UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Statistical Methods for Cohort Studies with Terminal Events

Permalink

https://escholarship.org/uc/item/1q35n80s

Author

Wang, Yue

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical Methods for Cohort Studies with Terminal Events

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Yue Wang

Dissertation Committee:
Professor Bin Nan, Chair
Professor Daniel L. Gillen
Professor Annie Qu

2023

DEDICATION

To Crystal, Kahlua and my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to give many thanks to Dr. Bin Nan for the guidance and support he has been giving me throughout my academic journey. His mentorship has been crucial in shaping my understanding and enhancing my skills in the field of statistics. I am truly blessed to have had such an exceptional advisor and I will always remember the countless hours we spent together discussing ideas and problems.

I am thankful to Dr. Dan Gillen and Dr. Annie Qu for serving on my dissertation committee and providing valuable feedback on my research. Special thanks are due to Dr. Dan Gillen. I have received many help and guidance from him for my survival and longitudinal lectures, as well as my third research topic.

I must thank my parents for all the love and caring they have been giving me, especially in the last three months before my final defense. Without them around, I wouldn't have been able to juggle all the things in my hand during the graduation season. Last and foremost, a big shoutout to my wife Crystal, who has always been by my side when I made all the important decisions of my life and career. Without her continuous encouragement, I couldn't imagine finishing my PhD degree.

VITA

Yue Wang

EDUCATION

Doctor of Philosophy in Statistics   **2023**
University of California, Irvine   *Irvine, California*

Master of Arts in Applied Statistics   **2018**
University of Michigan, Ann Arbor   *Ann Arbor, Michigan*

Bachelor of Science in Statistics   **2016**
Zhejiang University   *Hangzhou, China*

RESEARCH EXPERIENCE

Graduate Research Assistant   **2018–2023**
University of California, Irvine   *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant   **2021 and 2023**
University of California, Irvine   *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Kernel Estimation of Bivariate Time-varying Coefficient Model for Longitudinal Data with Terminal Event
Journal of the American Statistical Association
2023

An Improved Variable Selection Procedure for Adaptive Lasso in High-dimensional Survival Analysis.
Lifetime Data Analysis
2018

ABSTRACT OF THE DISSERTATION

Statistical Methods for Cohort Studies with Terminal Events

By

Yue Wang

Doctor of Philosophy in Statistics

University of California, Irvine, 2023

Professor Bin Nan, Chair

In a longitudinal cohort study, a group of subjects is chosen based on certain characteristics and then followed at routine intervals over time. At each visit, some measurements are recorded and the goal is oftentimes to model how they develop over time. During the follow-up, the collection of the data can be stopped by a terminal event. In this dissertation, we study several statistical methods for modeling longitudinal data while adjusting for the terminal event.

In Chapter II, we propose a nonparametric bivariate time-varying coefficient model for longitudinal measurements with the occurrence of a terminal event that is subject to right censoring. The time-varying coefficients capture the longitudinal trajectories of covariate effects along with both the followup time and the residual lifetime, in contrast to the existing work that either models longitudinal measures as a function of only the forward time or the backward time, or poses strong parametric assumptions. We consider a kernel smoothing method for estimating regression coefficients in our model and use cross-validation for bandwidth selection, applying undersmoothing in the final analysis to eliminate the asymptotic bias of the kernel estimator. We show that the kernel estimates follow a finite-dimensional normal distribution asymptotically under mild regularity conditions, and provide an easily computed sandwich covariance matrix estimator.

In Chapter III, we study the lifetime Medicare spending patterns of patients with end-stage renal disease (ESRD) stratified by waitlisting and kidney transplant. In addition to the terminal event, a non-terminal event that happened in the follow-up could also have an impact on the longitudinal measures. To study the heterogeneous Medicare cost trajectories across groups stratified by waitlisting and transplant, we proposed two models: a semiparametric regression model with both fixed and bivariate time-varying coefficients to compare unwaitlisted and waitlisted groups, and a bivariate time-varying coefficient model with different starting times (time since the first ESRD service and time since the kidney transplant) to compare untransplanted and transplanted groups. We use sandwich variance estimators to construct confidence intervals and validate inference procedures through simulations. Our analysis of the Medicare claims data reveals that waitlisting is associated with a lower daily medical cost at the beginning of ESRD service among waitlisted patients which gradually increases over time. Averaging over lifespan, however, there is no difference between waitlisted and unwaitlisted groups. A kidney transplant, on the other hand, reduces the medical cost significantly after an initial spike.

In Chapter IV, we study how the onset of a non-terminal events is associated with the terminal event. Existing methods under the framework of multi-state model or the semi-competing risks model rely on certain semiparametric assumptions for modeling the joint distribution of these two events, thus may subject to model misspecification and lack clear interpretations of their association. Moreover, they assume the independence between the censoring time C and both event times (S, T), which can be easily violated in realistic situations. We propose to estimate the onset of non-terminal event conditional on the terminal event time using a generalized Beran's estimator. We consider a left truncation time L in addition to the right censoring time C and only need a relaxed assumption of the independence between (L, C) and the non-terminal event time S conditional on the terminal event time T. Such estimates also enjoy more meaningful interpretations.

Chapter 1

Introduction

This dissertation primarily develops methods for cohort studies that incorporate terminal events. There are three different topics in this dissertation, and each focuses on a different problem with its own features.

1.1 Bivariate Time-Varying Coefficient Model Conditional on Death Time

There are two sets of widely used approaches for modelling longitudinal measures with a terminal event: the joint modeling approach using latent frailty and the marginal estimating equation approach using inverse probability weighting (IPW). Under the joint modeling framework, the survival time and the longitudinal process are assumed independent conditional on some latent random effects. Thorough reviews of this type of approach can be found in [45] and [41]. For the marginal estimating equation approach with IPW, readers can refer to [42]. However, they do not explicitly model the association between the terminal event time and the longitudinally measured response variable, and some approaches treat

the occurrence of death as "dropout", which implicitly defines the underlying stochastic processes of health status beyond death.

In this dissertation, we propose a nonparametric extension of [24]. In particular, regression coefficients are bivariate functions of both the chronological followup time t and the residual lifetime $T - t$ with unknown form, where t denotes the followup time and T denotes the terminal event time. Such a modelling strategy allows us to assess the varying effect of certain covariate when patients approach death. We estimate the regression coefficients using kernel smoothing and establish the asymptotic normality of kernel estimates together with convergence rates that depend on the bandwidth size. We also provide a consistent sandwich variance estimator that helps construct pointwise confidence bands.

1.2 Bivariate Functional Patterns of Lifetime Medicare Costs among ESRD patients

End-stage renal disease (ESRD) has become increasingly prevalent in the United States. It has long been observed that patients on the waiting list tend to have better health status compared to unwaitlisted patients (e.g.,[53]). However, there is fewer literature that studies medical spending aspects of waitlisting. We propose a semi-varying coefficient model to compare overall lifetime medical costs between waitlisted and unwaitlisted dialysis patients.

The survival benefit and economic implications of transplant itself have been extensively studied via a standard measure called incremental cost-effectiveness ratio (ICER) ([15]). The determination of ICER involves calculating the cumulative medical cost during the remaining lifetime, which is a single summary number to measure the costs of transplant and dialysis. Using a new regression model, called the mixed-varying coefficient model, we examine and compare the detailed medical cost trajectories of transplanted ESRD patients

2

during their lifespan to those on the waiting list.

1.3 Left Truncated Age of Disease Onset Data with Terminal Event

In classical medical studies, a common situation features two types of events: a non-terminal event at time S which can be a disease and a terminal event at time T which is usually death. It is often desired to understand the extent to which the two events associate with each other. Two widely used approaches in the existing literature are: (1) under the semicompeting risks framework, the joint distribution of these events is formulated via a gamma frailty model in the upper wedge where data are observable [6, 14]; and (2) an illness-death model is used with a shared frailty to incorporate the dependence structure [36, 55]. In either case, assumptions about certain semiparametric structure are made to model the joint distribution of the two events, which may lead to model misspecification. Moreover, both approaches assume the independence between the censoring time C and event times (S, T).

For a better understanding of the association between S and T and ease of interpretation, we propose to estimate the distribution of S conditional on T, i.e., we use T as a covariate. The conditional distribution can be estimated by an product-limit estimator widely known as "Beran's estimator", which can handle both right censoring and left truncation. This type of analysis can be shown to only rely on the independence of (L, C) and S conditional on T, where L is the left-truncation time, without posing any dependence structure, thus relaxes the assumption made in the existing literature.

Chapter 2

Kernel Estimation of Bivariate Time-varying Coefficient Model for Longitudinal Data with Terminal Event

2.1 Introduction

In longitudinal studies, it is often the case that the collection of repeated measurements is stopped by the occurrence of some terminal event, for example, death. There are two sets of widely used approaches for modelling longitudinal measures with a terminal event: the joint modeling approach using latent frailty and the marginal estimating equation approach using inverse probability weighting (IPW). Under the joint modeling framework, the survival time and the longitudinal process are assumed independent conditional on some latent random effects. Thorough reviews of this type of approach can be found in [45] and [41]. For

the marginal estimating equation approach with IPW, readers can refer to [42]. These ideas have also been applied to modeling recurrent events in the presence of a terminal event, see e.g., [22] and [16]. They may fall short in certain situations, however. First, as pointed out by [24], they do not explicitly model the association between the terminal event time and the longitudinally measured response variable, which is of primary interest in many applications. Second, in health studies where death is a terminal event, some approaches treat the occurrence of death as "dropout", either informative or non-informative, which implicitly defines the underlying longitudinally measured stochastic processes of health status beyond death. In other words, death causes "missing data" in such a view, which is questionable since death itself is a fundamental characteristic of health.

For these reasons, several reverse-time models have been considered in the recent literature. [4] considered a nonparametric approach for the mean of a reverse-time process. [27] considered a likelihood-based approach for the reverse-time model with applications to palliative care, with extension to a semiparametric approach introduced in [26]. [7] considered reverse alignment as a general technique for constructing models for survival processes and investigated several related statistical consequences. These methods model backward time with event time as the time origin, but lose the interpretation of chronological time effects that are of primary interest in conventional longitudinal studies. To keep the desired chronological time interpretation of regression coefficients and meanwhile to describe the effect of terminal event in longitudinal studies, [24] proposed a parametric nonlinear regression model conditional on the terminal event time which builds the residual lifetime into covariate effects. They showed that the complete case analysis that only uses data with uncensored event times is a valid approach, and proposed a two-stage approach that improves the efficiency of parameter estimates of the complete case analysis. But a parametric model can be easily misspecified, as we observe in our data example, and their two-stage method cannot handle time-varying covariates that occur overwhelmingly often in longitudinal studies.

In this article, we propose a nonparametric extension of [24]. In particular, regression coefficients are bivariate functions of both chronological followup time t and residual lifetime $T - t$ with unknown form, where t denotes the followup time and T denotes the terminal event time. Moreover, time-varying covariates are incorporated in our model. Such a modelling strategy allows us to assess the varying effect of certain covariate when patients approach death, which is of particular interest for the analysis of end-stage renal disease (ESRD) medical cost data. We estimate the regression coefficients using kernel smoothing and establish the asymptotic normality of kernel estimates together with convergence rate that depends on the bandwidth size. We also provide a consistent sandwich variance estimator that helps construct pointwise confidence bands.

The rest of the article is organized as follows. In Section 2 we introduce the time-varying coefficient model and the kernel estimating method with bandwidth determined via under-smoothing after cross-validation. We outline the asymptotic properties in Section 3. We provide simulations in Section 4 and the analysis of ESRD medical cost data in Section 5. We give a few concluding remarks in Section 6, and provide detailed proofs and additional numerical results in the online supplementary material.

2.2 Modeling Strategy and Estimating Method

2.2.1 Bivariate Time-Varying Coefficient Model

Let $Y(t)$ be a stochastic process denoting the response variable measured over time in a longitudinal study. Let $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))$ be p covariate processes. Note that we use bold letter to represent either a vector or a matrix in this article. Suppose the longitudinal cohort data consists of n independent copies of $(Y(t), \mathbf{X}(t))$, representing n individuals' observations in the study cohort, where the ith individual's data $(Y_i(t), X_{i1}(t), \ldots, X_{ip}(t))$

6

are measured at random time points τ_{ij}, $j = 1, \ldots, m_i$. Baseline covariates take constant values over time. We define $X_{i1}(t) \equiv 1$ for any i and t, which determines the intercept. Suppose each individual has m visits, where m is a finite number, but not all of them are observed because of early stopping due to terminal event or right censoring, which makes the number of actual visits varying among individuals. Specifically for subject i, denote the terminal event time as T_i and the right censoring time as C_i, then the number of visits of subject i is $m_i = \max\{j : j \leq m, \tau_{ij} \leq T_i \wedge C_i\}$, where $a \wedge b = \min\{a, b\}$. Denote the set of subjects whose terminal events are observed by $\mathcal{D} = \{i : T_i \leq C_i\}$.

We consider the following model for the longitudinal response variable Y_i observed at time τ_{ij}:

$$Y_i(\tau_{ij}) = \sum_{k=1}^{p} X_{ik}(\tau_{ij})\beta_k(\tau_{ij}, T_i - \tau_{ij}) + \varepsilon_i(\tau_{ij}), \tag{2.1}$$

where each $\varepsilon_i(t)$ is a zero-mean stochastic process with variance function $\sigma^2(t)$ and covariance function $\rho(t_1, t_2)$ for any $t_1 \neq t_2$. Assume all the quantities involved in this model are independent and identically distributed (i.i.d.) across individuals, which include $\{\tau_{ij}\}_{j=1}^{m}$, T_i, C_i, $\{X_{ik}(\cdot)\}_{k=1}^{p}$ and $\varepsilon_i(\cdot)$. Here i.i.d. is defined for processes on any finite index set. Suppressing the subscript i here without causing any confusion, we further assume that for each individual we have: (1) given $\tau_j = t$, $\varepsilon(\tau_j)$ has the same distribution as $\varepsilon(t)$ and is independent of T, C and $\{X_k(\tau_j)\}_{k=1}^{p}$; (2) given $(\tau_{j_1}, \tau_{j_2}) = (t_1, t_2)$, $(\varepsilon(\tau_{j_1}), \varepsilon(\tau_{j_2}))$ has the same distribution as $(\varepsilon(t_1), \varepsilon(t_2))$ and is independent of T, C, $\{X_k(\tau_{j_1})\}_{k=1}^{p}$ and $\{X_k(\tau_{j_2})\}_{k=1}^{p}$. In other words, data observed on a set of random times behave like observed on a set of constant times, which is commonly assumed in longitudinal data analysis. Because we do not estimate the survival function for the complete case analysis considered in this article, we do not need to assume T and C are independent or conditionally independent given covariates, which is a crucial assumption in traditional survival analysis.

Unlike the usual time-varying coefficient model for longitudinal data, a particularly important feature of model (2.1) is that the unknown coefficient $\beta_k(t, T - t)$ is allowed to be a bivariate function not only varying with time since entry, t (the usual setup, see e.g. [19]), but also varying with time from t to the terminal event, $T - t$ (also referred to as residual lifetime if T is death time). Unlike any conventional modeling strategy for longitudinal data with terminal event, allowing β_k to depend on $T - t$ directly captures the way in which impending failure modifies the effect of $X_{ik}(t)$. If none of the β_k, $k = 1, \ldots, p$, varies with $T - t$, then the above model (2.1) reduces to a standard time-varying coefficient model.

Model (2.1) extends [24] from a parametric model to a nonparametric model, from an intercept varying with $T - t$ only to all regression coefficients varying with both t and $T - t$, and from fixed baseline covariates only to time-varying covariates. The model also extends [30], who only considered a nonparametric intercept varying with $T - t$ without pursuing the asymptotic properties of their spline based estimating method. [25] considered a bivariate mean model, included no covariates and did not provide asymptotic results for their spline estimating method.

It becomes clear that model (2.1) is well-defined when T_i is observed, so all the observations collected from time at entry to T_i are complete data, whereas observations collected from time at entry to C_i before T_i are incomplete. This is another major distinction between a model that is conditional on T_i and conventional regression models for longitudinal data with terminal events. Since T_i is subject to right censoring, the problem determined by model (2.1) becomes a regression problem with censored covariate, for which the complete case analysis is a valid approach. This is the method we consider in this article for estimating unknown bivariate coefficient functions β_k, $k = 1, \ldots, p$. Including observations for censored individuals faces multifaceted difficulties, which will be discussed in the future work.

8

2.2.2　Bivariate Kernel Estimation

For any fixed point (t_0, s_0), we apply bivariate kernel smoothing to estimate $\boldsymbol{\beta}(t_0, s_0)$ by minimizing the following loss function with respect to b_k, $k = 1, \ldots, p$:

$$L_n(t_0, s_0) = \sum_{i \in \mathcal{D}} \sum_{j=1}^{m_i} \left(Y_{ij} - \sum_{k=1}^{p} X_{ik}(\tau_{ij}) b_k \right)^2 K \left(\frac{\tau_{ij} - t_0}{h}, \frac{T_i - \tau_{ij} - s_0}{h} \right), \qquad (2.2)$$

where $K : \mathbb{R}^2 \to \mathbb{R}$ is the kernel function, $h > 0$ is the bandwidth. There are two major distinctions between the resulting estimator from (2.2) and the estimator in [19]: First, the estimator in (2.2) involves terminal event time T_i and is based on complete data. Second, since β_k's are bivariate functions, a bivariate kernel function is used. Note that we use the same bandwidth for both time axes and ignore the off-diagonal element of the 2×2 bandwidth matrix in order to simplify the numerical implementation. Using multiple bandwidths requires multiple bandwidth selection procedures thus is more computationally cumbersome especially for large data sets. Later we show both theoretically and numerically that the kernel estimation using a common bandwidth performs satisfactorily.

Rewrite (2.2) into the following matrix form:

$$L_n(t_0, s_0) = \sum_{i \in \mathcal{D}} (\mathbf{Y}_i - \mathbf{X}_i \mathbf{b})^T \mathbf{K}_i(t_0, s_0; h) (\mathbf{Y}_i - \mathbf{X}_i \mathbf{b}),$$

where

$$\mathbf{X}_i = \begin{pmatrix} X_{i1}(\tau_{i1}) & \ldots & X_{ip}(\tau_{i1}) \\ \vdots & \ddots & \vdots \\ X_{i1}(\tau_{im_i}) & \ldots & X_{ip}(\tau_{im_i}) \end{pmatrix},$$

$\mathbf{K}_i(t_0, s_0; h)$ is a diagonal matrix with jth element given by $K((\tau_{ij} - t_0)/h, (T_i - \tau_{ij} - s_0)/h)/h^2$, and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$. We estimate the time-varying coefficients $\beta_k(t, s)$ by minimizing

$L_n(t_0, s_0)$ with respect to b_k, $k = 1, \ldots, p$, i.e.,

$$\widehat{\boldsymbol{\beta}}(t_0, s_0; h) = \arg\min_{\mathbf{b}} L_n(t_0, s_0),$$

which has a closed form solution given by

$$\widehat{\boldsymbol{\beta}}(t_0, s_0; h) = \left(\sum_{i\in\mathcal{D}} \mathbf{X}_i^T \mathbf{K}_i(t_0, s_0; h)\mathbf{X}_i\right)^{-1} \left(\sum_{i\in\mathcal{D}} \mathbf{X}_i^T \mathbf{K}_i(t_0, s_0; h)\mathbf{Y}_i\right). \tag{2.3}$$

The estimator (2.3) ignores the within-subject correlation following the working independence assumption, which was shown by [28] to be most efficient when a standard kernel is applied and the cluster size is finite. This counter-intuitive result was explained by [49] who also showed that higher efficiency could be achieved by using an alternative kernel method, which we do not pursue here because of both the numerical advantages of the working independence assumption and the efficiency result of [28] for using a standard kernel in (2.3).

2.2.3 Automatic Bandwidth Selection and Undersmoothing

A typical approach for automatic bandwidth selection is through K-fold cross-validation (CV). To keep the independence between training set and validation set, we partition the longitudinal data at the subject level such that all repeatedly measured observations of each subject belong to only one fold. The criterion for selecting bandwidth is to minimize the average predictive squared errors across all validation sets. In particular, let S_k, $k = 1, \ldots, K$, be the index set of subjects in the k-th fold, where $\cup_k S_k = \mathcal{D}$ and $S_k \cap S_l = \emptyset$ for any $k \neq l$, then the average predictive squared error criterion is given by

$$\mathrm{CV}(h) = \frac{1}{\sum_{i\in\mathcal{D}} m_i} \sum_{k=1}^{K} \sum_{i\in S_k} \sum_{j=1}^{m_i} \left[Y_{ij} - \mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(-k)}(\tau_{ij}, T_i - \tau_{ij}; h)\right]^2, \tag{2.4}$$

10

where $\widehat{\boldsymbol{\beta}}^{(-k)}$ represents the kernel estimator calculated by leaving out all observations in the k-th fold. Note that only the complete cases $\mathcal{D}$ are partitioned into folds. In practice, the criterion is minimized on a preselected grid of h.

Standard approaches to constructing nonparametric confidence bands for functions are complicated by the impact of bias. According to [17], bias decreases as the amount of statistical smoothing is reduced, which can be clearly seen from the asymptotic distributional results of kernel estimates. Therefore, one way of alleviating bias is to smooth the curve estimator less than would be optimal for point estimation. We choose to undersmooth by multiplying $n^{-\gamma}$ to the selected bandwidth using cross-validation for some $\gamma > 0$. We will see in Section 2.3 that undersmoothing still leads to the desirable asymptotic result as long as the undersmoothed bandwidth falls into the range specified by Condition 2 in Appendix 2.6.1.

2.2.4 A Special Case with Potentially Improved Efficiency

Under a special circumstance where only baseline covariates (or the so-called defined time-dependent covariates that are completely determined by baseline covariates and the time t) are of concern as in [24], or in an even less likely situation where time-varying covariates are of interest but the terminal event time only depends on baseline covariates, one might assume Gaussian error in model (2.1) and consider the following locally weighted pseudo likelihood function under working independence which includes both complete and censored data:

$$
\prod_{i=1}^{n} \left\{ \prod_{j=1}^{m_i} \left[\frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{1}{2\sigma_{ij}^2}(Y_{ij}-X_{ij}^T b)^2} \right]^{K(\frac{\tau_{ij}-t}{h}, \frac{T_i-\tau_{ij}-s}{h})} f_1(T_i|Z_i) \right\}^{\Delta_i}
$$
$$
\times \left\{ \int_{C_i}^{\infty} \prod_{j=1}^{m_i} \left[\frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{1}{2\sigma_{ij}^2}(Y_{ij}-X_{ij}^T b)^2} \right]^{K(\frac{\tau_{ij}-t}{h}, \frac{u-\tau_{ij}-s}{h})} dP(T_i \leq u|Z_i) \right\}^{1-\Delta_i} f_2(X_i, \boldsymbol{\tau}_i, Z_i).
$$

Note that, similar to [24], the assumption of conditional independent censoring is also needed to obtain the above locally weighted pseudo likelihood. In contrast, no assumption on the censoring mechanism is needed for the complete case analysis. In the above, Z_i is a vector of baseline covariates of subject i, $f_1(\cdot|Z_i)$ is the survival density given Z_i and f_2 is the joint density of X_i, $\boldsymbol{\tau}_i$ and Z_i. The nuisance parameter σ_{ij} denotes $\sigma(\tau_{ij})$ and can be estimated by the kernel estimator $\widehat{\sigma}_{ij}^2 = (nh)^{-1} \sum_{i'j'} \widehat{\varepsilon}_{i'j'}^2 K'((\tau_{i'j'} - \tau_{ij})/h)$, where K' is some univariate kernel function, and another nuisance parameter $P(T_i \leq u|Z_i)$ can be estimated using a proper survival model. The estimation of $\boldsymbol{\beta}(t, s)$ can be obtained by maximizing the above locally weighted pseudo likelihood with respect to b. A simulation study, summarized in the online supplementary material, shows that this method seems to yield valid results with improved efficiency over the complete cases analysis. Its theoretical justification, however, is beyond the scope of this work thus not pursued here, and as discussed in Section 6, several additional difficulties preclude the application of this approach to the case with time-varying covariates that are commonly observed in longitudinal studies.

2.3   Asymptotic Properties

2.3.1   Asymptotic Normality of $\widehat{\boldsymbol{\beta}}$

We consider the asymptotic joint normality of $\widehat{\boldsymbol{\beta}}$ at a finite number of pairs of distinct time points $\{(t_1, s_1), \ldots, (t_d, s_d)\}$. Under several mild regularity conditions given in the Appendix, we can show that the estimator (2.3) follows a multivariate normal distribution

as n approaches infinity. First we introduce some notation:

$$\mu_0 = \int K^2(x,y)dxdy,$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} \int x^2 K(x,y)dxdy & \int xyK(x,y)dxdy \\ \int yxK(x,y)dxdy & \int y^2 K(x,y)dxdy \end{pmatrix},$$

$$\boldsymbol{\eta}_j(t,s) = E\left[1(T \leq C)\mathbf{X}(\tau_j)\mathbf{X}(\tau_j)^T | \tau_j = t, T - \tau_j = s\right],$$

$$\boldsymbol{g}_{j,rk}(t,s) = \nabla\eta_{j,rk}(t,s)f_j(t,s)\nabla\beta_k(t,s)^T + \eta_{j,rk}(t,s)\nabla f_j(t,s)\nabla\beta_k(t,s)^T$$

$$+ \frac{1}{2}\eta_{j,rk}(t,s)f_j(t,s)\nabla^2\beta_k(t,s),$$

$$\boldsymbol{\Gamma}(t,s) = \begin{pmatrix} \sum_{j=1}^m \sum_{k=1}^p \langle \boldsymbol{\mu}_2, \boldsymbol{g}_{j,1k}(t,s) \rangle \\ \vdots \\ \sum_{j=1}^m \sum_{k=1}^p \langle \boldsymbol{\mu}_2, \boldsymbol{g}_{j,pk}(t,s) \rangle \end{pmatrix}_{p \times 1},$$

$$\boldsymbol{\Omega}(t,s) = \sum_{j=1}^m \boldsymbol{\eta}_j(t,s)f_j(t,s).$$

In the above, $\eta_{j,rk}(t,s)$ is the (r,k)-element of $p \times p$ matrix $\boldsymbol{\eta}_j(t,s)$; f_j denotes the joint density of τ_j and $T - \tau_j$; $\nabla\beta_k(t,s)$ is the 2×1 gradient of β_k at (t,s) and $\nabla^2\beta_k(t,s)$ is the 2×2 hessian matrix; $\langle \mathbf{A}, \mathbf{B} \rangle$ is the Frobenius inner product of matrices $\mathbf{A}$ and $\mathbf{B}$, i.e., $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^T\mathbf{B})$. Note that the subscript i is suppressed for random variables in above $\boldsymbol{\eta}_j(t,s)$ to simplify the notation since these are all defined for a generic subject and observations are assumed i.i.d.

Furthermore, the following quantities are defined at a set of finite number of pairs of distinct

time points $\{(t_1, s_1), \ldots, (t_d, s_d)\}$. Let $\mathbf{t} = (t_1, \ldots, t_d)^T$ and $\mathbf{s} = (s_1, \ldots, s_d)^T$. Define

$$\widehat{\boldsymbol{\beta}}(\mathbf{t}, \mathbf{s}; h) = \begin{pmatrix} \widehat{\boldsymbol{\beta}}(t_1, s_1; h) \\ \vdots \\ \widehat{\boldsymbol{\beta}}(t_d, s_d; h) \end{pmatrix}_{dp \times 1},$$

$$\boldsymbol{\beta}(\mathbf{t}, \mathbf{s}) = \begin{pmatrix} \boldsymbol{\beta}(t_1, s_1) \\ \vdots \\ \boldsymbol{\beta}(t_d, s_d) \end{pmatrix}_{dp \times 1},$$

$$\mathbf{B}(\mathbf{t}, \mathbf{s}) = \begin{pmatrix} \boldsymbol{\Omega}^{-1}(t_1, s_1)\boldsymbol{\Gamma}(t_1, s_1) \\ \vdots \\ \boldsymbol{\Omega}^{-1}(t_d, s_d)\boldsymbol{\Gamma}(t_d, s_d) \end{pmatrix}_{dp \times 1}, \tag{2.5}$$

$$\mathbf{V}(\mathbf{t}, \mathbf{s}) = \mu_0 \begin{pmatrix} \sigma^2(t_1)\boldsymbol{\Omega}^{-1}(t_1, s_1) & & \\ & \ddots & \\ & & \sigma^2(t_d)\boldsymbol{\Omega}^{-1}(t_d, s_d) \end{pmatrix}_{dp \times dp}. \tag{2.6}$$

Theorem 2.1. *For a finite integer d and fixed vectors $\mathbf{t}$ and $\mathbf{s}$ satisfying $t_l, s_l > 0$ for $l = 1, \ldots, d$ and $t_{l_1} \neq t_{l_2}$, $s_{l_1} \neq s_{l_2}$ when $l_1 \neq l_2$, under regularity conditions 1, 2a, 2b, and 3-7 given in Appendix 2.6.1, we have*

$$n^{1/2}h\left(\widehat{\boldsymbol{\beta}}(\mathbf{t}, \mathbf{s}; h) - \boldsymbol{\beta}(\mathbf{t}, \mathbf{s})\right) \to_d N\left(h_0^3 \mathbf{B}(\mathbf{t}, \mathbf{s}), \mathbf{V}(\mathbf{t}, \mathbf{s})\right) \tag{2.7}$$

as $n \to \infty$, where h_0 is defined as the limit of $n^{1/6}h$ in condition 2a.

