Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jing Yang Date

Flexible Association Methods for Bivariate Survival Data

By

Jing Yang
Doctor of Philosophy

Biostatistics

Limin Peng, Ph.D.
Advisor

David H. Howard, Ph.D.
Committee Member

Yijian Huang, Ph.D.
Committee Member

Amita K. Manatunga, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Flexible Association Methods for Bivariate Survival Data

By

Jing Yang
B.S., Beijing Normal University, 2009
M.S., Texas Tech University, 2011

Advisor: Limin Peng, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy
in Biostatistics
2016

Abstract

Flexible Association Methods for Bivariate Survival Data

By Jing Yang

Biomedical follow-up studies often involve multiple event times. The interrelationship among these event times is often of great scientific interest. In this dissertation, we focus on two scenarios involving multiple event times, semi-competing risks (Fine et al., 2001) and recurrent events.

The first project is to study the dependence structure between the nonterminal event and the terminal event in the semi-competing risks setting. We propose a new robust dependence measure without requiring distributional assumption, which can accommodate the exploration of the potential changing pattern of the dependence in the identifiable region of semi-competing risks data. We develop a nonparametric estimation procedure for the proposed measure by adopting a quantile regression framework. The estimation method can be readily extended to adjust for covariates. The proposed methods are evaluated by extensive simulation studies and an application to the Denmark diabetes registry data.

The second project is to develop a new nonparametric estimator of the dependence measure proposed in the first project. The new estimator can accommodate left truncation that occurs in semi-competing risks settings, requiring weaker constraints on the truncation mechanism. Asymptotic properties and inference procedures are established for the resulting estimator. We conduct simulation studies to assess the finite-sample performance of the new estimator. We also apply it to a Denmark diabetes registry dataset.

The third project is to explore the association between bivariate recurrent event processes under an observation window structure, which is motivated by the US Cystic Fibrosis Foundation Patient Registry (CFFPR) study. We propose a novel measure which can flexibly depict the association between two recurrent event processes. We further develop a regression framework for the proposed measure to allow for assessing whether and how the association is influenced by covariates. We establish the estimation procedure, which show promising results by some preliminary simulation studies. We also apply the proposed method to the CFFPR study.

Flexible Association Methods for Bivariate Survival Data

By

Jing Yang
B.S., Beijing Normal University, 2009
M.S., Texas Tech University, 2011

Advisor: Limin Peng, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy
in Biostatistics
2016

Acknowledgments

I would like to sincerely thank my advisor, Dr. Limin Peng, for her guidance, advice, and patience during my Ph.D. study. She has provided tremendous help, not only to my research by her expertise, but also to my personal development over the past five years.

I would also like to thank Dr. Yijian Huang, Dr. Amita Manatunga and Dr. David Howard for their time and effort in serving as my committee members. They provided very helpful suggestions that lead to substantial improvement in this dissertation.

I am also grateful to Kirk Easley for various consulting opportunities. I would like to extend my appreciation to all faculty and staff members in the Department of Biostatistics and Bioinformatics at Emory, and to all my friends, with whom I shared many happy moments over the past five years.

Finally, I would like to deeply thank my family for their unwavering support and unconditional love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

In biomedical follow-up studies, subjects may experience multiple events, which are often monitored for studying certain disease. The events can be of the same type, for example, a sequence of tumor recurrences or infection episodes. They can also be of different types, for example, complications causing diseases in different organs. The inter-relationship among these disease-related events often carry important information that can advance the understanding of disease progression. Thus, how to well assess the interplay among these events is of great scientific interest. In my dissertation, we focus on two scenarios involving multiple event times, semi-competing risks and recurrent events.

Semi-competing risks, termed by Fine et al. (2001), is a special structure of bivariate event times that consist of a nonterminal event (e.g., disease landmark) and a terminal event (e.g., death), with the characteristics that time to the nonterminal event can be censored by time to the terminal event, but not vice versa. The dependence between the nonterminal event and the terminal event can offer valuable insight on disease prognosis and thus poses an important problem to study. In the Denmark diabetes registry study (Andersen et al., 1993), for example, investigators have been interested in knowing how diabetic nephropathy (an indicator of kidney failure) influences mortality of diabetes, where time to diabetic nephropathy and time to death form a semi-competing risks structure. To address such an issue, as elaborated later in Section 1.2.1, a typical way is to employ a copula model linking the joint distribution and the marginal distributions of the two event times and let the association parameter capture the dependence structure. However, a main limitation of using such a copula based approach is that the dependence structure relies on the assumed relation between the joint distribution and its marginal distributions, which may be hard to verify based on the observed semi-competing risks data. This motives us to propose a new robust measure without any distributional assumption to capture

the dependence structure between the nonterminal event and the terminal event in the semi-competing setting. Also, left truncation on the terminal event is often encountered in observational studies. For example, only patients who lived long enough to enter the registry can provide data for the Denmark diabetes registry study. This thus motivates us to develop methods that can accommodate left truncation.

