}} + \widehat{p}_0 - p_0 + p_{\boldsymbol{\xi}} + p_0) - \log(p_{\boldsymbol{\xi}} + p_0) - \log(\widehat{p}_0/p_0)\right\} \xrightarrow{p} 0. \qquad (6.19)
\end{aligned}
$$

Therefore, we have

$$
\sup_{\boldsymbol{\xi}\in\Theta}|M_n(\boldsymbol{\xi}) - M(\boldsymbol{\xi})| \le \sup_{\boldsymbol{\xi}\in\Theta}|M_n(\boldsymbol{\xi}) - \widetilde{M}_n(\boldsymbol{\xi})| + \sup_{\boldsymbol{\xi}\in\Theta}|\widetilde{M}_n(\boldsymbol{\xi}) - M(\boldsymbol{\xi})| \xrightarrow{p} 0. \qquad (6.20)
$$

Next, by the concavity of the logarithm and the definition of $\widehat{\boldsymbol{\xi}}_n$, we have

$$
\mathbb{P}_n \log\left(\frac{\widehat{p}_{\widehat{\boldsymbol{\xi}}_n} + \widehat{p}_0}{2}\right) \ge \frac{1}{2}\mathbb{P}_n(\log \widehat{p}_{\widehat{\boldsymbol{\xi}}_n} + \log \widehat{p}_0) \ge \mathbb{P}_n \log \widehat{p}_0.
$$

Hence,

$$
M_n(\widehat{\boldsymbol{\xi}}_n) = \mathbb{P}_n \widehat{m}_{\widehat{\boldsymbol{\xi}}_n} \ge \mathbb{P}_n \widehat{m}_{\boldsymbol{\xi}_0} = M_n(\boldsymbol{\xi}_0). \qquad (6.21)
$$

In addition, the true value $\boldsymbol{\xi}_0$ is always the maximum point of $Pm_{\boldsymbol{\xi}}$ as by Jensen's inequality,

$$
M(\boldsymbol{\xi}) = P\log\left(\frac{p_{\boldsymbol{\xi}} + p_0}{2p_0}\right) \le \log P\left(\frac{p_0 + p_0}{2p_0}\right) = M(\boldsymbol{\xi_0}) = 0, \qquad (6.22)
$$

and the equality sign can be achieved only when $p_{\boldsymbol{\xi}} = p_0$. Thus, $\sup_{\boldsymbol{\xi}: d(\boldsymbol{\xi},\boldsymbol{\xi}_0)\ge\epsilon} M(\boldsymbol{\xi}) < M(\boldsymbol{\xi}_0)$ is confirmed. Combining Equations (6.20), (6.21), (6.22), and Theorem 5.7 of Van der Vaart (2000), we have $\|\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0\| \to 0$ in probability as $n \to \infty$. As the model is identifiable on the parameter set Θ, we conclude that $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$, and $\widehat{\pi}_n(s) \xrightarrow{p} \pi_0(s)$ uniformly in s.

Proof of convergence rate

We first derive the convergence rate of $\widetilde{\boldsymbol{\xi}}_n$, which is the counterpart of $\widehat{\boldsymbol{\xi}}_n$ with known weights using Theorem 8.1 in van der Vaart (2002) (equivalently, Theorem 3.4.4 in van der Vaart and Wellner (1996)). Consider the class of functions defined in (6.15), $\mathcal{M}_\delta = \{m_{\boldsymbol{\xi}} - m_{\boldsymbol{\xi}_0} : \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| < \delta\}$. The first condition of Theorem 8.1 in van der Vaart (2002) is satisfied under the Hellinger distance (Lemma 4.2 of van de Geer (2000)),

$$\begin{aligned} E(m_{\boldsymbol{\xi}} - m_{\boldsymbol{\xi}_0}) &= E \log \left(\frac{p_{\boldsymbol{\xi}} + p_0}{2p_0} \right) \\ &\leq -h^2\{(p_{\boldsymbol{\xi}} + p_0)/2, p_0\} \lesssim -h^2(p_{\boldsymbol{\xi}}, p_0), \end{aligned}$$

where the Hellinger distance h is defined in Section 3.3 and $x \lesssim y$ is a shorthand of $x \leq K_0 y$ for a constant K_0. Here the first inequality is by the connection of the Kullback-Leibler divergence to the Hellinger distance. The second inequality follows from the fact that the Hellinger distance between any pair of densities f_1 and f_2 is equivalent to the Hellinger distance between f_1 and $(f_1 + f_2)/2$ (page 328 of van der Vaart and Wellner (1996)).

