DESIGN AND ANALYSIS FOR TWO-PHASE STUDIES WITH SURVIVAL DATA

by

Xu Cao

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2024

Approved by:

Dr. Qingning Zhou

Dr. Yanqing Sun

Dr. Eliana Christou

Dr. Rob McGregor

ABSTRACT

XU CAO. Design and Analysis for Two-Phase Studies with Survival Data. (Under the direction of DR. QINGNING ZHOU)

Large cohort studies under simple random sampling could be prohibitive to conduct with a limited budget for epidemiological studies seeking to relate a failure time to some exposure variables that are expensive to obtain. In this case, two-phase studies are desirable. Failure-time-dependent sampling (FDS) is a commonly used cost-effective sampling strategy in such studies. To enhance study efficiency upon FDS, counting the auxiliary information of the expensive variables into both sampling design and statistical analysis is necessary.

Chapter 2 discusses the semiparametric inference for a two-phase failure-time-auxiliary-dependent sampling (FADS) design that allows the probability of obtaining the expensive exposures to depend on both the failure time and cheaply available auxiliary variables. To account for the sampling bias, we develop a semiparametric maximum pseudo-likelihood approach for inference and a nonparametric bootstrap procedure for variance estimation. The proposed estimator of regression coefficients is shown to be consistent and asymptotically normal. The simulation studies indicate that the proposed method works well in practical settings and is more efficient than other competing sampling schemes or methods. The analyses of two real data sets are provided for illustration.

In survival analysis, it's commonly assumed that all subjects in a study will eventually experience the event of interest. However, this assumption may not hold in various scenarios. For example, when studying the time until a patient progresses or relapses from a disease, those who are cured will never experience the event. These subjects are often labeled as "long-term survivors" or "cured", and their survival time is treated as infinite. When survival data include a fraction of long-term survivors, censored observations encompass both uncured individuals, for whom the event wasn't observed, and cured individuals who won't experience the event. Consequently, the cure status is unknown, and survival data comprise a mixture of cured and uncured individuals that can't be distinguished beforehand. Cure models are survival models designed to address this characteristic.

Chapter 3 considers the generalized case-cohort design for studies with a cure fraction. Under this design, the expensive covariates are measured only for a subset of the study cohort, called subcohort, and for all or a subset of the remaining subjects outside the subcohort who have experienced the event, called cases. We propose a two-step estimation procedure under the semiparametric transformation mixture cure models. We first develop a sieve maximum weighted likelihood method based only on the complete data and also devise an EM algorithm for implementation. We then update the resulting estimator via a working model between the outcome and cheap covariates or auxiliary variables using the full data. We show that the proposed estimator is consistent and asymptotically normal, regardless of whether the working model is correctly specified or not. We also propose a weighted bootstrap procedure for variance estimation. Extensive simulation studies demonstrate the superior performance of the proposed method in finite-sample. An application to the National Wilms' Tumor Study is provided for illustration.

A few directions for future research are discussed in Chapter 4.

DEDICATION

To my parents, Qingwen Cao and Guimei Wang, and to my beloved Yuhui Gong, who is statistically significant.

ACKNOWLEDGEMENTS

I extend my deepest and most sincere gratitude to my advisor, Dr. Qingning Zhou, for her exceptional guidance, generous support, and boundless patience throughout the development of this dissertation. Her continual encouragement and inspiring mentorship have been instrumental in completing this work and have set me on a promising path in my research endeavors.

Additionally, I am immensely grateful to my advisory committee members: Dr. Yanqing Sun, Dr. Eliana Christou, and Dr. Rob McGregor, for their invaluable insights, constructive feedback, and unwavering support throughout this journey.

I am also indebted to the faculty members of our department for their dedication to providing exceptional courses and their prompt assistance whenever needed during my studies. Furthermore, I express my gratitude to my friends, Jing Xu in particular, at UNC Charlotte for their companionship, which has enriched my experience and made my time here more vibrant and enjoyable.

Lastly, I dedicate this dissertation to my parents, whose love, support, and endless encouragement have shaped me into the person I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Survival Analysis

1.1.1 Survival Data

Survival data, also known as failure time data, is pervasive across various fields, including medicine, social sciences, and finance. This type of data tracks the time until a specific event occurs, such as death or machine failure. However, in real-world scenarios, the occurrence of the event of interest may not be observed for all study subjects due to incomplete follow-up, leading to censoring. Censoring comes in different forms, with right-censored data being among the most common. In right-censored data, the event of interest occurs after a certain observation period due to factors like the end of study or loss of participant follow-up. Another type of censoring is interval-censoring, where the exact event time is unknown but known to have occurred within a specific time frame, such as in periodic examinations or screenings.

One example of right censoring is the data set on incident diabetes from the Atherosclerosis Risk in Communities (ARIC) study. Each participant in this study was followed up every three years starting from 1987 and was examined for the events of interest at the follow-up visits. Subjects who did not develop diabetes before the end of the study or before they were lost to follow-up are classified as right-censored. In Chapter 2, we will study the association of high-sensitivity C-Reactive Protein (hs-CRP) level with time to incident diabetes after adjusting for other risk factors or confounding variables.

In survival analysis, it's commonly assumed that all subjects in a study will eventually experience the event of interest. However, this is not always the case. For example, when studying the time until a patient progresses or relapses from a disease, those who are cured will never experience the event. These subjects are often labeled as "long-term survivors" or "cured", and their survival time is treated as infinite. Since it's impractical to follow all individuals until they experience the event of interest, survival data typically involve right censoring, where only a lower bound of the survival time is known for some individuals. When survival data include a fraction of long-term survivors, censored observations encompass both uncured individuals, for whom the event wasn't observed, and cured individuals who won't experience the event. Consequently, the cure status is unknown, and survival data comprise a mixture of cured and uncured individuals that can't be distinguished beforehand.

A typical field in which the cure fraction of survival data is usually considered is cancer studies. One example is that in the National Wilms' Tumor Study on a rare childhood kidney cancer, we are aware that a certain number of patients will never experience a occurrence of the disease. In addition to contextual evidence supporting the existence of a cured fraction, the presence of a stable plateau in the Kaplan and Meier (1958) estimator of the survival function, alongside a considerable number of censored observations, suggests the existence of a cured fraction. Figure 3.1, illustrating this estimator for the time to relapse among patients with the kidney cancer (Breslow and Chatterjee, 1999), provides a compelling illustration of survival data with a cure fraction. In Chapter 3, we will investigate the relation between histology type and time to relapse among this kidney cancer patients with a cure fraction considered, while accounting for other risk factors or confounding variables.

1.1.2 Survival Models

In the statistical analysis of failure time data, the primary objective often involves estimating either the cumulative distribution function (CDF) or the survival function of the failure time. Here, we denote $F(t) = P(T \leq t)$ as the CDF of the failure time

T, and $S(t) = 1 - F(t)$ as the survival function.

When covariates are present, the primary interest often lies in examining their effect on the failure time. Regression analysis is commonly employed to quantify this effect or predict survival probabilities for new individuals. In this section, we will explore several semiparametric regression models frequently utilized in survival analysis, along with corresponding inference procedures. In what follows, let Z represent a vector of covariates, which may include variables such as treatment indicators, age, gender, and income. Additionally, let β denote the vector of regression parameters.

1.1.2.1 Proportional Hazards Model

The proportional hazards model, first proposed by Cox (1972), commonly known as the Cox model, describes that effect of covariates acts multiplicatively on the hazard function of the failure time T, and that the hazard functions under two different sets of covariates are proportional. Particularly, it assumes that the hazard function takes the form

$$\lambda(t|Z) = \lambda_0(t) \exp(z'\beta)$$

given $Z = z$, where $\lambda_0(t)$ is a baseline hazard function.

Over the past three decades, the Cox model has emerged as the most prevalent regression model in survival analysis. A key factor contributing to its widespread adoption is the availability of a simple and efficient inference method for β, known as the partial likelihood approach, tailored specifically for right-censored data (Cox (1975)). Unlike other methods, the partial likelihood function used in this procedure involves only β, eliminating the need to handle $\lambda_0(t)$. Andersen and Gill (1982) offered a simple and elegant proof of the asymptotic properties of the estimator for β using counting processes and martingale theory. Chapter 2 discusses the regression analysis of survival data using the Cox model in our studies.

1.1.2.2 Proportional Odds Model

The proportional odds (PO) model, first considered by McCullagh (1980), stands as another prevalent regression model in survival analysis. It can be represented by the equation:

$$\log(\frac{F(t|z)}{1-F(t|z)}) = \alpha_0(t) + z'\beta,$$

where $F(t|z)$ is the CDF of the failure time T give $Z = z$, and $\alpha_0(t)$ the the unknown baseline log-odds monotone increasing function. Bennett (1983) provided a non-parametric maximum likelihood estimation for the survival function. Brant (1990) proposed an approach to assess the goodness of fit of such models by comparing fits to the binary logistic models that are subsidiary to the overall model. Fagerland and Hosmer (2013) examined goodness-of-fit tests for this model. Mao and Wang (2010) introduced a semiparametric efficient estimation for a class of generalized proportional odds cure models.

1.1.2.3 Transformation Models

The regression models outlined above represent specific functional forms for the effect of covariates. However, there are instances where a more flexible model is desired. One such model is the linear transformation model, which defines the relationship between the failure time T and the covariates Z as follows:

$$u(T) = Z'\beta + \epsilon,$$

where u represents an unknown strictly increasing function, while ϵ follows a known distribution. This model results in various semiparametric models, contingent upon the specification of the distribution of ϵ. For instance, if ϵ follows an extreme value distribution, it leads to the proportional hazards model, whereas if ϵ follows the standard logistic distribution, it yields the proportional odds model. Among other researchers, Lu and Ying (2004) and Mao and Wang (2010) delved into the transformation model incorporating a cure fraction, wherein they assumed a class of linear transformation models for the survival time of uncured individuals.

Another transformation model proposed by Zeng and Lin (2006), offering considerable flexibility, assumes that the cumulative hazard function of T takes the form

$$\Lambda(t|z) = G(\Lambda_0(t)\exp(z'\beta))$$

given $Z = z$, where G is a prespecified strictly increasing function and Λ_0 is an unspecified nondecreasing function. This model exhibits high flexibility due to the different choices for G, allowing it to encompass numerous commonly employed models as special cases. For instance, in the Box-cox transformations,

$$G(x) = [(1+x)^\rho - 1]/\rho, \ \ \rho > 0,$$

where $\rho > 0$ corresponds to $G(x) = \log(1+x)$, or in the logarithmic transformations,

$$G(x) = \log(1+rx)/r,$$

where $r = 0$ corresponds to $G(x) = x$. By setting $\rho = 1$ or $r = 0$, we obtain the proportional hazards model, while setting $\rho = 0$ or $r = 1$ yields the proportional odds model. Furthermore, it is evident that under this model, T adheres to the linear transformation model $\log\Lambda_0(T) = Z'\beta + \log G^{-1}(\log\epsilon_0)$, where ϵ_0 follows a uniform distribution over [0, 1]. Zeng and Lin (2006) explored a more generalized version of this model for counting processes and investigated its maximum likelihood estimation under right-censoring. In Chapter 3, we incorporate the latter transformation model into our studies with a cure fraction.

1.1.2.4 Cure Models

Cure models, including the nonmixture cure model and the mixture cure model, are often explored in scenarios where survival data contain a mixture of cured and uncured individuals that cannot be distinguished beforehand, as detailed in the previous section.

The mixture cure model stands as a prevalent method for analyzing survival data that incorporates a cure fraction. It is a mixture of two separate regression models, one for the cure rate of the nonsusceptible population and another for the survival function of the susceptible population. Specifically, it is defined as

$$S(t|z) = p + (1 - p)S_0(t|z)$$

given $Z = z$, where p is the proportion of "long-term survivors" or "cured patient" and $S_0(t|z)$ the survival function for the susceptible individuals. Various methods have been considered for the conditional survival function for the uncured subjects. Farewell (1982) originally proposed parametric models. A semiparametric approach utilizing a Cox (1972) proportional hazards (PH) model was offered by Kuk and Chen (1992), Sy and Taylor (2000), and Kuk and Chen (2008).

The nonmixture cure model is an alternative approach for analyzing survival data with a cure fraction while preserving the proportional hazards structure across the entire population. It delineates the relationship between the failure time T and the covariates Z as follows:

$$S(t|z) = p^{F_0(t|z)} = \exp[\ln(p)F_0(t|z)],$$

where $F_0(t|z) = 1 - S_0(t|z)$ is the CDF of failure time T give $Z = z$. It offers a clear interpretation of how covariates impact the probability of cure, as demonstrated by Tsodikov (1998) and Tsodikov *et al.* (2003). Yakovlev and Tsodikov (1996), along with Chen *et al.* (1999), provided a biological derivation of this model. Chapter 3 discusses the semiparametric transformation mixture cure models in our study.

1.2 Two-Phase Sampling Designs

In many epidemiological studies, the outcomes of interest are times to failure events, such as cancer, heart disease, and HIV infection, and often much of the cost is spent on obtaining the measurements of the main exposure variables, e.g., biomarkers that rely on bioassay or genetic analysis to ascertain or medical records that require labor-intensive chart review. When the exposure variables are difficult or expensive to obtain, large cohort studies under simple random sampling could be prohibitive to conduct for investigators with a limited budget. Alternative cost-effective sampling designs together with efficient and robust inference procedures are desirable. Two-phase sampling designs are commonly used in practice to reduce cost and enhance study efficiency. In this section, we introduce several commonly used two-phase sampling designs.

1.2.1 Outcome-Dependent Sampling Designs

1.2.1.1 Case-control Studies

A case-control study is a type of observational study design utilized in epidemiology to explore the connection between exposure variables (such as risk factors or treatments) and outcomes of interest (such as diseases or health conditions). In such studies, researchers identify individuals with the outcome of interest (cases) and juxtapose them with individuals lacking the outcome (controls). Breslow (1982) delved into the design and analysis of case-control studies. In Vandenbroucke and Pearce (2012), case-control studies concerning disease incidence were discussed. Furthermore, Schildcrout and Rathouz (2010) proposed methods for the design and analysis of case-control and stratified case-control studies for binary outcomes.

1.2.1.2 ODS for Continuous Outcome

Outcome-dependent sampling (ODS), exemplified by the case-control study, is a retrospective sampling strategy that enhances study efficiency and reduces costs. It allows investigators to observe exposure based on value of the outcome as discussed in Weinberg and Wacholder (1993) and Whittemore (1997). Recent research, like Zhou *et al.* (2002) and Chatterjee *et al.* (2003), has extended ODS to encompass continuous outcomes. The core concept of this approach is to allocate resources to a subset of the population that provides the most informative data regarding the exposure-response relationship (e.g. Song *et al.* (2009); Zhou *et al.* (2011b)).

1.2.1.3 Case-cohort Studies

For analyzing failure time outcomes, Prentice (1986) proposed a case-cohort design, wherein costly exposure variables are gathered only for a randomly selected subset of the study cohort, termed the subcohort, along with all individuals who experience the failure event by a specified time, referred to as cases. Since its inception, the case-cohort design has garnered attention from various researchers, including Chen and Lo (1999), Cai and Zeng (2004), and Lu and Tsiatis (2006). The original case-cohort study is mainly used for rare events. When the failure event of interest is non-rare or moderately rare, Chen (2001), Cai and Zeng (2007), Kang and Cai (2009) among considered a generalized case-cohort design. This approach involves acquiring expensive exposure measurements for a subcohort and a selected subset, rather than all, of the remaining cases outside the subcohort. In Chapter 3, we explore a generalized case-cohort study within the framework of semiparametric transformation mixture cure models.

1.2.2 Outcome-Auxiliary-Dependent Sampling Designs

In practice, cheap auxiliary variables that are highly correlated with the expensive exposure variable are often available. The auxiliary variable is defined as the surrogate information that relates to the expensive variable but provides no additional information to the regression model when the expensive variable is known. For example, in the National Wilms' Tumor Study, it is of interest to evaluate the association of tumor histological type with time to disease relapse. The histological type can be examined by a local pathologist or an experienced pathologist from a central facility. The central examination tends to be more accurate but is more expensive and time-consuming. Thus, the central histological type can be treated as the expensive exposure variable, while the local type can serve as a cheap auxiliary variable. Among others, Wang and Zhou (2006), Wang *et al.* (2009), Wang and Zhou (2010), and Zhou *et al.* (2011a) considered two-phase designs that make use of auxiliary information. However, these works were focused on continuous and categorical outcomes. In Chapter 2, we develop a two-phase failure-time-auxiliary-dependent sampling (FADS) design.

1.3 Outline of the Dissertation

The remainder of this dissertation is organized as follows. In Chapter 2, we develop a two-phase failure-time-auxiliary-dependent sampling (FADS) design and propose a semiparametric maximum pseudo-likelihood method to analyze the resulting data. The two-phase FADS design allows the probability of obtaining the expensive exposure measurements at the second phase to depend on the values of the failure time and the auxiliary variable observed at the first phase. In addition, we propose a new semiparametric maximum pseudo-likelihood estimation method to reap the benefits gained by the two-phase FADS design, and develop a nonparametric bootstrap procedure for inference. The simulation studies indicate that the proposed method works well in practical settings and is more efficient than other competing sampling schemes or methods. The analyses of two real data sets are provided for illustration.

In Chapter 3, we discuss the generalized case-cohort design for studies with a cure fraction. Under this design, the expensive covariates are measured only for a subset of the study cohort, called subcohort, and for all or a subset of the remaining subjects outside the subcohort who have experienced the event, called cases. We propose a two-step estimation procedure under the semiparametric transformation mixture cure models. We first develop a sieve maximum weighted likelihood method based only on the complete data and also devise an EM algorithm for implementation. We then update the resulting estimator via a working model between the outcome and cheap covariates or auxiliary variables using the full data. We show that the proposed estimator is consistent and asymptotically normal, regardless of whether the working model is correctly specified or not. We also propose a weighted bootstrap procedure for variance estimation. Extensive simulation studies demonstrate the superior performance of the proposed method in finite-sample. An application to the National Wilms' Tumor Study is provided for illustration.

In Chapter 4, several directions for future research are described.

CHAPTER 2: SEMIPARAMETRIC INFERENCE FOR A TWO-PHASE FAILURE-TIME-AUXILIARY-DEPENDENT SAMPLING DESIGN

2.1 Introduction

As discussed in Chapter 1, in many epidemiological studies, when the exposure variables are difficult or expensive to obtain, large cohort studies under simple random sampling could be prohibitive to conduct for investigators with a limited budget. Alternative cost-effective sampling designs together with efficient and robust inference procedures are desirable. Two-phase study designs are commonly used in practice to reduce cost and enhance study efficiency. Typically, at the first phase of a two-phase study, a large random sample is drawn to collect the outcome and cheap covariates or auxiliary variables; at the second phase, the measurements of expensive covariates are obtained for a subset of the first-phase sample. There is an extensive literature on two-phase study designs, particularly on how to draw the second-phase sample.

For the failure time outcome, Prentice (1986) proposed a case-cohort design, where the expensive exposure variables are collected only for a simple random sample from the study cohort, called subcohort, and for all subjects who have experienced the failure event by a specified time, called cases. Since its proposal, the case-cohort design has been studied by many authors, including Chen and Lo (1999), Cai and Zeng (2004), Lu and Tsiatis (2006), Breslow and Wellner (2007), and Marti and Chavance (2011). The original case-cohort design is mainly used for rare events. When the failure event of interest is non-rare or not so rare, Chen (2001), Cai and Zeng (2007) and Kang and Cai (2009) among others considered a generalized case-cohort design, where the expensive exposure measurements are obtained for a subcohort and for a subset, instead of all, of the remaining cases outside the subcohort. Ding *et al.*

(2014) developed an outcome-dependent sampling (ODS) design by enriching a simple random sample with some selected cases that are believed to be more informative to the exposure-failure-time relationship based on the values of their observed failure times. These works did not incorporate auxiliary information into the study designs and analyses.

In practice, cheap auxiliary variables that are highly correlated with the expensive exposure variable are often available. In addition to the example of the National Wilms' Tumor Study mentioned in Chapter 1, another example arises from the Duke Lung Cancer Study that evaluates the effect of epidermal growth factor receptor (EGFR) genetic mutations on tumor response to EGFR-targeted therapy for patients with nonsmall cell lung cancer. Genetic assay on EGFR mutations is very expensive. The EGFR mutation score, a composite score indicating the likelihood of EGFR mutations estimated from patients baseline characteristics, provides auxiliary information about EGFR mutations. It is desirable to incorporate such available auxiliary information in designing and analyzing two-phase studies so as to further reduce cost and improve study efficiency. Among others, Wang and Zhou (2006), Wang *et al.* (2009), Wang and Zhou (2010), and Zhou *et al.* (2011a) considered two-phase designs that make use of auxiliary information. However, these works were focused on continuous and categorical outcomes.

In this chapter, we develop a two-phase failure-time-auxiliary-dependent sampling (FADS) design and propose a semiparametric maximum pseudo-likelihood method to analyze the resulting data. The two-phase FADS design allows the probability of obtaining the expensive exposure measurements at the second phase to depend on the values of the failure time and the auxiliary variable observed at the first phase. This design is innovative in that it allows readily available surrogate or auxiliary variables to be part of the design of an efficient study. It could play a significant role in many studies with a limited budget. In addition, we propose a new semiparametric maximum pseudo-likelihood estimation method to reap the benefits gained by the two-phase FADS design, and develop a nonparametric bootstrap procedure for inference. The proposed method can also be used to analyze data from the two-phase FDS design where the selection of the second-phase sample depends only on the failure time. In contrast to Ding *et al.* (2014), our method makes use of both first-phase and second-phase samples as well as the auxiliary information, thus yields more efficient estimation.

The remainder of this chapter is organized as follows. In Section 2.2, we describe the two-phase FADS design and present the proposed method for analyzing the resulting data. We also establish the asymptotic properties of the proposed estimator and develop a nonparametric bootstrap procedure for variance estimation. In Section 2.3, we conduct simulation studies to investigate the performance of the FADS design and the proposed method in finite samples and also to compare with the competing designs or methods. In Section 2.4, we illustrate the proposed design and method using the ARIC Study and the National Wilms' Tumor Study. We conclude with brief discussion in Section 2.5. The proofs of the asymptotic properties of the proposed estimator are given in the Appendix A.