Recurrent event data arise when the event of interest occurs repeatedly. Examples include repeated asthmatic attacks, recurrent infections and repeated hospitalizations. Often, a subject may experience more than one type of recurrent events, and the observation of these events are subject to an observation window that is from the start of follow-up to the last follow-up visit. In this setting, our interest is to assess the association between two recurrent event processes under a general window observation scheme. A motivating example is the US Cystic Fibrosis Foundation Patient Registry (CFFPR) study. Cystic Fibrosis (CF) is a lethal autosomal disease without known cure yet that commonly affects Caucasians due to mutation of CFTR gene. Pseudomonas aeruginosa (Pa) and Staphylococcus aureus (Sa) are two major pathogens of medical concerns for CF patients, and are often found to co-exist in the same niche influencing the CF pathogenesis. Recurrences of one type of pathogens may affect the risk of the other type, and thus there is an interest in the interplay occurring between the two. Investigators hope to know, for example, whether early recurrences of Sa infection would postpone the recurrences of Pa infection and how the interplay would be influenced when risk factors are involved. In CFFPR, not all CF children entered the registry right after birth. In fact, a large proportion of CF children delayed their entries due to late diagnosis of CF or some other reasons. As a result, the observations of Pa infection and Sa infection started from subject's first CFFPR visit and continued until the most recent follow-up. In other words, observations of recurrences of Pa infection and Sa infection are subject to an observation window. Without available records of Pa infection or Sa infection before registry entry, the nonzero lower bound

of the observation window can potentially complicate the analyses. To best of our knowledge, little has been done in literature to handel the association between two different types of recurrent events under an observation window structure. Therefore, we hope to fill in this gap by proposing a novel association measure and developing a regression framework for the new measure.

Throughout this dissertation research, we focus on developing methods to address the problems stated above for the semi-competing risks setting and the recurrent event setting. In the rest of this chapter, we present literature review separately on methods that study the dependence for semi-competing risks data and association for bivariate recurrent events data. An outline of this dissertation is given at the end of this chapter.

1.2 Literature Review

1.2.1 Existing work on dependence for semi-competing risks data

Let T_1 denote time to nonterminal event and T_2 denote time to terminal event.

In the literature tailored to semi-competing risks data, the dependence between the nonterminal event and the terminal event is often captured by the association parameter of a copula function, where the copula model is assumed for the joint distribution of (T_1, T_2) on the upper wedge $T_1 \leq T_2$. Fine et al. (2001) posited the Clayton (1978) copula and derived a closed-form estimator for the association parameter from a concordance estimating equation, which was determined as the ratio of concordant to discordant pairs. Their idea was based on the fact that the cross-ratio function was equal to the association parameter for the Clayton copula (Oakes, 1989). Wang (2003) subsequently studied the degree of dependence under a more general class of Archimedean copulas and suggested several estimating functions

for the association parameter. Lakhal et al. (2008) provided a general method for estimating the association parameter for Archimedean copulas. They also showed that the estimating functions provided by Fine et al. (2001) and Wang (2003) were their special cases.

For regression modelling, Ghosh (2006) extended the method of Fine et al. (2001) to association estimation across strata of one discrete covariate. Peng and Fine (2007) linked the joint distribution of (T_1, T_2) to its marginals through a known time-independent copula function but with an unknown time-varying association parameter, which accommodated more realistic scenarios that the dependence between T_1 and T_2 may change over time. Hsieh et al. (2008) generalized the method of Wang (2003) with covariates. Their approach allowed association parameter to vary in different subgroups, but required that covariates only took discrete values. More recently, Chen (2012) studied a nonparametric maximum likelihood approach under a general specification of the copula model.

While modeling the dependence structure between T_1 and T_2 based on a copula model is intuitive and useful, such an approach can impose some implicit limitations that may often be ignored. For example, it may be hard to verify the assumed relationship between the joint distribution of (T_1, T_2) and its marginal distributions, particularly with the observed semi-competing risks data. A similar problem also lies in the work of Shen and Thall (1998), in which a bivariate generalized von Morgenstern distribution that characterized the dependence by a single parameter was assumed. In addition, the interpretation of a copula parameter, constant or time-dependent, relies on the selection of the copula function. When there are covariates involved, a copula based approach is further prone to issues due to potential misspecifications of the marginal regression models for T_1 and T_2. All these considerations constitute the motivations of our first project, which proposes a new robust measure for the dependence structure between the nonterminal event and the terminal event

in the semi-competing risks setting.

1.2.2 Existing work on association for bivariate survival data

Association measures for bivariate failure times have been extensively studied. Existing measures such as correlation coefficient, Spearman's rho and Kendall's tau (Hougaard, 2000) are widely used and are designed to capture the association pattern over a whole study area. There are also measures that study the local association pattern, for example, cross ratio (Clayton, 1978; Oakes, 1982), local Kendall's tau (Oakes, 1989) and martingale covariance function (Prentice and Cai, 1992). Regression analysis has been studied for global measures (e.g., Hsu and Prentice, 1996; Therneau and Grambsch, 2000; Gorfine et al., 2006; Hsu et al., 2007; Gijbels et al., 2011; Veraverbeke et al., 2011) and local measures (e.g., Li et al., 2014). However, it is not straightforward to extend these methods to the bivariate recurrent event setting.