Lemma 8.6 of van der Vaart (2002) shows that the bracketing entropy of the class of functions $\mathcal{M}_\delta$ can be shown to be of order $1/\epsilon$, which implies

$$J_{[\,]}(\delta, \mathcal{F}, L_2) = \int_0^\delta \sqrt{K_1 \epsilon^{-1}} d\epsilon \lesssim \delta^{1/2},$$

where K_1 is a constant. Together with Lemma 8.2 of van der Vaart (2002), we have

$$E \sup_{f \in \mathcal{M}_\delta} |\mathbb{G}_n f| \lesssim \delta^{1/2} \left(1 + \frac{\delta^{1/2}}{\delta^2 \sqrt{n}} \right) = \delta^{1/2} + \frac{1}{\delta \sqrt{n}}.$$

Plugging in $\phi_n(\delta) = \sqrt{\delta} + 1/(\delta\sqrt{n})$ and $\delta_n = n^{-1/3}$ in Theorem 8.1 of van der Vaart (2002), we have

$$\frac{\phi_n(\delta_n)}{\delta_n^2} = \delta_n^{-3/2} + \frac{\delta_n^{-3}}{\sqrt{n}} = 2\sqrt{n}.$$

Then the rest of the conditions in Theorem 8.1 of van der Vaart (2002) are satisfied,

confirming the convergence rate is $n^{-1/3}$ for $\widetilde{\boldsymbol{\xi}}_n$. Combined with the asymptotic normality of $\widehat{w}_i$, we can show the convergence rate of $\widehat{\boldsymbol{\xi}}_n$ is $n^{-1/3}$ in the Hellinger distance.

Next, we follow the efficient score method by Bickel et al. (1993) to obtain the information bound of $\boldsymbol{\beta}$ given known weights. Note that model (3.1) belongs to the type II semiparametric regression model in Section 4.3 of Bickel et al. (1993). Let $\dot{\pi}(s) = d\pi(s)/ds$. By Proposition 4.3.2 of Bickel et al. (1993), $\dot{r}(\boldsymbol{X}|\boldsymbol{\beta}, \pi) = \partial\pi(\boldsymbol{\beta}^T\boldsymbol{X})/\partial\boldsymbol{\beta} = \dot{\pi}(\boldsymbol{\beta}^T\boldsymbol{X})\boldsymbol{X}$, and $\dot{I}(\boldsymbol{D}|\boldsymbol{\beta}, \pi) = \partial l_{\boldsymbol{\xi}}(\boldsymbol{D})/\partial\pi|_{\pi=\pi(\boldsymbol{\beta}^T\boldsymbol{X})} = w\{Y - \pi(\boldsymbol{\beta}^T\boldsymbol{X})\}/[\pi(\boldsymbol{\beta}^T\boldsymbol{X})\{1 - \pi(\boldsymbol{\beta}^T\boldsymbol{X})\}]$. Denote $\mathbb{B}_0$ as the sigma field generated by $\pi(\boldsymbol{\beta}^T\boldsymbol{X})$. Then the efficient score function of $\boldsymbol{\beta}$ takes the form $I^*(\boldsymbol{D}) = \widetilde{r}\dot{I}(\boldsymbol{D}|\boldsymbol{\beta}, \pi)$, where $\widetilde{r} = \dot{r}(\boldsymbol{X}|\boldsymbol{\beta}, \pi) - E_{\boldsymbol{\beta}_0,\pi_0}(\dot{r}(\boldsymbol{X}|\boldsymbol{\beta}, \pi)\dot{I}^2(\boldsymbol{D}|\boldsymbol{\beta}, \pi)|\mathbb{B}_0)/I(\boldsymbol{X})$, and $I(\boldsymbol{X}) = E_{\boldsymbol{\beta}_0,\pi_0}\{\dot{I}^2(\boldsymbol{D}|\boldsymbol{\beta}, \pi)|\mathbb{B}_0\}$. It follows that the information bound for $\boldsymbol{\beta}$ is $\Omega = E_{\boldsymbol{\beta}_0,\pi_0}\{I^*I^{*T}\}$.