2.2 Design and Method

2.2.1 Two-Phase FADS Design

Let $\tilde{T}$ denote the failure time, C the censoring time, $T = \min(\tilde{T},\ C)$ the observed time, $\Delta = I(\tilde{T} \leq C)$ the indicator of failure, $Y(t) = I(T \geq t)$ the at-risk process, X the expensive covariates, Z the other adjustment covariates that are available, and W the auxiliary variables of X. Assume that $\tilde{T}$ and C are conditionally independent given X and Z, and $\tilde{T}$ follows the proportional hazards model with the conditional cumulative hazard function given X and Z as follows:

$$\Lambda(t|X, Z) = \Lambda(t) \exp\{\beta_1' X + \beta_2' Z\}, \tag{2.1}$$

where $\Lambda(t)$ is an unspecified cumulative baseline hazard function and $\beta = (\beta_1, \beta_2')'$ is a p-dimensional vector of regression coefficients.

Let $\mathcal{T} \times \mathcal{W}$ denote the domain of (T, W). We partition $\mathcal{T}$ into J mutually exclusive and exhaustive strata: $A_j = (a_{j-1}, a_j]$, $j = 1, \ldots, J$, where $a_0, a_1, \ldots, a_J$ are known constants such that $0 = a_0 < a_1 < \cdots < a_{J-1} < a_J = \tau$ with τ being the length of study. We also partition $\mathcal{W}$ into L mutually exclusive and exhaustive strata: $B_l = (b_{l-1}, b_l]$, $l = 1, \ldots, L$, where $b_0, b_1, \ldots, b_L$ are known constants satisfying $-\infty = b_0 < b_1 < \cdots < b_{L-1} < b_L = \infty$. Thus, we have partitioned $\mathcal{T} \times \mathcal{W}$ into $J \times L$ mutually exclusive and exhaustive rectangles $A_j \times B_l$ for $j = 1, \ldots, J$ and $l = 1, \ldots, L$. For simplicity, we rewrite these rectangles as D_k for $k = 1, \ldots, K$ such that $\{D_k : k = 1, \ldots, K\} = \{A_j \times B_l : j = 1, \ldots, J \text{ and } l = 1, \ldots, L\}$ and $\mathcal{T} \times \mathcal{W} = \cup_{j=1}^{J} \cup_{l=1}^{L} A_j \times B_l = \cup_{k=1}^{K} D_k$, where $K = J \times L$. According to the failure status, we further partition the data into $(K + 1)$ strata:

$$S_k = D_k \cap \{\Delta = 1\},\ k = 1, \ldots, K \ \text{ and } \ S_{K+1} = \{\Delta = 0\}.$$

The proposed two-phase FADS design is as follows: at the first phase, we take a cohort of size N from the underlying population on which $\{T, \Delta, W, Z\}$ are observed; at the second phase, we observe X on a simple random sample (SRS), indexed by $\tilde{V}_0$, of the first-phase cohort and on supplemental samples, indexed by $\{\tilde{V}_1, \ldots, \tilde{V}_K\}$, taken by simple random sampling from the K failure strata $\{S_1, \ldots, S_K\}$ of the cohort that are outside $\tilde{V}_0$, respectively. Note that $\tilde{V}_j$ is allowed to be empty, meaning that no supplemental sample is taken from the stratum S_j. We refer to the data with X observed as the validation sample, indexed by $V = \cup_{k=0}^{K} \tilde{V}_k$, and the data without X observed as the nonvalidation sample, indexed by $\bar{V}$. Then the data structure for the two-phase FADS design can be written as:

$$\text{Nonvalidation Sample: } (T_i, \Delta_i, W_i, Z_i), \;\; i \in \bar{V}$$

$$\text{Validation Sample: } \begin{cases} (T_i, \Delta_i, X_i, W_i, Z_i), \;\; i \in \tilde{V}_0 \\ (T_i, \Delta_i, X_i, W_i, Z_i \,|\, (T_i, \Delta_i, W_i) \in S_k), \;\; i \in \tilde{V}_k, \;\; k = 1, \ldots, K. \end{cases}$$

The FADS design allows one to oversample certain segments of the population that are believed to be more informative to the relationship between T and X, such as the extreme strata of (T, W). It provides more flexibility than the FDS design by allowing the probability of obtaining the second-phase sample to depend on the auxiliary variable W as well. Figure 2.1 illustrates the proposed FADS design where $J = L = 3$ and at the second phase, in addition to SRS, four supplemental samples are taken from the four "corner" strata S_1, S_3, S_7 and S_9 that consist of the extreme values of (T, W).

2.2.2 Estimation and Inference

Now we consider the estimation and inference of the regression parameters β in model (2.1). The likelihood function based on the observed data can be written as

$$\begin{aligned} L(\beta, \Lambda, G) \;&=\; \left\{ \prod_{k=0}^{K} \prod_{i \in \tilde{V}_k} f_{\beta,\Lambda}(T_i, \Delta_i | X_i, Z_i) g(X_i | W_i, Z_i) \right\} \left\{ \prod_{k=1}^{K+1} \prod_{j \in \bar{V}_k} f_{\beta,\Lambda,G}(T_j, \Delta_j | W_j, Z_j) \right\} \\ &= \left\{ \prod_{k=0}^{K} \prod_{i \in \tilde{V}_k} f_{\beta,\Lambda}(T_i, \Delta_i | X_i, Z_i) g(X_i | W_i, Z_i) \right\} \left\{ \prod_{k=1}^{K+1} \prod_{j \in \bar{V}_k} \int f_{\beta,\Lambda}(T_j, \Delta_j | x, Z_j) dG(x | W_j, Z_j) \right\}, \end{aligned} \tag{2.2}$$

where for $k = 1, \ldots, K+1$, $\bar{V}_k = \bar{V} \cap S_k$ denotes the index set of the portion of the nonvalidation sample that belongs to the k-th stratum, $G(x|W, Z)$ and $g(x|W, Z)$ are the conditional distribution and density functions of X given (W, Z), respectively, and

$$f_{\beta,\Lambda}(T,\Delta|X,Z) \;=\; f_{\beta,\Lambda}(T|X,Z)^{\Delta}\bar{F}_{\beta,\Lambda}(T|X,Z)^{1-\Delta}$$

with $f_{\beta,\Lambda}(t|X,Z)$ and $\bar{F}_{\beta,\Lambda}(t|X,Z)$ being the conditional density and survival functions of $\tilde{T}$ given (X,Z) under the proportional hazards model (2.1), respectively.

We propose a two-step estimation procedure for β. We first obtain estimates of (G,Λ) and plug them into the likelihood function (2.2), and then estimate β by maximizing the pseudo-likelihood function. Specifically, let $U=(W,Z^*)$, where Z^* is an informative subset of Z in the sense that $G(x|W,Z)=G(x|U)$ almost surely. Without loss of generality, we assume that U is a d-dimensional vector of continuous variables. If U were discrete, then the kernel estimators below would become the empirical estimators. Note that we can write

$$G(x|u)=\sum_{k=1}^{K+1}\pi_k(u)G_k(x|u),$$

where $\pi_k(u)=P((T,\Delta,W)\in S_k\,|\,u)$ and $G_k(x|u)=P(X\leq x\,|\,u,(T,\Delta,W)\in S_k)$. For $k=1,\ldots,K+1$, we can estimate $\pi_k(u)$ and $G_k(x|u)$ by

$$\hat{\pi}_k(u)=\frac{\sum_{i=1}^{N} I((T_i,\Delta_i,W_i)\in S_k)\phi_h(U_i-u)}{\sum_{i=1}^{N}\phi_h(U_i-u)}$$

and

$$\hat{G}_k(x|u)=\frac{\sum_{i\in V_k} I(X_i\leq x)\phi_h(U_i-u)}{\sum_{i\in V_k}\phi_h(U_i-u)}\ ,$$

respectively, where $V_k=V\cap S_k$ denotes the index set of the validation sample that belongs to the k-th stratum and $\phi_h(\cdot)=\phi(\cdot/h)$ is a d-dimensional kernel function with bandwidth h. Note that the values of h in $\hat{\pi}_k$ and $\hat{G}_k$ can be different and dependent on k. We use the same notation only for simplicity. Furthermore, we estimate the cumulative hazard function $\Lambda(t)$ by the Breslow-Aalen estimator based on the SRS component $\tilde{V}_0$ of the validation sample,

$$\hat{\Lambda}(t) = \sum_{T_j \le t, j \in \tilde{V}_0} \frac{\Delta_j}{\sum_{l \in \tilde{V}_0} Y_l(T_j) \exp\left\{\hat{\beta}_{10}' X_l + \hat{\beta}_{20}' Z_l\right\}},$$

where $Y_l(t) = I(T_l \ge t)$ and $\hat{\beta}_0 = (\hat{\beta}_{10}', \hat{\beta}_{20}')'$ is the partial likelihood estimate of β based on $\tilde{V}_0$. We then plug $\hat{\pi}_k(u)$, $\hat{G}_k(x|u)$ and $\hat{\Lambda}(t)$ into the likelihood function (2.2) and obtain the following pseudo-log-likelihood function

$$\hat{l}(\beta, \hat{\Lambda}, \hat{G}) = \sum_{k=0}^{K} \sum_{i \in \tilde{V}_k} \log f_{\beta,\hat{\Lambda}}(T_i, \Delta_i | X_i, Z_i) + \sum_{k=1}^{K+1} \sum_{j \in \bar{V}_k} \log \hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j, \Delta_j | W_j, Z_j), \quad (2.3)$$

where

$$f_{\beta,\hat{\Lambda}}(T_i, \Delta_i | X_i, Z_i) = \left\{\hat{\lambda}(T_i) \exp\{\beta_1' X_i + \beta_2' Z_i\}\right\}^{\Delta_i} \exp\left\{ - \hat{\Lambda}(T_i) \exp\{\beta_1' X_i + \beta_2' Z_i\}\right\}$$

with $\hat{\lambda}(t)$ being the jump size of $\hat{\Lambda}(t)$ at time t, and

$$\hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j, \Delta_j | W_j, Z_j) = \sum_{r=1}^{K+1} \hat{\pi}_r(U_j) \frac{\sum_{l \in V_r} f_{\beta,\hat{\Lambda}}(T_j, \Delta_j | X_l, Z_j) \phi_h(U_l - U_j)}{\sum_{l \in V_r} \phi_h(U_l - U_j)}.$$

We then obtain the estimator of β, denoted by $\hat{\beta}$, by maximizing the pseudo-log-likelihood function (2.3). The asymptotic properties of $\hat{\beta}$ are established in the following theorems. The proofs of these theorems and the regularity conditions needed are given in the Appendix. Let $(\beta_0, \Lambda_0, G_0)$ denote the true values of (β, Λ, G). Define $n_V = |V|$, $n_k = |\tilde{V}_k|$ for $k = 0, \ldots, K$, and $N_k = |S_k|$ for $k = 1, \ldots, K+1$, where $|\cdot|$ denotes the size of a set. Assuming that as $N \to \infty$, $n_V/N \to \rho_V > 0$, $n_k/n_V \to \rho_k \ge 0$ for $k = 1, \ldots, K$, $n_0/n_V \to \rho_0 > 0$, and $N_k/N \to \gamma_k > 0$ for $k = 1, \cdots, K+1$. The theorems are stated below.

Theorem 1 (Consistency): Under Conditions (C1)-(C5) in the Appendix A, $\hat{\beta}$ is a consistent estimator of β_0 such that $\hat{\beta} \xrightarrow{P} \beta_0$.

Theorem 2 (Asymptotic Normality): Under Conditions (C1)-(C5) in the Appendix A, we have

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma(\beta_0)),$$

with

$$\Sigma(\beta_0) = I^{-1}(\beta_0)\left\{ I(\beta_0) + \Sigma_{\mathbb{G}}(\beta_0) + \sum_{k=1}^{K+1} \frac{\gamma_k^2}{\rho_k\rho_V + \gamma_k\rho_0\rho_V}\Sigma_k(\beta_0)\right\} I^{-1}(\beta_0),$$

where $I(\beta)$ is the information matrix of β with known (Λ_0, G_0),

$$\Sigma_{\mathbb{G}}(\beta) = \text{Var}\left\{ \int_0^\tau H(t;\beta)\mathbb{G}(t)d\Lambda_0(t)\right\}$$

with $\mathbb{G}$ being a mean zero Gaussian process and $H(t;\beta)$ defined in the Appendix,

$$\Sigma_k(\beta) = \text{Var}_k\left\{ \sum_{r=1}^{K+1} \big[\gamma_r(1-\rho_0\rho_V) - \rho_r\rho_V\big]\pi_r(U)E_r\big\{M_{X,U}(T,\Delta,W,Z;\beta,\Lambda_0)\big|U\big\}\right\},$$

in which $E_r(\cdot|U)$ denotes the conditional expectation given U and $(T,\Delta,W) \in S_r$, $\text{Var}_k(\cdot)$ denotes the variance given $(T,\Delta,W) \in S_k$, and

$$M_{X,U}(T,\Delta,W,Z;\beta,\Lambda) = \frac{\partial f_{\beta,\Lambda}(T,\Delta|X,Z)/\partial\beta}{f_{\beta,\Lambda,G_0}(T,\Delta|W,Z)} - \frac{\partial f_{\beta,\Lambda,G_0}(T,\Delta|W,Z)/\partial\beta}{\big[f_{\beta,\Lambda,G_0}(T,\Delta|W,Z)\big]^2} f_{\beta,\Lambda}(T,\Delta|X,Z).$$

2.2.3 Nonparametric Bootstrap

Since the asymptotic covariance matrix $\Sigma(\beta_0)$ is not easy to compute, we propose a nonparametric bootstrap procedure for estimating the variance of $\hat{\beta}$ (Efron, 1994). The details are given below. First, we sample N subjects from the original cohort with replacement to construct the new cohort. Second, we formulate the new validation sample and nonvalidation sample using subjects in the new cohort as follows: if a subject in the new cohort was a member in the SRS component of original validation sample, then the subject is allocated to the SRS of new validation sample; if a subject in the new cohort was a member in the supplemental components of original validation sample, then allocated to the supplements of new validation sample; if a subject in the new cohort was a member in the original nonvalidation sample, then allocated to the new nonvalidation sample. Lastly, we calculate the bootstrap estimate of β using the proposed method based on the new validation sample and nonvalidation sample. This procedure is repeated B times to obtain B bootstrap estimates of β, denoted by $\hat{\beta}^1, \ldots, \hat{\beta}^B$, whose sample variance is then used to estimate the variance of $\hat{\beta}$. We provide Figure 2.2 to illustrate the idea of our nonparametric bootstrap with the cohort size $N = 25$ and the same sampling scheme as in Figure 2.1.

2.3 Simulation Studies

In this section, we conduct simulation studies to evaluate the finite-sample performance of the proposed method and also compare it with some other designs or methods. We generate the failure time $\tilde{T}$ from the proportional hazards model $\Lambda(t|X, Z) = \Lambda(t)\exp\{\beta_1 X + \beta_2 Z\}$ with $X \sim N(0, 1)$, $Z \sim Ber(0.5)$, $(\beta_1, \beta_2) = (\log 2, -0.5)$ and $\Lambda(t) = t$. The censoring time C is simulated from the uniform distribution over $(0, \tau)$, where τ is the length of study determined by the desired rate of event, denoted by p(event) $= 40\%$ or 20%. Also, the auxiliary variable is generated as $W = X + e$, where e is independent of X and $e \sim N(0, \sigma_e^2)$ with $\sigma_e = 0.8$ or 0.5 yielding a correlation between X and W of 0.78 or 0.90, respectively. In the proposed estimation procedure, the bandwidth for kernel estimation of G_k and π_k are taken as $h_{N1}^{(k)} = \frac{1}{2}\hat{\sigma}_{Wk} n_{V_k}^{-1/3}$ and $h_{N2}^{(k)} = \frac{1}{2}\hat{\sigma}_{Wk} N_k^{-1/3}$, respectively, where n_{V_k} is the size of validation sample in the kth stratum, N_k is the size of kth stratum, and $\hat{\sigma}_{Wk}$ denotes the sample standard deviation of W within the k-th stratum, $k = 1, \ldots, K + 1$. The nonparametric bootstrap for variance estimation is conducted with $B = 200$ replicates.

Now we describe the study designs considered in the simulation studies. We first consider the proposed two-phase FADS design as follows. Recall that N is the cohort size, n_0 is the size of SRS, and n_k is the size of kth supplemental sample for $k = 1, \ldots, K$. As shown in Figure 2.1, we partition the domains of T and W into three mutually exclusive intervals with $J = L = 3$ and the cutoff points being $(30, 70)$-th percentiles of T and W, respectively. At the second phase, besides taking the SRS of size n_0, we draw four supplemental samples of sizes $\{n_1, n_3, n_7, n_9\}$ from the four "corner" failure strata $\{S_1, S_3, S_7, S_9\}$, respectively. Then the size of validation sample is $n_V = n_0 + n_1 + n_3 + n_7 + n_9$. The proposed method is used to analyze data from the FADS design. Recall that the auxiliary variable is generated as $W = X + e$ with $e \sim N(0, \sigma_e^2)$. We denote the proposed estimators corresponding to $\sigma_e = 0.8$ and 0.5 by $\hat{\beta}_{FADS_1}$ and $\hat{\beta}_{FADS_2}$, respectively.

For comparison, we also consider the two-phase FDS design, where the selection of the second-phase sample depends only on the failure time T rather than also on the auxiliary variable W. In particular, at the second phase, after taking the SRS of size n_0, we select two supplemental samples of sizes $n_1 + n_3$ and $n_7 + n_9$ from the two "tail" failure strata based on T, that is, $S_1 \cup S_2 \cup S_3$ and $S_7 \cup S_8 \cup S_9$, respectively. Then the size of the validation sample is $n_V = n_0 + n_1 + n_3 + n_7 + n_9$, the same as the size of the validation sample of the FADS design described above. We compare two methods for estimation under the FDS design. One is the estimated maximum semiparametric empirical likelihood method given by Ding *et al.* (2014), which does not utilize the nonvalidation sample and cannot incorporate the auxiliary information. The estimator obtained using this method is denoted by $\hat{\beta}_{ODS}$, since this sampling scheme is called outcome-dependent sampling (ODS) in Ding *et al.* (2014). Also, we apply the proposed method to analyze data from the FDS design. Unlike Ding *et al.* (2014), our method makes use of the nonvalidation sample and the auxiliary information. We denote the estimators corresponding to $\sigma_e = 0.8$ and 0.5 by $\hat{\beta}_{FDS_1}$ and $\hat{\beta}_{FDS_2}$, respectively. For comparison, we also consider the SRS design, where we take a simple random sample of the same size as the validation sample $n_V = n_0 + n_1 + n_3 + n_7 + n_9$. The corresponding maximum partial likelihood estimator is denoted by $\hat{\beta}_{SRS}$. In addition, we present the ideal case assuming that the expensive covariate X is available for the full cohort, and denote the corresponding maximum partial likelihood estimator as $\hat{\beta}_{FC}$. We consider different values of the sizes N, n_V, n_0 and n_k in the simulation studies. The results based on 500 replicates are given in Table 2.1.

In Table 2.1, "Bias" is calculated as the average of point estimates minus the true value, "SD" is the sample standard deviation of point estimates, "SE" is the average of standard error estimates, and "CP" is the coverage proportion of 95% confidence intervals, based on 500 simulations. From the results in Table 2.1, we can see that all estimators considered are virtually unbiased, the variance estimators accurately reflect the true variability, and the coverage probabilities of 95% confidence intervals are close to the nominal level. Also, for estimation of the regression coefficient β_1 of X, we observe that $\hat{\beta}_{FADS_2}$ is more efficient than $\hat{\beta}_{FADS_1}$, and $\hat{\beta}_{FDS_2}$ is more efficient than $\hat{\beta}_{FDS_1}$. This is expected since $\hat{\beta}_{FADS_2}$ and $\hat{\beta}_{FDS_2}$ correspond to a higher level of association between X and W compared with $\hat{\beta}_{FADS_1}$ and $\hat{\beta}_{FDS_1}$. Also, $\hat{\beta}_{FADS_1}$ and $\hat{\beta}_{FADS_2}$ are more efficient than $\hat{\beta}_{FDS_1}$ and $\hat{\beta}_{FDS_2}$, respectively, which implies that using auxiliary information in the study design helps to improve the estimation efficiency. Moreover, $\hat{\beta}_{FDS_1}$ and $\hat{\beta}_{FDS_2}$ are more efficient than $\hat{\beta}_{ODS}$, indicating that our estimation method gains efficiency over the method of Ding *et al.* (2014) by additionally utilizing the nonvalidation sample and auxiliary information. Lastly, as expected, the estimator $\hat{\beta}_{SRS}$ based on the SRS design is least efficient as it does not incorporate any information from the first-phase sample when selecting the second-phase sample. We also conducted simulation studies using the cutoff points at the $(15, 85)$-th percentiles of T and W, and the results are similar as above.

2.4 Real Data Analyses

In this section, we illustrate the two-phase FADS design and the proposed estimation method using two real data sets.

2.4.1 ARIC Study

We first apply the proposed design and method to a data set on incident diabetes from the Atherosclerosis Risk in Communities (ARIC) study. The ARIC study is a longitudinal and epidemiological cohort study consisting of 15972 men and women aged 45-64 at baseline, including both White and African American participants, recruited from four U.S. field centers: Forsyth County, North Carolina (Center F); Jackson, Mississippi (Center J); Minneapolis, Minnesota (Center M); and Washington County, Maryland (Center W) (The ARIC Investigators, 1989). Each participant was followed up every three years starting from 1987 and was examined for the events of interest at the follow-up visits. In this study, we are interested in evaluating the association of high-sensitivity C-Reactive Protein (hs-CRP) level with time to incident diabetes after adjusting for other risk factors or confounding variables. Specifically, the onset of incident diabetes is defined as a fasting glucose level of 140 mg/dL or above, a non-fasting glucose level of 200 mg/dL or above, self-reported physician diagnosis diabetes, or use of diabetic medications. We consider a categorical variable hs-CRP with four levels based on the quartiles of the continuous measure with three indicator variables, hs-CRP(C2), hs-CRP(C3) and hs-CRP(C4), defined for the second, third and fourth quartiles, respectively, with the first quartile being the reference level. The other variables considered in the model include race, gender, age, body mass index (BMI), field center, smoking status, drinking status, high-density lipoprotein (HDL) cholesterol level, and total cholesterol level. Since only Center W has both White and African American participants, we combine race and field center to generate a five-level variable with four indicators, including White in Center F, White in Center J, African American in Center M, and African American in Center W, where White in Center W is chosen as the reference level.