In the context of multi-type recurrent event data, main methods in the literature are regression analyses based on marginal models (e.g., Cai and Schaubel, 2004; Schaubel and Cai, 2005; Sun et al., 2009; Chen et al., 2012) in which dependence structures are left arbitrary, or conditional models (e.g., Abu-Libdeh et al., 1990; Cook et al., 2010) in which dependence structures are characterized by shared random effects. But methods that focus on handling the association between two different types of recurrent events are quite limited. Doss (1989) adapted Ripley's K measure (Ripley, 1976) to capture the association between bivariate point processes. Ventura et al. (2005) studied the dependence between two neurons by comparing the joint firing probability of two neurons spikes to the probability of firing predicted by independence. Both their methods did not account for censoring, and can not be easily adapted to survival settings. For survival data, Yan and Fine (2005) studied the time-varying association among multivariate continuously-observed temporal processes in

the regression setting. They proposed a GEE-type estimating equation, which was stratified based on the availability status of the response temporal processes at each time point. Their work provided some useful insight for modeling the association structure of bivariate recurrent events. That is, one may take the underlying counting processes of recurrent events as the response temporal processes. However, there is some challenge with using their estimating equation for the recurrent event setting considered here. This is because a non-zero lower bound of the observation window would result in the underlying counting processes of recurrent events unobservable at all time points. This prevents us from directly applying Yan and Fine (2005)'s method to our problem.

Most recently, Ning et al. (2015) proposed to capture the association between bivariate recurrent event processes by defining a rate ratio measure

$$\tilde{\rho}(s,t) = \frac{\lambda_{1|2}(s|t)}{\lambda_1(s)}, \quad s,t \geq 0,$$

interpreted as the additional probability for the occurrence of at least one event at time s in the first process due to the occurrence of the second type of event at time t, where $\lambda_{1|2}(s|t) = \lim_{\Delta\to 0^+} P\{N_1(s+\Delta) - N_1(s) > 0 | N_2(t+\Delta) - N_2(t) > 0\}/\Delta$ and $\lambda_1(s) = \lim_{\Delta\to 0^+} P\{N_1(s+\Delta) - N_1(s) > 0\}/\Delta$. Here $N_1(t)$ and $N_2(t)$ denote the number of type-1 and type-2 events that have occurred before time t, respectively. They modeled the rate ratio by a parametric function of time and developed a composite likelihood procedure for parameter estimation. However, their work did not consider any adjustments for covariates.

To best of our knowledge, there is little existing method for assessing the association between two different types of recurrent events under an observation window structure with covariates properly adjusted. Thus, we aim to propose a novel association measure and develop a regression framework for the new measure.

1.3 Outline

In Chapter 2, we introduce a new dependence measure well tailored to the semi-competing risks structure. Then we develop a simple nonparametric estimator, which requires that the gap time between truncation and censoring is independent of the truncation time itself. We present asymptotic studies of the proposed estimator as well as inference procedures. An extension to adjusting for covariates is subsequently discussed. Extensive simulation studies are conducted to evaluate the finite-sample performances of the proposed estimator. We illustrate the proposed method by applying to the Denmark diabetic registry data.

In Chapter 3, we propose a new estimator of the dependence measure proposed in Chapter 2. The new estimator can handle left truncation without requiring the strong assumption assumed by the estimator in Chapter 2. Asymptotic properties are established and simulation studies are conducted. The new proposal is also applied to the Denmark diabetic registry data.

In Chapter 4, we propose a novel measure that can flexibly depict the association between bivariate recurrent events processes. We further develop a regression framework for the proposed measure to allow for assessing how the association is in influenced by covariates. We propose an estimating procedure, utilizing stochastic integrals to facilitate computations. Our simulation studies suggest proper finite sample performance of the proposed method. We also apply the proposed method to a CFFPR dataset.

In Chapter 5, we discuss the plans and directions for future work.

Chapter 2

A New Flexible Dependence Measure for Semi-competing Risks Data

2.1 Proposed Dependence Measure

Let $Q_\tau(Y|A) \equiv \inf\{t : \Pr(Y \leq t|A) \geq \tau\}$ denote the τ-th quantile of Y given condition A holds. For the terminal event of interest, the quantile residual time at a given time point t_0 is defined as $Q_\tau(T_2 - t_0|T_2 > t_0)$.

To assess the dependence between T_1 and T_2, our basic idea is to compare the quantile residual time to the terminal event given the nonterminal event having occurred and that without the past occurrence of the nonterminal event. That is, we consider the cross quantile residual ratio (CQRR) defined as

$$CQRR(\tau; t_0) = \frac{Q_\tau(T_2 - t_0|T_2 > t_0, T_1 > t_0)}{Q_\tau(T_2 - t_0|T_2 > t_0, T_1 \leq t_0)}, \quad \tau \in (0,1), \quad t_0 > 0.$$