Last, following Liu and Qin (2018), we establish the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$ by treating the monotone function $\pi(\cdot)$ as a nuisance parameter. We first establish the asymptotic normality of $\widetilde{\boldsymbol{\beta}}_n$ which is the counterpart of $\widehat{\boldsymbol{\beta}}_n$ with known weights by verifying the conditions in Theorem 6.20 of van der Vaart (2002). To verify the no-bias condition, we can follow the steps in Section 9.3 of van der Vaart (2002). Let $\psi_{\boldsymbol{\beta},\pi}$ be the score function which is an approximation of the efficient score function I^* by using a least favorable submodel. Then we can verify the no-bias condition,

$$|P_{\boldsymbol{\beta}_0,\pi_0}\psi_{\boldsymbol{\beta}_0,\widehat{\pi}_n}| \lesssim \int_{\eta_1}^{\eta_2} |\widehat{\pi}_n - \pi_0|^2(s)ds = O_p(n^{-2/3}).$$

Note that the class of functions $\boldsymbol{x} \mapsto \boldsymbol{\beta}^T\boldsymbol{x}$ belong to the Donsker class by Lemma 6.11 in van der Vaart (2002), and the functions $s \mapsto \pi(s)$ belongs to the Donsker class by Theorem 2.7.5 in van der Vaart and Wellner (1996). It follows that the functions $\boldsymbol{x} \mapsto \psi_{\boldsymbol{\beta},\pi}$ belongs to the Donsker class by the permanence of the Donsker property (van der Vaart and Wellner (1996), Section 2.10). Therefore the Donsker condition is verified. Hence, by Theorem 6.20 of van der Vaart (2002), $\widetilde{\boldsymbol{\beta}}_n$ is asymptotically normal with a convergence rate of $\sqrt{n}$. Combining this result with the normality of the estimated weights, we have the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$.

6.3 Appendix for Aim 3

Table 6.12: Simulation results for Scenario 1-3: estimated area under the ROC curve and difference in restricted mean survival time between low-risk and high-risk groups estimated using the proposed method, the Cox model with time-varying covariates (COX), the Cox model with time-varying covariates and coefficients (VCOX), the partly conditional cox model (PC), and the partly conditional cox model with time-varying coefficients (VPC) on validation data. Number of replicates was 1000, $n = 400$.