After excluding subjects with diabetes at baseline or having missing values on covariates, the cohort for analysis consists of 9738 subjects. During the study period, 2135 subjects had developed diabetes, so the censoring rate is about 78%. There is not a well-defined auxiliary variable for hs-CRP. For illustration, we create an auxiliary variable $W = \text{hs-CRP} + e$ with $e \sim N(0, 5^2)$ such that the correlation between W and hs-CRP is around 0.80. The two-phase FADS design is implemented as in Figure 2.1. At the second phase, we take an SRS of size 800 and four supplemental samples with each of size 100 from the four "corner" failure strata, respectively. We apply the proposed method to analyze data from the FADS design and denote the estimator by FADS. We also consider the two-phase FDS design by selecting two supplemental samples of size 200 for each from the two "tail" failure strata based on T only. The estimator based on the proposed method is denoted by FDS. The estimator obtained using the method of Ding *et al.* (2014) is denoted by ODS. As in the simulation studies, for comparison, we also consider the estimator based on the SRS design, denoted by SRS, and the estimator based on the full cohort, denoted by FC. The analysis results are presented in Table 2.2. All methods suggest that hs-CRP, BMI, race-center, and HDL are significantly associated with the risk of diabetes. The FDS, FADS and FC methods also show that age and smoking status are significant. All methods except SRS indicate that total cholesterol is significant. The findings from FDS and FADS are consistent with those from FC. Also, FDS and FADS yield smaller standard errors than ODS and SRS as expected.

2.4.2 National Wilms' Tumor Study

We also apply the proposed method to a data set on Wilms' tumor, a rare childhood kidney cancer, from the National Wilms' Tumor Study (Breslow and Chatterjee, 1999). The data set includes 4028 patients from the third and fourth clinical trials of this study. It is of interest to assess the effects of tumor histological type, tumor stage, and age at diagnosis on time to disease relapse. The censoring rate is about 86%. The tumor histological type for each patient was examined by both a local pathologist and an experienced pathologist from a central facility. The latter examination tends to be more accurate but is more expensive and time-consuming. Although the central histological types are available for all patients in this data set, if the study investigators implemented a two-phase design by assessing the central histological types only on a small set of patients, the study cost would have been largely reduced.

We illustrate the two-phase FADS design and the proposed method using this data set. The local histological type can be used as the auxiliary variable W for the expensive central histological type X. Both W and X are binary and the missclassification rate is about 5%. The other adjustment covariates Z include tumor stage and age at diagnosis, where tumor stage is categorical with four stages and we define three indicator variables accordingly with the first stage being reference level. Since both X and W are binary, the kernel estimators $\hat{\pi}_k$ and $\hat{G}_k$ become empirical estimators when applying our method. We implement the two-phase FADS design as follows. The domain of T is partitioned into three mutually exclusive intervals, while the domain of W is naturally partitioned into two parts as W is binary. At the second phase, we take a simple random sample of size 400 and four supplemental samples with each of size 25 from the four "corner" failure strata as in Figure 2.1. We also consider the two-phase FDS design that selects two supplemental samples of size 50 for each from the two "tail" failure strata based on T only. We apply the proposed method to analyze data from the FADS and FDS designs, and denote the estimators by FADS and FDS, respectively. Similar to the ARIC study, we also consider the estimators SRS, ODS and FC for comparison. The analysis results are presented in Table 2.3. The tumor histological type and tumor stage are significantly associated with the risk of disease relapse based on all methods, while age at diagnosis is significant for FDS, FADS and FC but not for SRS or ODS. Thus, FDS and FADS yield the same findings as FC. Also, for all variables, FDS and FADS have smaller standard errors than SRS and ODS.

2.5 Conclusions

In this chapter, we developed a two-phase failure-time-auxiliary-dependent sampling (FADS) design for studies with expensive covariates and cheap surrogate or auxiliary variables. We proposed a new semiparametric maximum pseudo-likelihood method for inference and a nonparametric bootstrap procedure for variance estimation. The innovation of the proposed FADS design is that it allows the selection of second-phase sample to depend not only on the failure time but also on the readily available auxiliary variable, thus it provides more flexibility to oversample segments of the population that are believed to be more informative to the relationship between the failure time and the expensive covariate. The proposed method can reap the benefits gained by the FADS design and provide consistent estimates by accounting for the sampling bias. This new design and accompanying inference procedure could play a significant role in the success of many studies with a limited budget.

There are a few directions for future research. One is the selection of design parameters, such as the cutoff points and the allocation of sample sizes. In the simulation studies, we have tried the $(15, 85)$-th and $(30, 70)$-th percentiles as the cutoff points, and also tried allocating the sample sizes as $(n_0, n_k) = (400, 25)$ and $(300, 50)$. The estimation results are similar in these settings. Since the asymptotic variance given in Theorem 2 has a complex form, it is not straightforward to assess the effect of design parameters on the estimation efficiency in theory. There are also some practical considerations. For instance, if the cutoff points are too extreme and the censoring rate is high, the "corner" failure strata may not have enough subjects to be selected. The optimal choice of cutoff points and sample size allocation warrant future research. Another direction for research is the creation of an auxiliary variable when it is not available in practice. One possible way is to fit a predictive model for the expensive covariate using the SRS component $\tilde{V}_0$. The theoretical and numerical performance of this method warrants future research. Lastly, the proportional hazards model considered in this work, although widely used, may not hold in some applications. The proposed method can be extended without much effort to other models, such as the proportional odds model and the semiparametric transformation models.

Table 2.1: Simulation Results for the Estimation of β_1 and β_2

					$\beta_1 = \log 2$				$\beta_2 = -0.5$			
p(event)	N	n_V	(n_0, n_k)		Bias	SD	SE	CP	Bias	SD	SE	CP
40%	3000	500	(400, 25)	$\hat{\beta}_{SRS}$	0.002	0.079	0.077	0.946	-0.009	0.148	0.144	0.954
				$\hat{\beta}_{ODS}$	0.002	0.073	0.074	0.952	-0.009	0.142	0.144	0.947
				$\hat{\beta}_{FDS_1}$	0.004	0.066	0.064	0.950	0.000	0.112	0.113	0.954
				$\hat{\beta}_{FADS_1}$	0.003	0.063	0.063	0.948	0.000	0.111	0.113	0.950
				$\hat{\beta}_{FDS_2}$	0.008	0.054	0.054	0.954	0.001	0.111	0.112	0.962
				$\hat{\beta}_{FADS_2}$	0.007	0.053	0.053	0.948	-0.001	0.110	0.112	0.954
				$\hat{\beta}_{FC}$	-0.001	0.031	0.031	0.946	-0.004	0.057	0.059	0.968
	6000	1000	(800, 50)	$\hat{\beta}_{SRS}$	0.003	0.053	0.054	0.958	0.000	0.103	0.102	0.948
				$\hat{\beta}_{ODS}$	0.003	0.049	0.053	0.966	-0.001	0.105	0.102	0.942
				$\hat{\beta}_{FDS_1}$	0.005	0.046	0.044	0.945	0.005	0.079	0.079	0.941
				$\hat{\beta}_{FADS_1}$	0.004	0.044	0.043	0.953	0.006	0.079	0.079	0.949
				$\hat{\beta}_{FDS_2}$	0.007	0.038	0.037	0.937	0.004	0.079	0.079	0.951
				$\hat{\beta}_{FADS_2}$	0.006	0.038	0.037	0.953	0.005	0.079	0.079	0.947
				$\hat{\beta}_{FC}$	-0.001	0.021	0.022	0.952	-0.001	0.042	0.041	0.944
20%	5000	500	(400, 25)	$\hat{\beta}_{SRS}$	0.008	0.105	0.106	0.948	0.000	0.218	0.207	0.948
				$\hat{\beta}_{ODS}$	0.000	0.092	0.097	0.958	-0.011	0.185	0.190	0.955
				$\hat{\beta}_{FDS_1}$	0.010	0.088	0.085	0.954	0.006	0.140	0.141	0.950
				$\hat{\beta}_{FADS_1}$	0.009	0.086	0.085	0.956	0.008	0.138	0.141	0.958
				$\hat{\beta}_{FDS_2}$	0.014	0.073	0.072	0.949	0.004	0.141	0.143	0.949
				$\hat{\beta}_{FADS_2}$	0.013	0.074	0.072	0.945	0.005	0.141	0.143	0.954
				$\hat{\beta}_{FC}$	0.002	0.032	0.033	0.962	0.002	0.066	0.065	0.942
	10000	1000	(800, 50)	$\hat{\beta}_{SRS}$	0.006	0.078	0.075	0.952	-0.007	0.140	0.145	0.960
				$\hat{\beta}_{ODS}$	0.007	0.064	0.069	0.955	-0.006	0.117	0.135	0.976
				$\hat{\beta}_{FDS_1}$	0.005	0.058	0.058	0.945	0.004	0.095	0.099	0.961
				$\hat{\beta}_{FADS_1}$	0.004	0.058	0.058	0.942	0.004	0.095	0.098	0.966
				$\hat{\beta}_{FDS_2}$	0.008	0.050	0.049	0.949	0.002	0.097	0.100	0.957
				$\hat{\beta}_{FADS_2}$	0.007	0.049	0.049	0.951	0.002	0.097	0.100	0.960
				$\hat{\beta}_{FC}$	0.001	0.022	0.023	0.958	-0.001	0.045	0.046	0.948

Note: Bias, average estimate minus true value; SD, sample standard deviation; SE, average estimated standard error; CP, coverage proportion with 95% nominal level; SRS, the maximum partial likelihood method for the SRS design; ODS, the estimation method of Ding *et al.* (2014) for the FDS design; FDS, our estimation method for the FDS design; FADS, our estimation method for the FADS design; FC, the maximum partial likelihood method using the full cohort; SE, standard error estimates of $\hat{\beta}$. FDS_1 and FADS_1 correspond to $\sigma = 0.8$, yielding a correlation between X and W of 0.78. FDS_2 and FADS_2 correspond to $\sigma = 0.5$, yielding a correlation between X and W of 0.90.

Table 2.2: Analysis Results for the ARIC Study

	SRS		ODS		FDS		FADS		FC	
Variables	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE
hs-CRP(C2)	0.037	0.191	0.484	0.171	0.204	0.080	0.193	0.079	0.185	0.071
hs-CRP(C3)	0.548	0.185	0.956	0.173	0.514	0.079	0.543	0.079	0.512	0.068
hs-CRP(C4)	0.174	0.215	0.750	0.194	0.605	0.080	0.614	0.080	0.541	0.072
Gender	-0.033	0.154	0.186	0.136	-0.021	0.073	-0.022	0.073	-0.009	0.051
Age	0.017	0.013	0.017	0.009	0.029	0.013	0.029	0.013	0.021	0.004
BMI	0.084	0.012	0.078	0.014	0.074	0.009	0.072	0.009	0.072	0.004
White Center F	-0.048	0.202	0.160	0.176	-0.115	0.091	-0.051	0.092	-0.094	0.068
White Center J	0.277	0.188	0.402	0.172	0.158	0.086	0.185	0.085	0.159	0.064
African American Center M	0.636	0.239	0.992	0.196	0.481	0.102	0.506	0.099	0.468	0.073
African American Center W	1.305	0.445	1.608	0.535	0.374	0.180	0.422	0.178	0.343	0.150
Smoking Status	0.189	0.168	0.286	0.153	0.203	0.078	0.194	0.079	0.135	0.056
Drinking Status	-0.059	0.148	-0.115	0.135	-0.002	0.065	-0.027	0.065	0.002	0.049
HDL	-1.150	0.200	-1.248	0.134	-1.085	0.144	-1.110	0.143	-1.001	0.072
Total Cholesterol	0.045	0.060	0.115	0.051	0.064	0.028	0.075	0.029	0.079	0.020

Note: SRS, the maximum partial likelihood method for the SRS design; ODS, the estimation method of Ding *et al.* (2014) for the FDS design; FDS, our estimation method for the FDS design; FADS, our estimation method for the FADS design; FC, the maximum partial likelihood method using the full cohort; SE, standard error estimates of $\hat{\beta}$.

Table 2.3: Analysis Results for the National Wilms' Tumor Study

	SRS		ODS		FDS		FADS		FC	
Variables	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE
Histology	1.407	0.248	1.601	0.414	1.553	0.108	1.579	0.108	1.584	0.089
Stage II	0.665	0.316	0.991	0.654	0.625	0.268	0.605	0.268	0.667	0.122
Stage III	1.004	0.312	1.291	0.629	0.747	0.260	0.761	0.260	0.817	0.121
Stage IV	1.105	0.376	1.444	0.588	1.093	0.267	1.106	0.267	1.154	0.135
Age	0.052	0.038	0.060	0.051	0.065	0.027	0.061	0.027	0.068	0.015

Note: SRS, the maximum partial likelihood method for the SRS design; ODS, the estimation method of Ding *et al.* (2014) for the FDS design; FDS, our estimation method for the FDS design; FADS, our estimation method for the FADS design; FC, the maximum partial likelihood method using the full cohort; SE, standard error estimates of $\hat{\beta}$.

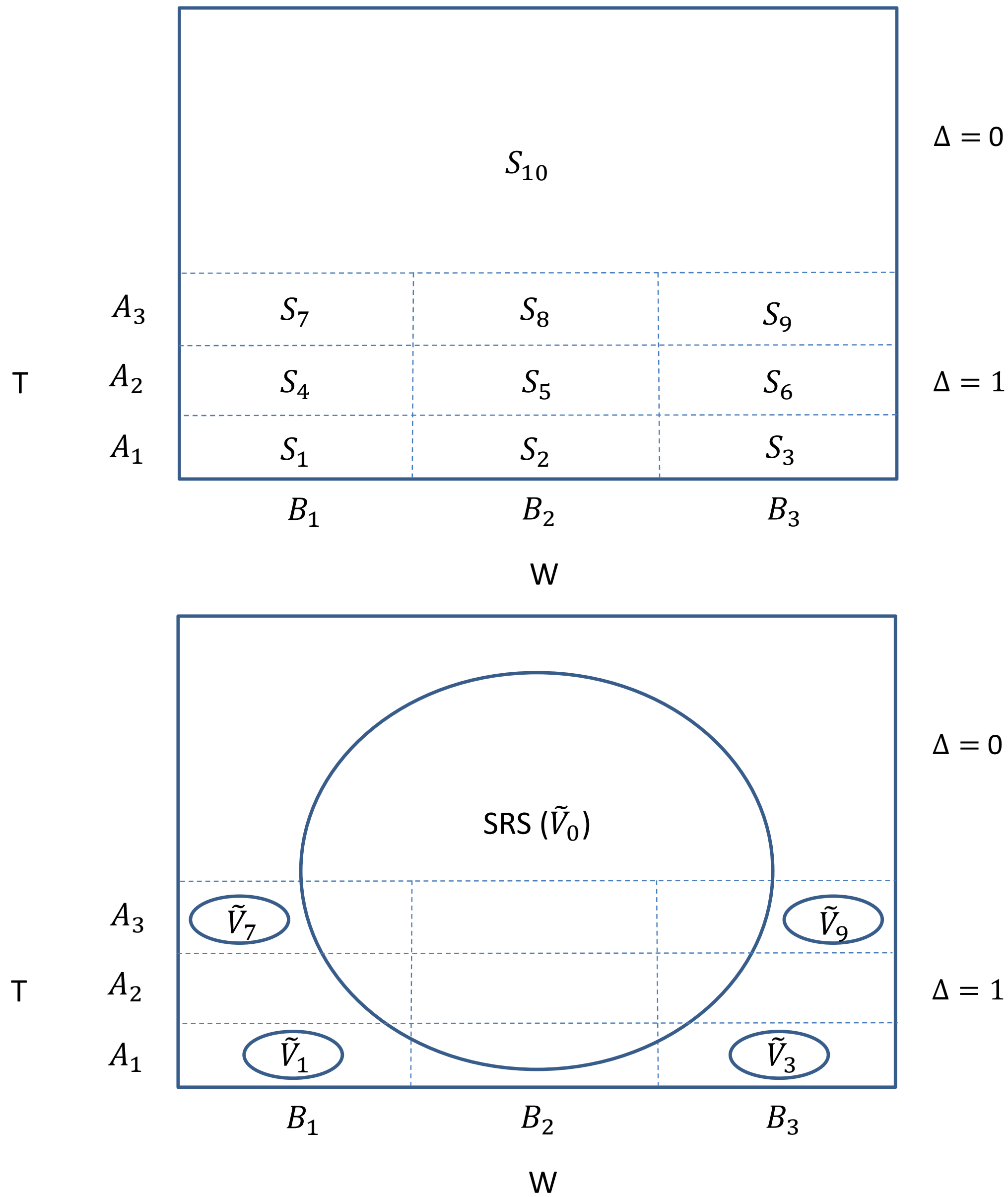

Figure 2.1: An illustration of the partitions $\{S_1, \ldots, S_{K+1}\}$ and the validation sample $\{\tilde{V}_0, \tilde{V}_1, \ldots, \tilde{V}_K\}$ under the two-phase FADS design: the domains of T and W are divided into three mutually exclusive intervals, respectively, with $J = L = 3$; four supplemental samples are selected from the four "corner" failure strata $\{S_1, S_3, S_7, S_9\}$, respectively.

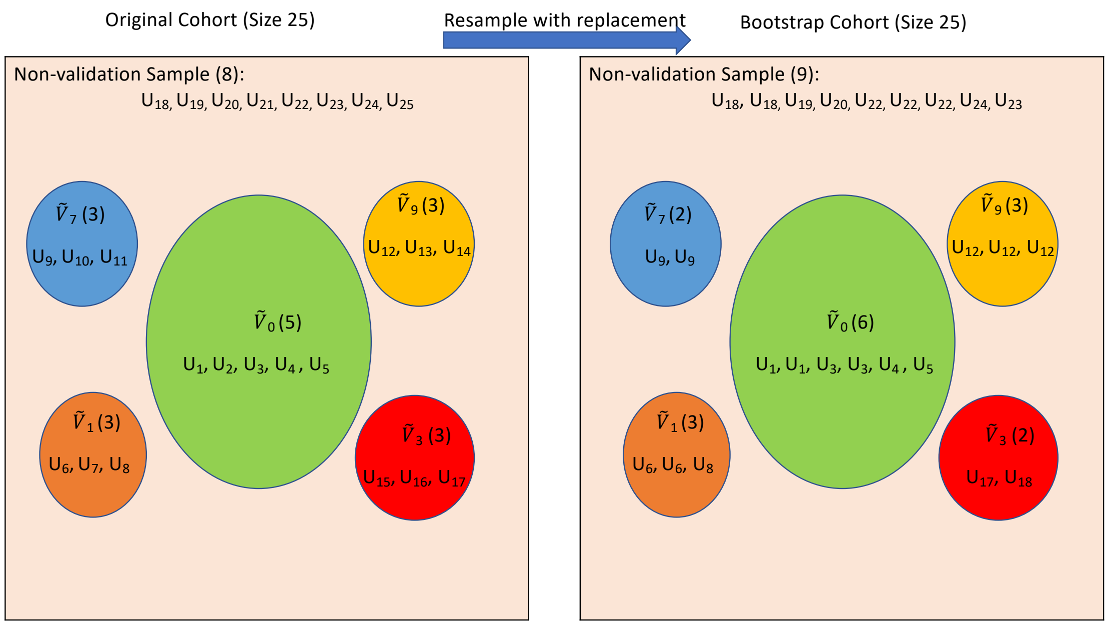

Figure 2.2: An illustration of the nonparametric bootstrap with the cohort size $N = 25$ and under the same sampling scheme as in Figure 2.1. U_i represents the ith subject in the original cohort, $i = 1, \ldots, 25$. The number in the parenthesis is the corresponding sample size.

CHAPTER 3: IMPROVING ESTIMATION EFFICIENCY FOR CASE-COHORT STUDIES WITH A CURE FRACTION

3.1 Introduction

As described in Section 1.1.1, in survival analysis, it is often assumed that all subjects in a study will eventually experience the event of interest. However, this assumption may not hold in various scenarios. For instance, when examining the time until a patient progresses or relapses from a disease, those who are cured will never experience the event. These individuals are frequently referred to as "long-term survivors" or "cured", and their survival time is regarded as infinite. Since it is impractical to track all individuals until they experience the event of interest, survival data typically involve right-censoring, where only a lower bound of the survival time is known for some individuals. When survival data include a fraction of long-term survivors, censored observations encompass both uncured individuals, for whom the event was not observed, and cured individuals who will not experience the event. Consequently, the cure status is unknown, and survival data comprise a mixture of cured and uncured individuals that cannot be distinguished beforehand. Cure models are survival models specifically designed to address this characteristic.

A typical field in which cure models are used is cancer studies. As the example given in Section 1.1.1, in the National Wilms' Tumor Study on a rare childhood kidney cancer, there is a certain number of patients will never experience the occurrence of the disease. Moreover, the presence of a stable plateau in the Kaplan and Meier (1958) estimator of the survival function, alongside a considerable number of censored observations, suggests the existence of a cured fraction. This observation, highlighted by Sy and Taylor (2000), serves as both an indicator and a prerequisite for cure models. This estimator for the time to relapse among patients with the kidney cancer (Breslow and Chatterjee, 1999), as shown in Figure 3.1, provides a compelling illustration of survival data with a cure fraction.