The above theorem shows that $n^{1/2}h(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ at different pairs of time points are asymptotically independent.

Directly using formulae (2.5) and (2.6) to estimate the asymptotic bias and variance is difficult due to their complicated forms. We will consider a simpler and numerically imple-

14

mentable consistent sandwich estimator for the asymptotic variance in the next subsection. We avoid estimating the bias via undersmoothing. Specifically, we eliminate the asymptotic bias by shrinking the selected bandwidth under cross-validation by a factor of $n^{-1/20}$, which shows satisfactory performance in simulations.

2.3.2 Sandwich Estimator and Pointwise Confidence Interval

With undersmoothing, the asymptotic bias in Theorem 2.1 disappears when n goes to infinity. Hence for any pair of time points in Theorem 2.1, denoted by (t_0, s_0), we only need to estimate the variance $\mathbf{V}(t_0, s_0)$ in order to construct the pointwise confidence interval. It turns out that the following sandwich estimator is a valid variance estimator:

$$\widehat{\mathbf{V}}(t_0, s_0) = nh^2 \left(\sum_{i \in \mathcal{D}} \mathbf{X}_i^T \mathbf{K}_{i0} \mathbf{X}_i \right)^{-1} \left(\sum_{i \in \mathcal{D}} \mathbf{X}_i^T \mathbf{K}_{i0} \widehat{\boldsymbol{\varepsilon}}_i \widehat{\boldsymbol{\varepsilon}}_i^T \mathbf{K}_{i0} \mathbf{X}_i \right) \left(\sum_{i \in \mathcal{D}} \mathbf{X}_i^T \mathbf{K}_{i0} \mathbf{X}_i \right)^{-1},$$

(2.8)

where $\widehat{\boldsymbol{\varepsilon}}_i$ is the residual vector for the i-th subject and $\mathbf{K}_{i0}$ is short for $\mathbf{K}_i(t_0, s_0; h)$. The elements of $\widehat{\boldsymbol{\varepsilon}}_i$ are calculated by

$$\widehat{\varepsilon}_{ij} = \widehat{\varepsilon}_i(\tau_{ij}) = Y_i(\tau_{ij}) - \sum_{k=1}^{p} X_{ik}(\tau_{ij}) \widehat{\beta}_k(\tau_{ij}, T_i - \tau_{ij}), \quad 1 \le j \le m_i.$$

The following theorem demonstrates the consistency of (2.8).

Theorem 2.2. *For the covariance matrix in Theorem 2.1, under regularity conditions 1, 2a, 2c, and 3-7 given in Appendix 2.6.1, we have*

$$\widehat{\mathbf{V}}(t_0, s_0) \to_p \mathbf{V}(t_0, s_0)$$

as $n \to \infty$, where $\mathbf{V}(t_0, s_0)$ is the corresponding diagonal block matrix in $\mathbf{V}(\mathbf{t}, \mathbf{s})$ at (t_0, s_0).

With this theorem, an approximate $1 - \alpha$ pointwise confidence interval of $\beta_k(t_0, s_0)$ without bias correction can be constructed as

$$\widehat{\beta}_k(t_0, s_0) \pm z_{\alpha/2} \left(\frac{1}{nh^2} \widehat{\mathbf{V}}(t_0, s_0)_{kk} \right)^{1/2}.$$

2.4   A Simulation Study

This section reports the numerical performance of the kernel estimator (2.3). For the simulation study, consider model (2.1) with $p = 3$ and $m = 20$, where the coefficients are the following functions:

$$\beta_1(x, y) = \frac{x}{4} \exp \left(-\frac{x^2 + y^2}{100} \right),$$

$$\beta_2(x, y) = \frac{1}{2} \left[\sin \left(\frac{2x}{5} \right) - \sin \left(\frac{y}{2} \right) \right],$$

$$\beta_3(x, y) = \cos \left(\frac{x^2 + y^2}{100} \right).$$

These functions are similar to those used in [54]. We generate visiting times in the following way: for subject i, the first visit time τ_{i1} is generated uniformly on $[0, 1]$, then τ_{ij}, $j > 1$, is generated independently from $\tau_{ij} - (j - 1) \sim \text{Beta}(\tau_{i1}/4\nu^2, (1 - \tau_{i1})/4\nu^2)$, where ν serves as an upper bound of standard deviation of the Beta distribution and is set to be 0.01. The generated interarrival time $\tau_{ij} - \tau_{i,j-1}$ falls into $[0, 2]$ with mean 1 and a very small variance. Thus the generated visiting schedule is approximately evenly spaced, mimicking a designed longitudinal study with annual visits. For covariates, X_1 is always 1, X_2 is generated from a standard normal distribution, and $X_3(t)$ is a mean-zero Gaussian process with covariance function $Cov(X_3(t), X_3(s)) = \exp(-(t - s)^2)$. Moreover, X_2 and $X_3(t)$ are correlated with covariance $Cov(X_2, X_3(t)) = 0.8 \exp(-t^2)$. Terminal event time T is generated from an

exponential distribution with rate $\exp(3X_{i2} + X_{i3}(0) - 5)$, then truncated at 15 and added 5. Thus $T \geq 5$ with probability 1. Censoring time follows a uniform distribution between 5 and $2T - 5$. This leads to dependent censoring and yields 50% censoring rate. The error term $\varepsilon_i(\tau_{ij})$ is generated by a nonhomogeneous Ornstein-Uhlenbeck (NOU) process $U_i(t)$ plus a random error. The NOU process satisfies $Var(U_i(t)) = \exp(1 - 0.1t)$ and $Corr(U_i(t_1), U_i(t_2)) = 0.5^{|t_1-t_2|}$, and the random error follows a standard normal distribution.

With this design, we simulate 1000 independent replications, each with a sample size $n = 4000$ that is about 10% of the sample size of the ESRD data analyzed in the next section. The kernel function is the density of a standard bivariate normal distribution truncated by a circle around $(0,0)$ which contains probability 0.95. To save the computing cost, we first run 5-fold cross-validation on 10 independent datasets with a grid search on $\{1.2^i : i = -10, ..., 10\}$, which yield an average bandwidth of 1. We then undersmooth it to obtain a bandwidth of $1 \times n^{-1/20} \approx 0.66$ and fix it for all the 1000 simulation replications. To achieve a better visual effect of bivariate functions, we plot a few slices of estimated coefficients. Specifically, we plot $\widehat{\beta}_k(t, T-t)$ varying with t at $T = 8$, 12, and 16, separately, which are the estimated covariate effects from the time of entry to the terminal event for individuals who died at time 8, 12 and 16. Among the 3×3 panels in Figure 2.1, each row represents a time-varying coefficient and each column represents one chosen value of T. There are 6 curves in each panel: the true function β_k (solid), the sample mean of estimators $\widehat{\beta}_k$ (long dashed), upper and lower 95% confidence bands calculated by the sample mean ± 1.96 times the sample standard deviation of the estimates (dashed), and the sample averages of upper and lower 95% confidence bands calculated using the sandwich variance estimates (dot-dashed). We can see that across all panels, the undersmoothing yields negligible biases, and the two types of confidence bands are nearly identical, indicating the validity of the proposed variance estimator.

To have a clear view of the performance of the 95% pointwise confidence intervals, in Figure 2.2 we further provide their coverage probabilities to the same coefficient curves depicted in

17
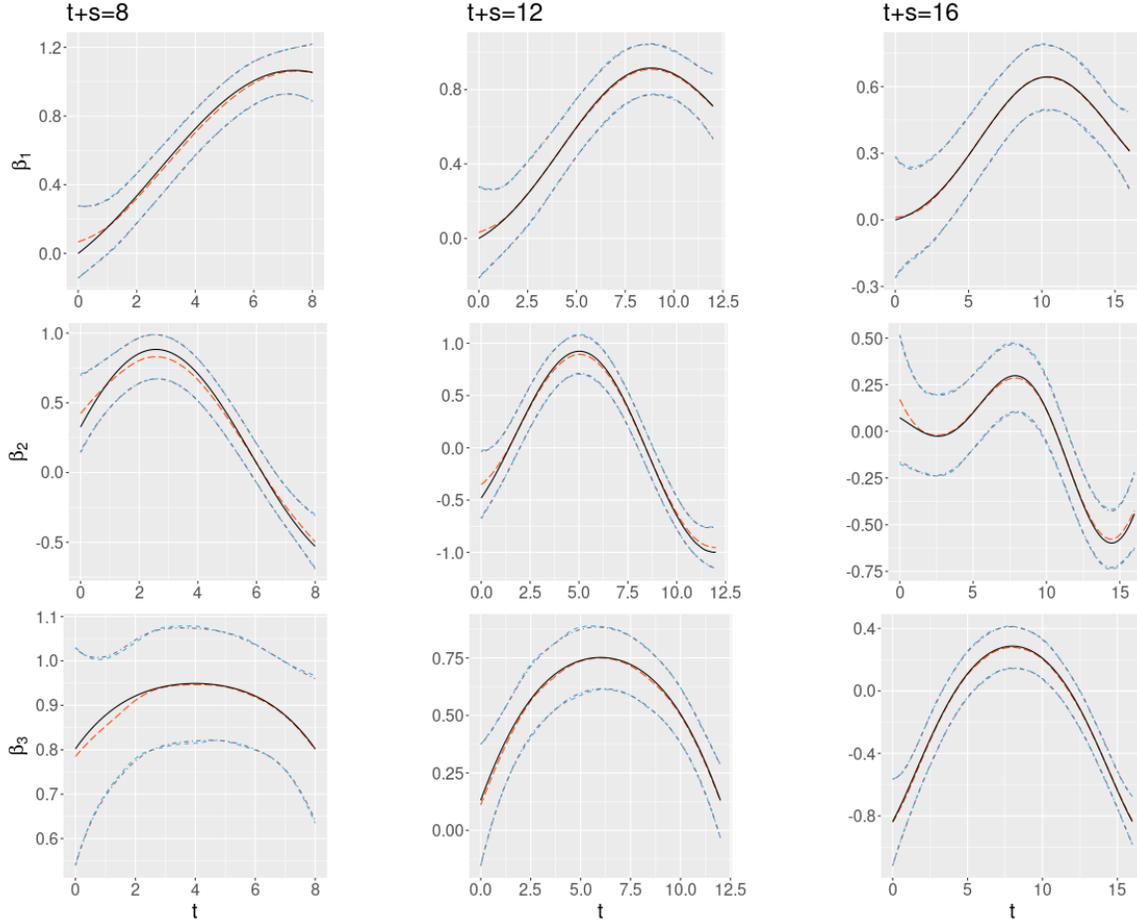
Figure 2.1: Mean of estimates and two types of 95% pointwise confidence bands at $T = 8$, 12, and 16.

Figure 2.1. From Figure 2.2 we see that the coverage probability is mostly around 95%, but can drop to near 85% on the boundaries or regions with large curvature of the coefficient due to relatively large biases of kernel smoothing in such regions.

At the request of an anonymous reviewer, we have implemented simulations with a much smaller sample size of $n = 400$, and provided results in Section D, Figure S2, of the online supplementary material. The overall performance is very similar, with a slightly larger bias for β_2 at $T = 8$ and, unsurprisingly, wider pointwise confidence bands.

We further verify the performance of hypothesis testing based on the asymptotic multivariate normal distribution given in Theorem 2.1. Such a test allows for comparing coefficient values
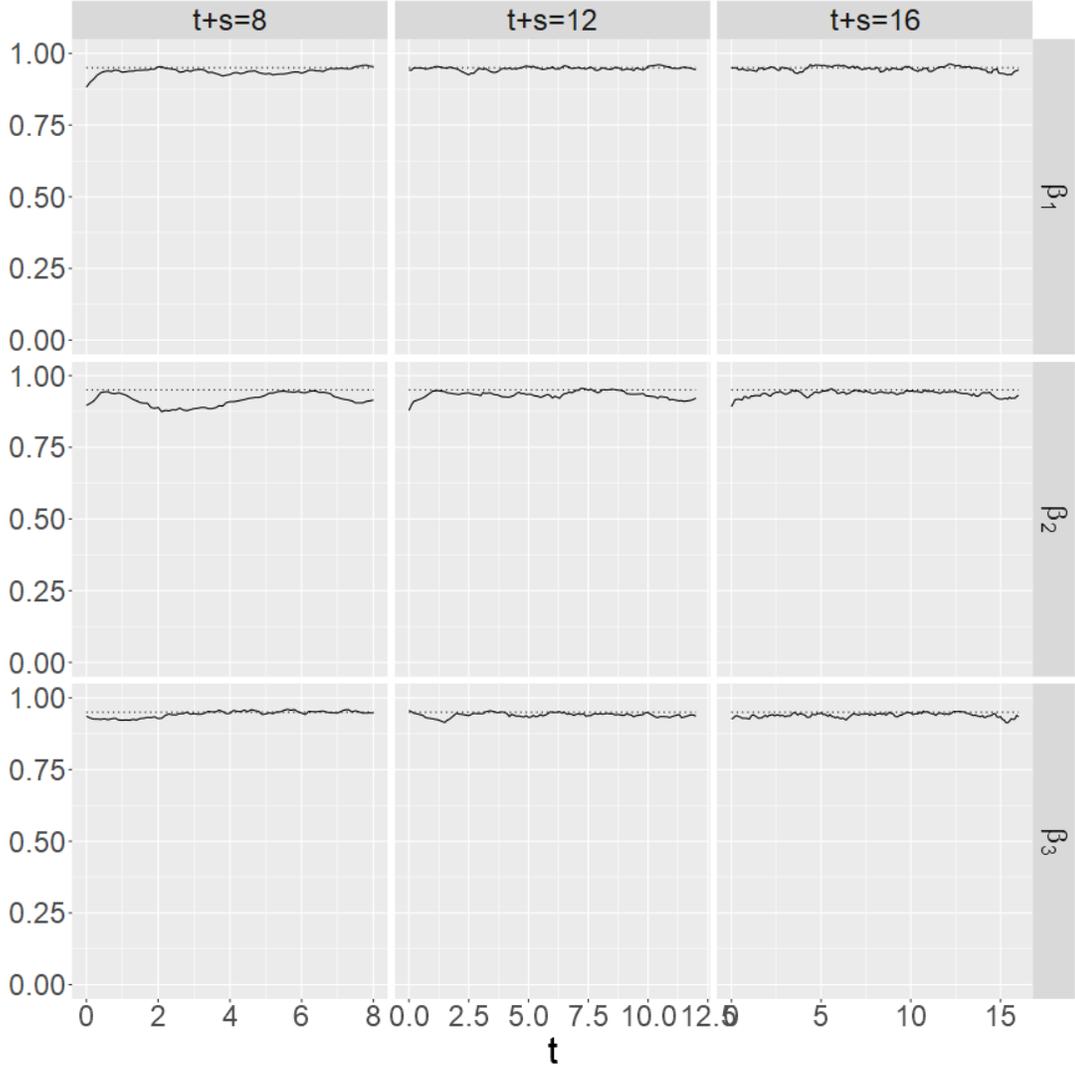
Figure 2.2: Coverage probabilities of 95% pointwise confidence intervals for β_1, β_2 and β_3. The dotted line shows the nominal level of 0.95.

at any two different pairs of time points. In order to also systematically evaluate the size of the test, we modify the simulation setup slightly by setting $\beta_3 = 0.5$ while keeping other simulation parameters unchanged, which creates a true null hypothesis of $\Delta\beta_3 = 0$. In Tables 1 and 2 we summarize the empirical power, or size when the null hypothesis is true, of a two-sided z-test. The empirical power is the rejection frequency of the test among 1000 simulation replications. In Table 1 we consider cases with the same failure time T but two different visit times t_1 and t_2; In Table 2 we consider the same visit time t but two different

failure times T_1 and T_2. We see from simulation results presented in both Tables 1 and 2 that empirical sizes of the tests are close to 0.05, showing the validity of the test. We also observe larger empirical powers for larger magnitudes of the coefficient differences.

t_1, t_2, T	$\Delta\beta_1$ (EP_1)	$\Delta\beta_2$ (EP_2)	$\Delta\beta_3$ (EP_3)
2, 4, 8	0.391 (0.980)	-0.146 (0.197)	0 (0.056)
2, 6, 8	0.670 (1.000)	-0.786 (1.000)	0 (0.054)
4, 6, 8	0.279 (0.905)	-0.640 (1.000)	0 (0.053)
2, 4, 12	0.273 (0.654)	0.610 (0.995)	0 (0.054)
2, 6, 12	0.553 (0.998)	0.616 (0.993)	0 (0.055)
4, 6, 12	0.281 (0.737)	0.006 (0.026)	0 (0.053)
2, 4, 16	0.134 (0.141)	0.038 (0.032)	0 (0.051)
2, 6, 16	0.317 (0.706)	0.214 (0.298)	0 (0.066)
4, 6, 16	0.183 (0.295)	0.176 (0.201)	0 (0.057)

Table 2.1: True difference $\Delta\beta_k = \beta_k(t_2, T - t_2) - \beta_k(t_1, T - t_1)$; empirical power EP_k for the hypothesis test $H_0 : \Delta\beta_k = 0$ vs $H_a : \Delta\beta_k \neq 0$, $k = 1, 2, 3$.

T_1, T_2, t	$\Delta\beta_1$ (EP_1)	$\Delta\beta_2$ (EP_2)	$\Delta\beta_3$ (EP_3)
8, 12, 2	-0.158 (0.272)	-0.637 (0.990)	0 (0.060)
8, 16, 2	-0.267 (0.564)	-0.872 (1.000)	0 (0.058)
12, 16, 2	-0.109 (0.139)	-0.235 (0.322)	0 (0.063)
8, 12, 4	-0.277 (0.685)	0.119 (0.218)	0 (0.058)
8, 16, 4	-0.524 (0.990)	-0.668 (1.000)	0 (0.055)
12, 16, 4	-0.247 (0.524)	-0.807 (1.000)	0 (0.048)
8, 12, 6	-0.275 (0.749)	0.765 (1.000)	0 (0.039)
8, 16, 6	-0.620 (1.000)	0.128 (0.210)	0 (0.064)
12, 16, 6	-0.345 (0.849)	-0.637 (0.995)	0 (0.059)

Table 2.2: True difference $\Delta\beta_k = \beta_k(t, T_2 - t) - \beta_k(t, T_1 - t)$; empirical power EP_k for the hypothesis test $H_0 : \Delta\beta_k = 0$ vs $H_a : \Delta\beta_k \neq 0$, $k = 1, 2, 3$.

2.5   The ESRD Medicare Data Analysis

We consider inpatient medical cost data of patients with end-stage renal disease (ESRD) from year 2007 to 2018 collected by the United States Renal Data System (USRDS). The longitudinal response is the daily inpatient cost paid by Medicare and the terminal event is death. The pattern of end-of-life Medicare cost has been identified in previous work. For

example, [4] showed an increasing and then decreasing pattern in Medicare costs before death among ovarian cancer patients. [29] found an increasing pattern in monthly outpatient EPO costs starting from 6 months prior to death and an initial jump since entry time, followed by a linear drop. When it comes to inpatient cost among ESRD patients, [24] established similar initial pattern as in [29] and an increasing then decreasing terminal pattern using a parametric model. Here we aim to investigate the patterns with a much larger sample size using our nonparametric modeling approach.

Following [24], we only include black and white patients who started their ESRD services in 2007 and were at least 65 years old when they started. We exclude patients who received kidney transplant because they could potentially have very different trajectories of inpatient costs. We also exclude patients who never had any hospitalization nor filled out the CMS Medical Evidence Report during the follow up. Instead of selecting a simple random sample of available ESRD patients for the analysis as in [24], all eligible patients are included in our analysis. Additionally, we are able to take advantage of the most updated data from USRDS, for which the follow-up ended on June 30th, 2018. We end up with a much larger sample size of 42,253 patients who died before the end of follow-up, much longer follow-up with an average of 34.6 months, and a very low censoring rate of only 3.74%. In the original data, a total cost is given for each hospitalization period. To convert the total cost into daily cost, we assume the cost rate is constant during each hospitalization. For example, if $10,000 is claimed for a patient during one hospitalization from May 1st to 20th, then the daily cost on May 10th is calculated to be $500. If the subject is not in hospital on a certain date, the cost is zero. We choose to select a 5% random sample of all days from entry to end of observation for each subject and calculate daily costs for these days. This greatly shortens the computation time for cross-validation and is valid because we do not require independence between observation time τ_j and time to death T.

We have considered the same set of covariates as in [24], which includes race, heart disease and

diabetes. Additionally, we include a binary covariate that indicates if Medicare is a secondary payer. Although most ESRD patients are eligible to apply for Medicare as their primary insurance payer, some are not immediately eligible for Medicare primary payer coverage at retirement because of their employment status and pre-existing primary insurance payers (e.g., group health plans). For this reason, this indicator is time-varying in the follow-up period. The estimated bivariate time-varying coefficients of the aforementioned covariates in the full model are displayed in Figure S1 of the online supplementary material, where the Bonferroni confidence band for each curve is constructed following [54] from the joint asymptotic distribution at 100 evenly spaced time points. These confidence bands serve as approximations of 95% global confidence bands, but can still be anti-conservative. From Figure S1 of the online supplementary material we see that the effect sizes of race, heart disease and diabetes variables are close to zero and their confidence bands contain zero over the entire support of $(\tau_{ij}, T_i - \tau_{ij})$. To capture the main pattern of Medicare payments, here we present the result of a reduced model that only includes the payer category:

$$\log(Y_{ij}/1000 + 1) = \beta_1(\tau_{ij}, T_i - \tau_{ij}) + \beta_2(\tau_{ij}, T_i - \tau_{ij})\text{Payer}_{ij} + \varepsilon_i(\tau_{ij}),$$

where log transformation $\log(Y/1000+1)$ corrects the highly skewed distribution of the daily Medicare payment Y.

In Figure 2.3, we plot the estimated curves and their confidence bands for β_1 and $\beta_1 + \beta_2$, respectively, obtained using a selected bandwidth of 12 days, where β_1 represents the log-transformed daily Medicare payment trajectory among ESRD patients when Medicare is the primary payer, and $\beta_1 + \beta_2$ corresponds to Medicare the secondary payer. The coefficient β_2 represents the Medicare payment difference between payer types, but plotting $\beta_1 + \beta_2$ provides a direct illustration of the medical cost pattern of Medicare payment as the secondary payer. Similar to how we display simulation results, we choose to only visualize $\beta_k(t, T - t)$ under several fixed values of T. Here we choose $T = 360$, 900, and 1440, corresponding to patients

who died roughly 1 year, 2.5 years and 4 years after entry, respectively.
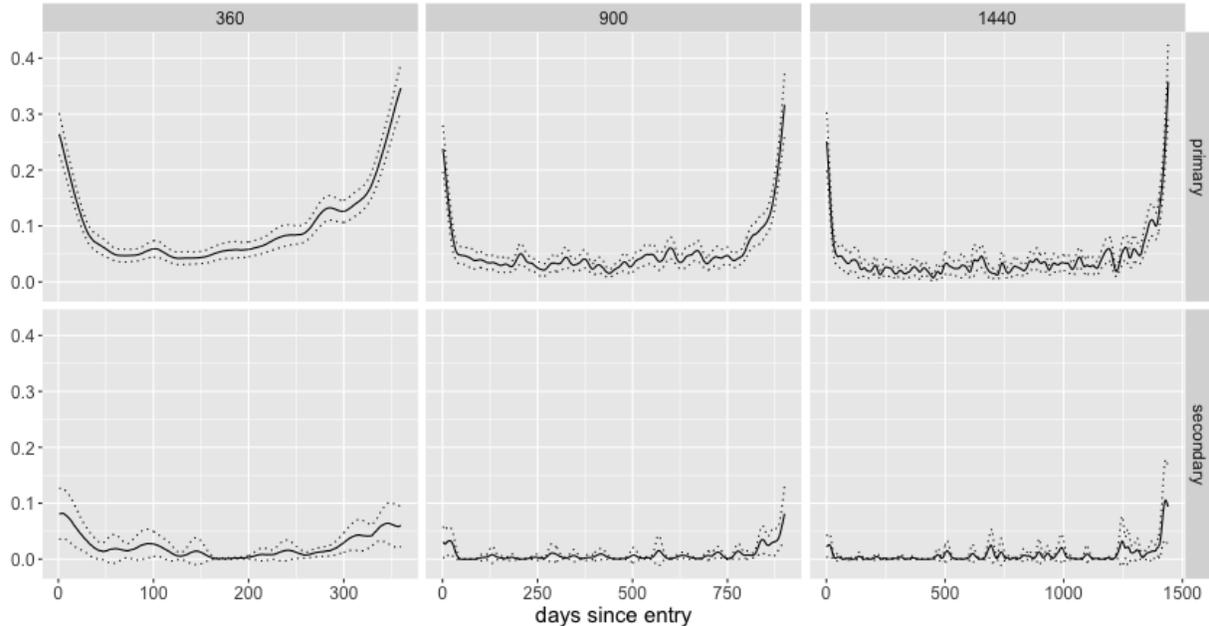
Figure 2.3: Estimated curves and their confidence bands for β_1 (Medicare as primary payer) and $\beta_1 + \beta_2$ (Medicare as secondary payer) under $T = 360, 900, 1440$.

From Figure 2.3 we see that the Medicare cost as primary payer starts to escalate from roughly 150 days prior to death, similar to the pattern observed in [24]. The peak value, however, is at the time of death, which is different to [24] where the peak was around three weeks before death. The initial pattern is different too: in our analysis, we find that the Medicare cost decreases drastically in the first two months after entry, then becomes stabilized overtime until close to death, whereas in [24], and in [29] as well, it increases first then decreases. Such differences are likely due to the restrictive parametric assumptions imposed in [24] and [29]. We also find that, when Medicare is secondary payer, the pattern of inpatient costs is similar but the magnitude is much smaller and at most times very close to zero. This is anticipated because a large portion of medical costs was paid first by some other insurance.

We further conduct formal statistical tests of Medicare payment patterns towards the end of life to confirm a key difference between our findings and [24]. These tests are based on

the finite-dimensional asymptotic independence and joint normality established in Theorem 2.1 for the intercept parameter β_1. We have showed in Section 2.4 that such a test has the correct power to detect the coefficient difference. In particular, for a given T, we consider three different residual lifetimes: 90 days, 30 days and 15 days prior to death. The tests compare β_1 at 90 versus 30 days and at 30 versus 15 days, respectively. Table 2.3 summarizes each estimated difference $\beta_1(t_2, T - t_2) - \beta_1(t_1, T - t_1)$, its 95% confidence interval, and the p-value of a one-sided test, showing significant evidence of a continuously increasing pattern of the intercept parameter β_1 towards the end of life for each of three different values of $T = 360$, 900, 1440 days. In other words, these tests reject the pattern obtained in [24] where the medical cost peaks and then declines during the last month before death.

It is well-known that kernel methods may yield larger biases at the boundaries. The boundary biases, however, diminish as the sample size increases (see 11) and can be further reduced by undersmoothing that is the approach we take in the ESRD data analysis. Although the choice of the undersmoothing factor $n^{1/20}$ seems working well in our simulations, in practice it is hard to determine what choice of bandwidth would result in a reasonable undersmoothing. However, given the fact that the estimated values of β_1 are very large with confidence intervals far from zero at the boundaries, it is reasonable to believe that the medical spending is much higher at the time when an ESRD patient becomes Medicare eligible and at the end of a patient's life.

It would be also of interest to test the medical spending pattern from a different angle, in other words, to compare β_1 with the same t but different T. Several such comparisons are shown in Table 4, where we see higher daily medical cost among subjects with shorter residual lifetime at the same time since enrollment. Such difference diminishes among longer survivors, however.

	$T - t_1 = 90, T - t_2 = 30$	$T - t_1 = 30, T - t_2 = 15$
$T = 360$	$0.067, (0.037, 0.098), 7.56 \times 10^{-6}$	$0.080, (0.041, 0.118), 2.61 \times 10^{-5}$
$T = 900$	$0.068, (0.035, 0.102), 3.78 \times 10^{-5}$	$0.071, (0.028, 0.113), 5.31 \times 10^{-4}$
$T = 1440$	$0.066, (0.024, 0.108), 1.11 \times 10^{-3}$	$0.079, (0.026, 0.132), 1.79 \times 10^{-3}$

Table 2.3: Estimated difference $\beta_1(t_2, T - t_2) - \beta_1(t_1, T - t_1)$, its 95% confidence interval, and p-value for the monotonicity test: $H_0 : \beta_1(t_1, T - t_1) = \beta_1(t_2, T - t_2)$ versus $H_a : \beta_1(t_1, T - t_1) < \beta_1(t_2, T - t_2)$.

	$T_1 = 900, T_2 = 360$	$T_1 = 1440, T_2 = 900$
$t = 90$	$0.014, (-0.004, 0.033), 6.56 \times 10^{-2}$	$0.006, (-0.018, 0.029), 3.15 \times 10^{-1}$
$t = 180$	$0.030, (0.012, 0.047), 4.67 \times 10^{-4}$	$0.008, (-0.007, 0.023), 1.42 \times 10^{-1}$
$t = 270$	$0.087, (0.065, 0.110), 1.02 \times 10^{-14}$	$-0.002, (-0.020, 0.015), 6.06 \times 10^{-1}$

Table 2.4: Estimated difference $\beta_1(t, T_2 - t) - \beta_1(t, T_1 - t)$, its 95% confidence interval, and p-value for the monotonicity test: $H_0 : \beta_1(t, T_1 - t) = \beta_1(t, T_2 - t)$ versus $H_a : \beta_1(t, T_1 - t) < \beta_1(t, T_2 - t)$.