Scenario		s	w	Proposed		COX		VCOX		PC		VPC	
				Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	$AUC_{s,w}$	0	0.6	.824	.021	.793	.027	.783	.084	.735	.039	.824	.021
		0	1.2	.820	.021	.790	.027	.779	.082	.732	.039	.820	.021
		1.2	0.6	.642	.145	.603	.146	.565	.147	.631	.146	.644	.146
		1.2	1.2	.663	.088	.621	.091	.579	.095	.652	.088	.667	.088
		2.4	0.6	.724	.104	.662	.111	.716	.106	.702	.108	.724	.104
		2.4	1.2	.754	.069	.683	.077	.746	.073	.729	.073	.756	.068
	$RMST_{diff}$	0		1.994	.189	1.797	.227	1.721	.544	1.357	.287	1.996	.191
		1.2		.740	.204	.543	.215	.344	.240	.685	.204	.754	.205
		2.4		.640	.154	.462	.163	.620	.158	.576	.157	.646	.153
2	$AUC_{s,w}$	0	0.6	.730	.026	.672	.035	.589	.123	.643	.042	.727	.027
		0	1.2	.793	.027	.734	.038	.612	.160	.701	.048	.790	.028
		1.2	0.6	.807	.040	.795	.043	.799	.043	.796	.044	.787	.048
		1.2	1.2	.859	.038	.843	.043	.848	.042	.846	.043	.837	.047
		2.4	0.6	.866	.054	.799	.087	.794	.090	.822	.079	.775	.105
		2.4	1.2	.893	.051	.816	.086	.812	.089	.843	.076	.791	.102
	$RMST_{diff}$	0		.953	.110	.824	.119	.420	.601	.742	.146	.946	.110
		1.2		1.365	.173	1.301	.180	1.324	.185	1.316	.185	1.282	.202
		2.4		.945	.151	.720	.232	.705	.250	.794	.220	.652	.275
3	$AUC_{s,w}$	0	0.6	.796	.023	.795	.023	.666	.159	.795	.023	.795	.023
		0	1.2	.816	.025	.815	.025	.677	.171	.815	.025	.816	.025
		1.2	0.6	.834	.038	.829	.039	.821	.042	.828	.040	.824	.041
		1.2	1.2	.895	.034	.888	.037	.879	.041	.887	.038	.881	.040
		2.4	0.6	.889	.051	.867	.063	.826	.089	.863	.066	.840	.083
		2.4	1.2	.926	.042	.898	.060	.850	.095	.894	.062	.868	.084
	$RMST_{diff}$	0		1.076	.105	1.075	.105	.581	.574	1.074	.104	1.076	.104
		1.2		1.308	.174	1.284	.176	1.257	.183	1.280	.175	1.266	.181
		2.4		.872	.129	.800	.167	.681	.227	.787	.171	.722	.212

Table 6.13: Simulation results for Scenario 1-3 when measurements were irregular: estimated area under the ROC curve and difference in restricted mean survival time between low-risk and high-risk groups estimated using the proposed method, the Cox model with time-varying covariates (COX), the Cox model with time-varying covariates and coefficients (VCOX), the partly conditional cox model (PC), and the partly conditional cox model with time-varying coefficients (VPC) on validation data. Number of replicates was 1000, $n = 400$.

Scenario		s	w	Proposed		COX		VCOX		PC		VPC	
				Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	$AUC_{s,w}$	0	0.6	.825	.021	.792	.028	.746	.130	.734	.038	.824	.021
		0	1.2	.820	.021	.788	.028	.742	.129	.731	.038	.820	.021
		1.2	0.6	.650	.144	.615	.154	.570	.158	.639	.150	.643	.147
		1.2	1.2	.670	.090	.631	.095	.581	.102	.659	.094	.664	.092
		2.4	0.6	.735	.103	.673	.112	.729	.108	.712	.109	.737	.104
		2.4	1.2	.753	.066	.687	.075	.747	.070	.729	.072	.756	.066
	$RMST_{diff}$	0		2.003	.194	1.791	.230	1.483	.837	1.356	.279	2.005	.193
		1.2		.745	.195	.562	.206	.346	.266	.691	.204	.712	.202
		2.4		.644	.151	.473	.161	.625	.153	.578	.159	.649	.149
2	$AUC_{s,w}$	0	0.6	.730	.025	.679	.034	.579	.126	.645	.043	.727	.026
		0	1.2	.794	.027	.743	.036	.601	.162	.704	.049	.791	.028
		1.2	0.6	.801	.042	.792	.046	.795	.045	.794	.047	.785	.050
		1.2	1.2	.851	.041	.839	.047	.844	.045	.844	.045	.836	.049
		2.4	0.6	.860	.056	.783	.091	.798	.087	.817	.081	.772	.102
		2.4	1.2	.886	.052	.795	.092	.816	.087	.836	.083	.789	.105
	$RMST_{diff}$	0		.958	.111	.844	.120	.379	.608	.750	.149	.952	.111
		1.2		1.331	.176	1.283	.199	1.307	.193	1.312	.195	1.278	.206
		2.4		.926	.160	.665	.240	.716	.248	.775	.230	.649	.283
3	$AUC_{s,w}$	0	0.6	.796	.022	.795	.022	.658	.163	.795	.022	.795	.022
		0	1.2	.817	.025	.816	.025	.668	.176	.816	.025	.817	.025
		1.2	0.6	.833	.036	.830	.038	.820	.043	.828	.039	.824	.040
		1.2	1.2	.893	.033	.889	.035	.876	.043	.887	.036	.881	.039
		2.4	0.6	.887	.051	.866	.063	.826	.091	.863	.065	.837	.085
		2.4	1.2	.922	.042	.898	.059	.851	.095	.894	.063	.864	.084
	$RMST_{diff}$	0		1.080	.100	1.078	.098	.557	.592	1.077	.099	1.079	.099
		1.2		1.304	.158	1.293	.162	1.255	.176	1.287	.163	1.270	.167
		2.4		.863	.135	.798	.169	.685	.230	.785	.178	.713	.218