In the presence of covariate information, the frequently employed cure regression models are the nonmixture cure model and the mixture cure model. The nonmixture cure model serves as a common approach in analyzing survival data incorporating a cure fraction, while maintaining the proportional hazards structure across the entire population. In addition, it offers a clear interpretation of how covariates impact the probability of cure, as demonstrated by Tsodikov (1998) and Tsodikov *et al.* (2003). Yakovlev and Tsodikov (1996), along with Chen *et al.* (1999), provided a biological derivation of this model. The mixture cure model is an alternative approach for analyzing survival data with a cure fraction. It is a mixture of two separate regression models, one for the cure rate of the nonsusceptible population and another for the survival function of the susceptible population. Various models have been considered for the conditional survival function for the uncured subjects. Farewell (1982) originally proposed parametric models. A semiparametric approach utilizing a Cox (1972) proportional hazards (PH) model was offered by Kuk and Chen (1992), Sy and Taylor (2000), and Kuk and Chen (2008), while a fully nonparametric estimation approach was presented in Patilea and Van Keilegom (2020). Lu and Ying (2004) and Mao and Wang (2010) investigated the transformation mixture cure model, wherein a class of linear transformation models is assumed for the survival time of uncured individuals. The mixture cure model with missing covariates was studied by Beesley *et al.* (2016). The advantage of mixture cure models compared to nonmixture cure models lies in their ability to account for the presence of both cured and uncured individuals within the population. While many studies have explored the mixture cure model, none have yet integrated it with two-phase studies which are desirable when working within constrained budgets.

In many epidemiological studies, the focus is on time-to-failure events like cancer, heart disease, and HIV infection. However, acquiring measurements for important exposure variables, e.g., biomarkers needing bioassays or genetic analyses, can be both difficult and expensive. This poses a challenge for investigators working with limited budgets. Consequently, cost-effective sampling designs and efficient inference procedures become crucial. To tackle this, two-phase studies are commonly employed in practice to minimize costs and enhance study efficiency. In the first phase, a large random sample is drawn to collect outcome data and less expensive covariates or auxiliary variables. Then, in the second phase, measurements of expensive covariates are obtained for a subset of the first-phase sample. A wealth of literature exists on two-phase study designs, with particular emphasis on selecting the second-phase sample.

For the failure time outcome, Prentice (1986) proposed a case-cohort design, where the expensive exposure variables are collected only for a simple random sample from the study cohort, called the subcohort, and for all subjects who have experienced the failure event by a specified time, called cases. Since its proposal, the case-cohort design has been extensively studied by many authors, including Chen and Lo (1999), Cai and Zeng (2004), Lu and Tsiatis (2006), Breslow and Wellner (2007), and Marti and Chavance (2011). The original case-cohort design is primarily used for rare events. When the failure event of interest is non-rare or not so rare, Chen (2001), Cai and Zeng (2007), and Kang and Cai (2009), among others, considered a generalized case-cohort design. In this design, the expensive exposure measurements are obtained for a subcohort and for a subset, instead of all, of the remaining cases outside the subcohort. In the area of the two-phase studies of survival data with a cure fraction, only Han and Wang (2020) and Xie *et al.* (2023) have explored the utilization of the nonmixture cure model within case-cohort studies. The disadvantage of nonmixture cure models is that they assume all individuals follow the same underlying survival function, which may not accurately represent the presence of a cure fraction in the population. This can lead to biased estimates when there is a substantial portion of individuals who are cured and will never experience the event of interest. Despite this, none of the existing two-phase studies have integrated the mixture cure model into their methodologies.

In this chapter, we consider the semiparametric transformation mixture cure models under a generalized case-cohort study. First, we create a sieve maximum weighted likelihood method using complete data and devise an EM algorithm to yield an inverse probability weighting (IPW) estimator. Then, we update the IPW estimator by incorporating a working model between the outcome with inexpensive covariates or auxiliary variables from the entire dataset. The fundamental idea behind the update approach is to identify a (asymptotically) mean-zero statistic, which is correlated with the original unbiased estimator. Then, we develop an update estimator by combining the original estimator with this statistic in an optimal linear manner. It has been demonstrated that the update estimator remains consistent and is at least as efficient as the original estimator. This methodology has found application in various scenarios involving incomplete or imprecise data (Chen and Chen (2000); Chen (2002); Wang and Wang (2015); Yan *et al.* (2017); Yang and Ding (2020)). In this chapter, we propose an update estimation procedure to enhance the efficiency of the IPW estimator for the semiparametric transformation mixture cure models under generalized case-cohort studies of survival data with cure fraction. This method is innovative in that the proposed update estimator is consistent and asymptotically at least as efficient as the complete data estimator, regardless of whether the working model is correctly specified or not. Specifically, we assume a working regression model for failure time given the inexpensive covariates and, if available, auxiliary variables. Subsequently, we fit this model to both the generalized case-cohort sample and the entire cohort data, resulting in two estimators with the same mean. By taking the difference between these estimators, we obtain a mean-zero statistic, which allows us to construct an update estimator through an optimal linear combination of the original IPW estimator with this statistic.

The rest of this chapter is structured as follows. Section 3.2 outlines the data structure, model assumptions and likelihood. Section 3.3 describes the proposed two-step estimation method and the asymptotic properties of the update estimator. Sections 3.4 and 3.5 present simulation studies and a data application, respectively. Section 3.6 concludes the paper with discussions on potential extensions or directions for future research.

3.2 Data, Model and Design

Assume that the disease could either be uncured ($A = 1$) or cured ($A = 0$). The time to disease is denoted by $\mathcal{T} = T_{\mathcal{A}} < \infty$ if $A = 1$ and $\mathcal{T} = T_{\mathcal{I}} = \infty$ if $A = 0$. Let $X = (U^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$ denote the vector of predictors, where U is the routine clinical prognostic factors considered as expensive covariates and Z is novel biologic markers treated as the other adjustment covariates that are available. Given $X = (U^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$, we assume the following semiparametric transformation mixture cure models:

$$\pi(X) = P(A = 1 \mid X) = g(\lambda_{\mathcal{I}} + \alpha_{\mathcal{I}}^{\mathrm{T}} U + \gamma_{\mathcal{I}}^{\mathrm{T}} Z), \tag{3.1}$$

where $g(x) = e^x/(1 + e^x)$; and

$$\mathcal{R}_{\mathcal{A},t}(X) = P(T_{\mathcal{A}} \leq t \mid X,\, A = 1) = 1 - \exp\big\{-\Lambda(t \mid X,\, A = 1)\big\}, \tag{3.2}$$

with the cumulative hazard function of $T_{\mathcal{A}}$ given by

$$\Lambda(t \mid X,\, A = 1) = G\Big(\Lambda_{\mathcal{A}}(t) \exp\big\{\alpha_{\mathcal{A}}^{\mathrm{T}} U + \gamma_{\mathcal{A}}^{\mathrm{T}} Z\big\}\Big),$$

where G is a prespecified increasing transformation function with $G(0) = 0$ and $G(\infty) = \infty$, and $\Lambda_{\mathcal{A}}(t) = \int_0^t \lambda_{\mathcal{A}}(s)\,ds$ with $\lambda_{\mathcal{A}}(\cdot) > 0$ being an unknown function. Note that $G(x) = x$ yields the proportional hazards model and $G(x) = \log(1+x)$ gives the proportional odds model. Denote the parameters by $\theta = (\alpha_{\mathcal{A}}, \gamma_{\mathcal{A}}, \lambda_{\mathcal{A}}(\cdot), \alpha_{\mathcal{I}}, \gamma_{\mathcal{I}}, \lambda_{\mathcal{I}})$.

Let C be the censoring time, $Y = \min(\mathcal{T}, C)$ and $\delta = I(\mathcal{T} \leq C)$. Then the observed data for a single subject consists of $O = (Y, \delta, X)$. Note that if $\delta = 1$, then $A = 1$; if $\delta = 0$, then A is unknown. Thus, the complete data likelihood function based on n i.i.d. observations $\mathcal{O} = \{O_1, \ldots, O_n\}$ can be written as

$$\begin{aligned} L_n(\theta) = \prod_{i=1}^{n} \Big\{ \pi(X_i)\, \lambda_{\mathcal{A}}(Y_i \mid X_i,\, A_i = 1)\, S_{\mathcal{A}}(Y_i \mid X_i,\, A_i = 1) \Big\}^{\delta_i} \\ \cdot \Big\{ \big(1 - \pi(X_i)\big) + \pi(X_i)\, S_{\mathcal{A}}(Y_i \mid X_i,\, A_i = 1) \Big\}^{1-\delta_i}, \end{aligned} \tag{3.3}$$

where the hazard and survival functions of $T_{\mathcal{A}}$ are given by

$$\lambda_{\mathcal{A}}(t \mid X_i,\, A_i = 1) = G'\Big(\Lambda_{\mathcal{A}}(t) \exp\big\{\alpha_{\mathcal{A}}^{\mathrm{T}} U_i + \gamma_{\mathcal{A}}^{\mathrm{T}} Z_i\big\}\Big) \lambda_{\mathcal{A}}(t) \exp\big\{\alpha_{\mathcal{A}}^{\mathrm{T}} U_i + \gamma_{\mathcal{A}}^{\mathrm{T}} Z_i\big\}$$

with G' being the first derivative of G, and

$$S_{\mathcal{A}}(t \mid X_i,\, A_i = 1) = \exp\Big\{ - G\Big(\Lambda_{\mathcal{A}}(t) \exp\big\{\alpha_{\mathcal{A}}^{\mathrm{T}} U_i + \gamma_{\mathcal{A}}^{\mathrm{T}} Z_i\big\}\Big)\Big\}.$$

A potential problem is that, in practice, the censoring time C is bounded, preventing the observation of cured subjects in the dataset. To address this issue, researchers often adopt the concept of "threshold", where if $\mathcal{T}$ is greater than the threshold, it implies $\mathcal{T} = +\infty$ as proposed by Taylor (1995). This widely accepted assumption is frequently applied in cure models literature. Consequently, when Y is observed to exceed the threshold, it is inferred that the individual is cured.

The generalized case-cohort study is conducted as follows. In the first phase, we use Bernoulli sampling to select a subset of the full cohort with a sampling probability of $q_1 \in (0,1)$. In the second phase, we conduct another Bernoulli sampling among the cases that were not selected in the first phase, with a sampling probability of $q_2 \in (0,1)$. We use g_1 to indicate the non-cases selected at phase I, and g_2 to indicate the cases selected at phase I and phase II. Thus, under our design, the probability of a subject being selected is

$$p_i = g_{1_i} \cdot q_1 + g_{2_i} \cdot [q_1 + (1-q_1)\cdot q_2],\ i = 1, \cdots, n.$$

3.3 Proposed Two-Step Estimation Method

3.3.1 Original Estimator

We first propose an EM algorithm and a sieve method for the original estimation. By treating A as a latent variable, we employ the inverse probability weighting to construct the complete data likelihood given by

$$\begin{aligned} L_n^c(\theta) = \prod_{i=1}^{n} &\Big\{\pi(X_i)^{A_i w_i}\big(1-\pi(X_i)\big)^{(1-A_i)w_i}\Big\} \\ &\Big\{\big(\lambda_{\mathcal{A}}(Y_i \mid X_i,\, A_i=1)\, S_{\mathcal{A}}(Y_i \mid X_i,\, A_i=1)\big)^{\delta_i A_i w_i} \cdot S_{\mathcal{A}}(Y_i \mid X_i,\, A_i=1)^{(1-\delta_i)A_i w_i}\Big\}, \end{aligned}$$

where the inverse probability weight $w_i = \frac{1}{p_i}$.

Note that the complete data log-likelihood is a linear function of A_i's, thus in the E-step, we calculate the conditional expectation of A_i given the observed data O_i as follows:

$$\begin{aligned} E(A_i \mid O_i) &= P(A_i = 1 \mid Y_i, \delta_i, X_i) \\ &= \delta_i + (1-\delta_i)\, P(A_i = 1 \mid Y_i, \delta_i = 0, X_i) \\ &= \delta_i + (1-\delta_i)\, \frac{\pi(X_i)\, S_{\mathcal{A}}(Y_i \mid X_i,\, A_i=1)}{\big(1-\pi(X_i)\big) + \pi(X_i)\, S_{\mathcal{A}}(Y_i \mid X_i,\, A_i=1)}. \end{aligned}$$

In the M-step, we maximize the following conditional expectation of the inverse probability weighted complete data log-likelihood given the observed data $\mathcal{O}$:

$$E\Big(\log L_n^c(\theta) \mid \mathcal{O}\Big) = \sum_{i=1}^{n} w_i \Big\{ E(A_i \mid O_i) \log \pi(X_i) + \big(1 - E(A_i \mid O_i)\big) \log \big(1 - \pi(X_i)\big) \Big\}$$
$$+ w_i \Big\{ \delta_i \log \big(\lambda_{\mathcal{A}}(Y_i \mid X_i,\, A_i = 1)\, S_{\mathcal{A}}(Y_i \mid X_i,\, A_i = 1)\big)$$
$$+ (1 - \delta_i) E(A_i \mid O_i) \log S_{\mathcal{A}}(Y_i \mid X_i,\, A_i = 1) \Big\}. \tag{3.4}$$

It is not easy to maximize (3.4) directly as it involves the unknown function $\lambda_{\mathcal{A}}(\cdot)$. To deal with this, we propose a sieve method based on B-splines. In particular, let $b_1, \ldots, b_{m_n}$ be a set of B-spline basis functions of order l over a knot sequence $0 = t_1 = \cdots = t_l < t_{l+1} < \cdots < t_{m_n} < t_{m_n+1} = \cdots = t_{m_n+l} = \tau$, where τ is the length of study. Define the sieve space

$$\mathcal{B}_n = \left\{ \lambda_{\mathcal{A}n}(t) = \sum_{j=1}^{m_n} \eta_j b_j(t) :\ M_n^{-1} \le \eta_j \le M_n \text{ for } j = 1, \ldots, m_n \right\}$$

for some diverging sequence M_n. Denote the Euclidean parameters by $\vartheta = (\alpha_{\mathcal{A}}, \gamma_{\mathcal{A}}, \alpha_{\mathcal{I}}, \gamma_{\mathcal{I}}, \lambda_{\mathcal{I}})$ and let $\mathcal{D}$ be a prespecified compact set in $\mathbb{R}^{2p+1}$ that denotes the parameter space for ϑ, where p is the dimension of X. The sieve MLE is defined by

$$\hat{\theta}_n = (\hat{\vartheta}_n, \hat{\lambda}_{\mathcal{A}n}) = \arg\max_{\vartheta \in \mathcal{D},\, \lambda_{\mathcal{A}} \in \mathcal{B}_n} L_n(\vartheta, \lambda_{\mathcal{A}}),$$

where $L_n(\vartheta, \lambda_{\mathcal{A}})$ is the observed data likelihood given by (3.3). The sieve MLE can be obtained via the EM algorithm described above. In the M-step, we employ the Nelder-Mead simplex algorithm built in the *optim* function in R to maximize (3.4) over $\mathcal{D} \times \mathcal{B}_n$. To ensure positivity of the spline coefficient η_j, we reparameterize it as $\exp(\eta_j^*)$, for $j = 1, \ldots, m_n$.

A potential issue arises due to the limited information available for distinguishing between cured and uncured individuals among censored subjects. Estimating the tail of the conditional survival function $S_{\mathcal{A}}(t|X)$ can be particularly challenging. Specially, if $S_{\mathcal{A}}(t|X)$ remains non-zero after $\max(\mathcal{T})$, it can lead to identifiability concerns. To address this, Taylor (1995) suggested considering $\max(\mathcal{T})$ as a threshold. They enforced $S_{\mathcal{A}}(t|X)$ to be zero beyond $\max(\mathcal{T})$ by setting $E(A_i|O_i)$ to 0 in the M-step of the EM algorithm for observations i with $\delta_i = 0$ and $Y_i > \max(\mathcal{T})$. This approach treats such observations as cured. Recall that when cured observations are present in survival data, $\lim_{t\to\infty} S(t|X) > 0$. In practice, this implies the Kaplan-Meier estimator of the survival function with a large plateau and a leveling-off to a value greater than 0 for $t > \max(\mathcal{T})$, as illustrated in Fig 2.1. This situation suggests that a significant number of observations in the plateau can reasonably be considered as cured, as proposed by Taylor (1995).

The inverse probability weighted (IPW) estimator is commonly known for its inefficiency. To enhance estimation efficiency, we adopt an update method that utilizes the available information in the full cohort by constructing a model relating inexpensive covariates or auxiliary variables to the failure time. This updated approach guarantees to be, at the very least, asymptotically as efficient as the original IPW estimator.

3.3.2 Update Estimator

Consider the working models given $X^* = (U^{*T}, Z^T)^T$, where U^* is the auxiliary variable of U, for the semiparametric transformation mixture cure models as follows:

$$\pi^*(X) = P(A = 1 \mid X^*) = g(\lambda_{\mathcal{I}}^* + \alpha_{\mathcal{I}}^{*\mathrm{T}} U^* + \gamma_{\mathcal{I}}^{*\mathrm{T}} Z), \tag{3.5}$$

where $g(x) = e^x/(1 + e^x)$; and

$$\mathcal{R}^*_{\mathcal{A},t}(X) = P(T_{\mathcal{A}} \le t \mid X^*,\, A = 1) = 1 - \exp\big\{ - \Lambda^*(t \mid X,\, A = 1)\big\}, \tag{3.6}$$

with the cumulative hazard function of $T_{\mathcal{A}}$ given by

$$\Lambda^*(t|X^*, A=1) = G(\Lambda^*_{\mathcal{A}}(t)\exp\{\alpha^{*T}_{\mathcal{A}}U^* + \gamma^{*T}_{\mathcal{A}}Z\}), \tag{3.7}$$

where $\Lambda^*_{\mathcal{A}}(t) = \int_0^t \lambda^*_{\mathcal{A}}(s)\,ds$ with $\lambda^*_{\mathcal{A}}(\cdot) > 0$ being an unknown function. Note that we can consider the working models given Z only, if U^* is not available. Denote the proposed update Euclidean estimators of parameters $\vartheta^* = (\alpha^*_{\mathcal{A}}, \gamma^*_{\mathcal{A}}, \alpha^*_{\mathcal{I}}, \gamma^*_{\mathcal{I}}, \lambda^*_{\mathcal{I}})$.

We first estimate the working semiparametric transformation mixture cure models (3.5) and (3.6) by using the EM algorithm and sieve method similarly as the above but with covariates U^* and Z instead. Let $\hat{\theta}^*_n = (\hat{\vartheta}^*_n, \hat{\lambda}^*_{\mathcal{A}_n})$ denote the sieve MLE of $\theta^* = (\vartheta^*, \lambda^*_{\mathcal{A}})$ based on the generalized case-cohort sample. Since U^* and Z are available for all subjects in the cohort, we can also obtain the sieve MLE of $\theta^* = (\vartheta^*, \lambda^*_{\mathcal{A}})$, denoted by $\bar{\theta}^*_n = (\bar{\vartheta}^*_n, \bar{\lambda}^*_{\mathcal{A}_n})$, based on the full cohort. Let $\Sigma = [\Sigma_{11}, \Sigma_{12}; \Sigma_{21}, \Sigma_{22}]$ be the covariance matrix of the limiting distribution of $(\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)^T, \sqrt{n}(\hat{\vartheta}^*_n - \bar{\vartheta}^*_n)^T)^T$, and let $\hat{\Sigma} = [\hat{\Sigma}_{11}, \hat{\Sigma}_{12}; \hat{\Sigma}_{21}, \hat{\Sigma}_{22}]$ denote a consistent estimator of Σ. We define the update estimator of ϑ as

$$\bar{\vartheta}_n = \hat{\vartheta}_n - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}(\hat{\vartheta}^*_n - \bar{\vartheta}^*_n).$$

We can show that the asymptotic covariance matrix of $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)$ is Σ_{11}, while the asymptotic covariance matrix of $\sqrt{n}(\bar{\vartheta}_n - \vartheta_0)$ is $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Thus, $\bar{\vartheta}_n$ is asymptotically at least as efficient as $\hat{\vartheta}_n$, and the theorem is stated as following.

3.3.3 Asymptotic Properties

Theorem 1: Under Conditions in the Appendix B,

$$\sqrt{n}(\bar{\vartheta}_n - \vartheta_0) = \sqrt{n}(\hat{\vartheta}_n - \vartheta_0) - \Sigma_{12}\Sigma_{22}^{-1}\sqrt{n}(\hat{\vartheta}^*_n - \bar{\vartheta}^*_n) + o_p(1) \to N(0, \Psi)$$

in distribution with $\Psi = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, where

$$\Sigma_{11} = I(\vartheta_0)^{-1} E\Big\{\frac{1}{p}[l(\vartheta_0, \Lambda_0; \mathcal{O})]^{\otimes 2}\Big\} I(\vartheta_0)^{-1},$$

$$\Sigma_{22} = I^*(\vartheta_0^*)^{-1} E\Big\{\frac{1-p}{p}[l^*(\vartheta_0^*, \Lambda_0^*; \mathcal{O}^*)]^{\otimes 2}\Big\} I^*(\vartheta_0^*)^{-1},$$

$$\Sigma_{12} = \Sigma_{21}^T = I(\vartheta_0)^{-1} E\Big\{\frac{1-p}{p} l(\vartheta_0, \Lambda_0; \mathcal{O}) l^*(\vartheta_0^*, \Lambda_0^*; \mathcal{O}^*)^T\Big\} I^*(\vartheta_0^*)^{-1},$$

and $l(\vartheta_0, \Lambda_0; \mathcal{O})$, $l^*(\vartheta_0^*, \Lambda_0^*; \mathcal{O}^*)$ and $I^*(\vartheta_0^*)$ are defined in the Appendix.

The proof of this theorem is sketched in the Appendix B. Note that there is no closed-form expression for Σ, and we estimate it using weighted bootstrap procedure as described in the next section.

3.3.4 Variance Estimation

We propose to estimate the covariance matrix Σ by using the weighted bootstrap method. Particularly, let $\{u_1, \cdots, u_n\}$ denote n independent realizations of a bounded positive random variable u satisfying $E(u) = var(u) = 1$. We use the exponential distribution with mean 1 in the simulation study and data application. Define the new weights $w_i^b = u_i w_i$ for $i = 1, \cdots, n$. Let $\hat{\theta}_n^b = (\hat{\vartheta}_n^b, \hat{\lambda}_{\mathcal{A}_n^b})$ be the sieve maximum weighted likelihood estimator that maximizes the new weighted log-likelihood function $l_n^{w^b}$ over $\mathcal{B}_n$, where $l_m^{w^b}$ is obtained by replacing w_i with w_i^b in l_n^w. We generate B samples of $\{u_1, \cdots, u_n\}$ and obtain the corresponding $\hat{\vartheta}_n^b$ as well as $\hat{\vartheta}_n^{*b}$ and $\bar{\vartheta}_n^{*b}$ similarly for $b = 1, \cdots, B$. Then we take $\hat{\Sigma}$ as the sample variance matrix of $(\sqrt{n}(\hat{\vartheta}_n^b - \vartheta_0)^T, \sqrt{n}(\hat{\vartheta}_n^{*b} - \bar{\vartheta}_n^{*b})^T)^T$. The asymptotic covariance matrix of $\sqrt{n}(\bar{\vartheta}_n - \vartheta_0)$ can be estimated by $\hat{\Sigma}_{11} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}$.