It is clear that a larger $CQRR(\tau; t_0)$, which reflects a larger difference in $Q_\tau(T_2 - t_0|T_2 > t_0, T_1 > t_0)$ and $Q_\tau(T_2 - t_0|T_2 > t_0, T_1 \leq t_0)$, indicates a larger impact of having $T_1 > t_0$ (versus $T_1 \leq t_0$) on the subsequent progression of T_2. Note that $CQRR(\tau; t_0)$ bears some similarity with the cross-ratio function in the semi-competing risks setting,

$$\frac{\lambda(t_2|T_1 = t_1)}{\lambda(t_2|T_1 > t_1)}, \quad t_1 \leq t_2,$$

where $\lambda(t_2|\cdot) = \frac{d}{d\epsilon} P(T_2 < t_2 + \epsilon | T_2 \geq t_2, \cdot)|_{\epsilon=0}$. Both of them assess the difference in the terminal event progression according to the timing of the nonterminating event. The distinction lies in that the cross-ratio function uses hazard functions to evaluate the progression of the terminating event, while the proposed $CQRR(\tau, ; t_0)$ adopts quantile residual time, which can be directly interpreted in the time scale. Like the cross-ratio function defined above, $CQRR(\tau; t_0)$ only concerns the joint distribution of (T_1, T_2) at the upper wedge (i.e. $T_1 \leq T_2$) and hence is nonparametrically identifiable with semi-competing risks data.

We further take a log transformation on $CQRR(\tau; t_0)$. Our proposed measure for

the dependence of semi-competing risks events is given by

$$LCQRR(\tau; t_0) = \log \left\{ \frac{Q_\tau(T_2 - t_0 | T_2 > t_0, T_1 > t_0)}{Q_\tau(T_2 - t_0 | T_2 > t_0, T_1 \leq t_0)} \right\}, \quad \tau \in (0,1), \quad t_0 > 0.$$

It is easy to interpret $LCQRR(\tau; t_0)$. For example, $LCQRR(\tau; t_0) > 0$ (< 0) suggests that the nonterminal event occurring before t_0 may be associated with a faster (or slower) progression to subsequent terminal event. The larger the magnitude of $LCQRR(\tau; t_0)$, the bigger the impact of having $T_1 \leq t_0$ on the residual lifetime for T_2. When T_1 and T_2 are independent, $LCQRR(\tau; t_0) = 0$ for any $\tau \in (0,1)$ and $t_0 > 0$. Examining $LCQRR(\tau; t_0)$ with different t_0's may help understand how the dependence between the nonterminal event and the terminal event evolves time. One may also vary the value of τ to evaluate the influence of T_1 on multiple segments of the residual time distribution of T_2.

2.2 Estimation and Inference Procedures

2.2.1 Data and notation

We begin with a formal introduction of data and notation. Let T_1 denote time to nonterminal event, T_2 denote time to terminal event, and C denote time to censoring, which is independent of (T_1, T_2). Without considering left truncation, the observed semi-competing risks data are $X = T_1 \wedge T_2 \wedge C$, $Y = T_2 \wedge C$, $\delta = I(T_1 < Y)$ and $\eta = I(T_2 < C)$, where $\wedge$ is the minimum operator.

With truncation, the observed data consist of n independent and identically distributed replicates of $(X^*, Y^*, \delta^*, \eta^*, L^*)$, denoted by $(X_i^*, Y_i^*, \delta_i^*, \eta_i^*, L_i^*)_{i=1}^n$, where $(X^*, Y^*, \delta^*, \eta^*, L^*)$ follows the conditional distribution of (X, Y, δ, η, L) given $Y > L$. We restrict L to be always less than C, meaning that censoring only occurs after sampling time. Such assumption has been imposed in much previous work, for exam-

ple, Wang (1991), Asgharian et al. (2002) and Li and Peng (2011). In addition, we assume that L is independent of (T_1, T_2) and $D = C - L$.

To simplify the presentation hereafter, we define additional notation, $\boldsymbol{A}^*(t_0) = (1, I(X^* > t_0))^T$, $\tilde{\boldsymbol{A}}^*(t_0) = (1, I(T_1^* > t_0))^T$, $\boldsymbol{A}(t_0) = (1, I(X > t_0))^T$ and $\tilde{\boldsymbol{A}}(t_0) = (1, I(T_1 > t_0))^T$. For a vector $\boldsymbol{v}$, we use $\boldsymbol{v}^{(l)}$ to denote the lth component of $\boldsymbol{v}$.

2.2.2 The proposed estimator

We first study the standard semi-competing risks setting without left truncation. To estimate $LCQRR(\tau; t_0)$, we consider a working quantile residual lifetime regression model, which takes the form,

$$Q_\tau(T_2 - t_0 | T_2 > t_0, I(T_1 > t_0)) = \exp\{\tilde{\boldsymbol{A}}(t_0)^T \boldsymbol{\beta}_0(\tau, t_0)\}, \tag{2.1}$$

where $\boldsymbol{\beta}_0(\tau, t_0)$ is a 2×1 vector of unknown coefficients. In model (2.1), $I(T_1 > t_0)$ serves as the only covariate, which is binary. Consequently, model (2.1) essentially does not impose any parametric assumptions. The coefficients, $\boldsymbol{\beta}_0^{(1)}(\tau, t_0)$ and $\boldsymbol{\beta}_0^{(2)}(\tau, t_0)$, correspond to $\log Q_\tau(T_2 - t_0 | T_2 > t_0, T_1 \leq t_0)$ and $\log Q_\tau(T_2 - t_0 | T_2 > t_0, T_1 > t_0) - \log Q_\tau(T_2 - t_0 | T_2 > t_0, T_1 \leq t_0)$ respectively. This indicates the equivalence between $LCQRR(\tau; t_0)$ and $\boldsymbol{\beta}_0^{(2)}(\tau, t_0)$. Therefore, estimating $\boldsymbol{\beta}_0^{(2)}(\tau, t_0)$ in the quantile regression framework leads to an estimator of $LCQRR(\tau; t_0)$.