Table 6.14: summary of number of visits in each simulation scenario.

Scenario	Median	Min	Max
1	2	1	10
2	2	1	10
3	2	1	11

Table 6.11: Simulation results under the NCC study under Scenario 4 (ESE is the empirical standard error; ASE is the average of estimated standard error; CP is the empirical coverage probability of the 95% confidence interval).

				Proposed method					IPW-based logistic regression					Conditional logistic regression				
N	n	PARA	True	Mean	Bias	ESE	ASE	CP	Mean	Bias	ESE	ASE	CP	Mean	Bias	ESE	ASE	CP
2500	700	β_1	.707	.678	-.029	.112	.129	.953	.687	-.020	.090	.086	.939	.684	-.023	.097	.087	.923
		β_2	.000	.007	.007	.125	.122	.925	.006	.006	.109	.108	.948	.008	.008	.126	.123	.934
		β_3	.707	.709	.002	.101	.116	.952	.707	.000	.085	.080	.908	.706	-.001	.092	.083	.898
		$\pi(.99)$	.001	.010	.009	.018	-	-	.033	.032	.005	-	-	-	-	-	-	-
		$\pi(1.41)$	.116	.103	-.013	.016	-	-	.064	-.052	.007	-	-	-	-	-	-	-
		$\pi(1.84)$	.105	.115	.010	.014	-	-	.120	.015	.013	-	-	-	-	-	-	-
5000	1400	β_1	.707	.699	-.008	.058	.081	.984	.697	-.010	.059	.060	.955	.697	-.010	.064	.060	.932
		β_2	.000	.003	.003	.084	.088	.948	.002	.002	.080	.077	.929	.000	.000	.095	.087	.912
		β_3	.707	.706	-.001	.056	.074	.984	.708	.001	.057	.058	.943	.705	-.002	.062	.059	.917
		$\pi(.99)$	.001	.004	.003	.008	-	-	.033	.032	.004	-	-	-	-	-	-	-
		$\pi(1.41)$	.116	.106	-.010	.010	-	-	.063	-.053	.005	-	-	-	-	-	-	-
		$\pi(1.84)$	.105	.112	.007	.008	-	-	.118	.013	.010	-	-	-	-	-	-	-

Bibliography

DI Abrams, AI Goldman, C Launer, JA Korvick, JD Neaton, LR Crane, M Grodesky, S Wakefield, K Muth, S Kornegay, et al. Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant of or have failed zidovudine therapy. *New England Journal of Medicine*, 330(10):657–662, 1994.

Sameer T Amin, David A Morrow, Eugene Braunwald, Sarah Sloan, Charles Contant, Sabina Murphy, and Elliott M Antman. Dynamic timi risk score for stemi. *Journal of the American Heart Association*, 2(1):e003269, 2013.

Per Kragh Andersen and Richard David Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 10:1100–1120, 1982.

Miriam Ayer, H Daniel Brunk, George M Ewing, William T Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pages 641–647, 1955.

Richard E Barlow, David J Bartholomew, JM Bremner, and H Daniel Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression.* Wiley New York, 1972.

Michael J Best and Nilotpal Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1-3):425–439, 1990.

Peter J Bickel, CAJ Klaassen, Y Ritov, and JA Wellner. *Efficient and adaptive inference in semiparametric models*. Johns Hopkins University Press, Baltimore, 1993.

Norman Breslow, W Ahrens, and I Pigeot. Handbook of epidemiology, 2005.