3.4 Simulation Studies

We conduct simulation studies to investigate the proposed method. We first generate the predictors from two different settings with cure rate about 60%: (I) $(U, Z)^T$

follow a bivariate normal distribution with mean zero and covariance matrix [1, 0.5; 0.5, 1], and the auxiliary variable $U^* = U + e$ with $e \sim N(0, \sigma^2)$ where $\sigma = 0.8$ or 1.7 corresponding to the correlation between U and U^* of about 77% or 51%, respectively; (II) $U \sim Ber(0.5)$, $Z \sim N(0, 1)$, and U^* generated from U with misclassfication rate 10% or 30%. Given $X = (U^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$, we generate A and $\mathcal{T}$ from the models (3.1) and (3.2) with the parameter values $\alpha_{\mathcal{A}} = \log(2)$, $\gamma_{\mathcal{A}} = -0.5$, $\alpha_{\mathcal{I}} = \log(2)$, $\gamma_{\mathcal{I}} = -0.5$, $\lambda_{\mathcal{I}} = -0.5$, and $\lambda_{\mathcal{A}}(t) = t + 1/2$. For the transformation function G in the model (3.2), we consider the class of logarithmic transformations $G(x) = \log(1 + rx)/r$ with $r = 0$ and 1, corresponding to the proportional hazards (PH) and proportional odds (PO) models, respectively. We generate the censoring time C from the following settings for the transformation models with $r = 0$ and 1, respectively, and then obtain $Y = \min(\mathcal{T}, C)$, $\delta = I(\mathcal{T} \leq C)$ and $C = \min\{\mathrm{Unif}(0, 5\tau/4), \tau\}$ with τ being the length of study. For the PH model, we set τ to be 4.5 and 11, yielding around censoring rate 66% and 61%, respectively. For the PO model, we set τ to be 11 and 50, yielding around censoring rate 67% and 62%, respectively. For the sieve estimation, we take the interior knot of B-spline to be the median of Y_i's and take the degree of B-spline basis functions to be 1. The subcohort is selected via independent Bernoulli sampling with success probability $q_1 = 0.2$ at phase I and $q_2 = 0.5$ at phase II. The weighted bootstrap procedure for variance estimation is based on 500 samples. We consider the sample size $n = 2000$. The simulation results are based on 1000 replicates.

Table 3.1 $\sim$ 3.4 give the estimation results for Euclidean parameters, including "Bias" calculated as the average point estimate minus the true value, "SSD" the sample standard deviation of point estimates, "ESE" the average of estimated standard errors, "CP" the coverage proportion of the 95% confidence interval based on normal approximation, and "RE" the relative efficiency of the update estimator compared the original estimator. First, one can see from Table 3.1 $\sim$ 3.4 that the bias is negligible, ESE is close to SSD, and CP is around 95% for both original and update estimators. Second, the update estimators are more efficient than the original estimators, and the more U and U^* are correlated, the better the efficiency gain is. Third, when the study is not sufficiently long, both the original and updated estimators exhibit greater efficiency when the threshold is applied compared to when it is not; however, when the study is sufficiently long, both the original and updated estimators exhibit similar performance whether the threshold is applied or not. According to the findings, it is evident that employing a threshold is at least as efficient as not using one. Therefore, we recommend utilizing the threshold in the data application.

3.5 Wilms' Tumor Study

We apply the proposed method to a data set on Wilms' tumor, a rare childhood kidney cancer, from the National Wilms' Tumor Study (Breslow and Chatterjee, 1999). The data set includes 4028 patients from the third and fourth clinical trials of this study. It is of interest to assess the effects of tumor histological type, tumor stage, and age at diagnosis on time to disease relapse. The censoring rate is about 86%. The tumor histological type for each patient was examined by both a local pathologist and an experienced pathologist from a central facility. The latter examination tends to be more accurate but is more expensive and time-consuming. Although the central histological types are available for all patients in this data set, if the study investigators implemented a two-phase design by assessing the central histological types only on a small set of patients, the study costs would have been largely reduced.

We illustrate the proposed method using this data set. The local histological type can be used as the auxiliary variable U^* for the expensive central histological type U, and the missclassification rate between U and U^* is about 5%. The adjustment available covariates Z include tumor stage and age at diagnosis, where tumor stage is categorical with four stages and we define three indicator variables accordingly with the first stage being reference level. The generalized case-cohort study is implemented as follows. The subcohort is selected via independent Bernoulli sampling with success probability $q_1 = 0.2$ at phase I and $q_2 = 0.5$ at phase II. The analysis results are presented in Tables 3.5∼ 3.6. Table 3.5 displays the outcomes obtained solely using the PH or PO model without considering the cure portion. Tables 3.6 presents the results derived from the proposed PH mixture cure model and the PO mixture cure model, respectively, with the threshold applied.

The data application results show that the update estimators are more efficient than the original estimators. Also, the proposed method yields smaller standard errors compared to those obtained when the cure fraction is not taken into account, so it is necessary to consider the cure fraction in the model. Moreover, the tumor histological type, tumor stage, and age exhibit significant associations with the risk of disease relapse based on the proposed method. In addition, Fig 2.1 and Fig 2.2 plot the estimated survival functions based on the PH mixture cure model and PO mixture cure model, respectively, and they confirm the our findings above.

3.6 Discussion

In this paper, we introduce a method to enhance estimation efficiency for case-cohort studies with a cure fraction. Our approach involves a novel two-step estimation procedure under semiparametric transformation mixture cure models for inference, coupled with a weighted bootstrap procedure for variance estimation. The key innovation of our method lies in its efficiency-enhancing update estimation for the semiparametric transformation mixture cure models under case-cohort studies. By leveraging information from the full cohort, our proposed method offers consistent estimates while accounting for sampling bias, and the update estimator is asymptotically at least as efficient as the original IPW estimator.

There are a few directions for future research. One is to theoretically exclude the “threshold” condition imposed in our study computations. This is prompted by the observations that the results are similar with or without threshold applied when the study is sufficiently long. Secondly, exploring an optimal working model could lead to further efficiency gains, despite the proposed working model works well in practice. Furthermore, there is potential for enhancing efficiency through improvements in sampling design. One promising approach is the adoption of a failure-time-dependent sampling design, as proposed by Ding *et al.* (2014). All three directions warrant further investigation.

Table 3.1: Simulation results from the PH model with (U, Z) from bivariate normal distribution

τ	σ	para	true	w/o threshold Bias	SSD	ESE	CP	RE	with threshold Bias	SSD	ESE	CP	RE
4.5		$\alpha_{\mathcal{A}}$	log(2)	0.038	0.082	0.080	0.931	1.000	0.005	0.075	0.073	0.943	1.000
		$\gamma_{\mathcal{A}}$	-0.5	-0.024	0.077	0.076	0.934	1.000	0.000	0.072	0.069	0.937	1.000
		$\alpha_{\mathcal{I}}$	log(2)	-0.033	0.129	0.126	0.939	1.000	0.013	0.119	0.118	0.955	1.000
		$\gamma_{\mathcal{I}}$	-0.5	0.023	0.119	0.119	0.951	1.000	-0.010	0.115	0.113	0.950	1.000
		$\lambda_{\mathcal{I}}$	-0.5	0.043	0.102	0.099	0.919	1.000	0.004	0.094	0.092	0.946	1.000
	1.7	update $\alpha_{\mathcal{A}}$	log(2)	0.037	0.079	0.076	0.923	1.077	0.004	0.072	0.068	0.934	1.085
		update $\gamma_{\mathcal{A}}$	-0.5	-0.025	0.067	0.067	0.929	1.321	-0.001	0.062	0.060	0.943	1.349
		update $\alpha_{\mathcal{I}}$	log(2)	-0.033	0.125	0.117	0.921	1.065	0.014	0.113	0.109	0.942	1.109
		update $\gamma_{\mathcal{I}}$	-0.5	0.025	0.092	0.090	0.935	1.673	-0.009	0.085	0.083	0.953	1.830
		update $\lambda_{\mathcal{I}}$	-0.5	0.038	0.071	0.068	0.906	2.064	-0.003	0.057	0.057	0.956	2.720
	0.8	update $\alpha_{\mathcal{A}}$	log(2)	0.036	0.076	0.073	0.918	1.164	0.003	0.069	0.065	0.937	1.181
		update $\gamma_{\mathcal{A}}$	-0.5	-0.024	0.065	0.064	0.931	1.403	-0.001	0.060	0.058	0.938	1.440
		update $\alpha_{\mathcal{I}}$	log(2)	-0.034	0.110	0.104	0.922	1.375	0.012	0.099	0.096	0.938	1.445
		update $\gamma_{\mathcal{I}}$	-0.5	0.025	0.086	0.083	0.934	1.915	-0.007	0.078	0.077	0.950	2.174
		update $\lambda_{\mathcal{I}}$	-0.5	0.036	0.068	0.065	0.914	2.250	-0.004	0.056	0.056	0.955	2.818
11		$\alpha_{\mathcal{A}}$	log(2)	0.003	0.068	0.065	0.931	1.000	0.003	0.068	0.065	0.930	1.000
		$\gamma_{\mathcal{A}}$	-0.5	0.001	0.063	0.062	0.955	1.000	0.001	0.063	0.061	0.955	1.000
		$\alpha_{\mathcal{I}}$	log(2)	0.008	0.113	0.109	0.944	1.000	0.009	0.113	0.109	0.942	1.000
		$\gamma_{\mathcal{I}}$	-0.5	-0.007	0.108	0.106	0.949	1.000	-0.007	0.108	0.106	0.949	1.000
		$\lambda_{\mathcal{I}}$	-0.5	0.006	0.089	0.086	0.939	1.000	0.006	0.089	0.086	0.939	1.000
	1.7	update $\alpha_{\mathcal{A}}$	log(2)	0.003	0.065	0.061	0.929	1.094	0.003	0.065	0.061	0.925	1.094
		update $\gamma_{\mathcal{A}}$	-0.5	0.000	0.056	0.054	0.946	1.266	0.000	0.056	0.053	0.946	1.266
		update $\alpha_{\mathcal{I}}$	log(2)	0.008	0.105	0.101	0.943	1.158	0.008	0.105	0.101	0.944	1.158
		update $\gamma_{\mathcal{I}}$	-0.5	-0.004	0.079	0.076	0.938	1.869	-0.004	0.079	0.076	0.939	1.869
		update $\lambda_{\mathcal{I}}$	-0.5	0.001	0.054	0.053	0.947	2.716	0.001	0.054	0.053	0.946	2.716
	0.8	update $\alpha_{\mathcal{A}}$	log(2)	0.004	0.063	0.059	0.932	1.165	0.005	0.063	0.059	0.931	1.165
		update $\gamma_{\mathcal{A}}$	-0.5	-0.001	0.055	0.052	0.941	1.312	-0.001	0.055	0.052	0.940	1.312
		update $\alpha_{\mathcal{I}}$	log(2)	0.006	0.092	0.089	0.942	1.509	0.006	0.092	0.089	0.939	1.509
		update $\gamma_{\mathcal{I}}$	-0.5	-0.002	0.073	0.071	0.936	2.189	-0.002	0.073	0.071	0.937	2.189
		update $\lambda_{\mathcal{I}}$	-0.5	0.000	0.053	0.052	0.944	2.820	0.000	0.053	0.052	0.947	2.820

Note: Bias, average estimate minus true value; SSD, sample standard deviation; ESE, average estimated standard error; CP, coverage proportion with 95% nominal level; RE, relative efficiency compared to the original estimators; $\tau = 4.5$ or 11, yielding around censoring rate 66% or 61%, respectively; $\sigma = 0.8$ or 0.7, corresponding to the correlation between U and U^* of about 77% or 51%, respectively.

Table 3.2: Simulation results from the PH model with $U \sim Ber(0.5)$, $Z \sim N(0, 1)$

				w/o threshold					with threshold				
τ	rate	para	true	Bias	SSD	ESE	CP	RE	Bias	SSD	ESE	CP	RE
4.5		$\alpha_{\mathcal{A}}$	log(2)	0.027	0.110	0.111	0.935	1.000	0.005	0.105	0.104	0.945	1.000
		$\gamma_{\mathcal{A}}$	-0.5	-0.018	0.059	0.058	0.932	1.000	-0.002	0.056	0.054	0.943	1.000
		$\alpha_{\mathcal{I}}$	log(2)	-0.033	0.190	0.191	0.946	1.000	0.009	0.185	0.185	0.949	1.000
		$\gamma_{\mathcal{I}}$	-0.5	0.022	0.107	0.103	0.933	1.000	-0.009	0.101	0.099	0.942	1.000
		$\lambda_{\mathcal{I}}$	-0.5	0.050	0.139	0.136	0.937	1.000	-0.001	0.132	0.129	0.940	1.000
	30%	update $\alpha_{\mathcal{A}}$	log(2)	0.022	0.108	0.104	0.935	1.037	0.001	0.103	0.098	0.935	1.039
		update $\gamma_{\mathcal{A}}$	-0.5	-0.020	0.051	0.047	0.911	1.338	-0.006	0.047	0.043	0.927	1.420
		update $\alpha_{\mathcal{I}}$	log(2)	-0.031	0.186	0.177	0.930	1.043	0.011	0.178	0.172	0.939	1.080
		update $\gamma_{\mathcal{I}}$	-0.5	0.027	0.065	0.063	0.919	2.710	-0.004	0.058	0.057	0.945	3.032
		update $\lambda_{\mathcal{I}}$	-0.5	0.048	0.112	0.107	0.922	1.540	-0.004	0.103	0.098	0.934	1.642
	10%	update $\alpha_{\mathcal{A}}$	log(2)	0.021	0.097	0.094	0.928	1.286	0.000	0.093	0.088	0.933	1.275
		update $\gamma_{\mathcal{A}}$	-0.5	-0.018	0.049	0.045	0.911	1.450	-0.004	0.046	0.041	0.924	1.482
		update $\alpha_{\mathcal{I}}$	log(2)	-0.033	0.147	0.142	0.938	1.671	0.008	0.143	0.137	0.937	1.674
		update $\gamma_{\mathcal{I}}$	-0.5	0.027	0.063	0.061	0.918	2.885	-0.003	0.056	0.055	0.949	3.253
		update $\lambda_{\mathcal{I}}$	-0.5	0.045	0.096	0.095	0.919	2.096	-0.005	0.088	0.086	0.939	2.250
11		$\alpha_{\mathcal{A}}$	log(2)	0.004	0.093	0.094	0.957	1.000	0.003	0.093	0.094	0.957	1.000
		$\gamma_{\mathcal{A}}$	-0.5	-0.005	0.053	0.049	0.934	1.000	-0.005	0.053	0.049	0.934	1.000
		$\alpha_{\mathcal{I}}$	log(2)	0.004	0.177	0.175	0.950	1.000	0.004	0.177	0.175	0.950	1.000
		$\gamma_{\mathcal{I}}$	-0.5	-0.006	0.094	0.093	0.945	1.000	-0.006	0.093	0.093	0.944	1.000
		$\lambda_{\mathcal{I}}$	-0.5	0.003	0.125	0.120	0.940	1.000	0.002	0.125	0.120	0.940	1.000
	30%	update $\alpha_{\mathcal{A}}$	log(2)	0.000	0.090	0.089	0.949	1.068	0.001	0.091	0.089	0.947	1.044
		update $\gamma_{\mathcal{A}}$	-0.5	-0.006	0.043	0.040	0.931	1.519	-0.007	0.043	0.040	0.932	1.519
		update $\alpha_{\mathcal{I}}$	log(2)	0.005	0.172	0.163	0.933	1.059	0.004	0.173	0.163	0.931	1.047
		update $\gamma_{\mathcal{I}}$	-0.5	-0.001	0.053	0.053	0.954	3.146	-0.002	0.053	0.053	0.954	3.079
		update $\lambda_{\mathcal{I}}$	-0.5	0.001	0.100	0.092	0.920	1.563	0.001	0.100	0.092	0.924	1.563
	10%	update $\alpha_{\mathcal{A}}$	log(2)	0.000	0.082	0.081	0.947	1.286	0.001	0.082	0.081	0.947	1.286
		update $\gamma_{\mathcal{A}}$	-0.5	-0.006	0.042	0.039	0.927	1.592	-0.006	0.042	0.038	0.924	1.592
		update $\alpha_{\mathcal{I}}$	log(2)	0.001	0.136	0.130	0.929	1.694	0.000	0.137	0.130	0.929	1.669
		update $\gamma_{\mathcal{I}}$	-0.5	-0.001	0.052	0.052	0.953	3.268	-0.001	0.052	0.052	0.955	3.199
		update $\lambda_{\mathcal{I}}$	-0.5	0.001	0.086	0.080	0.923	2.113	0.001	0.086	0.080	0.918	2.113

Note: Bias, average estimate minus true value; SSD, sample standard deviation; ESE, average estimated standard error; CP, coverage proportion with 95% nominal level; RE, relative efficiency compared to the original estimators; τ = 4.5 or 11, yielding around censoring rate 66% or 61%, respectively; rate, the missclassification rate between U and U^*.

Table 3.3: Simulation results from the PO model with (U, Z) from bivariate normal distribution

				w/o threshold					with threshold				
τ	σ	para	true	Bias	SSD	ESE	CP	RE	Bias	SSD	ESE	CP	RE
11		$\alpha_{\mathcal{A}}$	log(2)	0.008	0.116	0.120	0.952	1.000	-0.011	0.112	0.113	0.939	1.000
		$\gamma_{\mathcal{A}}$	-0.5	-0.001	0.109	0.114	0.953	1.000	0.012	0.106	0.109	0.945	1.000
		$\alpha_{\mathcal{I}}$	log(2)	-0.001	0.123	0.151	0.956	1.000	0.109	0.117	0.114	0.947	1.000
		$\gamma_{\mathcal{I}}$	-0.5	-0.001	0.118	0.142	0.955	1.000	-0.016	0.114	0.110	0.942	1.000
		$\lambda_{\mathcal{I}}$	-0.5	0.034	0.111	0.211	0.953	1.000	-0.017	0.093	0.090	0.930	1.000
	1.7	update $\alpha_{\mathcal{A}}$	log(2)	0.008	0.115	0.114	0.941	1.017	-0.008	0.112	0.107	0.934	1.000
		update $\gamma_{\mathcal{A}}$	-0.5	-0.001	0.095	0.097	0.944	1.316	0.009	0.093	0.091	0.939	1.299
		update $\alpha_{\mathcal{I}}$	log(2)	-0.001	0.118	0.142	0.954	1.087	0.016	0.112	0.106	0.945	1.091
		update $\gamma_{\mathcal{I}}$	-0.5	-0.002	0.092	0.115	0.962	1.645	-0.011	0.082	0.080	0.947	1.933
		update $\lambda_{\mathcal{I}}$	-0.5	0.018	0.106	0.183	0.961	1.097	-0.021	0.056	0.056	0.934	2.758
	0.8	update $\alpha_{\mathcal{A}}$	log(2)	0.009	0.109	0.107	0.944	1.133	-0.007	0.105	0.100	0.929	1.138
		update $\gamma_{\mathcal{A}}$	-0.5	-0.001	0.094	0.093	0.943	1.345	0.009	0.090	0.087	0.940	1.387
		update $\alpha_{\mathcal{I}}$	log(2)	-0.003	0.104	0.129	0.951	1.399	0.015	0.097	0.093	0.942	1.455
		update $\gamma_{\mathcal{I}}$	-0.5	0.000	0.085	0.106	0.965	1.927	-0.009	0.076	0.074	0.942	2.250
		update $\lambda_{\mathcal{I}}$	-0.5	0.018	0.103	0.173	0.957	1.161	-0.022	0.055	0.055	0.929	2.859
50		$\alpha_{\mathcal{A}}$	log(2)	0.001	0.108	0.106	0.947	1.000	0.005	0.106	0.104	0.951	1.000
		$\gamma_{\mathcal{A}}$	-0.5	0.003	0.098	0.102	0.956	1.000	0.001	0.097	0.101	0.957	1.000
		$\alpha_{\mathcal{I}}$	log(2)	-0.001	0.115	0.111	0.943	1.000	0.007	0.112	0.107	0.942	1.000
		$\gamma_{\mathcal{I}}$	-0.5	0.000	0.106	0.106	0.953	1.000	-0.005	0.105	0.103	0.952	1.000
		$\lambda_{\mathcal{I}}$	-0.5	0.030	0.096	0.102	0.949	1.000	0.005	0.086	0.084	0.943	1.000
	1.7	update $\alpha_{\mathcal{A}}$	log(2)	0.004	0.107	0.102	0.940	1.019	0.006	0.105	0.100	0.937	1.019
		update $\gamma_{\mathcal{A}}$	-0.5	0.002	0.084	0.086	0.957	1.361	0.000	0.083	0.085	0.957	1.366
		update $\alpha_{\mathcal{I}}$	log(2)	-0.001	0.107	0.104	0.941	1.155	0.006	0.104	0.099	0.938	1.160
		update $\gamma_{\mathcal{I}}$	-0.5	0.003	0.079	0.079	0.945	1.800	-0.002	0.077	0.075	0.944	1.860
		update $\lambda_{\mathcal{I}}$	-0.5	0.021	0.080	0.085	0.949	1.440	-0.001	0.053	0.052	0.952	2.633
	0.8	update $\alpha_{\mathcal{A}}$	log(2)	0.002	0.099	0.095	0.940	1.190	0.005	0.098	0.093	0.942	1.170
		update $\gamma_{\mathcal{A}}$	-0.5	0.003	0.082	0.083	0.956	1.428	0.001	0.081	0.082	0.960	1.434
		update $\alpha_{\mathcal{I}}$	log(2)	-0.002	0.094	0.092	0.948	1.497	0.005	0.090	0.087	0.935	1.549
		update $\gamma_{\mathcal{I}}$	-0.5	0.004	0.074	0.073	0.946	2.052	-0.001	0.071	0.069	0.949	2.187
		update $\lambda_{\mathcal{I}}$	-0.5	0.019	0.079	0.084	0.951	1.477	-0.001	0.051	0.051	0.949	2.844

Note: Bias, average estimate minus true value; SSD, sample standard deviation; ESE, average estimated standard error; CP, coverage proportion with 95% nominal level; RE, relative efficiency compared to the original estimators; $\tau = 11$ or 50, yielding around censoring rate 67% or 62%, respectively; $\sigma = 0.8$ or 0.7, corresponding to the correlation between U and U^* of about 77% or 51%, respectively.