2.6 Appendix

2.6.1 Regularity Conditions

Denote by C^q the class of functions with q-th order continuous derivatives. For the points $(t_l, s_l), l = 0, 1, ..., d$ in Thm 2.1 and 2.2, we need the following regularity conditions.

1. K is a probability density of the form $f(\|\mathbf{x}\|_2)$, where $f(\cdot)$ is of bounded variation on bounded support.

2. (a) $n^{1/6} h \to h_0 < \infty$, where $h_0 \geq 0$.

   (b) $nh^2 \to \infty$.

   (c) $n^{3/4} h^2 / \log n \to \infty$.

3. (a) For any j, k, l, $\eta_{j,kl}(x, y)$ is of class C^1 in a neighborhood of (t_l, s_l).

   (b) For any j, k, $E\left[X_k(\tau_j)^8 | \tau_j = x, T - \tau_j = y\right]$ is bounded in a neighborhood of (t_l, s_l).

25

(c) For any $j_1 \neq j_2$ and k, $E\left[X_k(\tau_{j_1})^8 | \tau_{j_1} = x, \tau_{j_2} = y, T - \tau_{j_1} = z\right]$ is bounded in a neighborhood of (t_l, t_l, s_l).

4. (a) For any j, f_j is of class C^1 in a neighborhood of (t_l, s_l).

   (b) For $j_1 \neq j_2$, $f_{\tau_{j_1}, \tau_{j_2}, T - \tau_{j_1}}$ is bounded in a neighborhood of (t_l, t_l, s_l).

5. For any k, β_k is of class C^2 in a neighborhood of (t_l, s_l).

6. (a) σ^2 is continuous at t_l.

   (b) $E\varepsilon(\cdot)^4$ is bounded in a neighborhood of t_l.

7. There exists j such that $\boldsymbol{\eta}_j(t_l, s_l)$ is positive definite and $f_j(t_l, s_l)$ is positive.

Remark: Most of the regularity conditions are direct extensions of those in [54] to the bivariate case. Specifically, Condition 1 ensures that K has a compact support on $\mathbb{R}^2$ and is symmetric, i.e.,

$$\iint xK(x, y)dxdy = 0, \quad \iint yK(x, y)dxdy = 0.$$

Conditions 2a and 2b, or 2a and 2c, together specify a range of feasible bandwidths, which justifies the use of undersmoothed bandwidth. Note that, pointed out by an anonymous reviewer, Condition 2c gives a more restrictive lower bound for bandwidth h and is only required for Theorem 2.2. For Theorem 2.1, more relaxed Condition 2b is sufficient. The exponent $3/4$ in Condition 2c is a result of finite moments of X and ε in Conditions 3b and 6b. A similar condition is given in [10] for the Nadaraya-Watson estimator. In particular, when Y is assumed to have a finite 8th order moment, which is a result of our conditions 3b, 3c and 6b, the bandwidth h in [10] needs to satisfy $(n/\log n)^{3/4}h \to \infty$ for a univariate kernel estimator. We would also like to point out that the assumptions of finite higher order moments for X and ε in Conditions 3b, 3c and 6b automatically hold for sub-Gaussian or

sub-exponential processes. Lastly, Condition 7 ensures that $\sum_{i \in \mathcal{D}} \mathbf{X}_i^T \mathbf{K}_i \mathbf{X}_i / n$ is invertible asymptotically. This is commonly assumed for regression models.

2.6.2 Technical Lemmas and Proofs

Lemma 2.1. *For any fixed n, let $X_{n,i}$, $i = 1, ..., n$, be i.i.d random variables with mean μ_n and variance σ_n^2, where $\sigma_n^2 = o(n)$. If $\mu_n = O(1)$, then $\sum_{i=1}^n X_{n,i}/n = O_p(1)$; if $\mu_n \to \mu$, then $\sum_{i=1}^n X_{n,i}/n \to_p \mu$.*

Proof. For the first part, by $\mu_n = O(1)$ we know there exist M and N such that $|\mu_n| \leq M$ when $n \geq N$. Thus for any $\varepsilon > 0$, when $n \geq N$ and $n \geq \sigma_n^2/\varepsilon M^2$, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_{n,i}\right| > 2M\right) \leq P\left(\left|\frac{1}{n}\sum_{i=1}^n X_{n,i} - \mu_n\right| > M\right) \leq \frac{\sigma_n^2}{nM^2} \leq \varepsilon.$$

This concludes the proof for the first part. For the second part, by $\mu_n \to \mu$ we know that for any $\varepsilon > 0$, there exists N' such that when $n \geq N'$, $|\mu_n - \mu| \leq \varepsilon/2$. Thus when $n \geq N'$, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_{n,i} - \mu\right| > \varepsilon\right) \leq P\left(\left|\frac{1}{n}\sum_{i=1}^n X_{n,i} - \mu_n\right| > \varepsilon/2\right) = \frac{4\sigma_n^2}{n\varepsilon^2} \to 0.$$

$\square$

Lemma 2.2. *Let $A_n(\mathbf{x})$ and $A(\mathbf{x})$ be symmetric matrix-valued functions, and $B_n(\mathbf{x})$ and $B(\mathbf{x})$ be matrix-valued functions of $\mathbf{x} \in \mathcal{X}$, where $B_n(\mathbf{x})$ and $B(\mathbf{x})$ may not be square matrices and can potentially be vectors. Let $\| \cdot \|_2$ denote the spectral norm, i.e., $\|A\|_2 = \sup_{\mathbf{x} \neq 0} \|A\mathbf{x}\|_2/\|\mathbf{x}\|_2$, and $\lambda_1(A(\mathbf{x}))$ be the smallest eigenvalue of $A(\mathbf{x})$. Suppose the following*

27

hold for a sequence $C_1, C_2, ...$ of subsets of $\mathcal{X}$:

$$\sup_{\mathbf{x} \in C_n} \|A_n(\mathbf{x}) - A(\mathbf{x})\|_2 \to_p 0,$$

$$\sup_{\mathbf{x} \in C_n} \|B_n(\mathbf{x}) - B(\mathbf{x})\|_2 \to_p 0,$$

$$\liminf_n \inf_{\mathbf{x} \in C_n} \lambda_1\big(A(\mathbf{x})\big) = a > 0,$$

$$\limsup_n \sup_{\mathbf{x} \in C_n} \|B(\mathbf{x})\|_2 = b < \infty.$$

Then we have

$$\sup_{\mathbf{x} \in C_n} \|A_n(\mathbf{x})^{-1} B_n(\mathbf{x}) - A(\mathbf{x})^{-1} B(\mathbf{x})\|_2 \to_p 0.$$

Proof. By definition, for any $\varepsilon > 0$ and $0 < \delta < a/2$, there exists N_1 such that, when $n > N_1$,

$$P\left(\sup_{\mathbf{x} \in C_n} \|A_n(\mathbf{x}) - A(\mathbf{x})\|_2 < \delta \right) > 1 - \varepsilon,$$

$$P\left(\sup_{\mathbf{x} \in C_n} \|B_n(\mathbf{x}) - B(\mathbf{x})\|_2 < \delta \right) > 1 - \varepsilon.$$

Moreover, there exists N_2 such that when $n > N_2$,

$$\inf_{\mathbf{x} \in C_n} \lambda_1\big(A(\mathbf{x})\big) > \frac{a}{2} \quad \text{and} \quad \sup_{\mathbf{x} \in C_n} \|B(\mathbf{x})\|_2 < b + 1.$$

When $n > \max(N_1, N_2)$, on the set where $\sup_{\mathbf{x} \in C_n} \|A_n(\mathbf{x}) - A(\mathbf{x})\|_2 < \delta$, we have for any $\mathbf{x} \in C_n$ and non-zero vector $\mathbf{y}$,

$$\|A_n(\mathbf{x})\mathbf{y} - A(\mathbf{x})\mathbf{y}\|_2 < \delta \|\mathbf{y}\|_2.$$

This further implies

$$\|A_n(\mathbf{x})\mathbf{y}\|_2 > \|A(\mathbf{x})\mathbf{y}\|_2 - \delta\|\mathbf{y}\|_2 > \left(\frac{a}{2} - \delta\right)\|\mathbf{y}\|_2,$$

which in turn gives

$$\sup_{\mathbf{x}\in C_n} \|A_n(\mathbf{x})^{-1}\|_2 \le \frac{2}{a - 2\delta}.$$

Thus we have

$$\sup_{\mathbf{x}\in C_n} \|A_n(\mathbf{x})^{-1}B_n(\mathbf{x}) - A(\mathbf{x})^{-1}B(\mathbf{x})\|_2$$

$$\le \sup_{\mathbf{x}\in C_n} \|A_n(\mathbf{x})^{-1}(B_n(\mathbf{x}) - B(\mathbf{x}))\|_2 + \sup_{\mathbf{x}\in C_n} \|A_n(\mathbf{x})^{-1}(A_n(\mathbf{x}) - A(\mathbf{x}))A(\mathbf{x})^{-1}B(\mathbf{x})\|_2$$

$$< \frac{2\delta}{a - 2\delta}\left(1 + \frac{2}{a}(b+1)\right)$$

with probability larger than $1 - 2\varepsilon$. The proof is complete. $\square$

Lemma 2.3. *If the class of functions $\mathcal{F}$ is Euclidean [37] for a constant envelope function F, $0 < F < \infty$, and g is a fixed function, then the class*

$$\mathcal{F}g = \{fg : f \in \mathcal{F}\}$$

is Euclidean for the envelope function $F|g|$.

Proof. For any probability measure Q such that $0 < QF|g| < \infty$, we can define R to be the probability measure with density $|g|/Q|g|$ with respect to Q, i.e., $dR = |g|dQ/Q|g|$. Clearly R is a valid measure because $0 < Q|g| < \infty$.

By the definition of Euclidean [37], we can find $n \le A\varepsilon^{-V}$ and $f_1, ..., f_n \in \mathcal{F}$ that makes the

29

following holds for any $f \in \mathcal{F}$:

$$\min_i R|f - f_i| \leq \varepsilon RF.$$

Rewriting this with respect to measure Q,

$$\min_i Q|fg - f_ig|/Q|g| \leq \varepsilon QF|g|/Q|g|.$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

2.6.3   Proof of Theorem 2.1

Define

$$\mathbf{A}_n(t, s; h) = \frac{1}{n} \sum_{i \in \mathcal{D}} \mathbf{X}_i^T \mathbf{K}_i(t, s; h) \mathbf{X}_i,$$

$$\mathbf{B}_n(t, s; h) = \frac{1}{n} \sum_{i \in \mathcal{D}} \mathbf{X}_i^T \mathbf{K}_i(t, s; h) \mathbf{Y}_i,$$

$$\mathbf{R}_n(t, s; h) = \mathbf{B}_n(t, s; h) - \mathbf{A}_n(t, s; h)\boldsymbol{\beta}(t, s),$$

then we have

$$\widehat{\boldsymbol{\beta}}(\mathbf{t}, \mathbf{s}; h) - \boldsymbol{\beta}(\mathbf{t}, \mathbf{s}) = \begin{pmatrix} \mathbf{A}_n(t_1, s_1; h) & & \\ & \ddots & \\ & & \mathbf{A}_n(t_d, s_d; h) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{R}_n(t_1, s_1; h) \\ \vdots \\ \mathbf{R}_n(t_d, s_d; h) \end{pmatrix}.$$

The proof proceeds by showing that $\mathbf{A}_n(t_l, s_l; h)$ converges to a fixed invertible matrix for $l = 1, ..., d$ and $\sqrt{nh^2} \boldsymbol{a}^T (\mathbf{R}_n(t_1, s_1; h)^T, ..., \mathbf{R}_n(t_d, s_d; h)^T)^T = \sqrt{nh^2} \sum_{l=1}^{d} \boldsymbol{a}_l^T \mathbf{R}_n(t_l, s_l; h)$ converges to a normal random variable for any non-zero vector $\boldsymbol{a} = (\boldsymbol{a}_1^T, ..., \boldsymbol{a}_d^T)^T$.

To show the consistency of $\mathbf{A}_n(t_l, s_l; h)$, we apply Lemma 2.1 by examining its expectation and variance. With some simple algebra and applying the dominated convergence theorem (DCT), we can show

$$EA_{n,k_1k_2}(t_l, s_l; h) = \sum_{j=1}^{m} \eta_{j,k_1k_2}(t_l, s_l) f_j(t_l, s_l) + o(1), \tag{2.9}$$

$$Var\left(A_{n,k_1k_2}(t_l, s_l; h)\right) = O\left(\frac{1}{nh^2}\right). \tag{2.10}$$

The detailed calculations of (2.9) and (2.10) will be provided later in this proof. Then by Lemma 2.1 and condition 2b, we have

$$\mathbf{A}_n(t_l, s_l; h) \to_p \mathbf{\Omega}(t_l, s_l). \tag{2.11}$$

To show $\sqrt{nh^2}\boldsymbol{a}^T(\mathbf{R}_n(t_1, s_1; h)^T, ..., \mathbf{R}_n(t_d, s_d; h)^T)^T$ converges to a normal random variable, we define

$$Z_{n,i}(\boldsymbol{a}) = \sum_{l=1}^{d} 1(i \in \mathcal{D}) \boldsymbol{a}_l^T \mathbf{X}_i^T \mathbf{K}_i(t_l, s_l; h) \left[\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}(t_l, s_l)\right]$$

so that $\boldsymbol{a}^T(\mathbf{R}_n(t_1, s_1; h)^T, ..., \mathbf{R}_n(t_d, s_d; h)^T)^T = \sum_{i=1}^{n} Z_{n,i}(\boldsymbol{a})/n$. We adopt Lyapunov's central limit theorem and verify the following Lyapunov's condition

$$\frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E\left|Z_{n,i}(\boldsymbol{a}) - EZ_{n,i}(\boldsymbol{a})\right|^{2+\delta} \to 0$$

for some $\delta > 0$ and $s_n^2 = \sum_{i=1}^{n} Var(Z_{n,i}(\boldsymbol{a}))$. To achieve this, we examine the expectation,

variance, and the $(2 + \delta)$-th central moment of $Z_{n,i}(\boldsymbol{a})$

$$EZ_{n,i}(\boldsymbol{a}) = h^2 \sum_{l=1}^{d} \boldsymbol{a}_l^T \boldsymbol{\Gamma}(t_l, s_l) + o(h^2), \tag{2.12}$$

$$Var(Z_{n,i}(\boldsymbol{a})) = \frac{\mu_0}{h^2} \sum_{l=1}^{d} \boldsymbol{a}_l^T \sigma^2(t_l) \boldsymbol{\Omega}(t_l, s_l) \boldsymbol{a}_l + o\left(\frac{1}{h^2}\right), \tag{2.13}$$

$$E|Z_{n,i}(\boldsymbol{a}) - EZ_{n,i}(\boldsymbol{\alpha})|^{2+\delta} = O\left(h^{-2-2\delta}\right). \tag{2.14}$$

The detailed calculations of (2.12)-(2.14) will be provided later in this proof. Thus we have $s_n^2 \geq Cn/h^2$ for some constant $C > 0$, and with large enough n,

$$\frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E\left|Z_{n,i}(\boldsymbol{a}) - EZ_{n,i}(\boldsymbol{a})\right|^{2+\delta} \leq C' \frac{n/h^{2+2\delta}}{n^{1+\delta/2}/h^{2+\delta}} = C' n^{-\delta/2} h^{-\delta} \to 0$$

by Condition 2b. Now we can claim

$$\frac{\sum_i Z_{n,i}(\boldsymbol{a}) - \sum_i EZ_{n,i}(\boldsymbol{a})}{\sqrt{\sum_{i=1}^{n} Var\left(Z_{n,i}(\boldsymbol{a})\right)}} \to_d N(0, 1).$$

By Cramer-Wold device, we have

$$\sqrt{nh^2}\mathbf{R}_n(\mathbf{t}, \mathbf{s}; h) \to_d N(h_0^3 \mathbf{B}^*(\mathbf{t}, \mathbf{s}), \mathbf{V}^*(\mathbf{t}, \mathbf{s})),$$

where

$$\mathbf{B}^*(\mathbf{t}, \mathbf{s}) = \begin{pmatrix} \boldsymbol{\Gamma}(t_1, s_1) \\ \vdots \\ \boldsymbol{\Gamma}(t_d, s_d) \end{pmatrix}, \quad \mathbf{V}^*(\mathbf{t}, \mathbf{s}) = \mu_0 \begin{pmatrix} \sigma^2(t_1)\boldsymbol{\Omega}(t_1, s_1) & & \\ & \ddots & \\ & & \sigma^2(t_d)\boldsymbol{\Omega}(t_d, s_d) \end{pmatrix}.$$

Further by Slutsky's theorem and (2.11), we have

$$\sqrt{nh^2}(\widehat{\boldsymbol{\beta}}(\mathbf{t}, \mathbf{s}; h) - \boldsymbol{\beta}(\mathbf{t}, \mathbf{s})) \to_d N(h_0^3 \mathbf{B}(\mathbf{t}, \mathbf{s}), \mathbf{V}(\mathbf{t}, \mathbf{s})).$$

Now we show the detailed calculations of (2.9)-(2.10) and (2.12)-(2.14). We first introduce the following notation

$$
\begin{aligned}
\boldsymbol{\phi}_{i,l} &= 1(i \in \mathcal{D})\mathbf{X}_i^T \mathbf{K}_i(t_l, s_l; h)\boldsymbol{\varepsilon}_i \\
&= \frac{1}{h^2} \sum_{j=1}^{m} 1(\tau_{ij} \le T_i \le C_i)\mathbf{X}_{ij} K\left(\frac{\tau_{ij} - t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h}\right)\varepsilon_{ij} \\
&\stackrel{d}{=} \sum_{j=1}^{m} \boldsymbol{\phi}_{ij,l},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\psi}_{i,l} &= 1(i \in \mathcal{D})\mathbf{X}_i^T \mathbf{K}_i(t_l, s_l; h) \begin{pmatrix} \mathbf{X}_{i1}^T \boldsymbol{\beta}(\tau_{i1}, T_i - \tau_{i1}) - \mathbf{X}_{i1}^T \boldsymbol{\beta}(t_l, s_l) \\ \vdots \\ \mathbf{X}_{im_i}^T \boldsymbol{\beta}(\tau_{im_i}, T_i - \tau_{im_i}) - \mathbf{X}_{im_i}^T \boldsymbol{\beta}(t_l, s_l) \end{pmatrix} \\
&= \frac{1}{h^2} \sum_{j=1}^{m} \left\{ 1(\tau_{ij} \le T_i \le C_i)\mathbf{X}_{ij} K\left(\frac{\tau_{ij} - t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h}\right) \right. \\
&\qquad \left. \times \sum_{k=1}^{p} X_{ijk}[\beta_k(\tau_{ij}, T_i - \tau_{ij}) - \beta_k(t_l, s_l)] \right\} \\
&\stackrel{d}{=} \sum_{j=1}^{m} \boldsymbol{\psi}_{ij,l},
\end{aligned}
$$

then we have $Z_{n,i}(\boldsymbol{a}) = \sum_{l=1}^{d} \boldsymbol{a}_l^T (\boldsymbol{\phi}_{i,l} + \boldsymbol{\psi}_{i,l})$. $\mathbf{X}_{ij}$ is short for $\mathbf{X}_i(\tau_{ij})$ and X_{ijk} is short for $X_{ik}(\tau_{ij})$, which will be used for the rest of the proof.

Proof of (2.9):

$$
\begin{aligned}
&EA_{n,k_1k_2}(t_l, s_l; h) \\
&= \frac{1}{h^2} \sum_{j=1}^{m} E\left[1(\tau_{ij} \le T_i \le C_i)X_{ijk_1}X_{ijk_2}K\left(\frac{\tau_{ij} - t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h}\right)\right] \\
&= \sum_{j=1}^{m} \iint \eta_{j,k_1k_2}(hx + t_l, hy + s_l)K(x, y)1(hy + s_l \ge 0)f_j(hx + t_l, hy + s_l)dxdy \\
&= \sum_{j=1}^{m} \eta_{j,k_1k_2}(t_l, s_l)f_j(t_l, s_l) + o(1).
\end{aligned}
$$

The last equality comes from DCT because η_{j,k_1k_2} and f_j are continuous by Condition 3a and 4a and K is bounded on bounded support by Condition 1.

Proof of (2.10):

$Var\left(A_{n,k_1k_2}(t_l,s_l;h)\right)$

$$= \frac{1}{nh^4}Var\left(\sum_{j=1}^{m}1(\tau_{ij}\le T_i\le C_i)X_{ijk_1}X_{ijk_2}K\left(\frac{\tau_{ij}-t_l}{h},\frac{T_i-\tau_{ij}-s_l}{h}\right)\right)$$

$$\le \frac{m}{nh^4}\sum_{j=1}^{m}E\left[1(\tau_{ij}\le T_i\le C_i)X_{ijk_1}X_{ijk_2}K\left(\frac{\tau_{ij}-t_l}{h},\frac{T_i-\tau_{ij}-s_l}{h}\right)\right]^2$$

$$= \frac{m}{nh^2}\sum_{j=1}^{m}\iint 1(hy+s_l\ge 0)\zeta_{j,k_1k_1k_2k_2}(hx+t_l,hy+s_l)K^2(x,y)f_j(hx+t_l,hy+s_l)dxdy$$

$$= O\left(\frac{1}{nh^2}\right),$$

where $\zeta_{j,k_1k_2k_3k_4}(x,y)=E[1(T_i\le C_i)X_{ijk_1}X_{ijk_2}X_{ijk_3}X_{ijk_4}|\tau_{ij}=x,T_i-\tau_{ij}=y]$. The inequality above is obtained following $(a_1+\cdots+a_k)^2\le n(a_1^2+\cdots+a_k^2)$, which is from Jensen's Inequality. The last equality again comes from DCT because $\zeta_{j,k_1k_2k_3k_4}$ is locally bounded by Condition 3b.

Proof of (2.12):

$$EZ_{n,i}(\boldsymbol{a})=\sum_{l=1}^{d}\boldsymbol{a}_l^T(E\boldsymbol{\phi}_{i,l}+E\boldsymbol{\psi}_{i,l})=\sum_{l=1}^{d}\boldsymbol{a}_l^TE\boldsymbol{\psi}_{i,l},$$

and we have

$$(E\boldsymbol{\psi}_{i,l})_{k_1}$$

$$= \frac{1}{h^2}\sum_{j=1}^{m} E\left\{ 1(\tau_{ij} \le T_i \le C_i) X_{ijk_1} K\left(\frac{\tau_{ij} - t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h} \right) \right.$$

$$\left. \times \sum_{k=1}^{p} X_{ijk}[\beta_k(\tau_{ij}, T_i - \tau_{ij}) - \beta_k(t_l, s_l)] \right\}$$

$$= \frac{1}{h^2}\sum_{j=1}^{m}\sum_{k=1}^{p} E\left\{ 1(\tau_{ij} \le T_i \le C_i) X_{ijk_1} X_{ijk} K\left(\frac{\tau_{ij} - t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h} \right) \right.$$

$$\left. \times [\beta_k(\tau_{ij}, T_i - \tau_{ij}) - \beta_k(t_l, s_l)] \right\}$$

$$= \sum_{j=1}^{m}\sum_{k=1}^{p} \iint K(x,y) 1(hy + s_l \ge 0) \eta_{j,k_1 k}(hx + t_l, hy + s_l)$$

$$\times [\beta_k(hx + t_l, hy + s_l) - \beta_k(t_l, s_l)] f_j(hx + t_l, hy + s_l) dx dy$$

$$= \sum_{j=1}^{m}\sum_{k=1}^{p} \iint K(x,y) 1(hy + s_l \ge 0)$$

$$\times \left[\eta_{j,k_1 k}(t_l, s_l) + hx \eta_{j,k_1 k}^{(1,0)}(\theta_1 hx + t_l, \theta_1 hy + s_l) + hy \eta_{j,k_1 k}^{(0,1)}(\theta_1 hx + t_l, \theta_1 hy + s_l) \right]$$

$$\times \left[hx \beta_k^{(1,0)}(t_l, s_l) + hy \beta_k^{(0,1)}(t_l, s_l) + h^2 xy \beta_k^{(1,1)}(\theta_2 hx + t_l, \theta_2 hy + s_l) \right.$$

$$\left. + \frac{h^2 x^2}{2} \beta_k^{(2,0)}(\theta_2 hx + t_l, \theta_2 hy + s_l) + \frac{h^2 y^2}{2} \beta_k^{(0,2)}(\theta_2 hx + t_l, \theta_2 hy + s_l) \right]$$

$$\times \left[f_j(t_l, s_l) + hx f_j^{(1,0)}(\theta_3 hx + t_l, \theta_3 hy + s_l) + hy f_j^{(0,1)}(\theta_3 hx + t_l, \theta_3 hy + s_l) \right] dx dy$$

$$= h^2 \sum_{j=1}^{m}\sum_{k=1}^{p} \langle \boldsymbol{\mu}_2, \boldsymbol{g}_{j,k_1 k}(t_l, s_l) \rangle + o(h^2)$$

$$= h^2 \boldsymbol{\Gamma}(t_l, s_l)_{k_1} + o(h^2),$$

where $\beta_k^{(p,q)} = \partial^{p+q}\beta_k/(\partial x^p \partial y^q)$ and $\theta_1, \theta_2, \theta_3 \in (0,1)$. The Taylor expansion is valid due to condition 3a, 4a and 5.