A main challenge with fitting model (2.1) is that the covariate $I(T_1 > t_0)$ is not always observed because T_1 is subject to censoring by both T_2 and C. Suppose there is no independent censoring by C, and then T_2 is fully observed. In this case, we see that $I(T_1 > t_0)$ is observed and equals $I(X > t_0)$ as long as $Y > t_0$. This suggests estimating $\boldsymbol{\beta}_0(\tau, t_0)$ by a stratified quantile regression analysis, which solves

the following estimating equation for $\boldsymbol{b} \in R^2$:

$$n^{-1/2} \sum_{i=1}^{n} I(Y_i > t_0) \boldsymbol{A}(t_0) \{ I[\log(Y_i - t_0) \leq \boldsymbol{A}(t_0)^T \boldsymbol{b}] - \tau \} = \mathbf{0}. \tag{2.2}$$

When T_2 is subject to independent censoring by C, we still have $I(T_1 > t_0) = I(X > t_0)$ given $Y > t_0$ and $\eta = 1$. This nice feature allows us to adapt existing methods for quantile residual lifetime model to handle the effect of censoring. Specifically, we can use a stratified version of Ma and Yin (2010)'s estimating equation, which takes the form,

$$n^{-1/2} \sum_{i=1}^{n} \frac{I(Y_i > t_0)\eta_i}{\hat{G}_c(Y_i)} \boldsymbol{A}_i(t_0) \{ I[\log(Y_i - t_0) \leq \boldsymbol{A}_i^T(t_0) \boldsymbol{b}] - \tau \} = \mathbf{0},$$

where $\hat{G}_c(\cdot)$ is the Kaplan-Meier estimate of the survival function of C.

When left truncation is present, we need to further modify the estimating equation (2.2) because $I(T_1 > t_0)$ may be missing and if observed, may not be randomly sampled. Our strategy is to weigh the observed data in an appropriate way such that the bias induced by truncation and censoring is corrected in the estimation of $\boldsymbol{\beta}_0(\tau, t_0)$. Let $D^* = C^* - L^*$. It is critical to note that under the independence between D and (T_1, T_2, L), the distributions of D and D^* are equivalent, and D^* is also independent of (T_1^*, T_2^*, L^*). This fact greatly facilitates the application of the inverse probability of censoring weighting (IPCW) in the present problem with truncated

data. Note that $I(Y^* > t_0)\eta^* \boldsymbol{A}^*(t_0) = I(T_2^* > t_0, T_2^* < C^*)\tilde{\boldsymbol{A}}^*(t_0)$, we can show that

$$
\begin{aligned}
&E\left\{ \frac{I(L^* \le t_0) I(Y^* > t_0)\eta^*}{G(Y^* - L^*)} \boldsymbol{A}^*(t_0)\{I[\log(Y^* - t_0) \le \boldsymbol{A}^{*T}(t_0)\boldsymbol{\beta}_0(\tau, t_0)] - \tau\} \right\} \\
&= E\left\{ \frac{I(L^* \le t_0) I(T_2^* > t_0, T_2^* < C^*)}{G(T_2^* - L^*)} \tilde{\boldsymbol{A}}^*(t_0)\{I[\log(T_2^* - t_0) \le \tilde{\boldsymbol{A}}^{*T}(t_0)\boldsymbol{\beta}_0(\tau, t_0)] - \tau\} \right\} \\
&= E\left\{ \frac{I(L^* \le t_0) I(T_2^* > t_0)\tilde{\boldsymbol{A}}^*(t_0)\{I[\log(T_2^* - t_0) \le \tilde{\boldsymbol{A}}^{*T}(t_0)\boldsymbol{\beta}_0(\tau, t_0)] - \tau\}}{G(T_2^* - L^*)} \right. \\
&\left. \times E[I(T_2^* - L^* < D^*) | T_1^*, T_2^*, L^*] \right\} \\
&= E\left\{ I(L^* \le t_0) I(T_2^* > t_0)\tilde{\boldsymbol{A}}^*(t_0)\{I[\log(T_2^* - t_0) \le \tilde{\boldsymbol{A}}^{*T}(t_0)\boldsymbol{\beta}_0(\tau, t_0)] - \tau\} \times \frac{G(T_2^* - L^*)}{G(T_2^* - L^*)} \right\} \\
&= c(t_0) E\left\{ I(T_2 > t_0)\tilde{\boldsymbol{A}}(t_0)\{I[\log(T_2 - t_0) \le \tilde{\boldsymbol{A}}^T(t_0)\boldsymbol{\beta}_0(\tau, t_0)] - \tau\} \right\} \\
&= 0,
\end{aligned}
$$