Tianxi Cai and Yingye Zheng. Resampling procedures for making inference under nested case–control studies. *Journal of the American Statistical Association*, 108(504):1532–1544, 2013.

Antiretroviral Therapy Cohort Collaboration et al. Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies. *The Lancet*, 372(9635):293–299, 2008.

Russell Davidson and James G MacKinnon. Bootstrap tests: How many bootstraps? *Econometric Reviews*, 19(1):55–68, 2000.

Bénédicte Delcoigne, Edoardo Colzani, Michaela Prochazka, Giovanna Gagliardi, Per Hall, Michal Abrahamowicz, Kamila Czene, and Marie Reilly. Breaking the matching in nested case–control data offered several advantages for risk estimation. *Journal of clinical epidemiology*, 82:79–86, 2017.

Jean-Marie Dufour and Jan F Kiviet. Exact inference methods for first-order autoregressive distributed lag models. *Econometrica*, pages 79–104, 1998.

Ruth Etzioni, Margaret Pepe, Gary Longton, Chengcheng Hu, and Gary Goodman. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*, 19(3):242–251, 1999.

Naseema Gangat, Domenica Caramazza, Rakhee Vaidya, Geeta George, Kebede Begna, Susan Schwager, Daniel Van Dyke, Curtis Hanson, Wenting Wu, Animesh Pardanani, et al. Dipss plus: a refined dynamic international prognostic scoring system for primary myelofibrosis that incorporates prognostic information from karyotype, platelet count, and transfusion status. *Journal of Clinical Oncology*, 29(4):392–397, 2011.

Anne I Goldman, Bradley P Carlin, Lawrence R Crane, Cynthia Launer, Joyce A Korvick, Lawrence Deyton, and Donald I Abrams. Response of cd4 lymphocytes and clinical consequences of treatment using ddi or ddc in patients with advanced hiv infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 11(2):161–169, 1996.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

Holly Janes and Margaret S Pepe. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics*, 64(1): 1–9, 2008.

M.C. Jones. The performance of kernel density functions in kernel distribution function estimation. *Statistics & Probability Letters*, 9(2):129 – 132, 1990. ISSN 0167-7152.

Daniel Keizman, Yu-Xiao Yang, Maya Gottfried, Hadas Dresler, Ilan Leibovitch, Kevin Haynes, Ronac Mamtani, and Ben Boursi. The association between age-related macular degeneration and renal cell carcinoma: A nested case–control study. *Cancer Epidemiology and Prevention Biomarkers*, 26(5):743–747, 2017.

Ryung S Kim. A new comparison of nested case–control and case–cohort designs and methods. *European journal of epidemiology*, 30(3):197–207, 2015.

KK Landry, KS Alexander, NA Zakai, SE Judd, DO Kleindorfer, VJ Howard, G Howard, and M Cushman. Association of stroke risk biomarkers with stroke symptoms: the reasons for geographic and racial differences in stroke cohort. *Journal of Thrombosis and Haemostasis*, 15(1):21–27, 2017.

Florian Leitenstorfer and Gerhard Tutz. Generalized monotonic regression based on b-splines with an application to air pollution data. *Biostatistics*, 8(3):654–673, 2006.

FDK Liddell, JC McDonald, DC Thomas, and Stella V Cunliffe. Methods of cohort analysis: Appraisal by application to asbestos mining. *Journal of the Royal Statistical Society. Series A (General)*, 140(4):469–491, 1977.

Hao Liu and Jing Qin. Semiparametric probit models with univariate and bivariate current-status data. *Biometrics*, 74(1):68–76, 2018.

Patrick Mair, Kurt Hornik, and Jan de Leeuw. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32 (5):1–24, 2009.

Marlena Maziarz, Patrick Heagerty, Tianxi Cai, and Yingye Zheng. On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics*, 73(1):83–93, 2017.

Marlena Maziarz, Yingye Zheng, and Marshall Brown. *partlyconditional: Partly Conditional Logistic and Cox models*, 2018. http://mdbrown.github.io/partlyconditional/, https://github.com/mdbrown/partlycondtional.

Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

GA McIntyre. A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, 3(4):385–390, 1952.