Proof of (2.13):

$$Var(Z_{n,i}(\boldsymbol{a})) = \sum_{l_1=1}^{d} \sum_{l_2=1}^{d} \boldsymbol{a}_{l_1}^T [Cov(\boldsymbol{\phi}_{i,l_1}, \boldsymbol{\phi}_{i,l_2}) + Cov(\boldsymbol{\psi}_{i,l_1}, \boldsymbol{\psi}_{i,l_2})] \boldsymbol{a}_{l_2}$$

$$= \sum_{l_1=1}^{d} \sum_{l_2=1}^{d} \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} \boldsymbol{a}_{l_1}^T [Cov(\boldsymbol{\phi}_{ij_1,l_1}, \boldsymbol{\phi}_{ij_2,l_2}) + Cov(\boldsymbol{\psi}_{ij_1,l_1}, \boldsymbol{\psi}_{ij_2,l_2})] \boldsymbol{a}_{l_2}.$$

For terms with $\boldsymbol{\phi}$, when $j_1 = j_2 = j$ and $l_1 = l_2 = l$,

$$[Cov(\boldsymbol{\phi}_{ij_1,l_1}, \boldsymbol{\phi}_{ij_2,l_2})]_{k_1,k_2}$$

$$= \frac{1}{h^4} E\left(1(\tau_{ij} \leq T_i \leq C_i) X_{ijk_1} X_{ijk_2} K^2\left(\frac{\tau_{ij} - t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h} \right) \varepsilon_{ij}^2 \right)$$

$$= \frac{1}{h^2} \iint 1(s_l + hy \geq 0) \eta_{j,k_1k_2}(t_l + hx, s_l + hy) K^2(x,y) \sigma^2(t_l + hx) f_j(t_l + hx, s_l + hy) dx dy$$

$$= \mu_0 \eta_{j,k_1k_2}(t_l, s_l) \sigma^2(t_l) f_j(t_l, s_l)/h^2 + o(1/h^2).$$

Here we additionally need Condition 6a to apply DCT. When $j_1 = j_2 = j$ and $l_1 \neq l_2$,

$$[Cov(\boldsymbol{\phi}_{ij_1,l_1}, \boldsymbol{\phi}_{ij_2,l_2})]_{k_1,k_2}$$

$$= \frac{1}{h^4} E\left(1(\tau_{ij} \leq T_i \leq C_i) X_{ijk_1} X_{ijk_2} \right.$$

$$\times K\left(\frac{\tau_{ij} - t_{l_1}}{h}, \frac{T_i - \tau_{ij} - s_{l_1}}{h} \right) K\left(\frac{\tau_{ij} - t_{l_2}}{h}, \frac{T_i - \tau_{ij} - s_{l_2}}{h} \right) \varepsilon_{ij}^2 \left. \right)$$

$$= \frac{1}{h^2} \iint 1(s_l + hy \geq 0) \eta_{j,k_1k_2}(t_l + hx, s_l + hy)$$

$$\times K(x,y) K(x + (t_{l_1} - t_{l_2})/h, y + (s_{l_1} - s_{l_2})/h)$$

$$\times \sigma^2(t_l + hx) f_j(t_l + hx, s_l + hy) dx dy$$

$$= o(1/h^2).$$

When $j_1 \neq j_2$ and $l_1 = l_2 = l$,

$$[Cov(\phi_{ij_1,l_1}, \phi_{ij_2,l_2})]_{k_1,k_2}$$

$$= \frac{1}{h^4} E\left(1(\tau_{ij_1}, \tau_{ij_2} \leq T_i \leq C_i) X_{ij_1 k_1} X_{ij_2 k_2} \right.$$

$$\left. \times K\left(\frac{\tau_{ij_1} - t_l}{h}, \frac{T_i - \tau_{ij_1} - s_l}{h}\right) K\left(\frac{\tau_{ij_2} - t_l}{h}, \frac{T_i - \tau_{ij_2} - s_l}{h}\right) \varepsilon_{ij_1} \varepsilon_{ij_2} \right)$$

$$= \frac{1}{h} \iiint 1(s_l + hz \geq 0) 1(s_l + hx + hz - hy \geq 0) \xi_{j_1 j_2, k_1 k_2}(t_l + hx, t_l + hy, s_l + hz)$$

$$\times K(x,z)K(y, x+z-y)\rho(t_l + hx, t_l + hy)f_{j_1,j_2}(t_l + hx, t_l + hy, s_l + hz)dxdydz$$

$$= O(1/h),$$

where $\xi_{j_1 j_2, k_1 k_2}(x,y,z) = E[1(T_i \leq C_i) X_{ij_1 k_1} X_{ij_2 k_2} | \tau_{ij_1} = x, \tau_{ij_2} = y, T_i - \tau_{ij_1} = z]$ and f_{j_1,j_2} is the joint density of τ_{ij_1}, τ_{ij_2} and $T_i - \tau_{ij_1}$. Note ξ, f_{j_1,j_2} and ρ are locally bounded by Condition 3c, 4b and 6a, respectively. When $j_1 \neq j_2$ and $l_1 \neq l_2$,

$$[Cov(\phi_{ij_1,l_1}, \phi_{ij_2,l_2})]_{k_1,k_2}$$

$$= \frac{1}{h^4} E\left(1(\tau_{ij_1}, \tau_{ij_2} \leq T_i \leq C_i) X_{ij_1 k_1} X_{ij_2 k_2} \right.$$

$$\left. \times K\left(\frac{\tau_{ij_1} - t_{l_1}}{h}, \frac{T_i - \tau_{ij_1} - s_{l_1}}{h}\right) K\left(\frac{\tau_{ij_2} - t_{l_2}}{h}, \frac{T_i - \tau_{ij_2} - s_{l_2}}{h}\right) \varepsilon_{ij_1} \varepsilon_{ij_2} \right)$$

$$= \frac{1}{h} \iiint 1(s_{l_1} + hz \geq 0) 1(t_{l_1} + s_{l_1} - t_{l_2} + hx + hz - hy \geq 0)$$

$$\times \xi_{j_1 j_2, k_1 k_2}(t_{l_1} + hx, t_{l_2} + hy, s_{l_1} + hz)K(x,z)K(y, x+z-y+(t_{l_1} + s_{l_1} - t_{l_2} - s_{l_2})/h)$$

$$\times \rho(t_{l_1} + hx, t_{l_2} + hy)f_{j_1,j_2}(t_{l_1} + hx, t_{l_2} + hy, s_{l_1} + hz)dxdydz$$

$$= O(1/h).$$

For terms with $\boldsymbol{\psi}$,

$$[Cov(\boldsymbol{\psi}_{ij_1,l_1}, \boldsymbol{\psi}_{ij_2,l_2})]^2_{k_1,k_2}$$

$$\leq [Var(\boldsymbol{\psi}_{ij_1,l_1})]_{k_1,k_1}[Var(\boldsymbol{\psi}_{ij_2,l_2})]_{k_2,k_2},$$

and we have,

$$[Var(\boldsymbol{\psi}_{ij,l})]_{k,k}$$

$$\leq \frac{1}{h^4} \sum_{k_1=1}^{p} \sum_{k_2=1}^{p} E\Bigg(1(\tau_{ij} \leq T_i \leq C_i) X_{ijk}^2 K^2 \left(\frac{\tau_{ij} - t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h}\right)$$

$$\times X_{ijk_1} X_{ijk_2} [\beta_{k_1}(\tau_{ij}, T_i - \tau_{ij}) - \beta_{k_1}(t_l, s_l)][\beta_{k_2}(\tau_{ij}, T_i - \tau_{ij}) - \beta_{k_2}(t_l, s_l)]\Bigg)$$

$$= \frac{1}{h^2} \sum_{k_1=1}^{p} \sum_{k_2=1}^{p} \iint 1(s_l + hy \geq 0) \zeta_{j,kkk_1k_2}(t_l + hx, s_l + hy) K^2(x, y)$$

$$\times \left[hx\beta_{k_1}^{(1,0)}(t_l + \theta_1' hx, s_l + \theta_1' hy) + hy\beta_{k_1}^{(0,1)}(t_l + \theta_1' hx, s_l + \theta_1' hy)\right]$$

$$\times \left[hx\beta_{k_2}^{(1,0)}(t_l + \theta_2' hx, s_l + \theta_2' hy) + hy\beta_{k_2}^{(0,1)}(t_l + \theta_2' hx, s_l + \theta_2' hy)\right] dxdy$$

$$= O(1).$$

Here Condition 3b, 4a and 5 are needed for DCT. In summary, we have

$$Var(Z_{n,i}(\boldsymbol{a})) = \frac{\mu_0}{h^2} \sum_{l=1}^{d} \sum_{j=1}^{m} \boldsymbol{a}_l^T \boldsymbol{\eta}_j(t_l, s_l) \sigma^2(t_l) f_j(t_l, s_l) \boldsymbol{a}_l + o\left(\frac{1}{h^2}\right).$$

Proof of (2.14):

$$E|Z_{n,i}(\boldsymbol{a}) - EZ_{n,i}(\boldsymbol{\alpha})|^{2+\delta}$$

$$\leq E\left(|Z_{n,1}(\boldsymbol{a})| + |EZ_{n,1}(\boldsymbol{a})|\right)^{2+\delta}$$

$$\lesssim (E|Z_{n,i}(\boldsymbol{a})|^{2+\delta} + |EZ_{n,i}(\boldsymbol{a})|^{2+\delta}).$$

The notation $\lesssim$ stands for less than up to a constant factor. The last inequality is from Jensen's Inequality. We already know $EZ_{n,i}(\boldsymbol{a}) = O(h^2)$ from the proof of (2.12). Further

we have

$$E|Z_{n,i}(\boldsymbol{a})|^{2+\delta} = E\left|\sum_{l=1}^{d}\sum_{j=1}^{m}\boldsymbol{a}_l^T(\boldsymbol{\phi}_{ij,l} + \boldsymbol{\psi}_{ij,l})\right|^{2+\delta}$$

$$\lesssim \sum_{l=1}^{d}\sum_{j=1}^{m}\left(E\left|\boldsymbol{a}_l^T\boldsymbol{\phi}_{ij,l}\right|^{2+\delta} + E\left|\boldsymbol{a}_l^T\boldsymbol{\psi}_{ij,l}\right|^{2+\delta}\right).$$

By similar calculations as before we can show

$$E\left|\boldsymbol{a}_l^T\boldsymbol{\phi}_{ij,l}\right|^{2+\delta} = h^{-4-2\delta}E\left|1(\tau_{ij} \leq T_i \leq C_i)\boldsymbol{a}_l^T\mathbf{X}_{ij}K\left(\frac{\tau_{ij}-t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h}\right)\varepsilon_{ij}\right|^{2+\delta}$$

$$= O(h^{-2-2\delta}),$$

$$E\left|\boldsymbol{a}_l^T\boldsymbol{\psi}_{ij,l}\right|^{2+\delta} = h^{-4-2\delta}E\left|1(\tau_{ij} \leq T_i \leq C_i)\boldsymbol{a}_l^T\mathbf{X}_{ij}K\left(\frac{\tau_{ij}-t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h}\right)\right.$$

$$\left.\times \sum_{k=1}^{p}X_{ijk}[\beta_k(\tau_{ij}, T_i - \tau_{ij}) - \beta_k(t_l, s_l)]\right|^{2+\delta}$$

$$\lesssim h^{-4-2\delta}\sum_{k=1}^{p}E\left|1(\tau_{ij} \leq T_i \leq C_i)\boldsymbol{a}_l^T\mathbf{X}_{ij}K\left(\frac{\tau_{ij}-t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h}\right)X_{ijk}\beta_k(\tau_{ij}, T_i - \tau_{ij})\right|^{2+\delta}$$

$$+ h^{-4-2\delta}\sum_{k=1}^{p}E\left|1(\tau_{ij} \leq T_i \leq C_i)\boldsymbol{a}_l^T\mathbf{X}_{ij}K\left(\frac{\tau_{ij}-t_l}{h}, \frac{T_i - \tau_{ij} - s_l}{h}\right)X_{ijk}\beta_k(t_l, s_l)\right|^{2+\delta}$$

$$= O(h^{-2-2\delta}),$$

for any $\delta \leq 2$. Condition 3b and 6b are needed to ensure the convergence under DCT. Now the proof of Theorem 3.1 in complete. $\square$

2.6.4   Proof of Theorem 2.2

By (2.11), we just need to show

$$\frac{h^2}{n}\sum_{i\in\mathcal{D}}\mathbf{X}_i^T\mathbf{K}_i\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i^T\mathbf{K}_i\mathbf{X}_i \to_p \sum_{j=1}^{m}\boldsymbol{\eta}_j(t_0, s_0)f_j(t_0, s_0)\sigma^2(t_0)\mu_0.$$

Notice that

$$
\left[\frac{h^2}{n} \sum_{i \in \mathcal{D}} \mathbf{X}_i^T \mathbf{K}_i \widehat{\boldsymbol{\varepsilon}}_i \widehat{\boldsymbol{\varepsilon}}_i^T \mathbf{K}_i \mathbf{X}_i \right]_{kl}
$$

$$
= \frac{h^2}{n} \sum_{i=1}^{n} \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} 1(\tau_{ij_1} \vee \tau_{ij_2} \le T_i \le C_i) X_{ik}(\tau_{ij_1}) X_{il}(\tau_{ij_2}) K_{ij_1} K_{ij_2},
$$

$$
\times \left\{ X_i(\tau_{ij_1})^T \left[\boldsymbol{\beta}(\tau_{ij_1}, T_i - \tau_{ij_1}) - \widehat{\boldsymbol{\beta}}(\tau_{ij_1}, T_i - \tau_{ij_1}) \right] + \varepsilon_i(\tau_{ij_1}) \right\}
$$

$$
\times \left\{ X_i(\tau_{ij_2})^T \left[\boldsymbol{\beta}(\tau_{ij_2}, T_i - \tau_{ij_2}) - \widehat{\boldsymbol{\beta}}(\tau_{ij_2}, T_i - \tau_{ij_2}) \right] + \varepsilon_i(\tau_{ij_2}) \right\},
$$

where $K_{ij} = h^{-2} K((\tau_{ij} - t_0)/h, (T_i - \tau_{ij} - s_0)/h)$. By the similar mean and variance calculations as in the proof of Theorem 2.1 following Lemma 2.1, under Conditions 3b, 3c and 6b we have

$$
\frac{h^2}{n} \sum_{i=1}^{n} \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} |X_{ik}(\tau_{ij_1}) X_{il}(\tau_{ij_2})| \| X_i(\tau_{ij_1}) X_i(\tau_{ij_2})^T \|_F K_{ij_1} K_{ij_2} = O_p(1),
$$

$$
\frac{h^2}{n} \sum_{i=1}^{n} \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} |X_{ik}(\tau_{ij_1}) X_{il}(\tau_{ij_2}) \varepsilon_i(\tau_{ij_2})| \| X_i(\tau_{ij_1}) \|_2 K_{ij_1} K_{ij_2} = O_p(1), \text{ and}
$$

$$
\frac{h^2}{n} \sum_{i=1}^{n} \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} 1(\tau_{ij_1} \vee \tau_{ij_2} \le T_i \le C_i) X_{ik}(\tau_{ij_1}) X_{il}(\tau_{ij_2}) \varepsilon_i(\tau_{ij_1}) \varepsilon_i(\tau_{ij_2}) K_{ij_1} K_{ij_2}
$$

$$
\to_p \sum_{j=1}^{m} \eta_{j,kl}(t_0, s_0) f_j(t_0, s_0) \sigma^2(t_0) \mu_0,
$$

where $\| \cdot \|_F$ denotes Frobenious norm. Thus it suffices to show

$$
\sup_{(t,s) \in \mathcal{S}_0(h)} \left\| \boldsymbol{\beta}(t, s; h) - \widehat{\boldsymbol{\beta}}(t, s; h) \right\|_2 \to_p 0, \tag{2.15}
$$

where

$$
\mathcal{S}_0(h) = \left\{ (t, s) : \left(\frac{t - t_0}{h}, \frac{s - s_0}{h} \right) \in \operatorname{supp} K \right\},
$$

and we divide the proof of (2.15) into four parts following Lemma 2.2.

(1) First we show

$$\sup_{(t,s)\in\mathcal{S}_0(h)} |A_n(t,s;h)_{kl} - A(t,s)_{kl}| \to_p 0 \qquad (2.16)$$

for some deterministic matrix $A(t,s)$. Consider the following class of functions

$$\mathcal{F}_{kl} = \left\{ \sum_{j=1}^m 1(\tau_j \leq T \leq C) X_k(\tau_j) X_l(\tau_j) K\left(\frac{\tau_j - t}{h^*}, \frac{T - \tau_j - s}{h^*}\right) : \right.$$
$$\left. h^* \in (0, \varepsilon^*), (t,s) \in \mathcal{S}_0(h^*) \right\},$$

where ε^* is defined below. This is equivalent to the following class of functions based on the fact that K is defined on a bounded support:

$$\mathcal{F}_{kl} = \left\{ \sum_{j=1}^m 1(\tau_j \leq T \leq C) X_k(\tau_j) X_l(\tau_j) K\left(\frac{\tau_j - t}{h^*}, \frac{T - \tau_j - s}{h^*}\right) \right.$$
$$\left. \cdot 1[(\tau_j, T - \tau_j) \in B_0(2\varepsilon^* R(K))] : h^* \in (0, \varepsilon^*), (t,s) \in \mathcal{S}_0(h^*) \right\},$$

with envelope function

$$F_{kl} = \sum_{j=1}^m |X_k(\tau_j) X_l(\tau_j)| \cdot \|K\|_\infty \cdot 1[(\tau_j, T - \tau_j) \in B_0(2\varepsilon^* R(K))],$$

where $R(K) = \sup_{\text{supp } K} \|\mathbf{x}\|_2$, $B_0(\varepsilon) = \{(t,s) : (t - t_0)^2 + (s - s_0)^2 \leq \varepsilon^2\}$, and ε^* is chosen such that $EF_{kl}^r < \infty$ for some $r > 1$. By Condition 3b, this can be done for any $r \in (1, 4]$. Next we take two steps to show that $\mathcal{F}_{kl}$ is Euclidean. First, notice that the class of kernel functions

$$\left\{ K\left(\frac{\tau_j - t}{h^*}, \frac{T - \tau_j - s}{h^*}\right) : h^* \in (0, \varepsilon^*), (t,s) \in \mathcal{S}_0(h^*) \right\}$$

is Euclidean under our Condition 1 by Lemma 22(i) in [37]. Second, by Lemma 17 in [37] and our Lemma 2.3, we know $\mathcal{F}_{kl}$ is Euclidean for F_{kl}. Define $\mathcal{G}^+$ as $\{g^+ : g \in \mathcal{G}\}$, where $g^+ = \max(g, 0)$. Then it is easy to show that $\mathcal{F}_{kl}^+$ is also Euclidean for F_{kl}.

Now we apply equation (2) in [40] for the class $\mathcal{F}_{kl}^+$ with $\gamma = h^2 \varepsilon$ and $\kappa(x) = x^r$ following his notation, which gives

$$\sup_{\mathcal{F}_{kl}^+} \frac{|P_n f - P f|}{\varepsilon(P_n f + P f) + \varepsilon h^2 (P_n F_{kl} + P F_{kl} + 1)} \leq 26 \tag{2.17}$$

with probability at least

$$1 - 32A(h^2\varepsilon)^{-V} \exp(-\varepsilon^2 h^2 n^{1-1/r}) - 4P F_{kl}^r 1(F_{kl}^r > n\varepsilon), \tag{2.18}$$

where P_n is the empirical measure of a random sample generated from the fixed probability measure P, and A and V are the Euclidean constants for F_{kl}. By some algebra, we can show $32A(h^2\varepsilon)^{-V} \exp(-\varepsilon^2 h^2 n^{1-1/r})$ converges to 0 when $h^2 n^{1-1/r}/\log n \to \infty$, which is guaranteed by Condition 2c. Further we know $P F_{kl}^r 1(F_{kl}^r > n\varepsilon) \to 0$ by DCT. These allow us to have (2.17) with arbitrarily large probability when n is large enough. Next we replace $P_n F_{kl}$ in the denominator of (2.17) by its upper bound. Since $P_n F_{kl} \to_p P F_{kl}$ by weak law of large numbers, we have $|P_n F_{kl} - P F_{kl}| < P F_{kl}$ with arbitrarily large probability when n is large enough. Thus (2.17) implies

$$\sup_{\mathcal{F}_{kl}^+} \frac{|P_n f - P f|}{\varepsilon(P_n f + P f) + \varepsilon h^2 (3 P F_{kl} + 1)} \leq 26 \tag{2.19}$$

with arbitrarily large probability when n is large enough.

Consider another class of functions:

$$\mathcal{F}_{kl,n} = \left\{ \sum_{j=1}^m 1(\tau_j \leq T \leq C) X_k(\tau_j) X_l(\tau_j) K\left(\frac{\tau_j - t}{h}, \frac{T - \tau_j - s}{h}\right) : (t, s) \in \mathcal{S}_0(h) \right\},$$

42

which is a subset of $\mathcal{F}_{kl}$ when n is large enough. By (2.19) we have with arbitrarily large probability,

$$\sup_{\mathcal{F}_{kl,n}^+} |h^{-2}P_n f - h^{-2}Pf| \leq \frac{26\varepsilon}{1 - 26\varepsilon}(\sup_{\mathcal{F}_{kl,n}^+} h^{-2}Pf + 3PF_{kl} + 1).$$

We just further need to show $\sup_{\mathcal{F}_{kl,n}^+} h^{-2}Pf = O(1)$ to conclude $\sup_{\mathcal{F}_{kl,n}^+} |h^{-2}P_n f - h^{-2}Pf| \to_p$ 0. When n is large enough, we have

$$\sup_{\mathcal{F}_{kl,n}^+} h^{-2}Pf$$

$$\leq \sup_{\mathcal{F}_{kl,n}} h^{-2}P|f|$$

$$= \sup_{(t,s)\in\mathcal{S}_0(h)} h^{-2}E \left| \sum_{j=1}^m 1(\tau_j \leq T \leq C)X_k(\tau_j)X_l(\tau_j)K\left(\frac{\tau_j - t}{h}, \frac{T - \tau_j - s}{h}\right) \right|$$

$$\leq \sum_{j=1}^m \sup_{(t,s)\in\mathcal{S}_0(h)} h^{-2}E \left[|X_k(\tau_j)X_l(\tau_j)| K\left(\frac{\tau_j - t}{h}, \frac{T - \tau_j - s}{h}\right) \right]$$

$$\leq \sum_{j=1}^m \int \sup_{(t,s)\in\mathcal{S}_0(h)} \phi(hx + t, hy + s; j, k, l)K(x, y)f_j(hx + t, hy + s)dxdy$$

$$= O(1)$$

by Condition 3b, here $\phi(x, y; j, k, l) = E\left[|X_k(\tau_j)X_l(\tau_j)| \big| \tau_j = x, T = x + y \right]$.

Similarly we can show $\sup_{\mathcal{F}_{kl,n}^-} |h^{-2}P_n f - h^{-2}Pf| \to_p 0$. Combining the two gives us

$$\sup_{(t,s)\in\mathcal{S}_0(h)} |A_n(t, s; h)_{kl} - EA_n(t, s; h)_{kl}|$$

$$= \sup_{\mathcal{F}_{kl,n}} \left| h^{-2}P_n f^+ - h^{-2}P_n f^- - h^{-2}Pf^+ + h^{-2}Pf^- \right|$$

$$\leq \sup_{\mathcal{F}_{kl,n}} \left| h^{-2}P_n f^+ - h^{-2}Pf^+ \right| + \sup_{\mathcal{F}_{kl,n}} \left| h^{-2}P_n f^- - h^{-2}Pf^- \right| \to_p 0.$$

What remains is to show

$$\sup_{(t,s)\in\mathcal{S}_0(h)} |EA_n(t,s;h)_{kl} - A(t,s)_{kl}| \to 0.$$

With $A(t,s) = \sum_{j=1}^m \boldsymbol{\eta}_j(t,s) f_j(t,s)$, when h is small enough, we have

$$\sup_{(t,s)\in\mathcal{S}_0(h)} |EA_n(t,s;h)_{kl} - A(t,s)_{kl}|$$

$$= \sup_{(t,s)\in\mathcal{S}_0(h)} \left| \sum_{j=1}^m \iint \eta_{j,kl}(hx+t, hy+s) K(x,y) f_j(hx+t, hy+s) dx dy \right.$$

$$\left. - \sum_{j=1}^m \iint \eta_{j,kl}(t,s) K(x,y) f_j(t,s) dx dy \right|$$

$$\leq \sum_{j=1}^m \iint K(x,y) \sup_{(t,s)\in\mathcal{S}_0(h)} \left| \eta_{j,kl}(hx+t, hy+s) f_j(hx+t, hy+s) \right.$$

$$\left. - \eta_{j,kl}(t,s) f_j(t,s) \right| dx dy$$

$$\to 0,$$

where the convergence holds because when n is large enough and $K(x,y) > 0$, both (t,s) and $(hx+t, hy+s)$ are within a neighborhood of (t_0, s_0) where η and f_j are continuous, then by the Heine–Cantor theorem we can bound $\eta_{j,kl}(hx+t, hy+s) f_j(hx+t, hy+s) - \eta_{j,kl}(t,s) f_j(t,s)$ by any arbitrary number.

(2) In this part we want to show

$$\sup_{(t,s)\in\mathcal{S}_0(h)} \|B_n(t,s;h) - B(t,s)\|_2 \to_p 0,$$

where

$$B(t,s) = \left[\sum_{j=1}^m \boldsymbol{\eta}_j(t,s) f_j(t,s) \right] \boldsymbol{\beta}(t,s).$$

44

Similarly as in part (1), we can show

$$\sup_{(t,s)\in\mathcal{S}_0(h)} \|B_n(t,s;h) - EB_n(t,s;h)\|_2 \to_p 0.$$

We omit the details to avoid duplications. Further by noticing that

$$EB_n(t,s;h) - B(t,s) = EA_n(t,s;h)\boldsymbol{\beta}(t,s) + ER_n(t,s;h) - A(t,s)\boldsymbol{\beta}(t,s),$$

we only need to show

$$\sup_{(t,s)\in\mathcal{S}_0(h)} \|ER_n(t,s;h)\|_2 \to 0,$$

which holds because for any k,

$$\sup_{(t,s)\in\mathcal{S}_0(h)} |ER_n(t,s;h)_k|$$

$$= \sup_{(t,s)\in\mathcal{S}_0(h)} \left| \sum_{j=1}^{m}\sum_{l=1}^{p} \iint 1(hy+s\geq 0)\eta_{j,kl}(hx+t,hy+s)K(x,y) \right.$$

$$\left. \times \left[\beta_l(hx+t,hy+s) - \beta_l(t,s)\right] f_j(hx+t,hy+s)dxdy \right|$$

$$\leq \sum_{j=1}^{m}\sum_{l=1}^{p} \iint K(x,y) \sup_{(t,s)\in\mathcal{S}_0(h)} \left[|\beta_l(hx+t,hy+s) - \beta_l(t,s)|\right.$$

$$\left. |\eta_{j,kl}(hx+t,hy+s)|f_j(hx+t,hy+s)\right]dxdy$$

$$\to 0,$$

where the convergence holds under Conditions 3a, 4a and 5.

(3) Under Condition 7, we know $\lambda_1(A(t_0,s_0))$ is strictly positive. Furthermore, under Condition 3a and 4a, $\lambda_1(A(\cdot,\cdot)) > 0$ is bounded away from 0 in a neighborhood of (t_0,s_0). Thus

we have

$$\liminf_n \inf_{(t,s)\in\mathcal{S}_0(h)} \lambda_1(A(t,s)) > 0.$$

(4) Whenever $\|B(t_0, s_0)\|_2 < \infty$, we have

$$\limsup_n \sup_{(t,s)\in\mathcal{S}_0(h)} \|B(t,s)\|_2 < \infty,$$

by Condition 3a, 4a and 5. Now the proof is complete. $\qquad\square$

2.6.5 Covariate Effects in the Full Model for ESRD Medicare Data Analysis

Figure 2.4 depicts the coefficient estimates in the the full model for the ESRD Medicare data analysis using a bandwidth of 10 days (cross-validated and undersmoothed):

$$\log(Y_{ij}/1000 + 1)$$
$$= \beta_1(\tau_{ij}, T_i - \tau_{ij}) + \beta_2(\tau_{ij}, T_i - \tau_{ij})\mathrm{Payer}_{ij} + \beta_3(\tau_{ij}, T_i - \tau_{ij})\mathrm{Race}_i$$
$$+ \beta_4(\tau_{ij}, T_i - \tau_{ij})\mathrm{Diabetes}_i + \beta_5(\tau_{ij}, T_i - \tau_{ij})\mathrm{Heart}_i + \varepsilon_i(\tau_{ij}).$$

Confidence Bands in Figure 2.4 are constructed following [54] via family-wise error control based on the asymptotic multivariate normal distribution given in Theorem 3.1 at 100 evenly spaced grid points for each curve, with a simplification where the gaps between grid points are bridged via linear interpolation. Such a modification yields wider than pointwise but still anti-conservative confidence bands comparing to global confidence bands. From Figure 2.4 we see that the estimated coefficients for the three baseline covariates (race, diabetes, and heart disease) are close to zero in magnitude and their simplified family-wise error

controlled confidence bands cover 0 everywhere. In other words, these coefficients are neither scientifically nor statistically significant. Thus we focus on the first two coefficients in our final analysis that is given in the main text.

2.6.6 Simulation for the Locally Weighted Pseudo Likelihood Approach under Working Independence

In this section, we adopt a simulation design similar to that in Section 2.4 with a smaller sample size of $n = 400$ to compare the pseudo likelihood approach described in Section 2.2.4 and the complete case estimator (2.3). In order to implement the pseudo likelihood approach, we generate conditionally independent terminal event time T and censoring time C from exponential distributions with rates $\exp(3X_{i2}+X_{i3}(0)-5)$ and $\exp(X_{i2}+3X_{i3}(0)-5)$, respectively, which are then truncated at 15 and added 5. Dropping unrelated factors, we have the following pseudo likelihood:

$$
\prod_{i=1}^{n}\left\{\prod_{j=1}^{m_i}\left[\frac{1}{\sqrt{2\pi}\sigma_{ij}}e^{-\frac{1}{2\sigma_{ij}^2}(Y_{ij}-X_{ij}^T b)^2}\right]^{K(\frac{\tau_{ij}-t}{h},\frac{T_i-\tau_{ij}-s}{h})}\right\}^{\Delta_i}
$$

$$
\times\left\{\int_{C_i}^{\infty}\prod_{j=1}^{m_i}\left[\frac{1}{\sqrt{2\pi}\sigma_{ij}}e^{-\frac{1}{2\sigma_{ij}^2}(Y_{ij}-X_{ij}^T b)^2}\right]^{K(\frac{\tau_{ij}-t}{h},\frac{u-\tau_{ij}-s}{h})}dP(T_i\leq u|Z_i)\right\}^{1-\Delta_i}.
$$

The pseudo likelihood estimating procedure involves two steps: (1) Estimating nuisance parameters $P(T_i \leq u|Z_i)$ and σ_{ij}. In this simulation study, we estimate $P(T_i \leq u|Z_i)$ by using stratified Kaplan-Meier estimates. Specifically, we first group subjects with similar baseline covariates $Z = (X_2, X_3(0))$ into 10 clusters using a k-means clustering, and then estimate the conditional survival distribution within each cluster. We estimate σ_{ij} by the kernel estimator $\widehat{\sigma}_{ij}^2 = (nh)^{-1}\sum_{i'j'}\widehat{\varepsilon}_{i'j'}^2 K((\tau_{i'j'} - \tau_{ij})/h)$, where h is the same bandwidth used for estimating β, and K is the standard Gaussian kernel truncated at ± 2. Then put

47

the estimates back into above pseudo likelihood to replace their corresponding unknown parameters. (2) Maximizing the log pseudo likelihood. We first find the complete case estimator, then use it as the initial point to run the Robbins-Monro algorithm for 1000 iterations with initial stepsize 0.01 to maximize the logarithm of the above pseudo likelihood functions with respect to b.