where $G(t) = P(D > t)$, $\alpha = P(Y > L)$ and $c(t_0) = P(L \le t_0)/\alpha$. These suggest estimating $\boldsymbol{\beta}_0(\tau, t_0)$ by solving the following estimating equation for $\boldsymbol{b}$:

$$
\boldsymbol{S}_n(\boldsymbol{b}, \tau, t_0) = \boldsymbol{0}, \tag{2.3}
$$

where

$$
\boldsymbol{S}_n(\boldsymbol{b}, \tau, t_0) = n^{-1/2} \sum_{i=1}^{n} \frac{I(L_i^* \le t_0) I(Y_i^* > t_0)\eta_i^*}{\hat{G}(Y_i^* - L_i^*)} \boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^* - t_0) \le \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}] - \tau\}.
$$

The resulting estimator is denoted by $\hat{\boldsymbol{\beta}}(\tau, t_0)$. Here, $\hat{G}(t)$ is the Kaplan-Meier estimator of $G(t)$ obtained from $(Y_i^* - L_i^*, 1 - \eta_i^*)_{i=1}^n$,

$$
\hat{G}(t) = \prod_{Y_i^* - L_i^* \le t} \left\{ 1 - \frac{\sum_{j=1}^n I(Y_j^* - L_j^* = Y_i^* - L_i^*, \eta_j^* = 0)}{\sum_{j=1}^n I(Y_i^* - L_i^* \le Y_j^* - L_j^*)} \right\}.
$$

Equation (2.3) can be easily solved given that it is a monotone estimating equation (Fygenson and Ritov, 1994). Specifically, following similar lines of Peng and

Fine (2009), we can transform the solution finding to equation (2.3) to locating the minimizer of the convex function $U_n(\boldsymbol{b}, \tau, t_0)$ given by

$$U_n(\boldsymbol{b}, \tau, t_0) = \sum_{i=1}^{n} I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* \left| \frac{\log(Y_i^* - t_0)}{\hat{G}(Y_i^* - L_i^*)} - \boldsymbol{b}^T \frac{\boldsymbol{A}_i^*(t_0)}{\hat{G}(Y_i^* - L_i^*)} \right| + \left| M - (2\tau - 1)\boldsymbol{b}^T \sum_{i=1}^{n} I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* \frac{\boldsymbol{A}_i^*(t_0)}{\hat{G}(Y_i^* - L_i^*)} \right|$$

where M is a sufficiently large positive number that can bound $\left|(2\tau - 1)\boldsymbol{b}^T \sum_{i=1}^{n} I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* \frac{\boldsymbol{A}_i^*(t_0)}{\hat{G}(Y_i^* - L_i^*)}\right|$. Minimization of the L_1-type function $U_n(\boldsymbol{b}, \tau, t_0)$ can be solved by using standard software, like the $rq()$ function in the contributed R package *quantreg*.

2.2.3 Asymptotic results

Given that the proposed estimator of $LCQRR(\tau; t_0)$ is the second element of $\hat{\boldsymbol{\beta}}(\tau, t_0)$, it suffices to derive the asymptotic properties of $\hat{\boldsymbol{\beta}}(\tau, t_0)$.

We assume the following regularity conditions:

C1. There exists $\nu > 0$ such that $P(D = \nu) > 0$ and $P(D > \nu) = 0$.

C2. (i) $0 < \tau_L \leq \tau_U \leq 1$; (ii) t_L and t_U are interior points of the support of X^*.

C3. (i) $\boldsymbol{\beta}_0(\tau, t_0)$ is Lipschitz continuous for $\tau \in [\tau_L, \tau_U]$ and $t_0 \in [t_L, t_U]$; (ii) $f(t|\tilde{\boldsymbol{A}}(t_0))$ is continuous and bounded above uniformly in t, t_0 and $\tilde{\boldsymbol{A}}(t_0)$, where $f(t|\tilde{\boldsymbol{A}}(t_0)) = dF(t|\tilde{\boldsymbol{A}}(t_0))/dt$ and $F(t|\tilde{\boldsymbol{A}}(t_0)) = E\{I(T_2 \leq t)|\tilde{\boldsymbol{A}}(t_0)\}$.

C4. For some $\rho_0 > 0$ and $c_0 > 0$, $\inf_{\boldsymbol{b} \in B(\rho_0), t_0 \in [t_L, t_U]} \text{eigmin} \boldsymbol{H}(\boldsymbol{b}, t_0) \geq c_0$, where $B(\rho) = \{\boldsymbol{b} \in R^2 : \inf_{\tau \in [\tau_L, \tau_U], t_0 \in [t_L, t_U]} ||\boldsymbol{b} - \boldsymbol{\beta}_0(\tau, t_0)|| \leq \rho\}$ and $\boldsymbol{H}(\boldsymbol{b}, t_0) = E[c(t_0)\tilde{\boldsymbol{A}}(t_0)^{\otimes 2} f(t_0 + \exp(\tilde{\boldsymbol{A}}^T(t_0)\boldsymbol{b})|\tilde{\boldsymbol{A}}^T(t_0)) \exp(\tilde{\boldsymbol{A}}^T(t_0)\boldsymbol{b})]$. Here $||\cdot||$ is the Euclidean norm and $\boldsymbol{u}^{\otimes 2} = \boldsymbol{u}\boldsymbol{u}^T$ for a vector $\boldsymbol{u}$.