Table 3.4: Simulation results from the PO model with $U \sim Ber(0.5)$, $Z \sim N(0,1)$

				w/o threshold					with threshold				
τ	rate	para	true	Bias	SSD	ESE	CP	RE	Bias	SSD	ESE	CP	RE
11		$\alpha_{\mathcal{A}}$	log(2)	0.019	0.175	0.174	0.944	1.000	-0.002	0.168	0.167	0.943	1.000
		$\gamma_{\mathcal{A}}$	-0.5	-0.012	0.092	0.090	0.938	1.000	0.002	0.089	0.086	0.939	1.000
		$\alpha_{\mathcal{I}}$	log(2)	-0.010	0.191	0.214	0.948	1.000	0.014	0.183	0.181	0.940	1.000
		$\gamma_{\mathcal{I}}$	-0.5	0.006	0.102	0.107	0.943	1.000	-0.012	0.097	0.096	0.945	1.000
		$\lambda_{\mathcal{I}}$	-0.5	0.055	0.146	0.166	0.931	1.000	-0.022	0.130	0.125	0.937	1.000
	30%	update $\alpha_{\mathcal{A}}$	log(2)	0.017	0.169	0.166	0.937	1.072	-0.002	0.163	0.160	0.947	1.062
		update $\gamma_{\mathcal{A}}$	-0.5	-0.013	0.073	0.069	0.929	1.588	0.000	0.070	0.066	0.941	1.617
		update $\alpha_{\mathcal{I}}$	log(2)	-0.007	0.184	0.197	0.949	1.078	0.015	0.174	0.169	0.938	1.106
		update $\gamma_{\mathcal{I}}$	-0.5	0.008	0.062	0.063	0.955	2.707	-0.008	0.056	0.056	0.958	3.000
		update $\lambda_{\mathcal{I}}$	-0.5	0.047	0.120	0.128	0.916	1.480	-0.022	0.101	0.096	0.931	1.657
	10%	update $\alpha_{\mathcal{A}}$	log(2)	0.018	0.148	0.147	0.946	1.398	-0.001	0.143	0.142	0.947	1.380
		update $\gamma_{\mathcal{A}}$	-0.5	-0.013	0.073	0.068	0.932	1.588	0.000	0.070	0.065	0.935	1.617
		update $\alpha_{\mathcal{I}}$	log(2)	-0.011	0.153	0.159	0.950	1.558	0.010	0.139	0.135	0.944	1.733
		update $\gamma_{\mathcal{I}}$	-0.5	0.009	0.061	0.060	0.954	2.796	-0.007	0.055	0.054	0.961	3.110
		update $\lambda_{\mathcal{I}}$	-0.5	0.048	0.104	0.111	0.918	1.971	-0.022	0.086	0.084	0.934	2.285
50		$\alpha_{\mathcal{A}}$	log(2)	0.002	0.157	0.157	0.950	1.000	-0.001	0.156	0.156	0.949	1.000
		$\gamma_{\mathcal{A}}$	-0.5	-0.004	0.084	0.081	0.937	1.000	-0.003	0.084	0.080	0.937	1.000
		$\alpha_{\mathcal{I}}$	log(2)	0.004	0.169	0.173	0.964	1.000	0.007	0.168	0.172	0.964	1.000
		$\gamma_{\mathcal{I}}$	-0.5	-0.004	0.093	0.091	0.943	1.000	-0.006	0.092	0.091	0.944	1.000
		$\lambda_{\mathcal{I}}$	-0.5	0.010	0.122	0.119	0.945	1.000	0.000	0.121	0.118	0.944	1.000
	30%	update $\alpha_{\mathcal{A}}$	log(2)	0.000	0.158	0.150	0.941	0.987	-0.001	0.150	0.149	0.943	1.082
		update $\gamma_{\mathcal{A}}$	-0.5	-0.006	0.068	0.062	0.923	1.526	-0.004	0.067	0.062	0.922	1.572
		update $\alpha_{\mathcal{I}}$	log(2)	0.003	0.162	0.161	0.950	1.088	0.006	0.162	0.161	0.943	1.075
		update $\gamma_{\mathcal{I}}$	-0.5	-0.001	0.053	0.052	0.952	3.079	-0.003	0.053	0.052	0.951	3.079
		update $\lambda_{\mathcal{I}}$	-0.5	0.008	0.097	0.091	0.937	1.582	-0.002	0.094	0.090	0.940	1.657
	10%	update $\alpha_{\mathcal{A}}$	log(2)	0.006	0.134	0.134	0.946	1.373	0.003	0.133	0.133	0.948	1.376
		update $\gamma_{\mathcal{A}}$	-0.5	-0.005	0.067	0.062	0.926	1.572	-0.004	0.066	0.061	0.922	1.620
		update $\alpha_{\mathcal{I}}$	log(2)	-0.001	0.133	0.128	0.945	1.615	0.001	0.132	0.128	0.946	1.620
		update $\gamma_{\mathcal{I}}$	-0.5	0.000	0.051	0.051	0.950	3.325	-0.002	0.051	0.051	0.948	3.254
		update $\lambda_{\mathcal{I}}$	-0.5	0.008	0.085	0.079	0.928	2.060	-0.002	0.083	0.078	0.933	2.125

Note: Bias, average estimate minus true value; SSD, sample standard deviation; ESE, average estimated standard error; CP, coverage proportion with 95% nominal level; RE, relative efficiency compared to the original estimators; $\tau = 11$ or 50, yielding around censoring rate 67% or 62%, respectively; rate, the missclassification rate between U and U^*.

Table 3.5: Analysis results for the National Wilms' Tumor Study under the PH or PO model assuming that there is not a cure fraction

	PH Model						PO Model					
Variables	$\hat{\beta}$	SE	p-value	update $\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value	update $\hat{\beta}$	SE	p-value
Histology	1.517	0.135	0.000	1.554	0.109	0.000	1.809	0.185	0.000	1.854	0.156	0.000
Stage II	0.787	0.175	0.000	0.686	0.124	0.000	0.874	0.200	0.000	0.773	0.143	0.000
Stage III	0.844	0.186	0.000	0.852	0.129	0.000	0.922	0.213	0.000	0.931	0.149	0.000
Stage IV	1.304	0.211	0.000	1.201	0.149	0.000	1.470	0.244	0.000	1.337	0.169	0.000
Age	0.041	0.026	0.117	0.076	0.016	0.000	0.053	0.029	0.064	0.087	0.019	0.000

Note: SE, standard error estimates of $\hat{\beta}$.

Table 3.6: Analysis results for the National Wilms' Tumor Study under the PH or PO mixture cure model with threshold applied

	PH Model						Cure Portion					
Variables	$\hat{\beta}$	SE	p-value	update $\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value	update $\hat{\beta}$	SE	p-value
Intercept	–	–	–	–	–	–	-3.014	0.173	0.000	-3.062	0.135	0.000
Histology	0.356	0.138	0.010	0.332	0.115	0.004	1.669	0.165	0.000	1.734	0.136	0.000
Stage II	-0.062	0.179	0.728	0.096	0.143	0.502	0.852	0.195	0.000	0.730	0.140	0.000
Stage III	0.029	0.184	0.873	0.126	0.159	0.428	0.861	0.202	0.000	0.857	0.142	0.000
Stage IV	0.500	0.186	0.007	0.560	0.146	0.000	1.290	0.229	0.000	1.152	0.156	0.000
Age	-0.047	0.023	0.041	-0.037	0.017	0.034	0.071	0.028	0.012	0.103	0.018	0.000
	PO Model						Cure Portion					
Variables	$\hat{\beta}$	SE	p-value	update $\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value	update $\hat{\beta}$	SE	p-value
Intercept	–	–	–	–	–	–	-3.016	0.174	0.000	-3.065	0.136	0.000
Histology	0.743	0.230	0.001	0.678	0.198	0.001	1.675	0.167	0.000	1.734	0.138	0.000
Stage II	-0.150	0.251	0.549	-0.011	0.207	0.957	0.858	0.197	0.000	0.757	0.140	0.000
Stage III	0.185	0.261	0.479	0.293	0.215	0.174	0.855	0.203	0.000	0.858	0.142	0.000
Stage IV	0.694	0.308	0.024	0.673	0.247	0.006	1.294	0.231	0.000	1.166	0.157	0.000
Age	-0.107	0.034	0.002	-0.103	0.026	0.000	0.073	0.028	0.010	0.107	0.019	0.000

Note: SE, standard error estimates of $\hat{\beta}$.

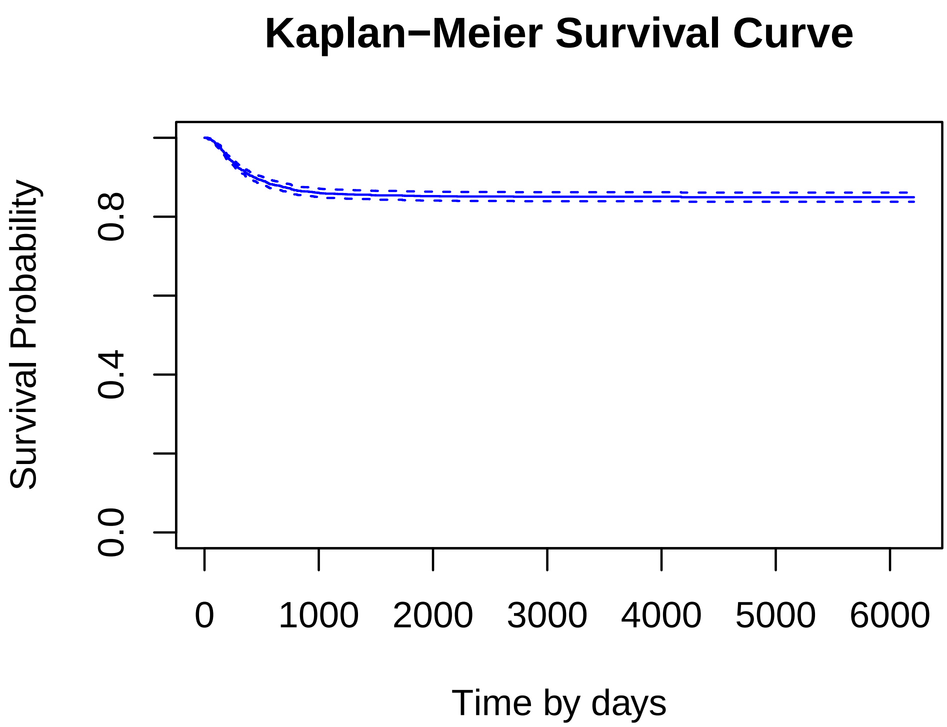

Figure 3.1: Kaplan-Meier estimators for the subjects of Wilms' Tumor Study

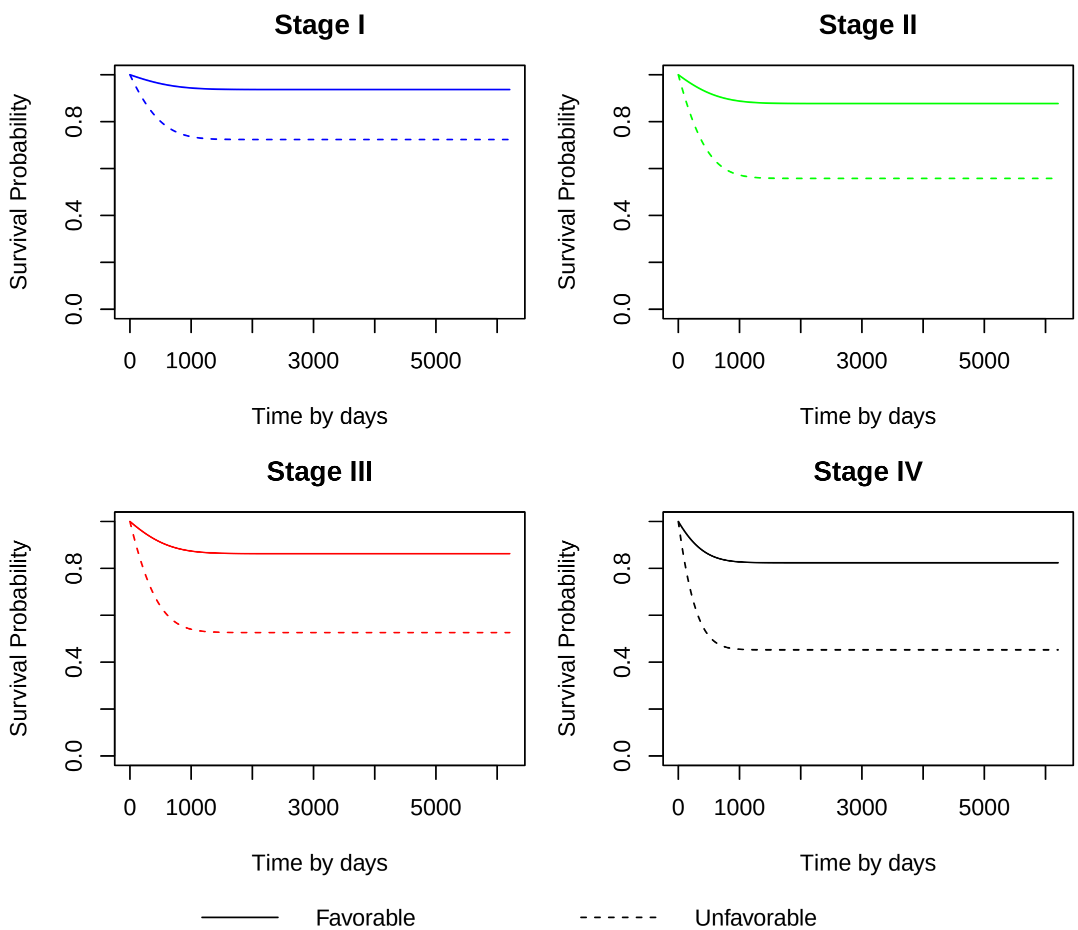

Figure 3.2: Estimated survival functions under the PH mixture cure model

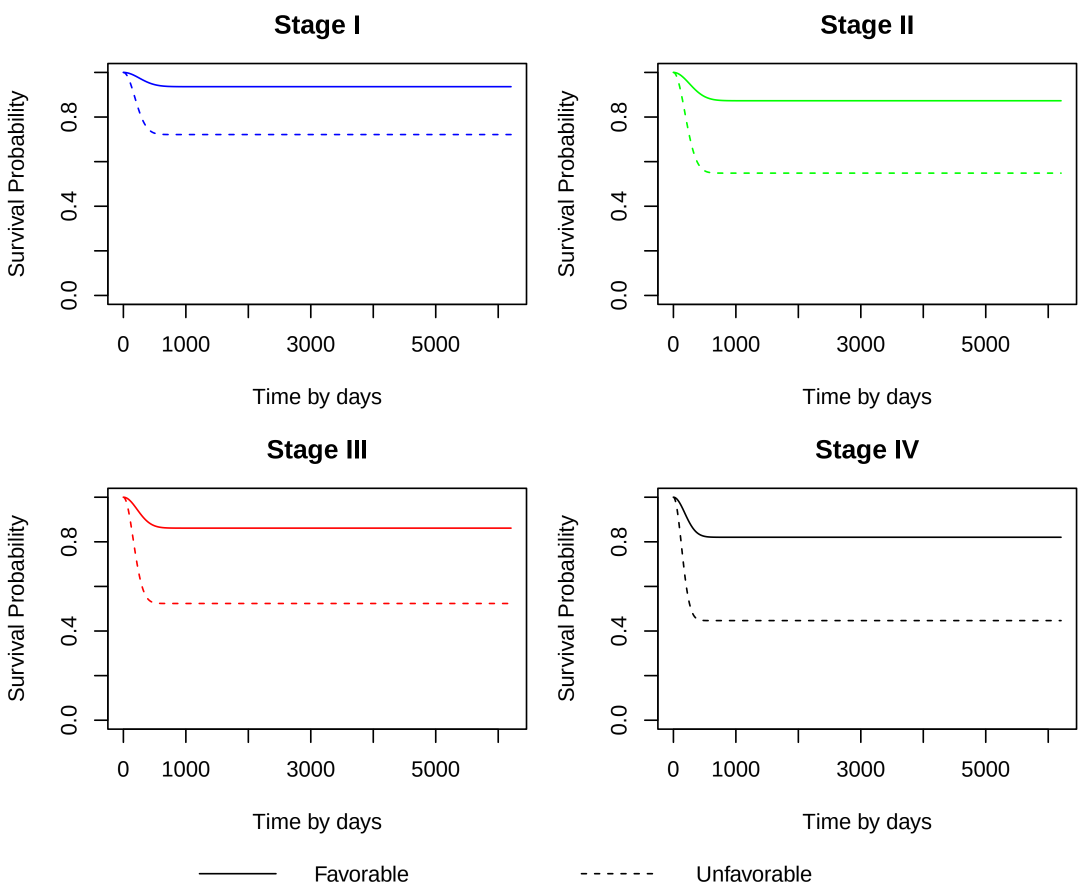

Figure 3.3: Estimated survival functions under the PO mixture cure model

CHAPTER 4: Future Research

In the context of the design and analysis of two-phase studies with survival data explored in this dissertation, several unresolved issues persist or demand more advanced methodologies. Within this chapter, we offer a concise overview of some of these issues, particularly those relevant to the investigations outlined in Chapters 2 and 3. Additionally, we highlight various directions for future research exploration.

In Chapter 2, we introduced a two-phase failure-time-auxiliary-dependent sampling (FADS) design for studies with expensive covariates and cheap surrogate or auxiliary variables. Additionally, we proposed a new semiparametric maximum pseudo-likelihood method for inference. While our simulations yielded promising results, several directions for future research merit exploration. One is the selection of design parameters, such as the cutoff points and the allocation of sample sizes. The complex form of the asymptotic variance provided in Theorem 2 renders it challenging to theoretically assess the impact of these parameters on estimation efficiency. The optimal choice of cutoff points and sample size allocation warrant future research. Another area ripe for exploration is the creation of auxiliary variables when unavailable in practice. One potential approach involves fitting a predictive model for the expensive covariate using the SRS component. Assessing the theoretical and numerical performance of this method represents a promising direction for future research. Lastly, while the proportional hazards model considered in this work is commonly employed, its applicability may be limited in certain scenarios. Extending the proposed method to alternative models, such as the proportional odds model and semiparametric transformation models, presents an opportunity for further advancement with minimal effort.

In Chapter 3, we presented a novel method aimed at enhancing estimation efficiency for case-cohort studies with a cure fraction. Our approach entails a two-step estimation procedure under semiparametric transformation mixture cure models, complemented by a weighted bootstrap procedure for variance estimation. Several directions for future research emerge. One such direction is to theoretically consider the exclusion of the "threshold" condition imposed in our study computations. This is motivated by the observation that similar results are obtained whether the threshold is applied or not, particularly in studies of sufficient duration. Additionally, exploring an optimal working model could potentially yield further efficiency gains, even though the proposed working model performs well in practice. Moreover, there is room for enhancing efficiency through advancements in sampling design. A promising approach in this regard is the adoption of a failure-time-dependent sampling design, as proposed by Ding *et al.* (2014). All three directions merit further investigation.

REFERENCES

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, **10**, 1100–1120.

Beesley, L. J., Bartlett, J. W., Wolf, G. T., and Taylor, J. M. G. (2016). Multiple imputation of missing covariates for the cox proportional hazards cure model. *Statistics in Medicine*, **35**, 4701–4717.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, **2**, 273–277.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, **46**, 1171–1178.

Breslow, N. (1982). Design and analysis of case-control studies. *Annual review of public health*, **3**, 29–54.

Breslow, N. E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Journal of the Royal Statistical Society, Series C*, **48**(4), 457–468.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, **34**(1), 86–102.

Cai, J. and Zeng, D. (2004). Sample size/power calculation for caseâcohort studies. *Biometrics*, **60**, 845–1057.

Cai, J. and Zeng, D. (2007). Power calculation for case–cohort studies with nonrare events. *Biometrics*, **63**(4), 1288–1295.

Chatterjee, N., Chen, Y.-H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, **98**(461), 158–168.

Chen, K. (2001). Generalized case-cohort sampling. *Journal of the Royal Statistical Society, Series B*, **63**(4), 791–809.

Chen, K. and Lo, S.-H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika*, **86**(4), 755–764.

Chen, M.-H., Joseph G., I., and Debajyoti, S. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**, 909–919.

Chen, Y.-H. (2002). Cox regression in cohort studies with validation sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 449â460.

Chen, Y.-H. and Chen, H. (2000). A unified approach to regression analysis under double- sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 449â460.

Cox, D. R. (1972). Regression models and life-tables. *J Royal Stat Soc, Ser B*, **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269â276.

Ding, J., Zhou, H., Liu, Y., Cai, J., and Longnecker, M. P. (2014). Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme. *Biostatistics*, **15**(4), 636–650.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, **89**, 463–475.

Fagerland, M. W. and Hosmer, D. W. (2013). A goodness-of-fit test for the proportional odds regression model. *Statistics in Medicine*, **32**, 2235–2249.

Fang, H.-B., Li, G., and Sun, J. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, **32**, 59–75.

Farewell, V. T. (1982). The use of a mixture model for the analysis of survival data with long-term survivors. *Biometrics*, **38**, 041â1046.

Foutz, R. V. (1977). On the unique solution to the likelihood equations. *Journal of the American Statistical Association*, **72**(357), 147–148.

Han, B. and Wang, X. (2020). Semiparametric estimation for the non-mixture cure model in case-cohort and nested case-control studies. *Computational Statistics & Data Analysis*, **144**.