Figure 2.5 shows the comparison between the complete case estimator and the estimator obtained from maximizing the pseudo likelihood. The solid curve represents the true coefficient values. The red middle, upper and lower curves are the mean and variability bands of the complete case estimator; The blue middle, upper and lower curves are the mean and 95% empirical confidence bands of the pseudo likelihood estimator. Notice the complete case estimator still works reasonably well at most time points for a much smaller sample size. In all plots, the confidence bands of the pseudo likelihood estimators are narrower, indicating improved efficiency by including censored subjects.
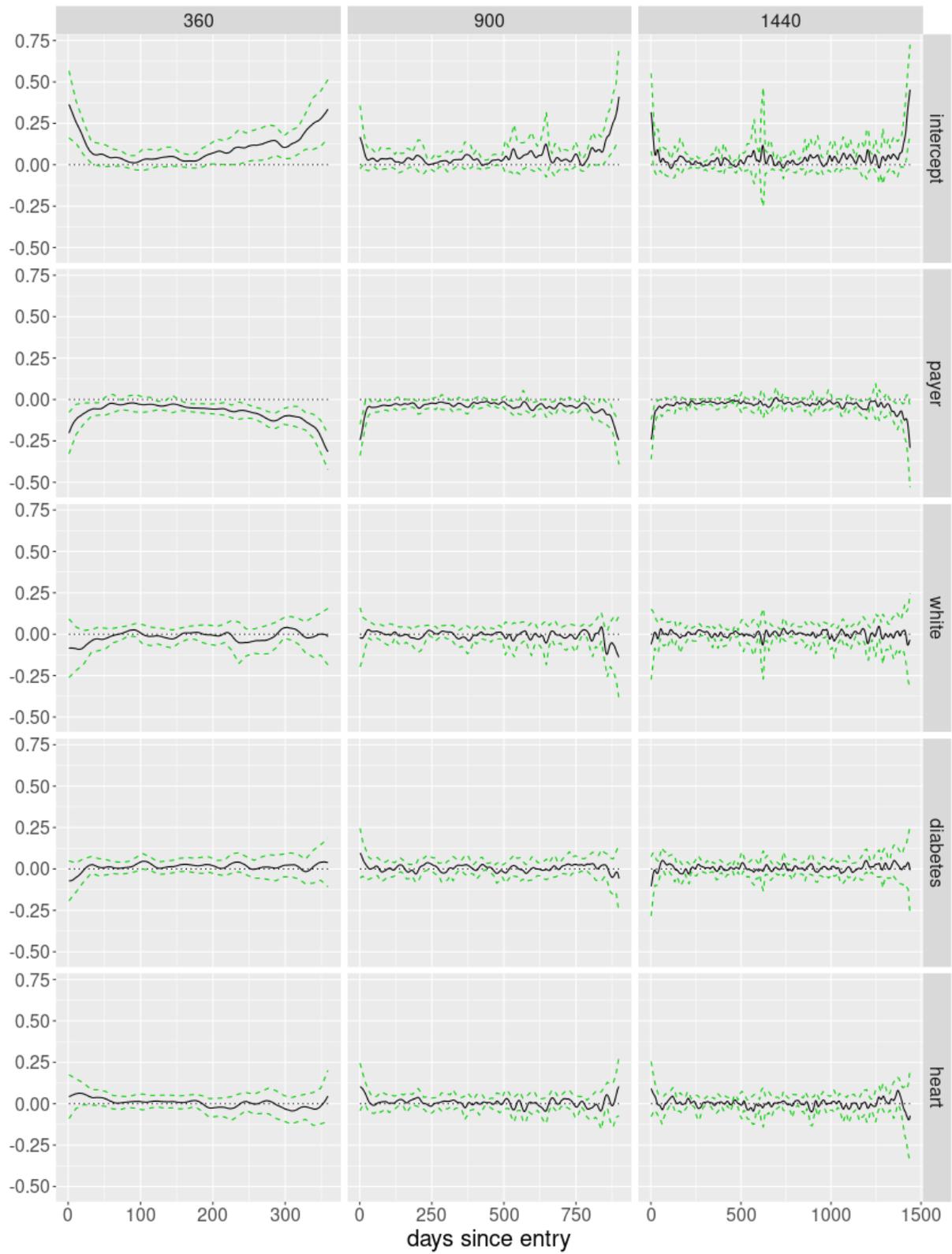
Figure 2.4: Coefficient estimates and their family-wise error controlled confidence bands for $T = 360, 900, 1440$ days.
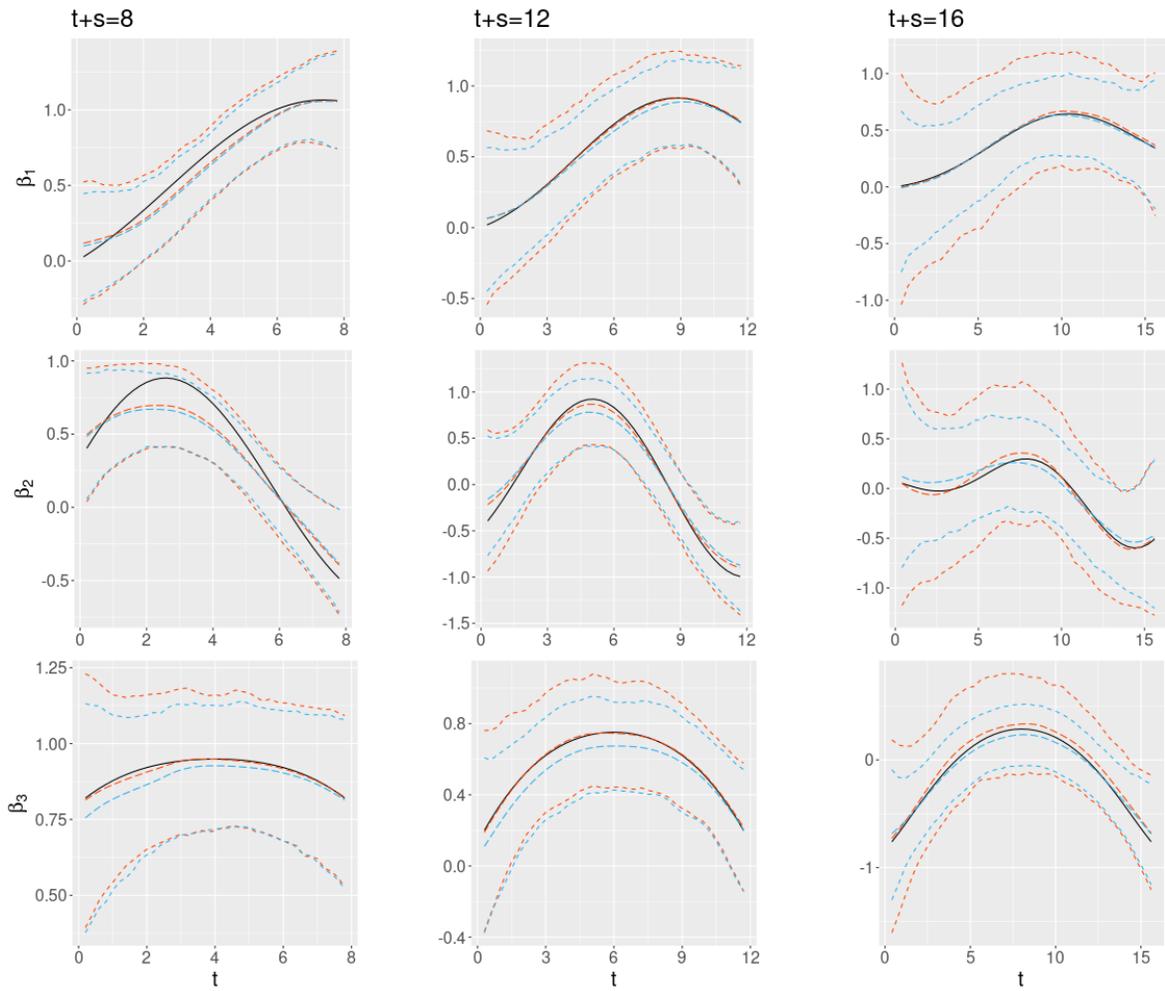
Figure 2.5: Mean estimates and empirical confidence bands at $T = 8, 12, 16$. Black: true coefficient. Red: complete case estimator; Blue: pseudo likelihood estimator.

Chapter 3

Bivariate Functional Patterns of Lifetime Medicare Costs among ESRD Patients

3.1 Introduction

End-stage renal disease (ESRD), also called end-stage kidney disease, is an advanced state of chronic kidney disease in which a patient's kidneys permanently stop functioning and the patient requires long-term dialysis or a kidney transplant to maintain life. ESRD has become increasingly prevalent in the United States. Over the last two decades in particular, the number of individuals with prevalent ESRD reached 808,330 in 2019, an increase of 107% from 2000 [47]. Compared to long-term dialysis, the survival benefit of kidney transplant has long been established [52, 33, 44]. The annual count of kidney transplants performed in patients undergoing dialysis has reached 24,511 in 2019, a 36.9% increase since 2010, and the number of newly waitlisted ESRD patients for kidney transplantation reached 28,553

in 2019. Despite an increase of transplantation, the percentage of prevalent ESRD patients with a kidney transplant remained unchanged at approximately 30%. Among the prevalent dialysis patients, 13.1% are still on the transplant waiting list by the end of 2019. With the high prevalence of ESRD patients, it is of tremendous interest to understand the medical spending patterns of patients who are on the waiting list and who have received transplants compared to unwaitlisted long-term dialysis patients.

It has long been observed that patients on the waiting list tend to have better health status compared to unwaitlisted patients. For example, [53] observed that the standardized mortality ratios for subgroups of the waitlisted patients are 38-58 percent lower than unwaitlisted patients. To be on the waitlist, an ESRD patient needs to obtain a referral to a local transplant center, where the patient will be evaluated by the transplant team for candidacy. Once becoming a candidate, the patient will be added to the waiting list. As of end of year 2022, there are 235 transplant centers in the United States. Although substantial variations in the practices of transplant evaluation have been reported [51], some common factors are considered across different transplant centers. For example, [32] found that 22% of the 59 transplant programs they surveyed have an absolute age cutoff in the range of 70 to 79. According to [34], 38% of the 202 programs they surveyed performed frailty assessments for candidacy. The evaluation practices based on these factors create disparities between patients on the waiting list and those who are not. While the existing literature on the waiting list focuses on the survival disparity, there is few literature that studies the economic implication of waitlisting. One of the goals of this paper is to study potential medical cost differences between unwaitlisted and waitlisted patients who have the same lifespan.

The survival and economic benefit of transplant itself has been extensively studied. A standard measure used in economic evaluation to compare healthcare interventions is the incremental cost-effectiveness ratio (ICER) [8]. Numerous articles have proposed ICER-based methods to study the cost-effectiveness of kidney transplantation compared to dialysis. See

[15] for a systematic review. The ICER is then compared to a willingness-to-pay threshold, for example, \$100K [35, 31, 3]. The determination of ICER involves calculating the cumulative medical cost during the remaining lifetime, which is a single summary number to measure the cost of transplant and dialysis. Using a new proposed method, we examine and compare the detailed medical cost trajectories of transplanted ESRD patients during their lifespan, and a new analysis approach is proposed for that matter.

Two analyses are considered. The first analysis is to compare overall medical costs between dialysis patients on the waiting list and those unwaitlisted, and the second analysis is between patients on the waiting list and those who later received kidney transplantation. The cost differences between the comparison groups are estimated using extensions of the bivariate time-varying coefficient model [50], with adjustment for demographics and comorbidities of ESRD patients.

3.2   ESRD Medicare Claims Data

In this paper we use the institutional Medicare claims data prepared by the United States Renal Data System (USRDS). The claims data covers 5 types of medical services: inpatient care, outpatient care, skilled nursing facility care, home health care and hospice care. We consider ESRD patients who started their first ESRD service in a 5-year window from the beginning of 2007 to the end of 2011 and were at least 65 years old at the first service. The follow-up ends at the end of 2017 and any patients still alive by then are excluded as death information is needed for the analysis. This type of analysis is commonly referred to as the complete case analysis and its validity under a longitudinal model conditional on survival time has been established by [24] and [50]. We only retain patients that had at most one kidney transplant during follow-up. As we will see later, restricting to patients with at most one transplant helps simplify the model interpretation and the estimation procedure.

According to [46], the claims database contains data for both Medicare as Primary Payer (MPP) and Medicare as Secondary Payer (MSP) patients, but it is impossible to determine the complete cost of care for ESRD patients with MSP coverage. For this reason, we only retain the days when patients were on MPP coverage. According to [47], Medicare ESRD expenditures for specific events or services (e.g., hospitalizations, medications) include only beneficiaries covered by traditional fee-for-service Medicare but not Medicare Advantage (MA) plans. For that reason, we cannot obtain the accurate individual spending amount for ESRD patients on MA plans and they are thus excluded. Patients with no BMI records and patients with missing gender information were also excluded for data completeness.

By applying all the aforementioned criteria to the USRDS dataset, we obtain 183157 patients. Among them, 175482 were never on the waiting list and never received transplant, 4933 were on the waiting list but never received transplant, and 2557 were placed on the waiting list and later received a transplant. The remaining 185 patients who were never on the waiting list but received a transplant may have different characteristics, thus are excluded from further analysis. For brevity, these three patient groups of interest are referred to respectively as unwaitlisted group, waitlisted (but not transplanted) group and transplanted group. In order to reduce the computing cost, we randomly sample 5000 patients from the unwaitlisted group, which also creates a balanced sample size comparing to the waitlisted group for the final analysis. Furthermore, for each patient we randomly select 10% of the days without replacement during the follow-up period and calculate the daily cost on the selected days. Patients with less than 10 selected days are excluded from the analysis, which leads to the final respective sample sizes of $n_1 = 4962$, $n_2 = 4918$ and $n_3 = 2548$ patients in the unwaitlisted, waitlisted and transplanted groups.

The daily cost of a specific day is calculated by summing up the average daily cost of each claim period that includes that day. For example, if one claim is from May 1st to 20th with a total of $1000 and another claim from May 11th to 30th with a total of $500, then the

daily cost of May 15th is \$75 (average daily cost of \$50 from the first claim and \$25 from the second). Daily cost variables include the total medical cost, which is the primary variable of interest, and five major medical cost components (inpatient, outpatient, skilled nursing care, home health care, and hospice care).

In addition to daily medical costs and death time, we also extract the following variables: demographic variables including age at the first ESRD service, sex and race; dialysis type for dialysis patients; health-related variables including an indicator variable for hypertension, an indicator variable for any of other comorbidities (congestive heart failure, atherosclerotic heart disease, other cardiac disease, cerebrovascular disease, peripheral vascular disease, diabetes, chronic obstructive pulmonary disease, cancer and toxic nephropathy) and BMI. The USRDS created a treatment history for all ESRD patients in the database. This history can be used to identify the treatment modality for any patient on a single day. The detailed and general treatment modality categories used by the USRDS are listed in Table 3.1. For our analysis, we use the general categories and create a binary variable that equals 1 if a patient is on peritoneal dialysis (PD) and 0 if on hemodialysis (HD). The observations on other treatment modalities other than transplant are excluded in the analysis. Dialysis type is the only time-varying covariate in our models.

We separate hypertension from the comorbidities because hypertension is very common among ESRD patients (87.3% among selected patients) and combining hypertension with other comorbidities will make the event almost always positive (96.7% among selected patients). Becasue 96.3% of the patients in the analysis dataset only have one BMI measure available and most of them (95.6%) are recorded at the baseline (the first time of receiving the ESRD service), we use the baseline BMI as a time-invariant covariate. For patients who do not have a baseline BMI measure, this variable is created by averaging all BMI values recorded in the follow-up period. We believe this causes minimal impact to the analysis because only a very small percent of patients do not have baseline measures and the BMI is

Type	Detail	General
1	Center hemodialysis (HD)	HD
2	Center Self HD	HD
3	Home HD	HD
4	HD training	HD
5	Continuous Ambulatory Peritoneal Dialysis (CAPD)	PD
6	CAPD training	PD
7	Continuous Cycling Peritoneal Dialysis (CCPD)	PD
8	CCPD training	PD
9	Other PD	PD
A	Uncertain dialysis	Unknown Dialysis
B	Discontinued dialysis	Discontinued Dialysis
T	Functioning transplant	Transplant
X	Lost-to-follow-up	Lost-to-follow-up
Z	Recovered function	Recovered Function

Table 3.1: Treatment modality categories created by the USRDS.

sparsely recorded with stable values over time for those with multiple measures.

For all three groups of patients, summary statistics of all the considered covariates and daily costs by types are provided in Table 3.4 in the Appendix.

3.3 Methods

We propose two models for the aforementioned two comparisons in Section 1, respectively. Let $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ denote index sets of individuals in the corresponding unwaitlisted, waitlisted and transplanted groups.

3.3.1 Model for Waitlisting

The first model is for the comparison of the unwaitlisted group and the waitlisted group, so the transplanted patients are exluded from this analysis. Let $Y_i^w(\tau_{ij})$ represent the log transformed total daily claim amount in \$100 unit, plus constant 1 to avoid log(0), on day

τ_{ij} since the first ESRD service and let T_i denote the day of death for the i-th patient, $i \in \mathcal{G}_1 \cup \mathcal{G}_2$, $j = 1, \ldots, m_i$. Hence $T_i - \tau_{ij}$ is the residual lifetime of the i-th patient at time τ_{ij}. We have

$$
\begin{aligned}
Y_i^w(\tau_{ij}) &= \beta_1(\tau_{ij}, T_i - \tau_{ij}) + \beta_2(\tau_{ij}, T_i - \tau_{ij})\text{PD}_{ij} + \alpha_1\text{Waitlisted}_i + \alpha_2\tau_{ij} \times \text{Waitlisted}_i \\
&\quad + \alpha_3\text{African American}_i + \alpha_4\text{Other Race}_i + \alpha_5\text{Female}_i + \alpha_6\text{Age}_i \qquad (3.1) \\
&\quad + \alpha_7\text{Hypertension}_i + \alpha_8\text{Other Comorbidities}_i + \alpha_9 I(25 \leq \text{BMI}_i < 30) \\
&\quad + \alpha_{10} I(\text{BMI}_i \geq 30) + \varepsilon_i^w(\tau_{ij}, T_i - \tau_{ij}).
\end{aligned}
$$

The reference group for race is white and the reference group for sex is male. Waitlisted is a binary time-invariant variable, which equals 1 if the patient was ever placed on the waiting list. The superscript w on Y and ε is used to differentiate with the other model we will introduce later, indicating that this model is for evaluating the cost difference between waitlisted and unwaitlisted groups. Note that we allow the independent and identically distributed error terms ε_i^w to be stochastic processes of both the time since the first service and the residual lifetime, which is based on our observation in the exploratory data analysis. The variance function of ε_i^w is denoted as $\sigma^2(t, s) = E[\varepsilon_i^w(\tau_{ij}, T_i - \tau_{ij})^2 | \tau_{ij} = t, T_i - \tau_{ij} = s]$. Observations are independent between patients but can be correlated within a patient which is captured by the auto-correlation of ε_i^w.

We did an exploratory data analysis using a model similar to (3.2) but with all coefficients to be bivariate functions of $(\tau_{ij}, T_i - \tau_{ij})$ (thus no separate interaction between Waitlisted and τ_{ij}), which has the same mean structure as the model proposed by [50] for analyzing longitudinal medical cost data with a terminal event. The variance structure extends from [50] by allowing σ^2 to depend on both times. Following the same proofs with slight modifications on conditions about ε_i^w, one can show that the asymptotic results about the estimation

and statistical inference in [50] still hold. We found that the coefficient of waitlisted variable slightly increases with τ_{ij}. Furthermore, except the intercept, the coefficient of PD variable and the coefficient of waitlisted variable, all other coefficients are very stable over time with minimal fluctuations. Such an observation leads to model (3.2) where β_1 and β_2 are bivariate time-varying and all α_k, $k = 1, \ldots, 10$, are constant coefficients. A model like this is referred to as the semi-varying coefficient model [56]. Here we extend the original semi-varying coefficient model by allowing the time-varying coefficients to be bivariate functions. Conditioning on both times as well as the baseline age enables us to compare unwaitlisted and waitlisted patients with the same lifespan and model how impending death influences medical costs. It was discovered in several previous work [29, 24, 50] that medical cost in dialysis patients tend to increase when patients approach death. By allowing β_1 and β_2 to also vary with residual lifetime $T - \tau_j$, we anticipate that model (1) captures the medical spending pattern better than its counterpart with univariate time-varying coefficients.

For estimating the time-invariant coefficients α_k, we adopt the profile weighted least squares (PWLS) method [12, 43]. To illustrate the estimating procedure, we use generic notation in the following:

$$Y_i^w(\tau_{ij}) = \mathbf{X}_i^w(\tau_{ij})^T \boldsymbol{\beta}(\tau_{ij}, T_i - \tau_{ij}) + \mathbf{Z}_i^w(\tau_{ij})^T \boldsymbol{\alpha} + \varepsilon_i^w(\tau_{ij}, T_i - \tau_{ij}), \tag{3.2}$$

where $\mathbf{X}_i^w(\tau_{ij})$ and $\mathbf{Z}_i^w(\tau_{ij})$ are the covariate vectors for subject i at time τ_{ij}. The PWLS estimator of $\boldsymbol{\alpha}$ in the above model is given by

$$\widehat{\boldsymbol{\alpha}} = \left[\mathbf{Z}^{wT}(\mathbf{I} - \mathbf{S})^T \mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{Z}^w \right]^{-1} \mathbf{Z}^{wT}(\mathbf{I} - \mathbf{S})^T \mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{Y}^w, \tag{3.3}$$

where $\mathbf{Z}^w = (\mathbf{Z}_1^{wT}, ..., \mathbf{Z}_{n_1+n_2}^{wT})^T$ with $\mathbf{Z}_i^w = (\mathbf{Z}_i^w(\tau_{i1})^T, ..., \mathbf{Z}_i^w(\tau_{im_i})^T)^T$, $\mathbf{Y}^w$ and $\mathbf{X}^w$ are defined similarly to $\mathbf{Z}^w$, $\mathbf{I}$ is the identity matrix, $\mathbf{S}$ is the kernel smoother matrix and $\mathbf{W}$ is a weight

matrix. Specifically, $\mathbf{S}$ can be expressed as:

$$
\mathbf{S} = \begin{pmatrix} \mathbf{X}_1^w(\tau_{11})^T \left[\mathbf{X}^{wT}\mathbf{K}^w(\tau_{11}, T_1 - \tau_{11})\mathbf{X}^w \right]^{-1} \mathbf{X}^{wT}\mathbf{K}^w(\tau_{11}, T_1 - \tau_{11}) \\ \vdots \\ \mathbf{X}_2^w(\tau_{21})^T \left[\mathbf{X}^{wT}\mathbf{K}^w(\tau_{21}, T_2 - \tau_{21})\mathbf{X}^w \right]^{-1} \mathbf{X}^{wT}\mathbf{K}^w(\tau_{21}, T_2 - \tau_{21}) \\ \vdots \end{pmatrix}_{(N_1+N_2)\times(N_1+N_2)} ,
$$

where $N_j = \sum_{i \in \mathcal{G}_j} m_i, j = 1, 2$, $\mathbf{K}^w(t, s)$ is a diagonal matrix with elements $K((\tau_{ij} - t)/h^w, (T_i - \tau_{ij} - s)/h^w)$, $K(\cdot, \cdot)$ is a bivariate kernel function and h^w is the bandwidth. We use a truncated Gaussian kernel function $K(x, y) = \exp(-(x^2 + y^2)/2) \cdot I(x^2 + y^2 < 6)$ throughout this article. The most efficient estimator among the PWLS estimators is the one using the inverse of the true covariance matrix of ε^w as the weight matrix $\mathbf{W}$. Once residuals $\widehat{\varepsilon}_{ij}^w$ are obtained, the diagonal elements of the covariance matrix can be estimated by the Nadaraya-Watson estimator:

$$
\widehat{\sigma}^2(t, s) = \frac{\sum_{i=1}^{n_1+n_2} \sum_{j=1}^{m_i} \widehat{\varepsilon}_{ij}^{w2} K\left(\frac{\tau_{ij}-t}{h^w}, \frac{T_i-\tau_{ij}-s}{h^w}\right)}{\sum_{i=1}^{n_1+n_2} \sum_{j=1}^{m_i} K\left(\frac{\tau_{ij}-t}{h^w}, \frac{T_i-\tau_{ij}-s}{h^w}\right)}. \tag{3.4}
$$

Note that this is a direct bivariate extension of the kernel estimator of the univariate variance function in [12]. A nonparametric estimation of the correlation of ε^w requires a trivariate kernel estimator which may suffer from curse of dimensionality. For this reason we use a working independence correlation by simply letting $\mathbf{W}$ be a diagonal matrix with elements $1/\widehat{\sigma}^2(\tau_{ij}, T_i - \tau_{ij})$.

The entire estimation procedure can be described in four steps. Firstly, estimate $\boldsymbol{\alpha}$ following (3.3) with $\mathbf{W} = \mathbf{I}$, denote the estimator as $\widehat{\boldsymbol{\alpha}}_{\mathbf{I}}$; Secondly, estimate $\boldsymbol{\beta}$ from the following

varying-coefficient model

$$Y_{ij}^{w*} \equiv Y_i^w(\tau_{ij}) - \mathbf{Z}_i^w(\tau_{ij})^T\widehat{\boldsymbol{\alpha}}_{\mathbf{I}} \tag{3.5}$$

$$= \mathbf{X}_i^w(\tau_{ij})^T\boldsymbol{\beta}(\tau_{ij}, T_i - \tau_{ij}) + \varepsilon_i^w(\tau_{ij}, T_i - \tau_{ij})$$

using the bivariate kernel approach of [50]:

$$\widehat{\boldsymbol{\beta}}(t, s; h^w) = \left[\mathbf{X}^{wT}\mathbf{K}^w(t, s)\mathbf{X}^w\right]^{-1}\mathbf{X}^{wT}\mathbf{K}^w(t, s)\mathbf{Y}^{w*}; \tag{3.6}$$

Thirdly, calculate the residuals

$$\widehat{\varepsilon}_{ij}^w = Y_{ij}^w - \mathbf{X}_i^w(\tau_{ij})^T\widehat{\boldsymbol{\beta}}(\tau_{ij}, T_i - \tau_{ij}; h^w) - \mathbf{Z}_i^w(\tau_{ij})^T\widehat{\boldsymbol{\alpha}}_{\mathbf{I}}$$

and obtain the diagonal weight matrix $\mathbf{W}$ by inverting (3.4); Lastly, estimate $\boldsymbol{\alpha}$ again using the diagonal weight matrix, denote the estimator as $\widehat{\boldsymbol{\alpha}}_{\mathbf{D}}$. As pointed out by [12], both $\widehat{\boldsymbol{\alpha}}_{\mathbf{I}}$ and $\widehat{\boldsymbol{\alpha}}_{\mathbf{D}}$ converge to $\boldsymbol{\alpha}$ with root-n rates that are faster than the nonparametric convergence rate of $\widehat{\boldsymbol{\beta}}$. As a result, the asymptotic bias and variance of $\widehat{\boldsymbol{\beta}}$ have the same form as those of a varying-coefficient model with true $\boldsymbol{\alpha}$ and are not affected by the weight matrix $\mathbf{W}$. So there is no advantage to update the estimator of $\boldsymbol{\beta}$ with $\widehat{\boldsymbol{\alpha}}_{\mathbf{D}}$, and we can adopt the inference procedure in [50] which eliminates the asymptotic bias via undersmoothing and estimates the variance by a sandwich estimator

$$\widehat{\mathrm{Var}}\left(\widehat{\boldsymbol{\beta}}(t, s; h^w)\right) \tag{3.7}$$

$$= \left[\mathbf{X}^{wT}\mathbf{K}^w(t, s)\mathbf{X}^w\right]^{-1}\left[\mathbf{X}^{wT}\mathbf{K}^w(t, s)\mathbf{R}\mathbf{K}^w(t, s)\mathbf{X}^w\right]\left[\mathbf{X}^{wT}\mathbf{K}^w(t, s)\mathbf{X}^w\right]^{-1},$$

where $\mathbf{R} = \operatorname{diag}(\widehat{\boldsymbol{\varepsilon}}_1^w \widehat{\boldsymbol{\varepsilon}}_1^{wT}, ..., \widehat{\boldsymbol{\varepsilon}}_n^w \widehat{\boldsymbol{\varepsilon}}_n^{wT})$ with $\widehat{\boldsymbol{\varepsilon}}_i^w = (\widehat{\varepsilon}_{i1}^w, ..., \widehat{\varepsilon}_{im_i}^w)^T$. A sandwich variance estimator is also derived for $\boldsymbol{\alpha}$ by [12]:

$$\widehat{\operatorname{Var}}(\widehat{\boldsymbol{\alpha}}) = \mathbf{D}^{-1} \mathbf{V} \mathbf{D}^{-1}, \tag{3.8}$$

where $\mathbf{D} = \mathbf{Z}^{wT}(\mathbf{I} - \mathbf{S})^T \mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{Z}^w$ and $\mathbf{V} = \mathbf{Z}^{wT}(\mathbf{I} - \mathbf{S})^T \mathbf{W} \mathbf{R} \mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{Z}^w$. The asymptotic normality of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ can be obtained following [12], which enables the construction of confidence intervals.

In the above estimating procedure, we follow [50] to use the same bandwidth h^w for both the forward follow-up time and the residual lifetime in the bivariate kernel function, and use the K-fold cross-validation (CV) for the bandwidth selection. Specifically, we randomly split the patients into K equal-sized sets and evaluate the following CV loss for each candidate of h^w:

$$CV(h^w) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{m_i} \left[Y_i^w(\tau_{ij}) - \mathbf{Z}_i^w(\tau_{ij})^T \widehat{\boldsymbol{\alpha}}_{\mathbf{I}}^{(-k)} - \mathbf{X}_i^w(\tau_{ij})^T \widehat{\boldsymbol{\beta}}^{(-k)}(\tau_{ij}, T_i - \tau_{ij}; h^w) \right]^2,$$

where C_k represents the indices of patients in the k-th set and $\cup_k C_k = \mathcal{G}_1 \cup \mathcal{G}_2$. Here $\widehat{\boldsymbol{\alpha}}_{\mathbf{I}}^{(-k)}$ and $\widehat{\boldsymbol{\beta}}^{(-k)}$ represent the estimates obtained from the dataset excluding patients in the k-th set. Once we have the optimal h^w that minimizes the CV loss, we follow the practice in [50] to reduce the bias via undersmoothing by a factor n^{θ^w} with a small $\theta^w > 0$. For simplicity, the same bandwidth is used for the estimation of σ^2 in (3.4).