Allison Meisner, Marco Carone, Margaret Pepe, and Kathleen F Kerr. Combining biomarkers by maximizing the true positive rate for a fixed false positive rate. 2017.

Hisashi Noma and Shiro Tanaka. Analysis of case-cohort designs with binary outcomes: Improving efficiency using whole-cohort auxiliary information. *Statistical methods in medical research*, 26(2):691–706, 2017.

Francesco Passamonti, Francisco Cervantes, Alessandro Maria Vannucchi, Enrica Morra, Elisa Rumi, Arturo Pereira, Paola Guglielmelli, Ester Pungolino, Marianna Caramella,

Margherita Maffioli, et al. A dynamic prognostic model to predict survival in primary myelofibrosis: a study by the iwg-mrt (international working group for myeloproliferative neoplasms research and treatment). *Blood, The Journal of the American Society of Hematology*, 115(9):1703–1708, 2010.

Rebecca Payne, Ming Yang, Yingye Zheng, Majken K Jensen, and Tianxi Cai. Robust risk prediction with biomarkers under two-phase stratified cohort design. *Biometrics*, 72(4):1037–1045, 2016.

Margaret S Pepe, Ziding Feng, Holly Janes, Patrick M Bossuyt, and John D Potter. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *Journal of the National Cancer Institute*, 100(20):1432–1438, 2008.

Margaret Sullivan Pepe, Ruth Etzioni, Ziding Feng, John D Potter, Mary Lou Thompson, Mark Thornquist, Marcy Winget, and Yutaka Yasui. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, 93(14):1054–1061, 2001.

R. L. Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11, 1986.

RL Prentice and NE Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

Jing Qin. *Biased Sampling, Over-identified Parameter Problems and Beyond*. Springer, 2017.

Jing Qin, Tanya P Garcia, Yanyuan Ma, Ming-Xin Tang, Karen Marder, and Yuanjia Wang. Combining isotonic regression and em algorithm to predict genetic risk under monotonicity constraint. *The annals of applied statistics*, 8(2):1182, 2014.

Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011.

T Robertson, F T Wright, and R L Dykstra. *Order restricted statistical inference*. Wiley, 1988.

Sherri Rose and Mark J Van der Laan. Why match? investigating matched case-control study designs with causal effect estimation. *The international journal of biostatistics*, 5(1), 2009.

Sven Ove Samuelsen. A psudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84(2):379–394, 1997.

Willi Sauerbrei, Patrick Royston, and Maxime Look. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal*, 49(3):453–473, 2007.

Kerstin Schütte, Christian Schulz, Alexander Link, and Peter Malfertheiner. Current biomarkers for hepatocellular carcinoma: Surveillance, diagnosis and prediction of prognosis. *World journal of hepatology*, 7(2):139, 2015.

Gary G Schwartz, Steinar Tretli, Linda Vos, and Trude E Robsahm. Prediagnostic serum calcium and albumin and ovarian cancer: A nested case-control study in the norwegian janus serum bank cohort. *Cancer epidemiology*, 49:225–230, 2017.

Weining Shen, Jing Ning, Ying Yuan, Anna S Lok, and Ziding Feng. Model-free scoring system for risk prediction with application to hepatocellular carcinoma study. *Biometrics*, 74(1):239–248, 2018.

Meredith S Shiels, Hormuzd A Katki, Allan Hildesheim, Ruth M Pfeiffer, Eric A Engels, Marcus Williams, Troy J Kemp, Neil E Caporaso, Ligia A Pinto, and Anil K Chaturvedi. Circulating inflammation markers, risk of lung cancer, and utility for risk stratification. *JNCI: Journal of the National Cancer Institute*, 107(10):djv199, 2015.

A Singal, ML Volk, A Waljee, R Salgia, P Higgins, MAM Rogers, and JA Marrero. Meta-analysis: surveillance with ultrasound for early-stage hepatocellular carcinoma in patients with cirrhosis. *Alimentary pharmacology & therapeutics*, 30(1):37–47, 2009.