Kang, S. and Cai, J. (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*, **96**(4), 887–901.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Kuk, Y. C. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**, 531â541.

Kuk, Y. C. and Chen, C.-H. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Ann Inst Stat Math*, **60**, 545â574.

Lu, W. and Tsiatis, A. A. (2006). Semiparametric transformation models for the case-cohort study. *Biometrika*, **93**(1), 207–214.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, **91**, 331–343.

Mao, M. and Wang, J.-L. (2010). Semiparametric efficient estimation for a class of generalized proportional odds cure models. *Journal of the American Statistical Association*, **105**, 302–311.

Marti, H. and Chavance, M. (2011). Multiple imputation analysis of case-cohort studies. *Statistics in Medicine*, **30**(13), 1595–1607.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **42**, 109–142.

Patilea, V. and Van Keilegom, I. (2020). A general approach for cure models in survival analysis. *Ann. Statist.*, **48(4)**, 2323–2346.

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

Schildcrout, J. S. and Rathouz, P. J. (2010). Longitudinal studies of binary response data following caseâcontrol and stratified caseâcontrol sampling: Design and analysis. *Biometrics*, **66**, 365â373.

Song, R., Zhou, H., and Kosorok, M. R. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, **96**(1), 221–228.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, **56**(1), 227–236.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, **51**, 899–907.

The ARIC Investigators (1989). The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. *American Journal of Epidemiology*, **129**(4), 687–702.

Tsiatis, A. A. (1981). A large sample study of cox's regression model. *The Annals of Statistics*, **9**(1), 93–108.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, **54**, 1508–1516.

Tsodikov, A., JG, I., and A. Y, Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**, 1063–1078.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vandenbroucke, J. P. and Pearce, N. (2012). Caseâcontrol studies: basic concepts. *International Journal of Epidemiology*, **41**, 1480–1489.

Wang, S. and Wang, C. (2001). A note on kernel assisted estimators in missing covariate regression. *Statistics and Probability Letters*, **55**(4), 439–449.

Wang, X. and Wang, Q. (2015). Semiparametric linear transformation model with differential measurement error and validation sampling. *Journal of Multivariate Analysis*, **141**, 67–80.

Wang, X. and Zhou, H. (2006). A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. *Biometrics*, **62**(4), 1149–1160.

Wang, X. and Zhou, H. (2010). Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. *Biometrics*, **66**(2), 502–511.

Wang, X., Wu, Y., and Zhou, H. (2009). Outcome-and auxiliary-dependent subsampling and its statistical inference. *Journal of Biopharmaceutical Statistics*, **19**(6), 1132–1150.

Weaver, M. A. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, **100**(470), 459–469.

Weinberg, C. R. and Wacholder, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika*, **80**(2), 461–465.

Whittemore, A. S. (1997). Multistage sampling designs and estimating equations. *Journal of the Royal Statistical Society: Series B*, **59**(3), 589–602.

Xie, P., Han, B. H., and Wang, X. W. (2023). Case-cohort studies for clustered failure time data with a cure fraction. *Statistical Papers*.

Yakovlev, A. and Tsodikov, A. (1996). *Stochastic models of tumor latency and their biostatistical applications*. World Scientific.

Yan, Y., Zhou, H., and Cai, J. (2017). Improving efficiency of parameter estimation in case-cohort studies with multivariate failure time data. *Biometrics*, **73**, 1042â1052.

Yang, S. and Ding, P. (2020). Journal of the american statistical association. *Biometrics*, **115**, 1540â1554.

Zeng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **69**, 507–564.

Zeng, D. and Lin, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, **93**, 627â640.

Zhou, H., Weaver, M., Qin, J., Longnecker, M., and Wang, M. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, **58**(2), 413–421.

Zhou, H., Wu, Y., Liu, Y., and Cai, J. (2011a). Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. *Biostatistics*, **12**(3), 521–534.

Zhou, H., Song, R., Wu, Y., and Qin, J. (2011b). Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome. *Biometrics*, **67**(1), 194–202.

Zhou, Q. and Wong, K. Y. (2023). Improving estimation efficiency of case-cohort study with interval-censored failure time data. *arXiv:2310.15070*.

Zhou, Q., Zhou, H., and Cai, J. (2017). Case-cohort studies with interval-censored failure time data. *Biometrika*, **104**, 17â29.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2

This appendix includes the proofs of Theorems 1 and 2 in chapter 2. In the following, we first present the regularity conditions and a useful lemma for the proofs.

A.1 Regularity Conditions

As in Section 2, we denote the true values of (β, Λ, G) by $(\beta_0, \Lambda_0, G_0)$ and define $n_V = |V|$, $n_k = |\tilde{V}_k|$ for $k = 0, \ldots, K$, and $N_k = |S_k|$ for $k = 1, \ldots, K+1$. The following conditions are needed to establish the asymptotic properties of $\hat{\beta}$.

(C1) β_0 lies in the interior of a known compact set $\mathcal{B}$ in $\mathbb{R}^p$, and $\Lambda_0(\cdot)$ is twice continuously differentiable with positive derivatives in $[0, \tau]$, where τ is the length of study.

(C2) The support of $(X', Z')'$ is bounded and not a proper subset of $\mathbb{R}^p$.

(C3) $\tilde{T}$ and C are conditionally independent given X and Z. Also, $P(T \geq \tau) > 0$.

(C4) As $N \to \infty$, $n_V/N \to \rho_V > 0$, $n_k/n_V \to \rho_k \geq 0$ for $k = 1, \ldots, K$, $n_0/n_V \to \rho_0 > 0$, and $N_k/N \to \gamma_k > 0$ for $k = 1, \cdots, K+1$.

(C5) $\phi(\cdot)$ is a d-dimensional α-th order bounded and symmetric kernel function with bounded support and $\int \phi^2 < \infty$, where d and α are positive integers. Also, $Nh^{2\alpha} \to 0$ and $Nh^{2d} \to \infty$ as $N \to \infty$.

A.2 A Useful Lemma

Lemma 1. Suppose that $\mathcal{F} = \{\xi(y,z,w,x;\beta,\Lambda) : \beta \in \mathcal{B}, \Lambda \in \mathcal{A}\}$ is a Donsker class of functions. Then

$$\sup_{\beta\in\mathcal{B}} \sup_{\Lambda\in\mathcal{A}} \left| \frac{\sum_{i\in V_k} \xi(y,z,w,X_i;\beta,\Lambda)\phi_h(U_i-u)}{\sum_{i\in V_k} \phi_h(U_i-u)} - \int_{\mathcal{X}} \xi(y,z,w,x;\beta,\Lambda)\, G(dx|u,(y,w)\in S_k) \right| = O_P(\eta_N),$$

where $\eta_N = \left[Nh^{2\alpha} + (Nh^{2d})^{-1}\right]^{1/2}$.

Proof: First we define

$$\mu_k(y,z,w,u;\beta,\Lambda) = \frac{1}{h^d\, n_{V_k}} \sum_{i\in V_k} \xi(y,z,w,X_i;\beta,\Lambda)\phi_h(U_i-u)$$

and

$$\nu_k(u) = \frac{1}{h^d\, n_{V_k}} \sum_{i\in V_k} \phi_h(U_i-u).$$

Since $\xi(y,z,w,x;\beta,\Lambda)$ belongs to a Donsker class, we have

$$\mu_k(y,z,w,u;\beta,\Lambda) \to \int_{\mathcal{X}} \xi(y,z,w,x;\beta,\Lambda)\, q(x,u|(y,w)\in S_k)\, dx,$$

almost surely, uniformly for all $\beta \in \mathcal{B}$ and $\Lambda \in \mathcal{A}$, where $q(x,u|(y,w)\in S_k)$ is the joint density function of (X,U) given $(Y,W) = (y,w) \in S_k$. By taking $\xi \equiv 1$, we have

$$\nu_k(u) \to \int_{\mathcal{X}} q(x,u|(y,w)\in S_k)\, dx,$$

almost surely. Hence, by the Slutsky's Theorem,

$$\sup_{\beta\in\mathcal{B}} \sup_{\Lambda\in\mathcal{A}} \left| \frac{\mu_k(y,z,w,u;\beta,\Lambda)}{\nu_k(u)} - \int_{\mathcal{X}} \xi(y,z,w,x;\beta,\Lambda)\, G(dx|u,(y,w)\in S_k) \right| \to 0,$$

almost surely. By the kernel estimation theory and Lemma 1 in Wang and Wang (2001), we can derive that

$$\sup_{\beta\in\mathcal{B}}\sup_{\Lambda\in\mathcal{A}}\left|\frac{\mu_k(y,z,w,u;\beta,\Lambda)}{\nu_k(u)}-\int_{\mathcal{X}}\xi(y,z,w,x;\beta,\Lambda)G(dx|u,(y,w)\in S_k)\right|=O_p(\eta_N),$$

which completes the proof.

A.3 Proof of Theorem 1

The full log-likelihood function based on data from the two-phase FADS design is given by

$$\begin{aligned}
l(\beta,\Lambda,G)=&\sum_{k=0}^{K}\sum_{i\in\tilde{V}_k}\Big[\log f_{\beta,\Lambda}(T_i,\Delta_i|X_i,Z_i)+\log g(X_i|W_i,Z_i)\Big]\\
&+\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\log f_{\beta,\Lambda,G}(T_j,\Delta_j|W_j,Z_j)\\
=&\sum_{k=0}^{K}\sum_{i\in\tilde{V}_k}\Big[\log f_{\beta,\Lambda}(T_i,\Delta_i|X_i,Z_i)+\log g(X_i|W_i,Z_i)\Big]\\
&+\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\log\Big[\int f_{\beta,\Lambda}(T_j,\Delta_j|x,Z_j)\,dG(x|W_j,Z_j)\Big].
\end{aligned}$$

The pseudo-log-likelihood function obtained by replacing G with its estimator $\hat{G}$ in the full log-likelihood can be written as

$$\begin{aligned}
\hat{l}(\beta,\Lambda,\hat{G})=&\sum_{k=0}^{K}\sum_{i\in\tilde{V}_k}\log f_{\beta,\Lambda}(T_i,\Delta_i|X_i,Z_i)+\sum_{k=0}^{K+1}\sum_{j\in\bar{V}_k}\log\hat{f}_{\beta,\Lambda,\hat{G}}(T_j,\Delta_j|W_j,Z_j)\\
=&\sum_{k=0}^{K}\sum_{i\in\tilde{V}_k}\Big[\Delta_i\big\{\log\lambda(T_i)+(\beta_1'X_i+\beta_2'Z_i)\big\}-\Lambda(T_i)\exp(\beta_1'X_i+\beta_2'Z_i)\Big]\\
&+\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\log\left[\sum_{r=1}^{K+1}\hat{\pi}_r(U_j)\frac{\sum_{l\in V_r}f_{\beta,\Lambda}(T_j,\Delta_j|X_l,Z_j)\,\phi_h(U_l-U_j)}{\sum_{l\in V_r}\phi_h(U_l-U_j)}\right]
\end{aligned}$$

Then the score function for β based on the full log-likelihood is given by

$$
\begin{aligned}
U_F(\beta, \Lambda, G) &= \frac{\partial l(\beta, \Lambda, G)}{\partial \beta} \\
&= \sum_{k=0}^{K} \sum_{i \in \tilde{V}_k} \left\{ \Delta_i \, [X_i', Z_i']' - \Lambda(T_i) \exp(\beta_1' X_i + \beta_2' Z_i) \, [X_i', Z_i']' \right\} \\
&\quad + \sum_{k=1}^{K+1} \sum_{j \in \bar{V}_k} \frac{\int \left[\partial f_{\beta,\Lambda}(T_j, \Delta_j | x, Z_j) / \partial \beta \right] dG(x | W_j, Z_j)}{\int f_{\beta,\Lambda}(T_j, \Delta_j | x, Z_j) \, dG(x | W_j, Z_j)}.
\end{aligned}
$$

The pseudo-score function for β based on the pseudo-log-likelihood has the form

$$
\begin{aligned}
U_F(\beta, \Lambda, \hat{G}) &= \frac{\partial \hat{l}(\beta, \Lambda, \hat{G})}{\partial \beta} \\
&= \sum_{k=0}^{K} \sum_{i \in \tilde{V}_k} \left\{ \Delta_i \, [X_i', Z_i']' - \Lambda(T_i) \exp(\beta_1' X_i + \beta_2' Z_i) \, [X_i', Z_i']' \right\} \\
&\quad + \sum_{k=1}^{K+1} \sum_{j \in \bar{V}_k} \frac{\sum_{r=1}^{K+1} \hat{\pi}_r(U_j) \frac{\sum_{l \in V_r} \left[\partial f_{\beta,\Lambda}(T_j, \Delta_j | X_l, Z_j) / \partial \beta \right] \phi_h(U_l - U_j)}{\sum_{l \in V_r} \phi_h(U_l - U_j)}}{\sum_{r=1}^{K+1} \hat{\pi}_r(U_j) \frac{\sum_{l \in V_r} f_{\beta,\Lambda}(T_j, \Delta_j | X_l, Z_j) \phi_h(U_l - U_j)}{\sum_{l \in V_r} \phi_h(U_l - U_j)}}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
\frac{1}{N} U_F(\beta, \hat{\Lambda}, \hat{G}) - \frac{1}{N} U_F(\beta, \Lambda_0, G_0) &= \left\{ \frac{1}{N} U_F(\beta, \hat{\Lambda}, \hat{G}) - \frac{1}{N} U_F(\beta, \hat{\Lambda}, G_0) \right\} \\
&\quad + \left\{ \frac{1}{N} U_F(\beta, \hat{\Lambda}, G_0) - \frac{1}{N} U_F(\beta, \Lambda_0, G_0) \right\}.
\end{aligned} \tag{A.1}
$$

We will show that both of the two terms on the right-hand side of (A.1) converge to 0. For the first term, by selecting a suitable function ξ in Lemma 1, we can prove that

$$
\frac{1}{N} U_F(\beta, \hat{\Lambda}, \hat{G}) - \frac{1}{N} U_F(\beta, \hat{\Lambda}, G_0) \xrightarrow{p} 0, \tag{A.2}
$$

uniformly for all $\beta \in \mathcal{B}$. In fact, define the class of functions

$$\begin{aligned}\mathcal{F}_\omega =&\Big\{ f_{\beta,\Lambda}(T,\Delta|X,Z) : \beta \in \mathcal{B},\, \Lambda \in BV_\omega[0,\tau] \Big\}\\ =&\Big\{ \big[\lambda(T)\exp(\beta_1'X+\beta_2'Z)\big]^\Delta \exp\{-\Lambda(T)\exp(\beta_1'X+\beta_2'Z)\} : \beta \in \mathcal{B},\, \Lambda \in BV_\omega[0,\tau] \Big\},\end{aligned}$$

where $BV_\omega[0,\tau]$ denotes the class of functions with the total variation in $[0,\tau]$ bounded by a given constant ω. Since X and Z are bounded, $\mathcal{F}_\omega$ is a Donkser class by Example 19.11 in van der Vaart (1998). Similarly, we can show that

$$\mathcal{F}_\omega' = \left\{ \frac{\partial f_{\beta,\Lambda}(T,\Delta|X,Z)}{\partial\beta} : \beta \in \mathcal{B},\, \Lambda \in BV_\omega[0,\tau] \right\},$$

is also a Donsker class. Note that $\hat{\pi}_r(u)$ is a consistent estimator of $\pi_r(u)$ and $\hat{\Lambda} \in BV_\omega[0,\tau]$, then (A.2) follows from Lemma 1.

For the second term on the right-hand side of (A.1), since $\hat{\Lambda}$ is a consistent estimator of Λ, by the continuous mapping theorem, we have

$$\frac{1}{N}U_F(\beta,\hat{\Lambda},G_0) - \frac{1}{N}U_F(\beta,\Lambda_0,G_0) \xrightarrow{p} 0,$$

uniformly for $\beta \in \mathcal{B}$. Thus, we have shown that (A.1) converges to 0 uniformly for $\beta \in \mathcal{B}$. Furthermore, by the strong law of large numbers,

$$\begin{aligned}\frac{1}{N}U_F(\beta,\Lambda_0,G_0) =&\rho_0\rho_V E\left\{ \frac{\partial f_{\beta,\Lambda_0}(T,\Delta|X,Z)/\partial\beta}{f_{\beta,\Lambda_0}(T,\Delta|X,Z)} \right\} + \sum_{k=1}^{K} \rho_k\rho_V E_k\left\{ \frac{\partial f_{\beta,\Lambda_0}(T,\Delta|X,Z)/\partial\beta}{f_{\beta,\Lambda_0}(T,\Delta|X,Z)} \right\}\\ &+ \sum_{k=1}^{K+1} \big[\gamma_k(1-\rho_0\rho_V) - \rho_k\rho_V\big] E_k\left\{ \frac{\partial f_{\beta,\Lambda_0,G_0}(T,\Delta|W,Z)/\partial\beta}{f_{\beta,\Lambda_0,G_0}(T,\Delta|W,Z)} \right\} + o_p(1).\end{aligned}$$

It then follows that $\frac{1}{N}U_F(\beta_0,\Lambda_0,G_0) \xrightarrow{p} 0$. Since (A.1) converges to 0, we have

$$\frac{1}{N}U_F(\beta_0,\hat{\Lambda},\hat{G}) \xrightarrow{p} 0. \tag{A.3}$$

Similarly as showing that (A.1) converges to 0, we can prove that

$$\frac{1}{N}\frac{\partial U_F(\beta,\hat{\Lambda},\hat{G})}{\partial\beta}-\frac{1}{N}\frac{\partial U_F(\beta,\Lambda_0,G_0)}{\partial\beta}\xrightarrow{p}0,$$

uniformly for all $\beta\in\mathcal{B}$, as $N\to\infty$. Further note that uniformly for $\beta\in\mathcal{B}$,

$$-\frac{1}{N}\frac{\partial U_F(\beta,\Lambda_0,G_0)}{\partial\beta}\xrightarrow{p}I(\beta),$$

where $I(\beta)$ is the information matrix of β with known (Λ_0,G_0), given by

$$\begin{aligned}I(\beta)=&-\rho_0\rho_V E\left\{\frac{\partial^2\log f_{\beta,\Lambda_0}(T,\Delta|X,Z)}{\partial\beta\partial\beta'}\right\}-\sum_{k=1}^{K}\rho_k\rho_V E_k\left\{\frac{\partial^2\log f_{\beta,\Lambda_0}(T,\Delta|X,Z)}{\partial\beta\partial\beta'}\right\}\\&-\sum_{k=1}^{K+1}\left[\gamma_k(1-\rho_0\rho_V)-\rho_k\rho_V\right]\left\{\frac{\partial^2\log f_{\beta,\Lambda_0,G_0}(T,\Delta|Z,W)}{\partial\beta\partial\beta'}\right\}.\end{aligned}$$

Then we have

$$-\frac{1}{N}\frac{\partial U_F(\beta,\hat{\Lambda},\hat{G})}{\partial\beta}\xrightarrow{p}I(\beta),\tag{A.4}$$

uniformly for $\beta\in\mathcal{B}$. Therefore, combining (A.3) and (A.4), it follows from Foutz (1977) and Weaver and Zhou (2005) that $\hat{\beta}$ is a consistent estimator of β_0.

A.4 Proof of Theorem 2

To account for the variability induced by using $\hat{\Lambda}$ and $\hat{G}$ in the pseudo-likelihood function, we decompose the pseudo-score function $U_F(\beta,\hat{\Lambda},\hat{G})$ into three terms as

$$\begin{aligned}\frac{1}{\sqrt{N}}U_F(\beta,\hat{\Lambda},\hat{G})=&\frac{1}{\sqrt{N}}U_F(\beta,\Lambda_0,G_0)\\&+\left\{\frac{1}{\sqrt{N}}U_F(\beta,\hat{\Lambda},G_0)-\frac{1}{\sqrt{N}}U_F(\beta,\Lambda_0,G_0)\right\}\\&+\left\{\frac{1}{\sqrt{N}}U_F(\beta,\hat{\Lambda},\hat{G})-\frac{1}{\sqrt{N}}U_F(\beta,\hat{\Lambda},G_0)\right\}.\end{aligned}\tag{A.5}$$

In the following, we will derive the limiting distribution for each of the three terms on the right-hand side of (A.5) and also show that the three terms are asymptotically independent.

The first term of (A.5) is given by

$$\frac{1}{\sqrt{N}} U_F(\beta, \Lambda_0, G_0) = \frac{1}{\sqrt{N}} \sum_{k=0}^{K} \sum_{i \in \tilde{V}_k} \frac{\frac{\partial}{\partial \beta} f_{\beta,\Lambda_0}(T_i, \Delta_i | X_i, Z_i)}{f_{\beta,\Lambda_0}(T_i, \Delta_i | X_i, Z_i)} + \frac{1}{\sqrt{N}} \sum_{k=1}^{K+1} \sum_{j \in \bar{V}_k} \frac{\frac{\partial}{\partial \beta} f_{\beta,\Lambda_0,G_0}(T_j, \Delta_j | Z_j, W_j)}{f_{\beta,\Lambda_0,G_0}(T_j, \Delta_j | Z_j, W_j)}. \tag{A.6}$$

It is easy to show that

$$\frac{1}{\sqrt{N}} U_F(\beta_0, \Lambda_0, G_0) \xrightarrow{d} \mathcal{N}(0, I(\beta_0)), \tag{A.7}$$

where $I(\beta)$ is the information matrix of β with known (Λ_0, G_0) and is defined in the proof of Theorem 1.