61

3.3.2 Model for Transplantation

To evaluate the medical cost associated with transplantation among waitlisted patients, we propose the following model:

$$
\begin{aligned}
Y_i^t(\tau_{ij}) = {} & \xi_1(\tau_{ij}, T_i - \tau_{ij}) + \xi_2(\tau_{ij}, T_i - \tau_{ij})\text{African American}_i \\
& + \xi_3(\tau_{ij}, T_i - \tau_{ij})\text{Other Race}_i + \xi_4(\tau_{ij}, T_i - \tau_{ij})\text{Female}_i + \xi_5(\tau_{ij}, T_i - \tau_{ij})\text{Age}_i \\
& + \xi_6(\tau_{ij}, T_i - \tau_{ij})\text{Hypertension}_i + \xi_7(\tau_{ij}, T_i - \tau_{ij})\text{Other Comorbidities}_i \quad (3.9) \\
& + \xi_8(\tau_{ij}, T_i - \tau_{ij})I(25 \leq \text{BMI}_i < 30) + \xi_9(\tau_{ij}, T_i - \tau_{ij})I(\text{BMI}_i \geq 30) \\
& + \gamma(\tau_{ij} - S_i, T_i - \tau_{ij})I(\tau_{ij} \geq S_i) + \varepsilon_i^t(\tau_{ij}, T_i - \tau_{ij}), \, i \in \mathcal{G}_2 \cup \mathcal{G}_3,
\end{aligned}
$$

where S_i represents the day of receiving a kidney transplant and $Y_i^t(\tau_{ij})$ represents the daily claim amount on day τ_{ij} after the same transformation that yields $Y_i^w(\tau_{ij})$. The superscript t on Y and ε indicates that this model focuses on the comparison between transplanted and waitlisted patient groups. The set of coefficients ξ_k, $k = 1, ..., 9$, vary with both the time since the first ESRD service and the residual lifetime, whereas the coefficient γ varies with the time since kidney transplant and the residual lifetime. The coefficient γ is the difference of the medical cost of a transplanted patient since the day of receiving a kidney transplant to the medical cost that the patient would have had spent if the patient were still on the waitlist, which is the parameter of interest in this analysis. Unlike the other bivariate time-varying coefficients, γ kicks in at time S_i. For this reason, we refer to this model as the mixed-time varying-coefficient model. Notice that the binary variable indicating dialysis type is no longer included in model (3.10) because transplanted patients no longer need dialysis after transplant. The model is again conditional on the death time T_i, which allows us to examine the difference in the longitudinal cost trajectories of waitlisted and transplanted patients from the time of transplantation to the end of life.

An implicit assumption in model (3.10) is, conditional on the lifetime and all other covariates, the mean medical cost for transplanted patients prior to receiving transplantation is the same as that for patients who never received transplants. We conducted an exploratory data analysis to verify this assumption. Specifically, we fitted a bivariate time-varying coefficient model conditional on the death time and all other covariates to compare the medical costs of transplanted patients during their pre-transplant time and untransplanted patients. The results indicate similar medical costs for the two groups, implying that the diversion does not happen until the time of transplant.

For notational simplicity, consider a mixed-time varying-coefficient model in the following form:

$$Y_i^t(\tau_{ij}) = \mathbf{X}_i^t(\tau_{ij})^T \boldsymbol{\xi}(\tau_{ij}, T_i - \tau_{ij}) + I(\tau_{ij} \geq S_i)\gamma(\tau_{ij} - S_i, T_i - \tau_{ij}) + \varepsilon_i^t(\tau_{ij}, T_i - \tau_{ij}). \quad (3.10)$$

For the waitlisted group, the model reduces to the following bivariate varying coefficient model

$$Y_i^t(\tau_{ij}) = \mathbf{X}_i^t(\tau_{ij})^T \boldsymbol{\xi}(\tau_{ij}, T_i - \tau_{ij}) + \varepsilon_i^t(\tau_{ij}, T_i - \tau_{ij}), \quad i \in \mathcal{G}_2. \quad (3.11)$$

Thus the regression coefficients $\boldsymbol{\xi}$ can be estimated by applying the method of [50] using the waitlisted group data only. Although the estimation of $\boldsymbol{\xi}$ may be improved by also including the pre-transplant measurements in $\mathcal{G}_3$, using only measurements in $\mathcal{G}_2$ keeps the independence between $\widehat{\boldsymbol{\xi}}$ and data in $\mathcal{G}_3$, thus helps develop the asymptotic results of $\widehat{\boldsymbol{\gamma}}$ that is given below. This suggests a two-stage estimation procedure. Firstly, choose a bandwidth h_1^t and estimate $\boldsymbol{\xi}$ in model (3.11) for the waitlisted group:

$$\widehat{\boldsymbol{\xi}}(t, s) = \left[(\mathbf{X}_{\mathcal{G}_2}^t)^T \mathbf{K}_{\mathcal{G}_2}^t(t, s) \mathbf{X}_{\mathcal{G}_2}^t \right]^{-1} \mathbf{X}_{\mathcal{G}_2}^{tT} \mathbf{K}_{\mathcal{G}_2}^t(t, s) \mathbf{Y}_{\mathcal{G}_2}^t,$$

where $\mathbf{X}_{\mathcal{G}_2}^t$ and $\mathbf{Y}_{\mathcal{G}_2}^t$ represents design matrix and response vector of the waitlisted group only and $\mathbf{K}_{\mathcal{G}_2}^t(t,s)$ is a diagonal matrix with elements $K((\tau_{ij} - t)/h_1^t, (T_i - \tau_{ij} - s)/h_1^t)$. Secondly, for the transplanted group, (i) evaluate the predicted cost of each patient using the estimates obtained in the first stage $\mathbf{X}_i^t(\tau_{ij})^T\widehat{\boldsymbol{\xi}}(\tau_{ij}, T_i - \tau_{ij}; h_1^t)$; (ii) subtract the predicted cost from the observed cost; and (iii) choose a bandwidth h_2^t and estimate γ from the following model

$$
\begin{aligned}
Y_{ij}^{t*} &\equiv Y_i^t(\tau_{ij}) - \mathbf{X}_i^t(\tau_{ij})^T\widehat{\boldsymbol{\xi}}(\tau_{ij}, T_i - \tau_{ij}; h_1^t) \\
&= \gamma(\tau_{ij} - S_i, T_i - \tau_{ij}) + \varepsilon_i^t(\tau_{ij}, T_i - \tau_{ij})
\end{aligned}
\tag{3.12}
$$

using post-transplant observations (hence $\tau_{ij} \geq S_i$). The estimator is given by

$$
\widehat{\gamma}(t,s) = \frac{\sum_{i \in \mathcal{G}_3} \sum_{\tau_{ij} \geq S_i} Y_{ij}^{t*} K\left(\frac{\tau_{ij} - S_i - t}{h_2^t}, \frac{T_i - s}{h_2^t}\right)}{\sum_{i \in \mathcal{G}_3} \sum_{\tau_{ij} \geq S_i} K\left(\frac{\tau_{ij} - S_i - t}{h_2^t}, \frac{T_i - s}{h_2^t}\right)}.
\tag{3.13}
$$

If $\widehat{\boldsymbol{\xi}}(\tau_{ij}, T_i - \tau_{ij}; h_1^t)$ is replaced by true $\boldsymbol{\xi}(\tau_{ij}, T_i - \tau_{ij})$, then the coefficient γ in (3.13) is the same coefficient γ in (3.10). Using a similar feature of the estimation for model (3.2), we can eliminate the impact of the estimation error of $\widehat{\boldsymbol{\xi}}$ in (3.13) to the estimation of γ asymptotically by choosing a bandwidth that yields faster convergence rate in the above first-step estimator. When this is done properly with details given below, the asymptotic bias of $\widehat{\gamma}$ disappears and the asymptotic normality of $\widehat{\gamma}$ can be obtained following [50] with an asymptotic variance unaffected by the first step estimation of $\boldsymbol{\xi}$.

For bandwidth selection, we choose the first step bandwidth h_1^t using the following CV criterion:

$$
CV(h_1^t) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{m_i} \left[Y_i^t(\tau_{ij}) - \mathbf{X}_i^t(\tau_{ij})^T\widehat{\boldsymbol{\xi}}^{(-k)}(\tau_{ij}, T_i - \tau_{ij}; h_1^t) \right]^2,
\tag{3.14}
$$

where $\cup_k C_k = \mathcal{G}_2$. To ensure a faster convergence rate, we do not undersmooth the first

step estimates. Let h_{10}^t be the optimal bandwidth that minimizes the CV criterion (3.14). It should achieve a balance between the bias and variance of $\widehat{\boldsymbol{\xi}}$ at a rate of $O(n_2^{-1/6})$ based on Theorem 3.1 of [50]. As a result, $\widehat{\boldsymbol{\xi}}$ converges at a rate of $O(n_2^{-1/3})$.

For the second step estimation, we use the following CV criterion:

$$CV(h_2^t) = \sum_{k=1}^K \sum_{i \in C_k} \sum_{\tau_{ij} \geq S_i} \left[Y_{ij}^{t*} - \widehat{\gamma}^{(-k)}(\tau_{ij} - S_i, T_i - \tau_{ij}; h_2^t) \right]^2 , \qquad (3.15)$$

where $\cup_k C_k = \mathcal{G}_3$ and Y_{ij}^{t*} are calculated using h_{10}^t. Let h_{20}^t be the optimal bandwidth that minimizes criterion (3.15). We then undersmooth the second step estimator by choosing a smaller bandwidth $h_2^{t*} = h_{20}^t/n_3^{\theta^t}$, where $\theta^t > 0$. We undersmooth more aggressively here by letting $\theta^t > \theta^w$ because of the slower than root-n convergence rate of $\widehat{\boldsymbol{\xi}}$. Given that the sample sizes n_2 and n_3 are with the same order, the rate of convergence of $\widehat{\boldsymbol{\xi}}$ is faster than that of $\widehat{\gamma}$, which eliminates the impact of the first step estimation on the second step estimation asymptotically. Then we can estimate the variance of $\widehat{\gamma}$ by

$$\widehat{\text{Var}}\left(\widehat{\gamma}(t, s; h_2^{t*}) \right) = \frac{\sum_{i \in \mathcal{G}_3} \left[\sum_{\tau_{ij} \geq S_i} \widehat{\varepsilon}_{ij}^t K\left(\frac{\tau_{ij} - S_i - t}{h_2^{t*}}, \frac{T_i - s}{h_2^{t*}} \right) \right]^2}{\left[\sum_{i \in \mathcal{G}_3} \sum_{\tau_{ij} \geq S_i} K\left(\frac{\tau_{ij} - S_i - t}{h_2^{t*}}, \frac{T_i - s}{h_2^{t*}} \right) \right]^2}, \qquad (3.16)$$

where the residuals are computed by

$$\widehat{\varepsilon}_{ij}^t = Y_i^t(\tau_{ij}) - \mathbf{X}_i^t(\tau_{ij})^T \widehat{\boldsymbol{\xi}}(\tau_{ij}, T_i - \tau_{ij}; h_{10}^t) - I(\tau_{ij} \geq S_i)\widehat{\gamma}(\tau_{ij} - S_i, T_i - \tau_{ij}; h_2^{t*}).$$

It is fairly straightforward to extend model (3.10) for multiple kidney transplants. For example, we can simply add another varying-coefficient term $\zeta(\tau_{ij} - R_i, T_i - \tau_{ij})$ where R_i represents the time at the second kidney transplant, if it ever happens. This adds a third step to the estimation procedure, which could complicate the bandwidth selections in all three steps and may suffer from the shrinking sample size of patients who received two or

more transplants.

3.4 Simulation

In this section we conduct two simulation experiments to examine the estimation and inference procedures proposed for the two models.

3.4.1 Simulation for the Semi-Varying Coefficient Model

For model (3.2) we consider two covariates with constant coefficients and two covariates (including the intercept) with bivariate time-varying coefficients. Specifically, we generate Z_1^{w*} and Z_2^{w} from a bivariate normal distribution with standard normal marginal distributions and a correlation coefficient of 0.5, then transform Z_1^{w*} by $Z_1^{w} = I(Z_1^{w*} > 0)$ that mimics the Waitlisted variable in model (3.2); X_1^{w} is always 1; and X_2^{w} is a time-varying binary covariate generated in the following: at $t_0 = 0$, $X_2^{w}(0)$ follows a Bernoulli distribution with probability 0.5, then $X_2^{w}(t)$ alternates between 0 and 1 at time points $t_1, t_2, ...$ where $t_j - t_{j-1}$ follows an exponential distribution with rate 0.05. This time-varying binary covariate mimics the variable PD in model (3.2). Death time T is generated by adding constant 1 to an exponentially distributed random variable with rate $\exp(0.5Z_1^{w} + 0.2Z_2^{w} - 3.5)$. The added constant 1 avoids small T values happening before the first observation time τ_1. The censoring time C follows a uniform distribution between 20 and 40, which yields a censoring rate of 34%, which is higher than the actual censoring rate of around 10% among ESRD patients who started first service in 2007-2011 and at least 65 years old. The uniform distribution mimics the real data because we only retain ESRD patients who started first service in 2007-2011 and the follow-up ends at the end of 2017. For observation times, we follow the same procedure in [50]. That is, we generate τ_1 from a uniform distribution between 0 and 1. Observation

	$\widehat{\boldsymbol{\alpha}}_{\mathbf{I}}$	$\widehat{\boldsymbol{\alpha}}_{\mathbf{D}}$
MSE	$(15.56 \cdot 10^{-4},\ 4.04 \cdot 10^{-4})$	$(14.34 \cdot 10^{-4},\ 3.76 \cdot 10^{-4})$
Coverage	$(0.954,\ 0.952)$	$(0.936,\ 0.938)$

Table 3.2: MSE and CI coverage probability for two PWLS estimators of $\boldsymbol{\alpha}$.

times τ_{ij}, $j > 1$, is generated independently from $\tau_{ij} - (j-1) \sim \text{Beta}(\tau_{i1}/4\nu^2, (1-\tau_{i1})/4\nu^2)$, where ν serves as an upper bound of standard deviation of the Beta distribution and is set to be 0.01. The generated interarrival time $\tau_{ij} - \tau_{i,j-1}$ has mean 1 and a small variance. Observation times beyond $\min(T, C)$ are dropped. The error term $\varepsilon(\tau_j, T_i - \tau_{ij})$ is generated by a nonhomogeneous Ornstein-Uhlenbeck (NOU) process $U(\tau_j)$ plus a random error. In particular, the NOU process satisfies $\text{Var}(U(t)) = \exp(1 - 0.1t)$ and $\text{Corr}(U(t), U(s)) = 0.5^{|t-s|}$, and the random error follows a standard normal distribution. The longitudinal outcome $Y^w(\tau_j)$ is generated according to model (3.2) with $\alpha_1 = 1$, $\alpha_2 = 2$,

$$\beta_1(t, s) = \frac{t}{4} \exp\left(-\frac{t^2 + s^2}{100}\right), \quad \text{and} \quad \beta_2(t, s) = \frac{1}{2}\left[\sin\left(\frac{2t}{5}\right) - \cos\left(\frac{s}{2}\right)\right].$$

We simulate 500 datasets following the simulation design. For each dataset we generate $n = 1000$ subjects and only keep the subjects with $T \leq C$. To save the computing cost, we conduct 5-fold CV for only 10 datasets , which yields an average bandwidth of 1.55. We then undersmooth it to obtain a bandwidth of $1.55 \times 660^{-0.05} \approx 1.12$, where 660 is the average number of complete cases, and fix it for all the 500 datasets. In Table 3.2, we report the mean squared error (MSE) of the PWLS estimators $\widehat{\boldsymbol{\alpha}}_{\mathbf{I}} = (\widehat{\alpha}_{1,\mathbf{I}}, \widehat{\alpha}_{2,\mathbf{I}})^T$ and $\widehat{\boldsymbol{\alpha}}_{\mathbf{D}} = (\widehat{\alpha}_{1,\mathbf{D}}, \widehat{\alpha}_{2,\mathbf{D}})^T$, as well as the coverage probabilities of the 95% confidence intervals constructed using their respective sandwich variance estimators (3.8). From Table 3.2 we see that $\widehat{\boldsymbol{\alpha}}_{\mathbf{D}}$ improves the MSE of $\widehat{\boldsymbol{\alpha}}_{\mathbf{I}}$ by 8% and 7% respectively and the sandwich variance estimators yield coverage probabilities close to the nominal level for both estimators.

We further examine the estimators of the time-varying coefficients $\widehat{\boldsymbol{\beta}}$. Here we follow the same way of [50] to visualize a bivariate function. That is, we fix T and plot $\widehat{\beta}_k(t, T - t)$
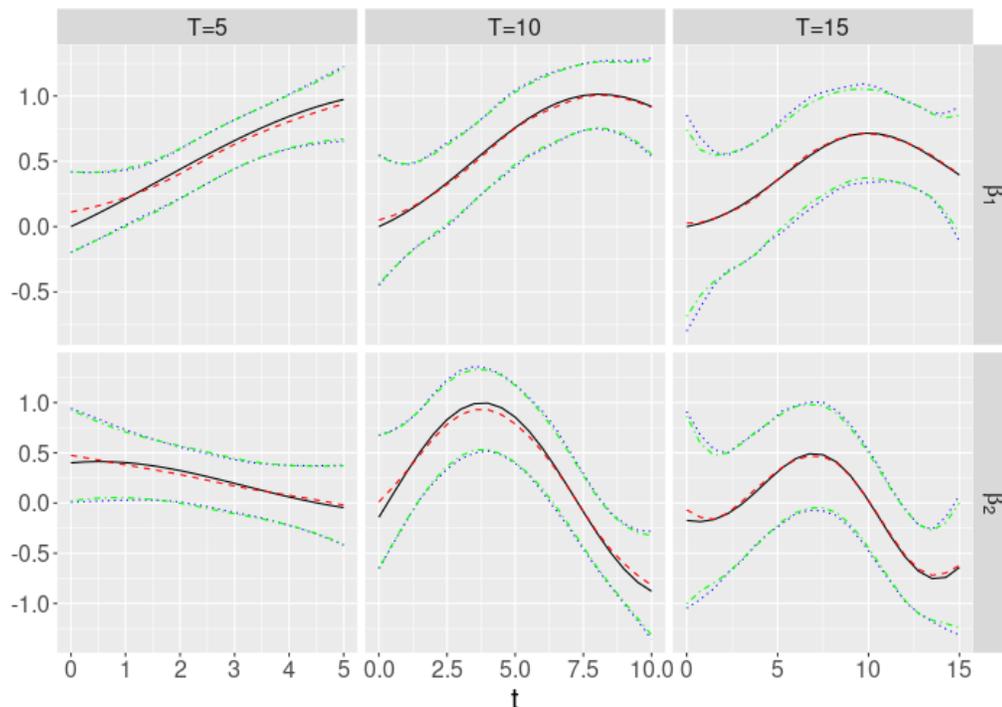
Figure 3.1: True $\beta_k(t, T-t)$ vs sample mean of $\widehat{\beta}_k(t, T-t)$, $k = 1, 2$ and 95% empirical bands vs mean of 95% confidence bands at $T = 5, 10, 15$.

varying with t. Among the 2×3 panels in Figure 3.1, each row represents a time-varying coefficient and each column represents a value of T. Within each panel, we plot the true coefficient curves β_k (black solid line), the sample mean of the estimates $\widehat{\beta}_k$ (red dashed line), the 95% empirical bands calculated by the mean $\pm$ 1.96 times the standard deviation of $\widehat{\beta}_k$ (blue dotted lines) and the mean of 95% confidence bands calculated using the sandwich estimator (green dash-dot lines). We can see across all panels that the estimators $\widehat{\beta}_k$ have negligible biases relative to their variabilities. Moreover, the empirical 95% bands overlap well with the mean of 95% confidence intervals, which indicates the validity of the sandwich variance estimator (3.8).

3.4.2 Simulation for the Mixed-Time Varying-Coefficient Model

For the mixed-time varying-coefficient model (3.10), we conduct a simulation study as follows: X_1^t is always 1, X_2^t and X_3^t are generated in the same way as Z_1^w and Z_2^w in the previous subsection. We do not simulate time-varying covariate as there is no such covariate in the proposed model (3.10) for the ESRD data. Both the death time T and the censoring time C are generated similarly as in the previous subsection, where T is 1 plus a random number generated from an exponential distribution with rate $\exp(0.5X_2^t + 0.2X_3^t - 3.5)$ and C is uniform between 20 and 40, yielding a censoring rate of around 34%. The transplant time S is generated in the following. When T is smaller than 20, we generate a Bernoulli random variable that takes the value of 1 with a probability of 0.3, and if it is 1, S is generated under a uniform distribution between 0 and T; When T is larger than or equal to 20, we generate a Bernoulli random variable that takes the value of 1 with a probability of 0.7, and if it is 1, S is generated under a uniform distribution between 0 and 20. This design mimics the situation where a patients who lives longer since the onset of ESRD has a larger probability of receiving a transplant. Under this design, the percentage of transplanted patients is around 40% among deceased patients. We follow the same simulation design in the previous subsection to generate the visiting times τ_{ij} and the error term $\varepsilon(\tau_j, T_i - \tau_{ij})$. The longitudinal response is generated from model (3.10) with the following coefficients:

$$
\begin{aligned}
\xi_1(t, s) &= \frac{t}{4} \exp\left(-\frac{t^2 + s^2}{100}\right), \\
\xi_2(t, s) &= \exp\left(-\frac{ts}{t^2 + s^2}\right), \\
\xi_3(t, s) &= \cos\left(\frac{t^2 + s^2}{100}\right), \\
\gamma(t, s) &= \frac{1}{2}\left[\sin\left(\frac{2t}{5}\right) - \cos\left(\frac{s}{2}\right)\right].
\end{aligned}
$$

Following the simulation design, we simulate 500 datasets each with $n = 1000$ and then only keep the subjects with $T \leq C$. For bandwidth selection, we first run a 5-fold CV on untransplanted subjects for 10 datasets using (3.14), which yields an average bandwidth of 2.1. For each dataset, we use the bandwidth 2.1 to estimate $\boldsymbol{\xi}$ and obtain the partial residuals for post-transplant observations of transplanted subjects using (3.13). We then run another 5-fold CV on transplanted subjects for the same 10 datasets using (3.15). The second CV yields an average bandwidth of 2.2, which is later undersmoothed to be $2.2 \times n^{-0.1} \approx 1.1$. Notice that we only shrink the second bandwidth for undersmoothing with a larger factor compared to the previous subsection, which leads to a faster convergence rate for $\widehat{\boldsymbol{\xi}}$ in order to achieve a desirable asymptotic behavior of $\widehat{\gamma}$.

Our main interest is to estimate γ. The performance of the bivariate time-varying coefficient estimator $\widehat{\gamma}$ given $T - S$ is presented in Figure 3.2. Within each panel, we plot the true coefficient curve γ (black solid line), the sample mean of the estimator $\widehat{\gamma}$ (red dashed line), the 95% empirical confidence bands calculated by the mean $\pm$ 1.96 times the standard deviation of $\widehat{\gamma}$ (blue dotted lines) and the mean of 95% confidence bands calculated using the sandwich estimator in (3.16) (green dash-dot lines). The bias is larger than that in the previous subsection, which is likely due to the first step estimation error, but still relatively small compared to the variability of $\widehat{\gamma}$. The empirical bands overlap well with the mean of the confidence intervals, which shows that the variance is approximately well by the sandwich estimator.
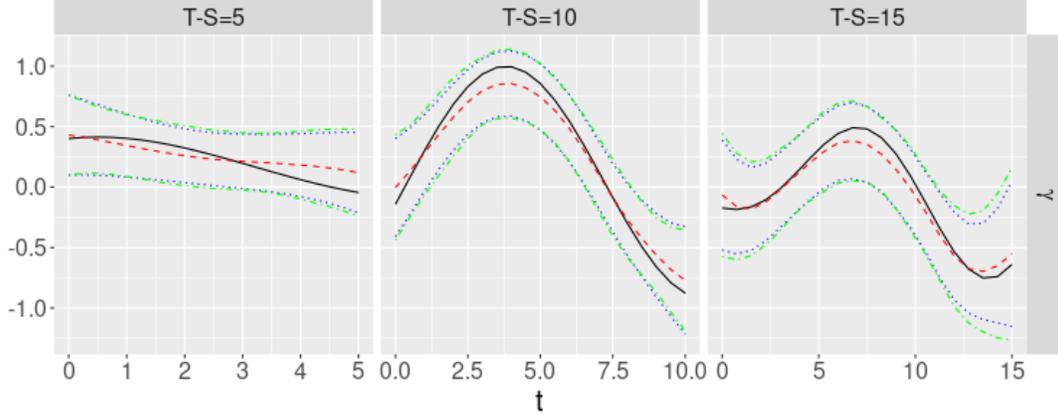
Figure 3.2: True $\gamma(t, T - S - t)$ vs sample mean of $\widehat{\gamma}(t, T - S - t)$ and 95% empirical bands vs mean of 95% confidence bands at $T - S = 5, 10, 15$.

3.5 ESRD Data Analysis

3.5.1 Cost Difference Associated with Waitlisting

In this subsection, we report the results of estimating coefficients in model (3.2), where the constant coefficients are estimated by the reweighted PWLS estimator (3.3) and the time-varying coefficients are estimated by the bivariate kernel estimator (3.6). Implementing the CV criterion in Section 3.1 yields a bandwidth of 30, which is then shrunken to be $30/(n_1 + n_2)^{0.05} \approx 19$ for undersmoothing. Coefficients α_1 and α_2 are of great interest, which represent the overall medical cost difference at the time of the first ESRD service and the increase in cost difference over time between unwaitlisted and waitlisted patients with the same lifespan and the same age of onset of ESRD, adjusting for all other covariates. The point estimates of all the constant coefficients $\boldsymbol{\alpha}$ are listed in Table 3.3 together with their 95% confidence intervals and p-values calculated from the sandwich variance estimator (3.8). From the table we see that waitlisting is significantly associated with a lower daily medical cost at the beginning of ESRD service among waitlisted patients, but the medical cost gradually increases over time. Averaging over the lifespan, however, the waitlisting effect, modeled by only including the main effect of Waitlisted variable in model (3.2), disappears

71

Variable	Point estimate $\widehat{\alpha} \cdot 10^{-2}$	95% CI$\cdot 10^{-2}$	P-value
Waitlisted	-3.359	(-4.701, -2.016)	$< 10^{-4}$
Time×Waitlisted	1.217	(0.846, 1.588)	$< 10^{-4}$
Race: black	0.311	(-0.798, 1.420)	0.582
Race: other	-1.395	(-3.408, 0.618)	0.174
Sex: female	2.183	(1.320, 3.047)	$< 10^{-4}$
Age	0.066	(-0.021, 0.153)	0.137
Hypertension	-1.455	(-2.683, -0.228)	0.020
Other comorbidities	2.033	(1.036, 3.031)	$< 10^{-4}$
$25 \leq$ BMI < 30	1.646	(0.588, 2.704)	0.002
BMI ≥ 30	3.955	(2.879, 5.032)	$< 10^{-4}$

Table 3.3: Re-weighted estimates of the constant coefficients. The variable Time is the time since first ESRD service in year (365 days).

(results not shown). Female gender, comorbidities except hypertension and higher BMI are all significantly associated with higher amount of daily Medicare claims. Hypertension is associated with a lower amount of daily Medicare claims, and race and age are not significant.

It is also of interest to examine $\widehat{\beta}_k$, $k = 1, 2$. Similar to how we visualize the bivariate curves in the simulations, we draw $\widehat{\beta}_k(t, T - t)$ as univariate functions of t for three values of T at 500, 1250 and 2000 days, respectively, in Figure 3.3. These numbers are roughly the quantiles of patients' lifetime durations from the onset of ESRD. In the figure we can see that the overall medical cost trajectory is U-shaped. This is true for both unwaitlisted and waitlisted patient groups because the waitlisting effects (both main effect and interaction with time) are of a much smaller magnitude comparing to $\widehat{\beta}_1$. A similar pattern was also observed by [50] for the inpatient cost of ESRD patients who never received transplantation. Compared to HD patients, PD patients tend to have a lower medical cost during early days since the onset of ESRD, but this difference gradually disappears as time goes by.