Define $N_i^G(t) = I(Y_i^* - L_i^* \leq t, \eta_i^* = 0)$, $Y_i(t) = I(Y_i^* - L_i^* \geq t)$, $y(t) = P(Y^* - L^* \geq t)$, $\lambda^G(t) = \lim_{\Delta \to 0} P(Y^* - L^* \in (t, t+\Delta) | Y^* - L^* \geq t)/\Delta$, $\Lambda^G(t) = \int_0^t \lambda^G(s) ds$, and $M_i^G(t) = N_i^G(t) - \int_0^\infty Y_i(s) d\Lambda^G(s)$. Let $\boldsymbol{w}(\boldsymbol{b}, \tau, t_0, t) = E\{\boldsymbol{A}^*(t_0) Y(t) I(L^* \leq t_0) I(Y^* > t_0) \eta^* \{I[\log(Y^* - t_0) \leq \boldsymbol{A}^{*T}(t_0)\boldsymbol{b}] - \tau\} G(Y^* - L^*)^{-1}\}$, $\boldsymbol{\zeta}_i(\tau, t_0) = \boldsymbol{\xi}_{1,i}(\tau, t_0) - \boldsymbol{\xi}_{2,i}(\tau, t_0)$, where $\boldsymbol{\xi}_{1,i}(\tau, t_0) = I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* \boldsymbol{A}_i^*(t_0) \{I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0) \boldsymbol{\beta}_0(\tau, t_0)] - \tau\} G(Y_i^* - L_i^*)^{-1}$ and $\boldsymbol{\xi}_{2,i}(\tau, t_0) = \int_0^\infty \boldsymbol{w}(\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0, s) \frac{dM_i^G(s)}{y(s)}$, $i = 1, ..., n$.

We have following theorems:

Theorem 2.2.1. *Under conditions C1–C4,*

$$\lim_{n \to \infty} \sup_{\tau \in [\tau_L, \tau_U], t_0 \in [t_L, t_U]} ||\hat{\boldsymbol{\beta}}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)|| \to_p 0.$$

Theorem 2.2.2. *Under conditions C1–C4, $\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)\}$ weakly converge to a mean zero Gaussian process with covariance matrix given by*

$$\boldsymbol{\Phi}(\tau', t_0', \tau, t_0) = \boldsymbol{H}\{\boldsymbol{\beta}_0(\tau', t_0'), t_0'\}^{-1} E\{\boldsymbol{\zeta_1}(\tau', t_0') \boldsymbol{\zeta_1}(\tau, t_0)^T\} [\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau, t_0), t_0\}^{-1}]^T,$$

where $\tau, \tau' \in [\tau_L, \tau_U]$ and $t_0, t_0' \in [t_L, t_U]$.

Theorem 2.2.1 implies that the proposed estimator of $LCQRR(\tau; t_0)$ is uniformly consistent in $\tau \in [\tau_L, \tau_U]$ and $t_0 \in [t_L, t_U]$. Theorem 2.2.2 presents a closed form expression for the asymptotic distribution of the proposed estimator of $LCQRR(\tau; t_0)$. Detailed proofs of Theorem 2.2.1 and 2.2.2 are provided in Section 2.7 Appendix.

2.2.4 Inference procedures

The asymptotic covariance matrix of $\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)\}$ involves unknown density functions. It is straightforward to use bootstrapping procedures or adapt re-

sampling approaches, such as Parzen, Wei, and Ying (1994) and Jin, Ying, and Wei (2001), to estimate the asymptotic covariance without requiring density estimation. Alternatively, we can also derive a consistent plug-in estimate for the covariance matrix following the lines of Peng and Fine (2009). The specific procedure follows.

1. Calculate $\hat{\boldsymbol{\Sigma}}(\tau, t_0, \tau, t_0) = n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\zeta}}_i(\tau, t_0)^{\otimes 2}$, where

$$\hat{\boldsymbol{\zeta}}_i(\tau, t_0) = \frac{I(L_i^* \le t_0, Y_i^* > t_0)\eta_i^*}{\hat{G}(Y_i^* - L_i^*)}\boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^* - t_0) \le \boldsymbol{A}_i^{*T}(t_0)\hat{\boldsymbol{\beta}}(\tau, t_0)] - \tau\}$$
$$- I(\eta_i^* = 0)\Bigg(\sum_{j=1}^{n}\boldsymbol{A}_j^*(t_0)I(Y_j^* - L_j^* \ge Y_i^* - L_i^*)I(L_j^* \le t_0, Y_j^* > t_0)\eta_j^*$$
$$\times\{I[\log(Y_j^* - t_0) \le \boldsymbol{A}_j^{*T}(t_0)\hat{\boldsymbol{\beta}}(\tau, t_0)] - \tau\}\{\hat{G}(Y_j^* - L_j^*)\}^{-1}\Big/\sum_{j=1}^{n}I(Y_j^* - L_j^* \ge Y_i^* - L_i^*)\Bigg).$$

2. Use spectral decomposition to find a symmetric matrix $\boldsymbol{E}_n(\tau, t_0)$ such that $\hat{\boldsymbol{\Sigma}}(\tau, t_0, \tau, t_0) = \boldsymbol{E}_n^2(\tau, t_0)$.