Amit G Singal, Hari S Conjeevaram, Michael L Volk, Sherry Fu, Robert J Fontana, Fred Askari, Grace L Su, Anna S Lok, and Jorge A Marrero. Effectiveness of hepatocellular carcinoma surveillance in patients with cirrhosis. *Cancer Epidemiology and Prevention Biomarkers*, 2012.

Steven J Skates, Nora Horick, Yinhua Yu, Feng-Ji Xu, Andrew Berchuck, Laura J Havrilesky, Henk WA De Bruijn, Ate GJ Van Der Zee, Robert P Woolas, Ian J Jacobs, et al. Preoperative sensitivity and specificity for early-stage ovarian cancer when combining cancer antigen ca-125ii, ca 15-3, ca 72-4, and macrophage colony-stimulating factor using mixtures of multivariate normal distributions. *Journal of clinical oncology*, 22(20):4059–4066, 2004.

Ewout W Steyerberg, Frank E Harrell Jr, Gerard JJM Borsboom, MJC Eijkemans, Yvonne Vergouwe, and J Dik F Habbema. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, 54(8):774–781, 2001.

Nathalie C Støer and Sven Ove Samuelsen. Inverse probability weighting in nested case-control studies with additional matching—a simulation study. *Statistics in medicine*, 32(30):5328–5339, 2013.

Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL `https://CRAN.R-project.org/package=survival`. version 2.38.

Terry M. Therneau and Patricia M. Grambsch. *Modeling survival data: extending the Cox model*. Springer-Verlag New York, 2000.

R Houston Thompson, Bradley C Leibovich, Christine M Lohse, John C Cheville, Horst Zincke, Michael L Blute, and Igor Frank. Dynamic outcome prediction in patients with clear cell renal cell carcinoma treated with radical nephrectomy: the d-ssign score. *The Journal of urology*, 177(2):477–480, 2007.

Lu Tian, Lihui Zhao, and LJ Wei. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15(2):222–233, 2014.

Ilker Unal. Defining an optimal cut-point value in roc analysis: an alternative approach. *Computational and mathematical methods in medicine*, 2017, 2017.

Hajime Uno, Tianxi Cai, Lu Tian, and Lee-Jen Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and LJ Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

U.S. Cancer Statistics Working Group. U.S. cancer statistics data visualizations tool, based on november 2018 submission data (1999-2016). `www.cdc.gov/cancer/dataviz`, 2019.

S. A. van de Geer. *Applications of Empirical Process Theory*. Cambridge University Press, 2000.

Marc P van der Schroeff, Ewout W Steyerberg, Marjan H Wieringa, Ton PM Langeveld, Jan Molenaar, and Robert J Baatenburg de Jong. Prognosis: a variable parameter. dynamic prognostic modeling in head and neck squamous cell carcinoma. *Head & neck*, 34(1):34–41, 2012.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Aad W van der Vaart. *Lectures on Probability Theory and Statistics Part iii: Semi-parameric statistics*. Springer, 2002.

Aad W van der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, New York, 1996.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

Hans van Houwelingen and Hein Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.

Sholom Wacholder, Mitchell H Gail, David Pee, and Ron Brookmeyer. Alternative variance and efficiency calculations for the case-cohort design. *Biometrika*, 76(1):117–123, 1989.

Qingxiang Yan, Leonidas E Bantis, Janet L Stanford, and Ziding Feng. Combining multiple biomarkers linearly to maximize the partial area under the roc curve. *Statistics in medicine*, 37(4):627–642, 2018.

Ehsan Zamanzade and M Mahdizadeh. Using ranked set sampling with extreme ranks in estimating the population proportion. *Statistical methods in medical research*, page 0962280218823793, 2019.

Ehsan Zamanzade and Michael Vock. Variance estimation in ranked set sampling using a concomitant variable. *Statistics & Probability Letters*, 105:1–5, 2015.

D. Zeng and D. Y. Lin. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, 102:1387–1396, 2007.

Zheng Zhang, Ying Lu, and Lu Tian. On feature ensemble optimizing the sensitivity and partial roc curve. *Statistica Sinica*, 29:1395–1418, 2019.

Yingye Zheng and Patrick J Heagerty. Partly conditional survival models for longitudinal data. *Biometrics*, 61(2):379–391, 2005.