Because of the logarithmic transformation on the claim amount, the estimates can not be straightforwardly interpreted as the increment of daily Medicare claims in dollars associated with one-unit increase in the corresponding variable. The estimates, however, can be translated into the increment over a reference value. Taking the estimate of α_1 as an example,
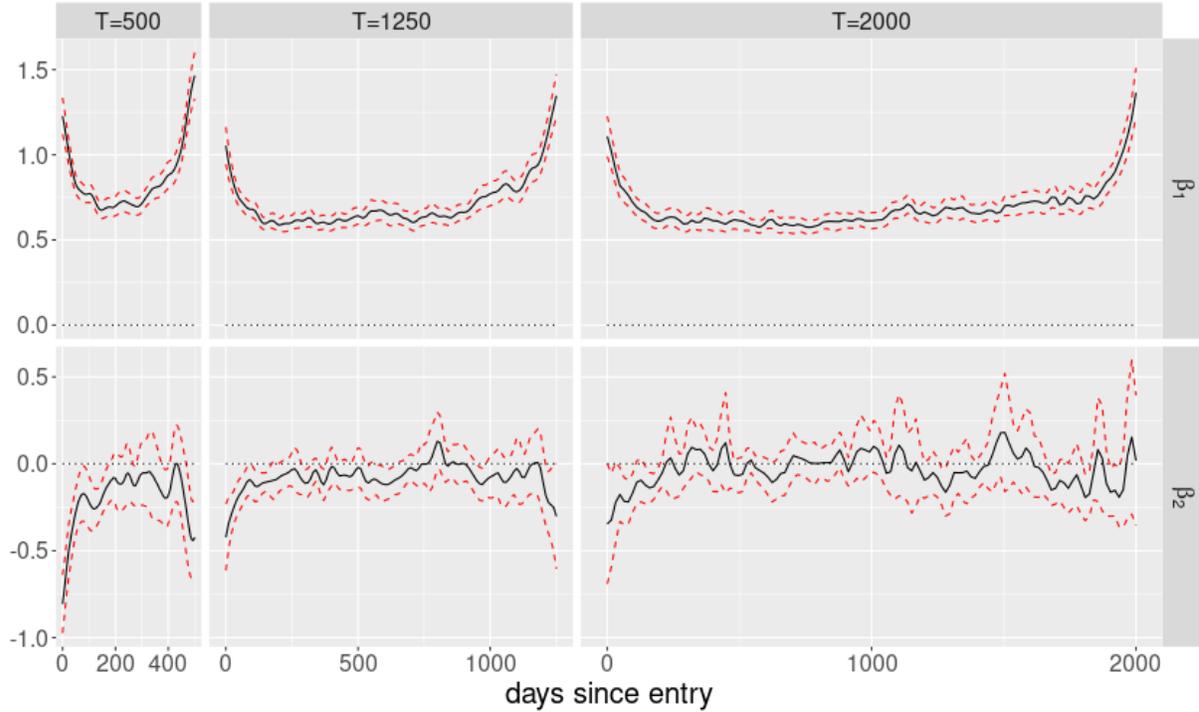
Figure 3.3: $\widehat{\beta}_k(t, T - t), k = 1, 2$ (solid curves) and their confidence bands (dashed curves) at $T = 500, 1250, 2000$ days.

it can be interpreted in the following way: compared to an unwaitlisted patient for whom Medicare pays \$100 on the day of his/her first service (this is within a normal range as the mean daily claim amount is \$192 and the median is \$85), a waitlisted patient with the same values of all other variables claims \$6.61 less on the day of the first service. This is obtained from

$$\log\left(\frac{100 - 6.61}{100} + 1\right) - \log\left(\frac{100}{100} + 1\right) \approx \widehat{\alpha}_1.$$

Similar calculations can be done for other reference values and other variables.
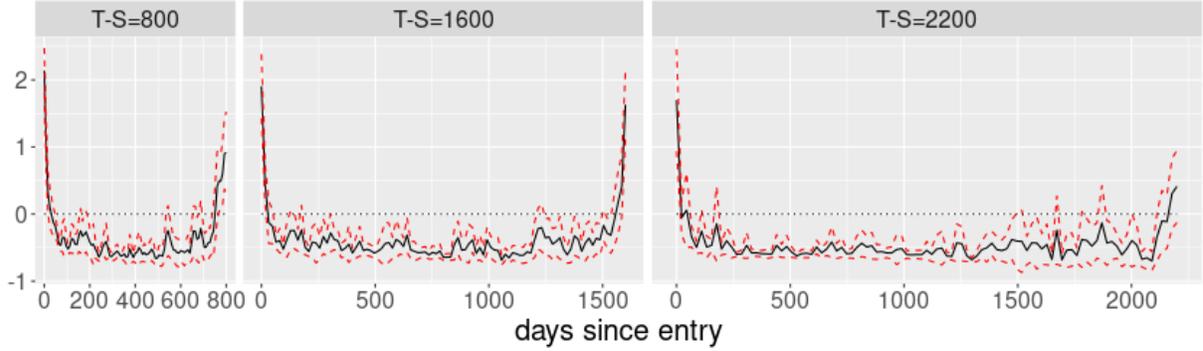
Figure 3.4: $\widehat{\gamma}(t, T-S-t)$ (solid curves) and its confidence bands (dashed curves) at $T-S = 800, 1600, 2200$ days.

3.5.2 Cost Difference Associated with Kidney Transplant

Now we apply the two-stage estimation and inference procedure described in Subsection 3.2 to model (3.10). For bandwidth selection in the first stage, we conduct a 5-fold CV using the data of waitlisted patients. The selected bandwidth in this stage is 300 that is used for the estimation without any shrinkage. In the second stage, we conduct a 5-fold CV using the data of transplanted patients. The selected second-step bandwidth is 20 and then shrunken by $n_3^{0.1}$ for undersmoothing, which yields a bandwidth of 9.1.

The estimator $\widehat{\gamma}$ for the total claim amount are shown in Figure 3.4 with their pointwise confidence bands calculated from the sandwich variance estimator (3.16). The three selected $T-S$ values are roughly quantiles among the transplanted patients. Note that $\gamma(t, T-S-t)$ represents the cost difference between transplanted and waitlisted ESRD patients during their lifespan since the time of transplantation adjusted by all other covariates in model (3.10). From Figure 3.4 we see that there is an initial spike of added total Medicare cost for transplanted patients after transplantation comparing to waitlisted patients and then the total medical spending of transplanted patients quickly drops below the level of waitlisted patients. There is an uptick of total medical spending among transplanted patients towards their end of life compared to waitlisted patients.

74

To better understand the overall cost difference between waitlisted and transplanted ESRD patients, we dive into details by examining five major types of the Medicare claims, which are inpatient, outpatient, skilled nursing, home health, and hospice costs. Replacing the total cost in model (3.10) by each individual type of Medicare cost, we run five separate analyses using the same two-stage estimation method. For each individual type of cost, we conduct bandwidth selection, estimation and inference independently. For bandwidth selection in both stages, we follow the same procedure described in Section 3.2 and carried out a 5-fold CV. The selected first-stage bandwidth values are (60, 1000, 200, 300, 200) respectively and are used directly for corresponding estimation without any shrinkage. The selected second-step bandwidth values are (14, 150, 110, 100, 30) respectively and then shrunken by $n_3^{0.1}$ for undersmoothing.

The estimates of $\widehat{\gamma}$ for all five types of services are presented in Figure 3.5 together with corresponding pointwise confidence bands calculated from their sandwich variance estimators. There are several interesting findings from Figure 3.5: (1) Inpatient cost is the one that causes the initial spike after transplantation in the total medical cost among transplanted patients. The inpatient cost difference between transplanted and waitlisted patients quickly disappears after the initial spike. (2) Outpatient cost of transplanted patients is consistently lower than that of untransplanted patients and the difference is stable over the entire lifespan since transplantation. This is primarily due to patients stopping their dialysis services after transplantation because the majority of outpatient spending is dialysis-related, and explains the lower total cost among transplanted patients after the initial spike. Also, this implies that the kidney transplant could significantly reduce the overall lifetime medical cost for ESRD patients who have a relatively long survival after receiving the kidney transplantation. (3) There is no significant difference in skilled nursing care. (4) There is a significantly higher daily spending in home health care during the first a few months after transplantation. (5) There is a significantly higher daily spending among transplanted patients in hospice service before death. Note that although the total daily cost almost equals the sum of these 5 types

of daily costs, the 5 separate groups of curves do not simply add up to their corresponding curves in Figure 3.4 due to the logarithmic transformation.

3.6   Appendix: Summary Statistics of ESRD Data Sets

Here we provide some summary statistics for the covariates and response variables in the analysis data sets with three groups (unwaitlisted, waitlisted, transplanted) of ESRD patients, respectively. The first row contains the mean and the standard deviation (SD) of the within-subject number of selected days. The covariates are divided into four types: subject-level continuous variables (rows 1-5), subject-level categorical variables (rows 6-12), observation-level continuous variables (rows 13-18), and observation-level categorical variables (rows 19-20). For a subject- or observation-level continuous variable, its mean and standard deviation are calculated over all subjects or all observations within a particular group. For a subject- or observation-level categorical variable, the number and percentage of subjects or observations in each category are calculated within a particular group.
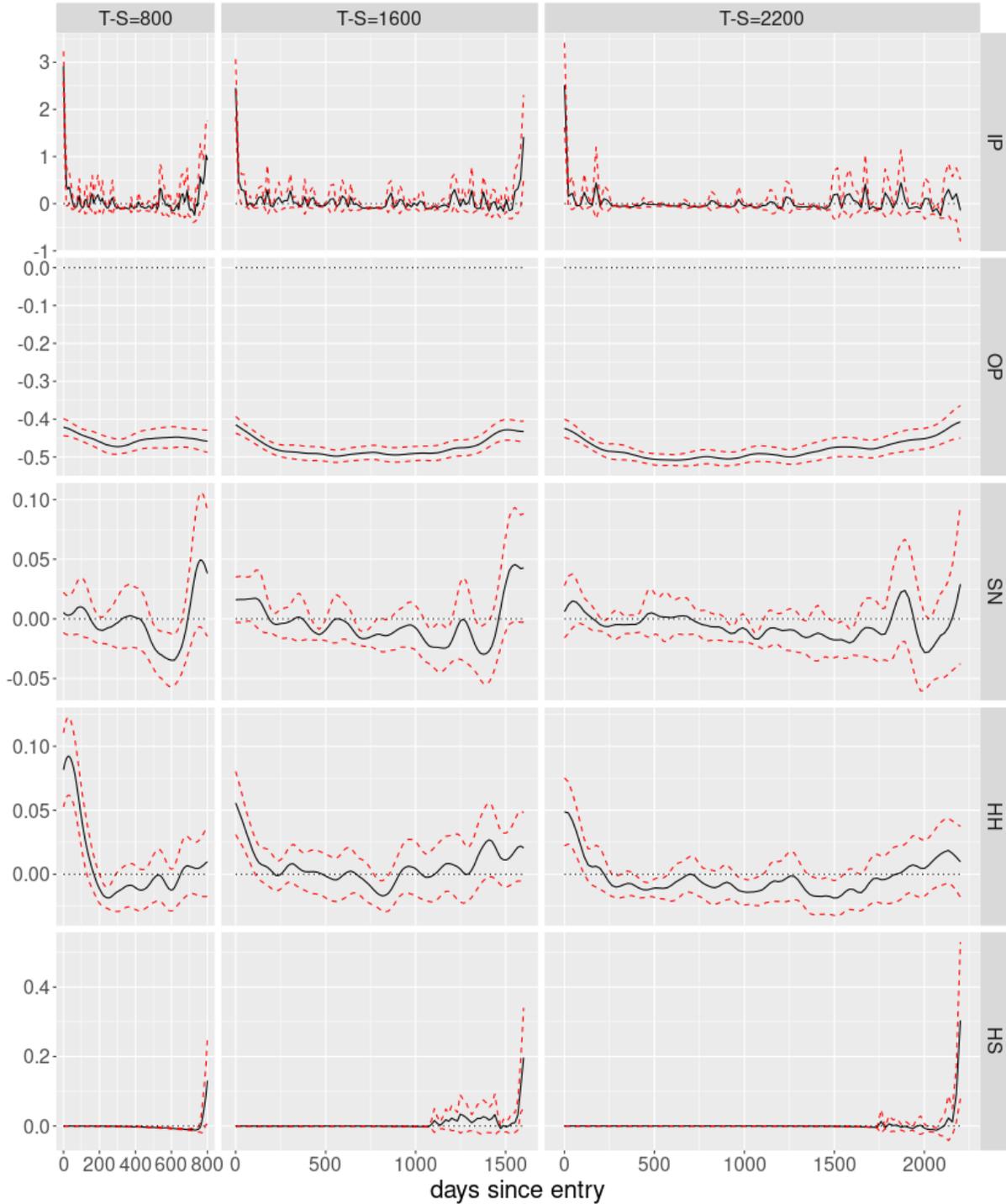
Figure 3.5: $\widehat{\gamma}(t, T - S - t)$ and its confidence bands at $T - S = 800, 1600, 2200$ days for five different types of costs (IP - inpatient, OP - outpatient, SN - skilled nursing, HH - home health, HS - hospice).

row		Unwaitlisted ($n_1 = 4962$)				Waitlisted ($n_2 = 4918$)				Transplanted ($n_3 = 2548$)			
		Mean	SD	Size	Percentage	Mean	SD	Size	Percentage	Mean	SD	Size	Percentage
1	Number of sampled days	80.8	78.6	–	–	152.1	86.6	–	–	187.2	96.6	–	–
2	Number of days from first service to death	914.4	835.2	–	–	1703.6	864.8	–	–	2199.5	916.2	–	–
3	Number of days from first service to transplant	–	–	–	–	–	–	–	–	654.9	599.2	–	–
4	Age at first service (year)	76.4	7.1	–	–	69.3	4.0	–	–	69.6	3.8	–	–
5	BMI	28.1	7.2	–	–	28.6	6.3	–	–	28.5	6.1	–	–
6	Sex: Male	–	–	2684	54.1	–	–	3208	65.2	–	–	1731	67.9
7	Sex: Female	–	–	2278	45.9	–	–	1710	34.8	–	–	817	32.1
8	Race: White	–	–	3750	75.6	–	–	3546	72.1	–	–	2073	81.4
9	Race: Black	–	–	997	20.1	–	–	1020	20.7	–	–	330	13.0
10	Race: Other	–	–	215	4.3	–	–	352	7.2	–	–	145	5.7
11	Hypertension	–	–	4255	85.8	–	–	4345	88.3	–	–	2253	88.4
12	Other comorbidities	–	–	4308	86.8	–	–	3858	78.4	–	–	1747	68.6
13	Daily cost: Inpatient ($)	94.7	507.2	–	–	81.6	535.3	–	–	72.9	493.5	–	–
14	Daily cost: Outpatient ($)	86.4	234.0	–	–	89.4	221.5	–	–	39.9	179.8	–	–
15	Daily cost: Skilled nursing ($)	17.2	82.4	–	–	8.2	60.4	–	–	5.9	53.0	–	–
16	Daily cost: Home health ($)	8.0	25.5	–	–	4.9	21.4	–	–	4.0	19.4	–	–
17	Daily cost: Hospice ($)	1.3	17.9	–	–	0.5	12.5	–	–	0.7	14.0	–	–
18	Daily cost: Total ($)	207.5	557.1	–	–	184.7	577.0	–	–	123.3	525.2	–	–
19	Number of days on HD	–	–	379345	94.6	–	–	650051	86.9	–	–	–	–
20	Number of days on PD	–	–	21480	5.4	–	–	98137	13.1	–	–	–	–

Table 3.4: Summary statistics of the sampled data sets.

78

Chapter 4

A Generalized Product-limit Estimator for Truncated and Censored Data With Terminal Event

4.1 Introduction

In longitudinal cohort studies, participants are oftentimes followed for a period of time, during which multiple events may occur for each person. One common situation features two types of events: a non-terminal event at time S which in many cases is the onset of some disease, and a terminal event at time T which is usually death. An example of the non-terminal event is the onset of Alzheimer's disease. This type of two events data where the terminal event can censor the non-terminal event but not vice versa is referred to as the semi-competing risks (SCR) data.

Under this framework, it is often desirable to understand to what extent the two events are associated and to evaluate how the association varies with some covariates. Two widely used

approaches in the existing literature are: (1) under the semicompeting risks framework, the joint distribution of these events is formulated via a gamma frailty model in the upper wedge where data are observable [6, 14]; and (2) an illness-death model is used with a shared frailty to incorporate the dependence structure [36, 55]. In either case, assumptions about certain semiparametric structure are made to model the joint distribution of the two events, which may lead to model misspecification. Moreover, both approaches assume the independence between the censoring time C and event times (S, T).

We aim to relax those assumptions. To be more specific, we focus on the conditional distribution of S given T that is used as a covariate. This becomes an estimation problem with a censored covariate, for which the complete case analysis is a valid approach. In other words, we remove the record when T is censored and only keep the record in which T is observed, yielding the so-called "complete data". In such a complete data set, it is possible for S to be censored by T, which corresponds to the real life situation where patients die free of disease. Left truncation of the non-terminal event could also happen, under which situation the cases where $S < L$ are not observed. With right censoring and left truncation, the conditional distribution $P(S \leq s | T = t)$ can be estimated by a kernel weighted product-limit estimator widely known as the "Beran's estimator". This type of analysis based on complete data can be shown to only rely on the independence of (L, C) and S conditional on T, thus relaxing the assumption made in the existing literature.

One also needs to develop a reliable bandwidth selection approach for the Beran's estimator, which is crucial for any kernel estimation. [23] gave a theoretical formula for selecting the optimal bandwidth. However, it is extremely difficult to use as it contains multiple unknown functions. [39] proposed to minimize a mean integrated squared error (MISE) approximated by bootstrap resamples. This approach requires a pilot bandwidth. Although the paper provided a formula for it after conducting a preliminary analysis, the formula is not justified universally. In this article, we propose a data-driven bandwidth selection procedure based

on the C-index [18], which requires no pilot bandwidth and is easily implementable and computationally efficient.

The work is organized as follows. In Section 4.2, we will introduce the Beran's estimator, which uses T as a covariate and deals with right censoring and left truncation for S. The Nadaraya-Watson (NW) estimator will serve as a reference for comparison when there is no left truncation. We will introduce a new bandwidth selection approach, which is based on the C-index. A Greenwood variance estimator is also introduced to construct confidence intervals for the Beran's estimator. In Section 4.3, we conduct simulations under several designs to demonstrate the validity of the Beran's estimator and to compare its performance to the NW estimator when there is no left truncation. In Section 4.4, we apply the proposed method to The 90+ Study.

4.2   Estimation under Right Censoring and Left Truncation

We first introduce some notation. Denote the individual non-terminal event time as S_i and the terminal event time as T_i that can censor S_i (denoting $S_i = \infty$ for mathematical convenience in this situation) but not vice versa. Moreover, there exists a random right censoring time, denoted as C_i, which can censor both S_i and T_i or T_i only. There also exists a left truncation time L_i such that an individual is not observed if S_i, T_i or C_i happened before L_i. Denote the distribution functions of S_i, T_i, C_i and L_i as F_S, F_T, F_C and F_L, respectively. We further denote $a_L = \inf\{x : F_L(x) > 0\}$, $a_S = \inf\{x : F_S(x) > 0\}$ and $b_L = \inf\{x : F_L(x) = 1\}$. Under left truncation and right censoring, the observable random

variables are defined as follows:

$$X_i = T_i \wedge C_i, \quad \eta_i = I(T_i \leq C_i), \quad Y_i = S_i \wedge X_i, \quad \delta_i = I(S_i \leq X_i),$$

where $a \wedge b$ represents the minimum of a and b. The observations can thus be described as $\{(L_i, X_i, \eta_i, Y_i, \delta_i) : i = 1, ..., n, Y_i \geq L_i\}$.

For the methods we are going to introduce, we focus on estimating the conditional distribution $P(S \leq s | T = t)$. We denote this distribution by $F(s|t)$. When $a_S < a_L$, $F(s|t)$ is not identifiable because we do not know the probability of S being left truncated. A common practice under this situation is to switch the objective to estimating $P(S \leq s | T = t, S > l)$, where $l > a_L$ [48]. We will denote this conditional distribution as $F(s|t, l)$. Note when $a_S > a_L$ we can choose an l such that $a_S > l$ and then $F(s|t, l)$ becomes the same as $F(s|t)$. In other words, this is the situation that $F(s|t)$ is identifiable.

Since in real life S cannot happen beyond T, $F(s|t, l)$ only has a valid interpretation when $s \leq t$. Thus $F(s|t, l)$ can be constructed by two parts: the first part is a continuous sub-distribution, which stops at t; the second part is a probability mass, which is placed at ∞ by convention. Since the continuous sub-distribution does not reach 1 at t, we can normalize it by $F(t|t, l)$ to make it a valid distribution function. Denote $F(s|t, l)/F(t|t, l)$ by $F_1(s|t, l)$ that can be interpreted as the distribution of S conditional on death at $T = t$ and disease happens after l but before t. On the other hand, the probability mass $1 - F(t|t, l)$, denoted by $F_2(t|l)$, represents the probability of dying at t disease free. Therefore, $F(s|t, l)$ gives both $F_1(s|t, l)$ and $F_2(t|l)$ and vice versa.

4.2.1 A Simple Case Without Left Truncation

We first consider the case that S is never left truncated. In other words, we have $b_L \leq a_S$. In this situation, $F(s|t)$ is identifiable and can be estimated using a simple Nadaraya-Watson estimator:

$$\widehat{F}_{NW}(s|t) = \frac{\sum_{i=1}^{n} I(Y_i \leq s)\eta_i \delta_i K_h(X_i - t)}{\sum_{i=1}^{n} \eta_i K_h(X_i - t)}, \tag{4.1}$$

where $K_h(x) = K(x/h)/h$, K is a kernel function and $h > 0$ is the bandwidth. Throughout this article, we use a truncated Gaussian kernel $K(x) = \exp(-x^2/2)I(|x| < 3)$.

Note that $\widehat{F}_{NW}(s|t)$ only uses the complete data, i.e., $\eta = 1$. It can be shown that $\widehat{F}_{NW}(s|t)$ is a consistent estimator of $P(S \leq s|T = t, T \leq C, T \geq L, S \geq L)$ [9]. If we further assume the independence of S and (L, C) given T, together with $b_L \leq a_S$, we can show that $\widehat{F}_{NW}(s|t)$ converges to $F(s|t)$ in probability.

Also note that we do not require the independence of T and (L, C), which is typically required for the semi-competing risks model under left truncation and right censoring (see e.g. [21]). The reason is that we do not need to estimate the distribution of T. The price we pay is the loss of efficiency as the subset of data with $\eta = 0$ is removed.

4.2.2 A General Case With Left Truncation

When there is left truncation for S, i.e., $a_S < b_L$, we focus on estimating $F(s|t, l)$. As we discussed earlier, a special case is when $a_S > a_L$, where we can choose an l such that $F(s|t, l)$ becomes $F(s|t)$. On the other hand, when $a_S \leq a_L$, $F(s|t)$ cannot be identified and we turn to estimating $F(s|t, l)$. The following discussion does not differentiate between the two situations and focuses on the estimation of $F(s|t, l)$.

Under left truncation, the Nadaraya-Watson estimator (4.1) cannot be used to estimate $F(s|t, l)$. It cannot estimate $F(s|t)$ either because the condition $S \geq L$ in $P(S \leq s|T = t, T \leq C, T \geq L, S \geq L)$ (see [9]) cannot be removed when there is left truncation. For this reason, we resort to generalizing the Beran's estimator that is a kernel weighted product-limit estimator of the conditional survival function given a covariate for right censored data [2]. A generalization of the kernel weighted product-limit estimator for left truncated and right censored survival data given a covariate was considered by [20], where an almost sure asymptotic representation was also established. We generalize the estimator further to the case that T is the covariate that is subject to right censoring, which gives

$$\widehat{F}_{PL}(s|t, l) = 1 - \prod_{l < s_j \leq s} \left(1 - \frac{\sum_{i=1}^{n} I(Y_i = s_j) \eta_i \delta_i K_h(X_i - t)}{\sum_{i=1}^{n} I(L_i \leq s_j \leq Y_i) \eta_i K_h(X_i - t)} \right), \tag{4.2}$$

where the subscript PL stands for "product-limit" and s_j's are the distinct disease onset times. Similar to the Nadaraya-Watson estimator (4.1), $\widehat{F}_{PL}(s|t, l)$ only uses complete data. Under the independence of S and (L, C) given T and other regularity conditions, the asymptotic normality of the generalized Beran's estimator is given by [20]:

$$\sqrt{\sum_{i=1}^{n} \eta_i h} \left[\widehat{F}_{PL}(s|t, l) - F(s|t, l) \right] \xrightarrow{d} N \left(0, \sigma^2(s|t, l) \right), \tag{4.3}$$

where

$$\sigma^2(s|t, l) = \frac{\mu_2}{\mu_1^2} (1 - F(s|t, l))^2 \int_l^s \frac{dH(u|t)}{C^2(u|t)} du / m(t),$$

$$H(s|t) = P(Y \leq s, \delta = 1 | X = t, \eta = 1, Y \geq L),$$

$$C(s|t) = P(L \leq s < Y | X = t, \eta = 1, Y \geq L),$$

$$\mu_1 = \int K(z) dz,$$

$$\mu_2 = \int K^2(z) dz$$

and m is the conditional density of X given $(Y \geq L, \eta = 1)$.

4.2.3 Greenwood Variance Estimator

In this section, we introduce a plug-in variance estimator for the generalized Beran's estimator (4.2), where the asymptotic variance in (4.3) can be estimated by first estimating the unknown components of $\sigma^2(s|t, l)$:

$$\widehat{H}(s|t) = \frac{\sum_{i=1}^{n} I(Y_i \leq s)\delta_i \eta_i K_h(X_i - t)}{\sum_{i=1}^{n} \eta_i K_h(X_i - t)},$$

$$\widehat{C}(s|t) = \frac{\sum_{i=1}^{n} I(L_i \leq s < Y_i)\eta_i K_h(X_i - t)}{\sum_{i=1}^{n} \eta_i K_h(X_i - t)},$$

$$\widehat{m}(t) = \frac{\sum_{i=1}^{n} \eta_i K_h(X_i - t)}{\mu_1 \sum_{i=1}^{n} \eta_i}.$$

Then the asymptotic variance $\sigma^2(s|t, l)$ can be estimated by replacing $F(s|t, l)$, $m(t)$, $H_1(u|t)$ and $C^2(u|t)$ with $\widehat{F}_{PL}(s|t, l)$, $\widehat{m}(t)$, $\widehat{H}_1(u|t)$ and $\widehat{C}(u|t)\widehat{C}(u - |t)$, respectively. The resulting variance estimator is

$$\begin{aligned}
\widehat{Var}&(\widehat{F}_{PL}(s|t, l)) \\
&= \frac{\mu_2}{\sum_{i=1}^{n} \eta_i h \mu_1^2 \widehat{m}(t)} \left(1 - \widehat{F}_{PL}(s|t, l)\right)^2 \int_l^s \frac{d\widehat{H}(u|t)}{\widehat{C}(u|t)\widehat{C}(u - |t)} du \\
&= \left(1 - \hat{F}_{PL}(s|t)\right)^2 \\
&\quad \times \int_l^s \frac{d\sum_{i=1}^{n} 1(Y_i \leq u)\delta_i \eta_i K^*\left(\frac{X_i-t}{h}\right)}{\left[\sum_{i=1}^{n} 1(L_i \leq u < Y_i)\eta_i K^*\left(\frac{X_i-t}{h}\right)\right] \left[\sum_{i=1}^{n} 1(L_i < u \leq Y_i)\eta_i K^*\left(\frac{X_i-t}{h}\right)\right]}, \quad (4.4)
\end{aligned}$$

where $K^*(x) = K(x) \cdot \mu_1/\mu_2$. This formula is the analog of Greenwood's formula, an variance estimator of the Kaplan-Meier estimator, with a weight $K^*\left(\frac{X_i-t}{h}\right)$ for each subject.

4.2.4 Bandwidth Selection

For the bandwidth selection, we adopt a cross-validation (CV) criterion based on Harrell's concordance index, or C-index [18]. C-index is a widely used metric for the evaluation of prediction accuracy of survival models and was first proposed for the Cox model specifically. C-index is defined as the percentage of concordant pairs among comparable pairs within the dataset. A pair of subjects (i, j) is "comparable" if we can determine which of them (i or j) was the first to experience an event from the data. A comparable pair (i, j) is also "concordant" if the subject who experiences the earlier event is identified as having the greater prognostic score, e.g., $\hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$ from the Cox model, and "discordant" otherwise.

The C-index has not been directly used to evaluate the generalized Beran's estimator since the estimator does not generate a "prognostic score". We borrow the concept by defining "concordant" as subject i having a longer predicted survival than subject j with a probability of more than 0.5 when subject i indeed outlives subject j. Thus the redefined C-index can be expressed as

$$\text{C-index} = \frac{\sum_{i,j} \delta_i I(Y_i < Y_j) I(\hat{p}_{ji} > 0.5)}{\sum_{i,j} \delta_i I(Y_i < Y_j)}, \tag{4.5}$$

where i and j satisfy $\eta_i = \eta_j = 1$ and $Y_i, Y_j > l$. The numerator and the denominator represent the number of concordant and the number of comparable pairs, respectively. The predicted probability of subject j staying disease-free longer than subject i, $\hat{p}_{ji}$, is calculated in the following:

$$\hat{p}_{ji} \stackrel{d}{=} \widehat{P}(S_j > S_i | T_i = t_i, T_j = t_j, S_i > l, S_j > l)$$
$$= \int_l^\infty \left(1 - \widehat{F}_{PL}(u|t_i, l)\right) d\widehat{F}_{PL}(u|t_j, l) + 0.5 \left(1 - \widehat{F}_{PL}(\infty|t_i, l)\right) \left(1 - \widehat{F}_{PL}(\infty|t_j, l)\right).$$

Since $\widehat{F}_{PL}$ is a step function, the integral is in fact evaluated as a finite sum. It is well known that when the last event is censored, the Kaplan-Meier estimator does not go to 0, which is a property shared by product-limit estimators including the generalized Beran's estimator. When this happens, we assign the remaining probability mass to infinity, i.e., $\widehat{P}(S = \infty | T = t, S > l) = \lim_{s \to \infty} (1 - \widehat{F}_{PL}(s|t, l))$, or simply $1 - \widehat{F}_{PL}(\infty|t, l)$. Note that when both S_i and S_j are estimated to take infinite value with positive probabilities, they are treated as equal and thus only half of multiplied probabilities is added to $\hat{p}_{ij}$. Mathematically, adding the term ensures $\hat{p}_{ij} + \hat{p}_{ji} = 1$.