For the second term of (A.5), note that

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}U_F(\beta,\hat{\Lambda},G_0)-\frac{1}{\sqrt{N}}U_F(\beta,\Lambda_0,G_0)\\
=&\frac{1}{\sqrt{N}}\Bigg\{\sum_{k=0}^{K}\sum_{i\in\tilde{V}_k}\left[\frac{\partial}{\partial\beta}\log f_{\beta,\hat{\Lambda}}(T_i,\Delta_i|X_i,Z_i)-\frac{\partial}{\partial\beta}\log f_{\beta,\Lambda_0}(T_i,\Delta_i|X_i,Z_i)\right]\\
&\quad+\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\int_{\mathcal{X}}\left[\frac{\partial}{\partial\beta}\log f_{\beta,\hat{\Lambda}}(T_j,\Delta_j|x,Z_j)-\frac{\partial}{\partial\beta}\log f_{\beta,\Lambda_0}(T_j,\Delta_j|x,Z_j)\right]dG_0(x|W_j,Z_j)\Bigg\}\\
=&\frac{1}{\sqrt{N}}\Bigg\{\sum_{k=0}^{K}\sum_{i\in\tilde{V}_k}\left\{\left[-\hat{\Lambda}(T_i)+\Lambda_0(T_i)\right]\exp(\beta_1'X_i+\beta_2'Z_i)\,[X_i',Z_i']'\right\}\\
&\quad+\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\int_{\mathcal{X}}\left[-\hat{\Lambda}(T_j)+\Lambda_0(T_j)\right]\exp(\beta_1'x+\beta_2'Z_j)\,\left[x',Z_j'\right]'dG_0(x|W_j,Z_j)\Bigg\}\\
=&\frac{1}{\sqrt{N}}\Bigg\{\sum_{k=0}^{K}\sum_{i\in\tilde{V}_k}\left\{\int_0^\tau -I(t\le T_i)\,d(\hat{\Lambda}(t)-\Lambda_0(t))\exp(\beta_1'X_i+\beta_2'Z_i)\,[X_i',Z_i']'\right\}\\
&\quad+\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\int_{\mathcal{X}}\int_0^\tau -I(t\le T_j)\,d(\hat{\Lambda}(t)-\Lambda_0(t))\exp(\beta_1'x+\beta_2'Z_j)\,\left[x',Z_j'\right]'dG_0(x|W_j,Z_j)\Bigg\}\\
=&\int_0^\tau -\frac{|V|}{\sqrt{N\cdot n_0}}\sum_{k=0}^{K}\frac{n_k}{|V|}\frac{1}{n_k}\sum_{i\in\tilde{V}_k}\left\{I(t\le T_i)\exp(\beta_1'X_i+\beta_2'Z_i)\,[X_i',Z_i']'\right\}\\
&\quad-\frac{|\bar{V}|}{\sqrt{N\cdot n_0}}\sum_{k=1}^{K+1}\frac{N_k-n_k}{|\bar{V}|}\frac{1}{N_k-n_k}\sum_{j\in\bar{V}_k}I(t\le T_j)\Bigg\{\int_{\mathcal{X}}\exp(\beta_1'x+\beta_2'Z_j)\,\left[x',Z_j'\right]'\\
&\qquad dG_0(x|W_j,Z_j)\Bigg\}d\sqrt{n_0}(\hat{\Lambda}(t)-\Lambda_0(t))\\
=&\int_0^\tau -\frac{\rho_V}{\sqrt{\rho_0}}\sum_{k=0}^{K}\rho_k E_k\left\{I(t\le T)\exp(\beta_1'X+\beta_2'Z)\,[X',Z']'\right\}\\
&\quad-\frac{1-\rho_V}{\sqrt{\rho_0}}\int_0^\tau\sum_{k=1}^{K+1}\frac{\gamma_k-\rho_V\rho_k}{1-\rho_V}E_k\left\{I(t\le T)\int_{\mathcal{X}}\exp(\beta_1'x+\beta_2'Z)\,[x',Z']'\,dG_0(x|W,Z)\right\}\\
&\qquad d\sqrt{n_0}(\hat{\Lambda}(t)-\Lambda_0(t))+o_p(1).
\end{aligned}
$$

Let $H(t;\beta)$ denote the integrand in the last equation above. Then we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}U_F(\beta,\hat{\Lambda},G_0)-\frac{1}{\sqrt{N}}U_F(\beta,\Lambda_0,G_0)\\
=&\int_0^\tau H(t;\beta)\,d\sqrt{n_0}(\hat{\Lambda}(t)-\Lambda_0(t))+o_p(1)\\
=&\sqrt{n_0}\Big\{\zeta(\hat{\Lambda};\beta)-\zeta(\Lambda_0;\beta)\Big\}+o_p(1).
\end{aligned} \tag{A.8}$$

where

$$\zeta(\Lambda;\beta)=\int_0^\tau H(t;\beta)\,d\Lambda(t).$$

As shown in Tsiatis (1981), $\sqrt{n_0}(\hat{\Lambda}-\Lambda_0)$ converges weakly to a mean zero Gaussian process $\mathbb{G}$. Then by Theorem 20.8 (Delta method) in van der Vaart (1998), we have

$$\sqrt{n_0}\Big\{\zeta(\hat{\Lambda};\beta)-\zeta(\Lambda_0;\beta)\Big\}\xrightarrow{d}\zeta'_{\Lambda_0}(\mathbb{G};\beta),$$

where

$$\begin{aligned}
\zeta'_{\Lambda_0}(\mathbb{G};\beta)&=\frac{\partial}{\partial\epsilon}\int_0^\tau H(t;\beta)\,(1+\epsilon\mathbb{G}(t))\,d\Lambda_0(t)\Big|_{\epsilon=0}\\
&=\int_0^\tau H(t;\beta)\,\mathbb{G}(t)\,d\Lambda_0(t).
\end{aligned}$$

Then by (A.8) and Slutsky's Theorem, we obtain

$$\frac{1}{\sqrt{N}}U_F(\beta_0,\hat{\Lambda},G_0)-\frac{1}{\sqrt{N}}U_F(\beta_0,\Lambda_0,G_0)\xrightarrow{d}\mathcal{N}(0,\Sigma_{\mathbb{G}}(\beta_0)), \tag{A.9}$$

where

$$\Sigma_{\mathbb{G}}(\beta)=\mathrm{Var}\Big\{\int_0^\tau H(t;\beta)\,\mathbb{G}(t)\,d\Lambda_0(t)\Big\}.$$

For the third term of (A.5), similar to Zhou *et al.* (2011a), we have

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}U_F(\beta,\hat{\Lambda},\hat{G})-\frac{1}{\sqrt{N}}U_F(\beta,\hat{\Lambda},G_0)\\
=&\frac{1}{\sqrt{N}}\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\left\{\frac{\frac{\partial}{\partial\beta}\hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j,\Delta_j|W_j,Z_j)}{\hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j,\Delta_j|W_j,Z_j)}-\frac{\frac{\partial}{\partial\beta}f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)}{f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)}\right\}\\
=&\frac{1}{\sqrt{N}}\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\left\{\frac{\frac{\partial}{\partial\beta}\hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j,\Delta_j|W_j,Z_j)}{f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)}-\frac{\frac{\partial}{\partial\beta}f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)}{[f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)]^2}\hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j,\Delta_j|W_j,Z_j)\right\}\\
&\qquad\times\left\{\frac{f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)}{\hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j,\Delta_j|W_j,Z_j)}\right\}\\
=&\frac{1}{\sqrt{N}}\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\left\{\frac{\frac{\partial}{\partial\beta}\hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j,\Delta_j|W_j,Z_j)}{f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)}-\frac{\frac{\partial}{\partial\beta}f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)}{[f_{\beta,\hat{\Lambda},G_0}(T_j,\Delta_j|W_j,Z_j)]^2}\hat{f}_{\beta,\hat{\Lambda},\hat{G}}(T_j,\Delta_j|W_j,Z_j)\right\}\\
&+O_p(\eta_N)\\
=&\frac{1}{\sqrt{N}}D_F(\beta,\hat{\Lambda},\hat{G})+O_p(\eta_N),
\end{aligned}
$$

where the second last equality can be shown by Lemma 1 as in the proof of consistency and the summation in the second last equation is denoted by $D_F(\beta,\hat{\Lambda},\hat{G})$. We will establish the weak convergence of $\frac{1}{\sqrt{N}}D_F(\beta,\hat{\Lambda},\hat{G})$. Note that

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}D_F(\beta,\hat{\Lambda},\hat{G})\\
=&\frac{1}{\sqrt{N}}\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\sum_{r=1}^{K+1}\hat{\pi}_r(U_j)\frac{\sum_{i\in V_r}M_{X_i,U_i}(T_j,\Delta_j,W_j,Z_j;\beta,\hat{\Lambda})\phi_h(U_i-U_j)}{\sum_{i\in V_r}\phi_h(U_i-U_j)}\\
=&\frac{1}{\sqrt{N}}\sum_{r=1}^{K+1}\sum_{i\in V_r}\sum_{k=1}^{K+1}\sum_{j\in\bar{V}_k}\frac{N_r(U_j)}{n_{V_r}(U_j)}\frac{n_{\bar{V}_k}(U_j)}{N(U_j)}\frac{M_{X_i,U_i}(T_j,\Delta_j,W_j,Z_j;\beta,\hat{\Lambda})\phi_h(U_i-U_j)}{n_{\bar{V}_k}(U_j)}\\
=&\frac{1}{\sqrt{N}}\sum_{r=1}^{K+1}\frac{\gamma_r}{\rho_r\rho_V+\gamma_r\rho_0\rho_V}\sum_{i\in V_r}\sum_{k=1}^{K+1}\big[\gamma_k(1-\rho_0\rho_V)-\rho_0\rho_V\big]\pi_k(U_i)E_k\big\{M_{X_i,U_i}(T,\Delta,W,Z;\beta,\Lambda_0)\big|U_i\big\}\\
&+o_p(1)
\end{aligned}
$$

where

$$N(U_j) = \sum_{i=1}^{N} \phi_h(U_i - U_j), \quad N_r(U_j) = \sum_{i \in S_r} \phi_h(U_i - U_j),$$

$$n_{V_r}(U_j) = \sum_{i \in V_r} \phi_h(U_i - U_j), \quad n_{\bar{V}_k}(U_j) = \sum_{i \in \bar{V}_k} \phi_h(U_i - U_j),$$

$$M_{X_i,U_i}(T,\Delta,W,Z;\beta,\Lambda) = \frac{\frac{\partial}{\partial\beta} f_{\beta,\Lambda}(T,\Delta|X_i,Z)}{f_{\beta,\Lambda,G_0}(T,\Delta|W,Z)} - \frac{\frac{\partial}{\partial\beta} f_{\beta,\Lambda,G_0}(T,\Delta|W,Z)}{[f_{\beta,\Lambda,G_0}(T,\Delta|W,Z)]^2} f_{\beta,\Lambda}(T,\Delta|X_i,Z).$$

By Liapounov's Central Limit Theroem and Cramér-Wold Theorem, we have

$$\frac{1}{\sqrt{N}} D_F(\beta_0, \hat{\Lambda}, \hat{G}) \xrightarrow{d} \mathcal{N}\left(0,\ \sum_{k=1}^{K+1} \frac{\gamma_k^2}{\rho_k\rho_V + \gamma_k\rho_0\rho_V} \Sigma_k(\beta_0)\right), \tag{A.10}$$

where

$$\Sigma_k(\beta) = \mathrm{Var}_k \left\{ \sum_{r=1}^{K+1} \big[\gamma_r(1-\rho_0\rho_V) - \rho_r\rho_V\big] \pi_r(U) E_r\big\{M_{X,U}(T,\Delta,W,Z;\beta)\big|U\big\} \right\}.$$

We have derived the limiting distribution for each of the three terms on the right-hand side of (A.5). Now we will show that the three terms are asymptotically independent of each other. Since $\frac{1}{\sqrt{N}} D_F(\beta_0, \hat{\Lambda}, \hat{G})$ can be considered as a function of $\{X_i, U_i; i \in V\}$ for large N, it is asymptotically independent of the second term in (A.6) that is based on the nonvalidation sample $\bar{V}$. We use $\frac{1}{\sqrt{N}} U_F^1(\beta, \Lambda_0, G_0)$ to denote the first term of (A.6). Then

$$
\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{\sqrt{N}}D_F(\beta_0,\hat{\Lambda},\hat{G}),\ \frac{1}{\sqrt{N}}U_F^1(\beta_0,\Lambda_0,G_0)\right)\\
=&\frac{1}{N}\mathrm{Cov}\left(\sum_{r=1}^{K+1}\frac{\gamma_r}{\rho_r\rho_V+\gamma_r\rho_0\rho_V}\sum_{i\in V_r}g_1(X_i,U_i;\beta_0,\Lambda_0),\ \sum_{k=0}^{K}\sum_{i\in\tilde{V}_k}g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)\right)\\
=&\frac{1}{N}\sum_{r=1}^{K}\frac{\gamma_r}{\rho_r\rho_V+\gamma_r\rho_0\rho_V}\sum_{i\in V_r}\mathrm{Cov}\Big(g_1(X_i,U_i;\beta_0,\Lambda_0),\ g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)\Big)\\
=&\frac{1}{N}\sum_{r=1}^{K}\frac{\gamma_r}{\rho_r\rho_V+\gamma_r\rho_0\rho_V}\sum_{i\in V_r}\Big\{E\big(g_1(X_i,U_i;\beta_0,\Lambda_0)g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)\big)\\
&\qquad-E\big(g_1(X_i,U_i;\beta_0,\Lambda_0)\big)E\big(g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)\big)\Big\}\\
=&\frac{1}{N}\sum_{r=1}^{K}\frac{\gamma_r}{\rho_r\rho_V+\gamma_r\rho_0\rho_V}\sum_{i\in V_r}\Big\{E\big[E(g_1(X_i,U_i;\beta_0,\Lambda_0)g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)|X_i,Z_i,W_i)\big]\\
&\qquad-E\big(g_1(X_i,U_i;\beta_0,\Lambda_0)\big)E\big[E(g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)|X_i,Z_i)\big]\Big\}\\
=&\frac{1}{N}\sum_{r=1}^{K}\frac{\gamma_r}{\rho_r\rho_V+\gamma_r\rho_0\rho_V}\sum_{i\in V_r}\Big\{E\big[g_1(X_i,U_i;\beta_0,\Lambda_0)E(g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)|X_i,Z_i))\big]\\
&\qquad-E(g_1(X_i,U_i;\beta_0,\Lambda_0))E\big[E(g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)|X_i,Z_i)\big]\Big\},
\end{aligned}
$$

where

$$
g_1(X_i,U_i;\beta,\Lambda)=\sum_{k=1}^{K+1}\big[\gamma_k(1-\rho_0\rho_V)-\rho_0\rho_V\big]\pi_k(U_i)E_k\big\{M_{X_i,U_i}(T,\Delta,W,Z;\beta,\Lambda)\big|U_i\big\}
$$

and

$$
g_2(T_i,\Delta_i,X_i,Z_i;\beta,\Lambda)=\frac{\frac{\partial}{\partial\beta}f_{\beta,\Lambda}(T_i,\Delta_i|X_i,Z_i)}{f_{\beta,\Lambda}(T_i,\Delta_i|X_i,Z_i)}.
$$

Since $E\big(g_2(T_i,\Delta_i,X_i,Z_i;\beta_0,\Lambda_0)|X_i,Z_i\big)=0$, we have $\frac{1}{\sqrt{N}}D_F(\beta_0,\hat{\Lambda},\hat{G})$ and $\frac{1}{\sqrt{N}}U_F^1(\beta_0,\Lambda_0,G_0)$ are asymptotically uncorrelated and, since they are asymptotically normal, thus independent. Hence, $\frac{1}{\sqrt{N}}D_F(\beta_0,\hat{\Lambda},\hat{G})$ and $\frac{1}{\sqrt{N}}U_F(\beta_0,\Lambda_0,G_0)$ are asymptotically independent. Similarly, we can also prove that $\frac{1}{\sqrt{N}}D_F(\beta_0,\hat{\Lambda},\hat{G})$ and $\frac{1}{\sqrt{N}}U_F(\beta_0,\hat{\Lambda},G_0)-$

$\frac{1}{\sqrt{N}}U_F(\beta_0, \Lambda_0, G_0)$ are asymptotically independent. In addition,

$$\text{Cov}\left(\frac{1}{\sqrt{N}}U_F(\beta_0, \Lambda_0, G_0),\ \frac{1}{\sqrt{N}}U_F(\beta_0, \hat{\Lambda}, G_0) - \frac{1}{\sqrt{N}}U_F(\beta_0, \Lambda_0, G_0)\right) \tag{A.11}$$
$$= \frac{1}{N}\text{Cov}\left(U_F(\beta_0, \Lambda_0, G_0),\ U_F(\beta_0, \hat{\Lambda}, G_0)\right) - \frac{1}{N}\text{Cov}\left(U_F(\beta_0, \Lambda_0, G_0),\ U_F(\beta_0, \Lambda_0, G_0)\right)$$

By the convergence of $\hat{\Lambda}$ shown in Tsiatis (1981) and the Delta method, (A.11) is equal to zero. Hence, $\frac{1}{\sqrt{N}}U_F(\beta_0, \Lambda_0, G_0)$ and $\frac{1}{\sqrt{N}}U_F(\beta_0, \hat{\Lambda}, G_0) - \frac{1}{\sqrt{N}}U_F(\beta_0, \Lambda_0, G_0)$ are asymptotically independent.

We have shown that the three terms in (A.5) are asymptotically independent. Combining (A.7), (A.9) and (A.10), we obtain

$$\frac{1}{\sqrt{N}}U_F(\beta_0, \hat{\Lambda}, \hat{G}) \xrightarrow{d} \mathcal{N}\left(0,\ I(\beta_0) + \Sigma_{\mathbb{G}}(\beta_0) + \sum_{k=1}^{K+1} \frac{\gamma_k^2}{\rho_k\rho_V + \gamma_k\rho_0\rho_V}\Sigma_k(\beta_0)\right). \tag{A.12}$$

Using the first-order Taylor series expansion of the pseudo-score function $U_F(\beta_0, \hat{\Lambda}, \hat{G})$ around the true parameter β_0, we have

$$\sqrt{N}(\hat{\beta} - \beta_0) = \left[-\frac{1}{N}\frac{\partial U_F(\beta^*, \hat{\Lambda}, \hat{G})}{\partial \beta'}\right]^{-1}\left[\frac{1}{\sqrt{N}}U_F(\beta_0, \hat{\Lambda}, \hat{G})\right], \tag{A.13}$$

where β^* is on the line segment between $\hat{\beta}$ and β_0. By (A.4) and consistency of $\hat{\beta}$, it is easy to show that as $N \to \infty$,

$$\left[-\frac{1}{N}\frac{\partial U_F(\beta^*, \hat{\Lambda}, \hat{G})}{\partial \beta'}\right]^{-1} \xrightarrow{p} I^{-1}(\beta_0). \tag{A.14}$$

Combining (A.12), (A.13) and (A.14), we have

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma(\beta_0)),$$

which completes the proof.

APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 3

Let $\theta_0 = (\vartheta_0, \Lambda_0)$ denote the true value of $\theta = (\vartheta, \Lambda)$ in models (3.1) and (3.2). We first present consistency and asymptotic normality for the orginal estimator $\hat{\theta}_n$. The regularity conditions needed are described as follows:

(C1) Let τ be a time point satisfying $E(\delta I\{Y \geq \tau\}) > 0$.

(C2) Suppose that $\int_0^\infty P(C > t) d\Lambda_0(t) < \infty$.

(C3) ϑ_0 is an interior point of a compact set $\mathcal{D}$ in $\mathbb{R}^{2p+1}$ that denotes the parameter space for ϑ, where p is the dimension of X.

(C4) $\Lambda_0(\cdot)$ is continuously differentiable up to order d, where d is positive integer, with strictly positive derivative $\lambda_0(\cdot)$ on $[0, \tau]$ and $\Lambda_0'(t) > 0$ for $t \in [0, \tau]$.

(C5) Let $X = (U^T, Z^T)^T$. The distribution of X has a bounded support in $\mathbb{R}^{2p+1}$. If $aX + b = 0$, then $a = 0$ and $b = 0$. For some $\kappa > 0$, $a^T var(X) a \geq \kappa a^T E(XX^T) a$ for all $a \in \mathbb{R}^{2p+1}$.

(C6) The transformation function G is three-times continuously differently with $G(0) = 0$, $G'(0) > 0$ and $\{1 + G'(x)\} e^{-G(x)} \leq c_1 (1 + x)^{-\nu_0}$ for some constant $\nu_0 > 0$ and $c_1 > 0$.

(C7) The degree of Bernstein polynomials satisfies $m = o(n^\nu)$ with $1/(2d) < \nu < 1/2$, and $M_n = O(n^a)$ with $a > 0$ controlling the size of the sieve space $\mathcal{B}_n$.

(C8) θ_0^* is the unique value of θ^* that minimizes $KL(\theta^*)$, and ϑ_0^* is an interior point of a compact set $\mathcal{D}^*$ in $\mathbb{R}^{2p+1^*}$. $\Lambda_0^*(\cdot)$ is continuously differentiable up to order d with strictly positive derivative $\lambda_0^*(\cdot)$ and $\Lambda_0^*(\tau) < \infty$.

(C9) Let $X^* = (U^{*T}, Z^T)^T$. The distribution of X^* has a bounded support in $\mathbb{R}^{2p+1^*}$. If $aX^* + b = 0$, then $a = 0$ and $b = 0$. For some $\kappa^* > 0$, $a^T var(X^*) a \geq \kappa^* a^T E(X^* X^{*T}) a$ for all $a \in \mathbb{R}^{2p+1^*}$.

Once the conditions (C1)-(C7) hold, according to Fang *et al.* (2005), Zeng and Lin (2007) and Zhou *et al.* (2017) $\hat{\vartheta}_n$ is a consistent and asymptotically normal estimator of ϑ such as

$$\hat{\vartheta}_n \to \vartheta_0$$

in probability, and

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \to N(0, \Sigma_{11})$$

in distribution.

Next, we prove the consistency and asymptotic normality for the update estimator $\bar{\vartheta}_n$. Let $\theta_0^* = (\vartheta_0^*, \Lambda_0^*)$ be the value of $\theta^* = (\vartheta^*, \Lambda^*)$ that minimizes the Kullback-Leibler divergence given by

$$KL(\theta^*) = E\big\{ \log \big(\frac{L(\mathcal{O}^*)}{L(\theta^*|\mathcal{O}^*)}\big)\big\},$$

where $L(\theta^*|\mathcal{O}^*)$ denotes the likelihood function at θ^* under the working models (3.5) and (3.6) based on the data $\mathcal{O}^* = \{Y_1, \cdots, Y_n, \delta_1, \cdots, \delta_n, X^*\}$, and $L(\mathcal{O}^*)$ denotes the true likelihood of $\mathcal{O}^*$. According to Zhou and Wong (2023), the consistency and asymptotic normality of $\bar{\vartheta}_n$ can be established.