For bandwidth selection, we consider using the revised C-index (4.5) in each fold of the cross-validation (CV). The following criterion is defined for K-fold CV:

$$CV(h) = \frac{\sum_{k=1}^{K} \sum_{i,j \in \mathcal{C}_k} \delta_i I(Y_i < Y_j) I(\hat{p}_{ji}^{(-k)} > 0.5)}{\sum_{k=1}^{K} \sum_{i,j \in \mathcal{C}_k} \delta_i I(Y_i < Y_j)},$$

where $\mathcal{C}_k$'s are disjoint and $\cup_k \mathcal{C}_k = \mathcal{C}$ is index set of all complete cases with $Y > l$. Note that when evaluating on the k-th fold, all the other $K - 1$ folds are used to calculate $\hat{p}_{ji}^{(-k)}$. The numbers of concordant and comparable pairs across all folds are then summed up to form the numerator and the denominator of the CV criterion, respectively.

4.3   Simulation Study

In this section, we conduct simulations that mimic The 90+ Study data to examine the performance of estimators (4.1) and (4.2) respectively. We first generate T^*, L^* and C^* from a zero-mean three-dimensional multivariate normal distribution. The covariance matrix of the multivariate normal distribution has diagonal elements 1 and off-diagonal elements 0.5.

Then generate T, L and C as follows:

$$T = 20\Psi^{-1}\left(\Phi(T^*)\right) + 90, \quad L = 10\Phi(L^*) + a_L, \quad C = 20\Psi^{-1}\left(\Phi(C^*)\right) + 100,$$

where Φ and Ψ are the cumulative distribution functions of standard normal distribution and exponential distribution with rate parameter 1, respectively, and the constant a_L determines the lower bound of the left truncation time distribution. Thus marginally T and C follow shifted exponential distributions and L follows a uniform distribution. Such a design allows us to generate correlated T, L and C. Lastly we generate S by adding a constant 80 to randomly generated values from an exponential distribution with rate $2/(T - 70)$.

For such generated data we have $a_S = 80$ and $b_L = a_L + 10$. Here we consider three different values of a_L: 69, 79 and 89. For each setting we visualize our estimators for three different T values: 95, 100 and 110. These death times are roughly the quartiles of the distribution of T. In the first setting where $a_L = 69$, we have $b_L < a_S$ which yields data without the occurrence of left truncation. From our previous discussion, both estimators (4.1) and (4.2) can be used to estimate $F(s|t)$. In the second setting where $a_L = 79$, we have $a_L < a_S < b_L$. In this case only estimator (4.2) can be used to estimate $F(s|t)$ since estimator (4.1) is biased. In the third setting where $a_L = 89$, we have $a_S < a_L$, which means $F(s|t)$ is not identifiable anymore. We thus switch target to estimating $F(s|t,l)$ for $l = 89.1$ using estimator (4.2). For each of the three settings we conduct a simulation study to illustrate these aforementioned points. Further in the third setting, we will illustrate the validity of the confidence intervals constructed using the Greenwood variance estimator. Under each setting, we simulate 100 datasets, each containing 1000 subjects. We then apply left truncation to the datasets and use the bandwidth selection approach we introduced earlier. The estimators calculated using the undersmoothed bandwidths will be displayed. Specifically, we undersmooth the CV-selected bandwidths by $n^{0.05}$, where n is the average number of complete cases under each simulation setting.
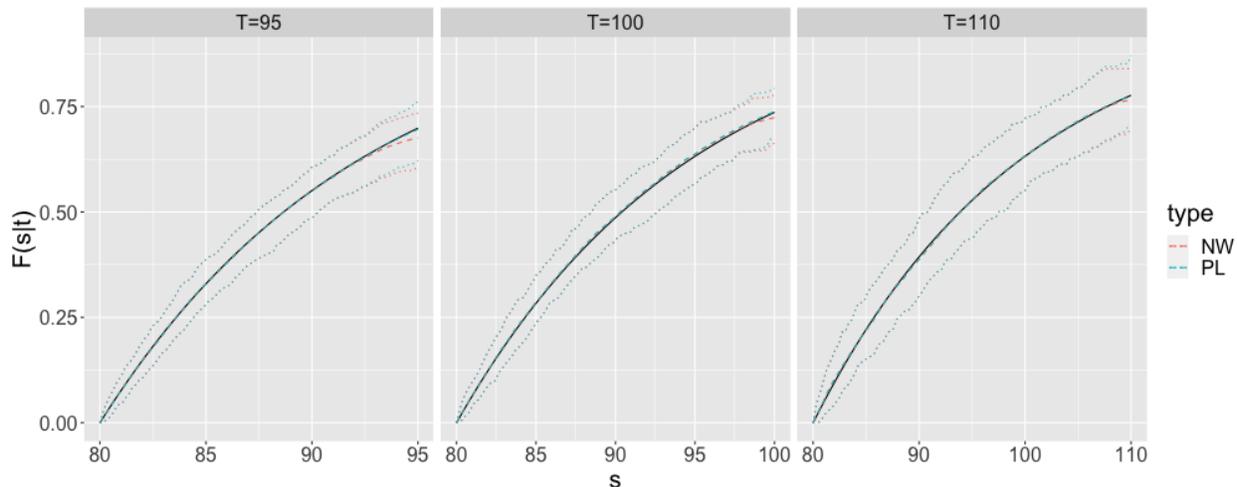
Figure 4.1: Means and confidence bands for $\widehat{F}_{NW}(s|t)$ and $\widehat{F}_{PL}(s|t)$ under $T = 95$, 100 and 110.

4.3.1 First Design: $a_L = 69$

In the first design, we examine the two estimators $\widehat{F}_{NW}(s|t)$ and $\widehat{F}_{PL}(s|t)$ as well as their variabilities. In Figure 4.1, each of the three subfigures corresponds to a specific value of t. Within each subfigure, the black curve represents the true values of $F(s|t)$. The red and blue dashed curves represent the means of $\widehat{F}_{NW}(s|t)$ and $\widehat{F}_{PL}(s|t)$ of the 100 replications, respectively. The confidence bands are the 2.5% and 97.5% quantiles of the estimates. Under this setting, the two estimators perform very similarly in terms of variability and both have negligible biases.

4.3.2 Second Design: $a_L = 79$

In the second design, we present the same set of subfigures as in the first setting. It can be seen that the two estimators have very similar varibilities for all values of s. However, in terms of bias, $\widehat{F}_{NW}(s|t)$ completely misses the truth.
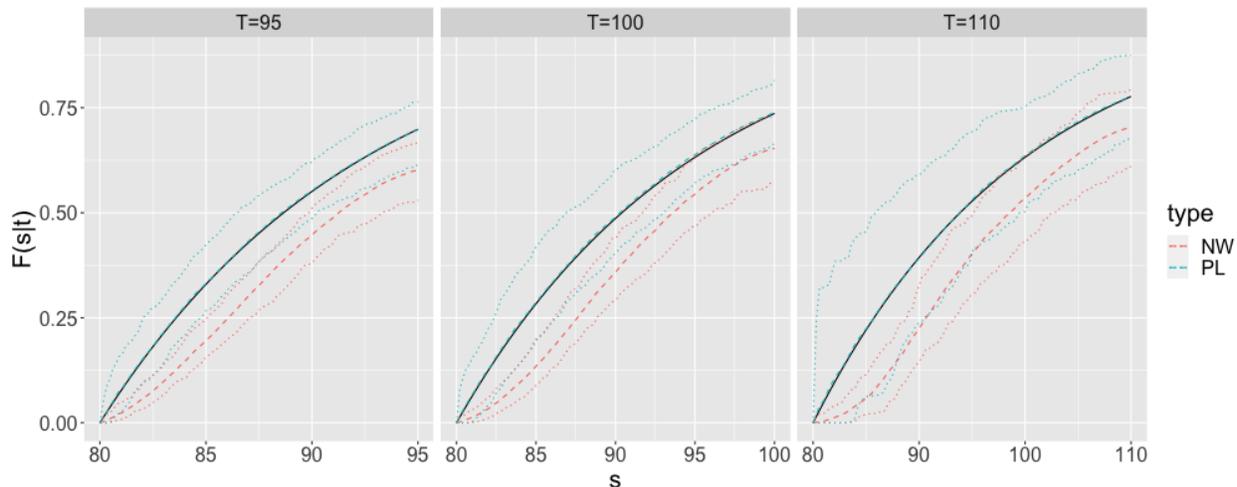
Figure 4.2: Means and confidence bands for $\widehat{F}_{NW}(s|t)$ and $\widehat{F}_{PL}(s|t)$ under $T = 95$, 100 and 110.

4.3.3   Third Design: $a_L = 89$

In the third design, instead of estimating $F(s|t)$, we switch to estimating $F(s|t, l)$ where we choose $l = 89.1$ to meet the requirement of $l > a_L$. From Figure 4.3 we see that the means of $\widehat{F}_{PL}(s|t, l)$ are very close to their true values with negligible biases.

We present the coverage probabilities of confidence intervals calculated from the Greenwood's formula (4.4) in Figure 4.4. The nominal level of 0.95 is shown in the figure as the dotted horizontal line. Across all subfigures, coverage probabilities start from 0 at $s = l$ and then quickly rise to around 0.95. The low coverage towards the left end is not unique to the generalized Beran's estimator which was reported before for the Kaplan-Meier estimator [13]. This can be seen easily prior to the first observed event time, where $\widehat{F}_{PL}(s|t, l) = 0$ with a confidence interval $[0, 0]$. Clearly the confidence interval fails to cover the true value $F(s|t, l)$ as long as $F(s|t, l)$ is non-zero. The same issue can happen during the first a few observed event times. Ad hoc fixes were proposed to resolve this issue, e.g., [1] and [13]. However, since our primary objective is to obtain valid estimator $\widehat{F}_{PL}(s|t, l)$, we do not pursue those methods in this work.
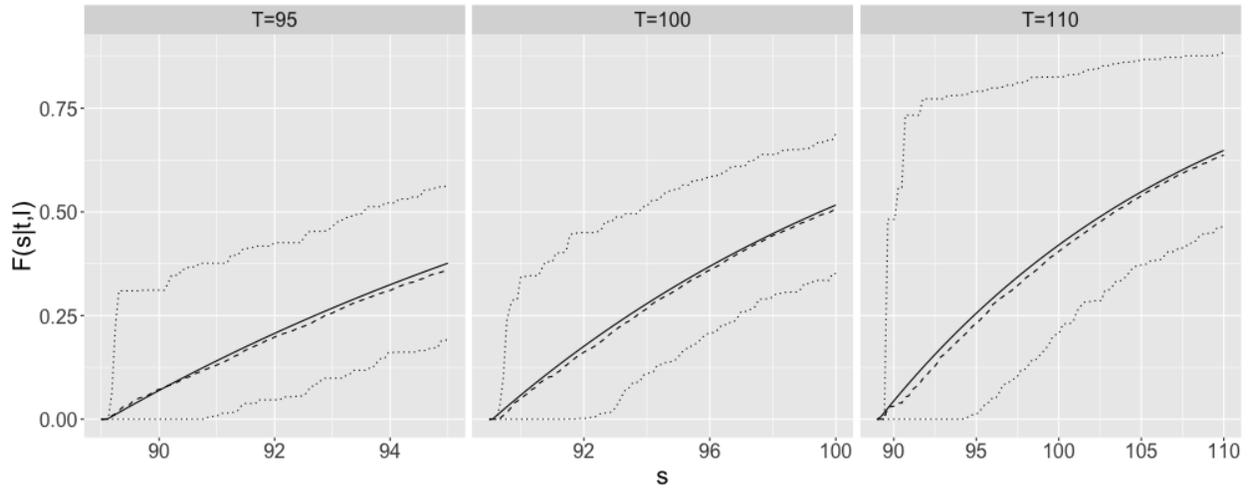
Figure 4.3: Mean and confidence bands for $\widehat{F}_{PL}(s|t,l)$ under $T = 95$, 100 and 110.
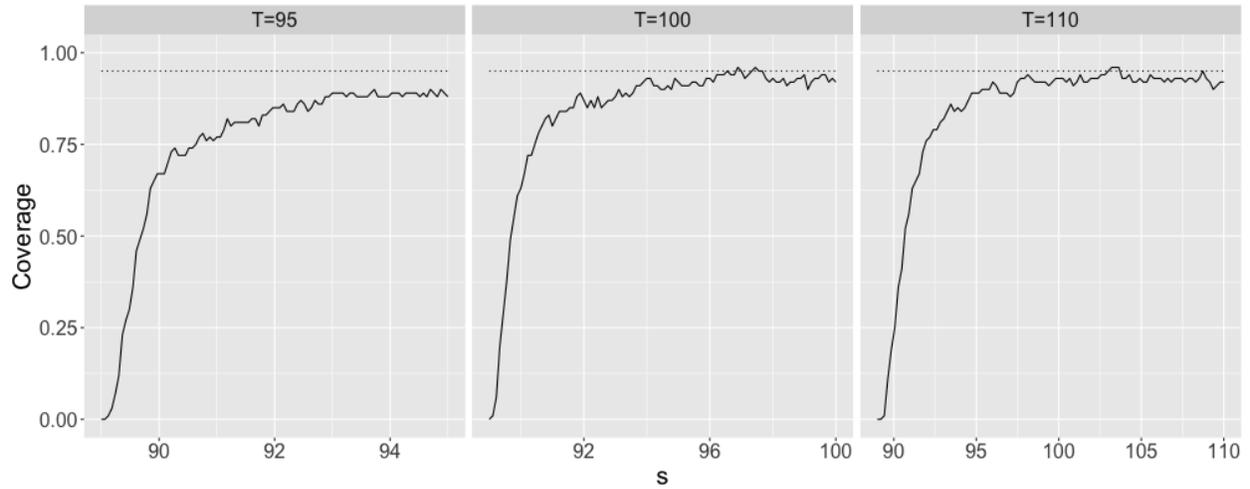
Figure 4.4: Coverage probabilities of Greenwood confidence intervals for $\widehat{F}_{PL}(s|t,l)$ under $T = 95$, 100 and 110.

4.4   The 90+ Study

We apply our approach to the data obtained from The 90+ Study. See [5] for a detailed description of the study. The participants were from two cohorts, one is the Leisure World Cohort Study [38], where everyone was recruited from a retirement community named Leisure World, and one consisted of volunteers recruited later. As of April 2023, the number of participants reached 921 after excluding those with no initial or follow-up evaluation and those already demented at enrollment. For this analysis, we only use data from participants who are already deceased. A breakdown by cohort and gender of the deceased participants are displayed in Table 4.1.

	Male	Female	Total
Leisure World	166	416	582
Volunteer	94	103	197
Total	260	519	779

Table 4.1: Number of participants by gender and cohort.

In Figure 4.5, we provide the estimated lifetime dementia distribution for all participants as well as the four groups stratified by gender and cohort. The estimator is calculated using the undersmoothed bandwidths (by a factor of $779^{0.05}$). Three different death ages are conditioned on: $T = 95$, 100 and 103. It can be seen that the dementia distributions are approximately uniform, particularly for all participants, given that dementia does happen in the lifetime.

We further display the lifetime dementia probability as a function of death age in Figure 4.6. It can be seen that male participants from the volunteer cohort has a relatively low lifetime dementia probability compared to all other groups conditional on death around age 96. However, the difference tends to vanish for the oldest old (more than 100 years old). Specifically, the lifetime dementia probabilities conditional on death age at 92, 94, 96, 98, 100, 102 are 0.13, 0.33, 0.49, 0.61, 0.70, 0.81, respectively. This shows that dementia is
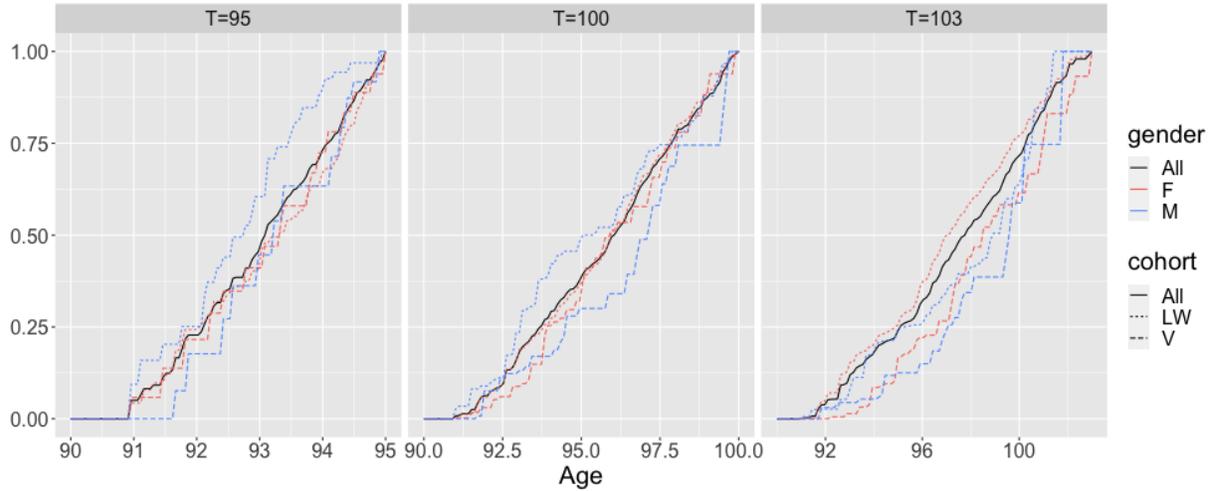
Figure 4.5: Estimated dementia onset distributions for all The 90+ Study participants and within four stratified groups who developed dementia during lifetime conditional on $T = 95$, 100 and 103.

extremely prevalent among the oldest old.

To summarize, the probability of having dementia during lifetime is increasing when people live longer; Among those who had developed dementia, the onset of the disease can happen at any age equally likely.
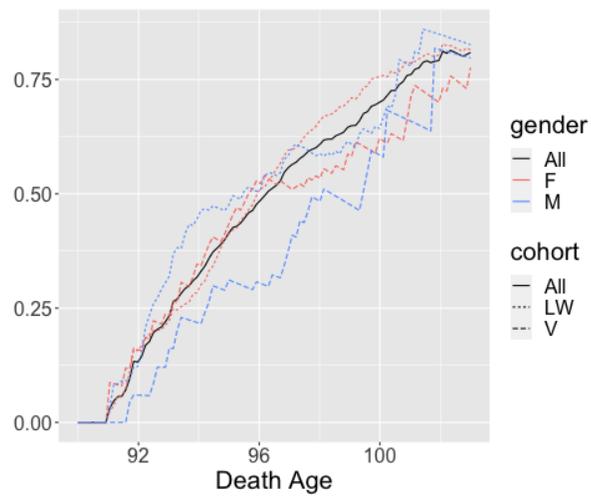
Figure 4.6: Lifetime dementia probabilities for all The 90+ Study participants and within four stratified groups conditional on $T = 95$, 100 and 103.

Bibliography

[1] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer US, New York, NY, 1993.

[2] R. Beran. Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley, 1981.

[3] D. Cameron, J. Ubels, and F. Norström. On What Basis Are Medical Cost-effectiveness Thresholds Set? Clashing Opinions and an Absence of Data: A Systematic Review. *Global health action*, 11(1):1447828, 2018.

[4] K. C. G. Chan and M.-C. Wang. Backward Estimation of Stochastic Processes With Failure Events as Time Origins. *The Annals of Applied Statistics*, 4(3):1602, 2010.

[5] M. M. Corrada, R. Brookmeyer, A. Paganini-Hill, D. Berlau, and C. H. Kawas. Dementia incidence continues to increase with age in the oldest old: The 90+ study. *Annals of Neurology*, 67(1):114–121, Jan. 2010.

[6] R. DAY, J. BRYANT, and M. LEFKOPOULOU. Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. *Biometrika*, 84(1):45–56, Mar. 1997.

[7] W. Dempsey and P. McCullagh. Survival Models and Health Sequences. *Lifetime Data Analysis*, 24(4):550–584, 2018.

[8] M. F. Drummond, M. J. Sculpher, K. Claxton, G. L. Stoddart, and G. W. Torrance. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford university press, 2015.

[9] U. Einmahl and D. M. Mason. An Empirical Process Approach to the Uniform Consistency of Kernel-Type Function Estimators. *Journal of Theoretical Probability*, 13(1):1–37, Jan. 2000.

[10] U. Einmahl and D. M. Mason. Uniform in Bandwidth Consistency of Kernel-Type Function Estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

[11] J. Fan, N. E. Heckman, and M. P. Wand. Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. *Journal of the American Statistical Association*, 90(429):141–150, Mar. 1995.

[12] J. Fan, T. Huang, and R. Li. Analysis of Longitudinal Data With Semiparametric Estimation of Covariance Function. *Journal of the American Statistical Association*, 102(478):632–641, 2007.

[13] M. P. Fay, E. H. Brittain, and M. A. Proschan. Pointwise confidence intervals for a survival distribution with small samples or heavy censoring. *Biostatistics*, 14(4):723–736, Sept. 2013.

[14] J. P. Fine, H. Jiang, and R. Chappell. On semi-competing risks data. *Biometrika*, 88(4):907–919, Dec. 2001.

[15] R. Fu, N. Sekercioglu, W. Berta, and P. C. Coyte. Cost-effectiveness of Deceased-donor Renal Transplant Versus Dialysis to Treat End-stage Renal Disease: A Systematic Review. *Transplantation direct*, 6(2):e522, 2020.

[16] D. Ghosh and D. Y. Lin. Marginal Regression Models for Recurrent and Terminal Events. *Statistica Sinica*, 12(3):663–688, 2002.

[17] P. Hall. Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density. *The Annals of Statistics*, 20(2):675–694, 1992.

[18] F. E. Harrell, Jr, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, May 1982.

[19] D. R. Hoover, J. A. Rice, C. O. Wu, and L.-P. Yang. Nonparametric Smoothing Estimates of Time-varying Coefficient Models With Longitudinal Data. *Biometrika*, 85(4):809–822, 1998.

[20] C. Iglesias Pérez and W. González Manteiga. Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *Journal of Nonparametric Statistics*, 10(3):213–244, Jan. 1999. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10485259908832761.

[21] H. Jiang, J. P. Fine, and R. Chappell. Semiparametric Analysis of Survival Data with Left Truncation and Dependent Right Censoring. *Biometrics*, 61(2):567–575, June 2005.

[22] J. D. Kalbfleisch, D. E. Schaubel, Y. Ye, and Q. Gong. An Estimating Function Approach to the Analysis of Recurrent and Terminal Events. *Biometrics*, 69(2):366–374, 2013.

[23] I. V. Keilegom, M. G. Akritas, and N. Veraverbeke. Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics & Data Analysis*, 35(4):487–500, Feb. 2001.

[24] S. Kong, B. Nan, J. D. Kalbfleisch, R. Saran, and R. Hirth. Conditional Modeling of Longitudinal Data With Terminal Event. *Journal of the American Statistical Association*, 113(521):357–368, 2018.

[25] L. Li, C.-H. Wu, J. Ning, X. Huang, Y.-C. T. Shih, and Y. Shen. Semiparametric Estimation of Longitudinal Medical Cost Trajectory. *Journal of the American Statistical Association*, 113(522):582–592, 2018.

[26] Z. Li, H. Frost, T. D. Tosteson, L. Zhao, L. Liu, K. Lyons, H. Chen, B. Cole, D. Currow, and M. Bakitas. A Semiparametric Joint Model for Terminal Trend of Quality of Life and Survival in Palliative Care Research. *Statistics in Medicine*, 36(29):4692–4704, 2017.

[27] Z. Li, T. D. Tosteson, and M. A. Bakitas. Joint Modeling Quality of Life and Survival Using a Terminal Decline Model in Palliative Care Studies. *Statistics in Medicine*, 32(8):1394–1406, 2013.

[28] X. Lin and R. J. Carroll. Nonparametric Function Estimation for Clustered Data When the Predictor is Measured without/with Error. *Journal of the American statistical Association*, 95(450):520–534, 2000.

[29] L. Liu, R. A. Wolfe, and J. D. Kalbfleisch. A Shared Random Effects Model for Censored Medical Costs and Mortality. *Statistics in Medicine*, 26(1):139–155, 2007.

[30] X. Lu, B. Nan, P. Song, and M. Sowers. Longitudinal Data Analysis With Event Time as a Covariate. *Statistics in Biosciences*, 2(1):65–80, 2010.

[31] G. Machnicki, L. Seriai, and M. A. Schnitzler. Economics of Transplantation: A Review of the Literature. *Transplantation reviews*, 20(2):61–75, 2006.

[32] D. A. Mandelbrot, A. Fleishman, J. R. Rodrigue, S. P. Norman, and M. Samaniego. Practices in the Evaluation of Potential Kidney Transplant Recipients Who Are Elderly: A Survey of US Transplant Centers. *Clinical transplantation*, 31(10):e13088, 2017.

[33] A. Massie, X. Luo, E. Chow, J. Alejo, N. Desai, and D. Segev. Survival Benefit of Primary Deceased Donor Transplantation With High-KDPI Kidneys. *American journal of transplantation*, 14(10):2310–2316, 2014.

[34] M. A. McAdams-DeMarco, S. E. V. P. Rasmussen, N. M. Chu, D. Agoons, R. F. Parsons, T. Alhamad, K. L. Johansen, S. G. Tullius, R. Lynch, M. N. Harhay, et al. Perceptions and Practices Regarding Frailty in Kidney Transplantation: Results of a National Survey. *Transplantation*, 104(2):349, 2020.

[35] J. Menzin, L. M. Lines, D. E. Weiner, P. J. Neumann, C. Nichols, L. Rodriguez, I. Agodoa, and T. Mayne. A Review of the Costs and Cost Effectiveness of Interventions in Chronic Kidney Disease: Implications for Policy. *Pharmacoeconomics*, 29(10):839–861, 2011.

[36] G. G. Nielsen, R. D. Gill, P. K. Andersen, and T. I. A. Sørensen. A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. *Scandinavian Journal of Statistics*, 19(1):25–43, 1992. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

[37] D. Nolan and D. Pollard. U-Processes: Rates of Convergence. *The Annals of Statistics*, pages 780–799, 1987.

[38] A. Paganini-Hill, R. K. Ross, and B. E. Henderson. Prevalence of chronic disease and health practices in a retirement community. *Journal of Chronic Diseases*, 39(9):699–707, Jan. 1986.

[39] R. Peláez, R. Cao, and J. M. Vilar. Bootstrap Bandwidth Selection and Confidence Regions for Double Smoothed Default Probability Estimation. *Mathematics*, 10(9):1523, Jan. 2022. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[40] D. Pollard. Uniform Ratio Limit Theorems for Empirical Processes. *Scandinavian Journal of Statistics*, pages 271–278, 1995.

[41] D. Rizopoulos. *Joint Models for Longitudinal and Time-to-event Data: With Applications in R*. CRC press, 2012.

[42] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the american statistical association*, 90(429):106–121, 1995.

[43] S.-L. Shen, J.-L. Cui, C.-L. Mei, and C.-W. Wang. Estimation and Inference of Semi-varying Coefficient Models With Heteroscedastic Errors. *Journal of Multivariate Analysis*, 124:70–93, 2014.

[44] M. Tonelli, N. Wiebe, G. Knoll, A. Bello, S. Browne, D. Jadhav, S. Klarenbach, and J. Gill. Systematic Review: Kidney Transplantation Compared With Dialysis in Clinically Relevant Outcomes. *American journal of transplantation*, 11(10):2093–2109, 2011.

[45] A. A. Tsiatis and M. Davidian. Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. *Statistica Sinica*, 14(3):809–834, 2004.

[46] United States Renal Data System. *Researcher's Guide to the USRDS Database*. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2019.

[47] United States Renal Data System. *USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States*. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2022.

[48] M.-C. Wang, N. P. Jewell, and W.-Y. Tsai. Asymptotic Properties of the Product Limit Estimate Under Random Truncation. *The Annals of Statistics*, 14(4), Dec. 1986.

[49] N. Wang. Marginal Nonparametric Kernel Regression Accounting for Within-subject Correlation. *Biometrika*, 90(1):43–52, 2003.

[50] Y. Wang, B. Nan, and J. D. Kalbfleisch. Kernel Estimation of Bivariate Time-Varying Coefficient Model for Longitudinal Data with Terminal Event. *Journal of the American Statistical Association*, (in press), 2023.

[51] A. M. Whelan, K. L. Johansen, T. Copeland, C. E. McCulloch, D. Nallapothula, B. K. Lee, G. R. Roll, M. R. Weir, D. B. Adey, and E. Ku. Kidney Transplant Candidacy Evaluation and Waitlisting Practices in the United States and Their Association With Access to Transplantation. *American Journal of Transplantation*, 22(6):1624–1636, 2022.

[52] R. Wolfe, K. McCullough, and A. Leichtman. Predictability of Survival Models for Waiting List and Transplant Patients: Calculating LYFT. *American Journal of Transplantation*, 9(7):1523–1527, 2009.

[53] R. A. Wolfe, V. B. Ashby, E. L. Milford, A. O. Ojo, R. E. Ettenger, L. Y. Agodoa, P. J. Held, and F. K. Port. Comparison of Mortality in All Patients on Dialysis, Patients on Dialysis Awaiting Transplantation, and Recipients of a First Cadaveric Transplant. *New England journal of medicine*, 341(23):1725–1730, 1999.

[54] C. O. Wu, C.-T. Chiang, and D. R. Hoover. Asymptotic Confidence Regions for Kernel Smoothing of a Varying-coefficient Model With Longitudinal Data. *Journal of the American statistical Association*, 93(444):1388–1402, 1998.

[55] J. Xu, J. D. Kalbfleisch, and B. Tai. Statistical Analysis of Illness-Death Processes and Semicompeting Risks Data. *Biometrics*, 66(3):716–725, Sept. 2010.

[56] W. Zhang, S.-Y. Lee, and X. Song. Local Polynomial Fitting in Semivarying Coefficient Model. *Journal of Multivariate Analysis*, 82(1):166–188, 2002.