3. Calculate $\boldsymbol{D}_n(\tau, t_0) = [\boldsymbol{S}_n^{-1}\{\boldsymbol{e}_{n,1}(\tau, t_0), \tau, t_0\} - \hat{\boldsymbol{\beta}}(\tau, t_0), \boldsymbol{S}_n^{-1}\{\boldsymbol{e}_{n,2}(\tau, t_0), \tau, t_0\} - \hat{\boldsymbol{\beta}}(\tau, t_0)]$, where $\boldsymbol{e}_{n,j}$ is the jth column of $\boldsymbol{E}_n(\tau, t_0)$, and $\boldsymbol{S}_n^{-1}\{\boldsymbol{e}, \tau, t_0\}$ is defined as the solution to $\boldsymbol{S}_n(\boldsymbol{b}, \tau, t_0) - \boldsymbol{e} = 0$.

4. A consistent estimate for the asymptotic covariance matrix of $\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)\}$ is given by

$$n\boldsymbol{D}_n(\tau', t_0')\boldsymbol{E}_n^{-1}(\tau', t_0')\hat{\boldsymbol{\Sigma}}(\tau', t_0', \tau, t_0)\boldsymbol{E}_n^{-1}(\tau, t_0)\boldsymbol{D}_n^T(\tau, t_0).$$

   In the special case that $\tau' = \tau$ and $t_0' = t_0$, a consistent estimate for the asymptotic variance matrix is simplified as $n\{\boldsymbol{D}_n^{\otimes 2}(\tau, t_0)\}$.

We can also develop second-stage inferences following the lines of Peng and Fine (2009). For example, we can summarize $LCQRR(\tau; t_0)$ over $t_0 \in [t_L, t_U]$ by $\Omega_\tau = \frac{1}{t_U - t_L}\int_{t_L}^{t_U}\boldsymbol{\beta}_0^{(2)}(\tau, t_0)dt_0$, which may be consistently estimated by $\hat{\Omega}_\tau =$

$\frac{1}{t_U-t_L}\int_{t_L}^{t_U}\hat{\boldsymbol{\beta}}^{(2)}(\tau,t_0)dt_0$. We can show that the limiting distribution of $\sqrt{n}(\hat{\Omega}_\tau-\Omega_\tau)$ is a mean zero normal distribution, the variance of which may be consistently estimated by $n\hat{\sigma}^2_{\Omega_\tau}$, where $\hat{\sigma}^2_{\Omega_\tau}$ equals the (2,2) element of $\frac{1}{n^2}\sum_{i=1}^n\{\frac{1}{t_U-t_L}\int_{t_L}^{t_U}\sqrt{n}\boldsymbol{D}_n(\tau,t_0)\boldsymbol{E}_n^{-1}(\tau,t_0)\hat{\boldsymbol{\zeta}}_i(\tau,t_0)dt_0\}^{\otimes 2}$. This result naturally renders a Wald-type test, $T_{\Omega_\tau}=\hat{\Omega}_\tau/\hat{\sigma}_{\Omega_\tau}$, for the null hypothesis $H_{01}: LCQRR(\tau;t_0)=0,\ \ t_0\in[t_L,t_U]$. That is, we reject H_{01} when $|T_{\Omega_\tau}|>100(1-\alpha/2)$th percentile of $N(0,1)$ distribution, where α is the desired significance level. Similar results can be obtained for the overall summary and testing of $LCQRR(\tau;t_0)$ over $\tau\in[\tau_L,\tau_U]$, corresponding to $\Omega_{t_0}=\frac{1}{\tau_U-\tau_L}\int_{\tau_L}^{\tau_U}\boldsymbol{\beta}_0^{(2)}(\tau,t_0)d\tau$, and $H_{03}:\ LCQRR(\tau;t_0)=0,\tau\in[\tau_L,\tau_U]$ respectively.

We can also test the constancy of $LCQRR(\tau;t_0)$ over t_0 or τ. For example, a null hypothesis of interest may take the form, $H_{02}: LCQRR(\tau;t_0)=C_\tau,\ t_0\in[t_L,t_U]$, where C_τ is an unspecified constant and may change with τ. Let $\Xi(\tau,t_0)$ denote a known weight function satisfying $\Xi(\tau,t_0)\geq 0$ and $\int_{t_L}^{t_U}\Xi(\tau,t_0)dt_0=1$. If H_{02} holds, then $\int_{t_L}^{t_U}\Xi(\tau,t_0)\boldsymbol{\beta}_0^{(2)}(\tau,t_0)dt_0-\Omega_\tau=[\int_{t_L}^{t_U}\Xi(\tau,t_0)dt_0-1]C_\tau=0$. This motivates us to construct a test statistic for H_{02} based on $\Gamma_\tau=\sqrt{n}\{\int_{t_L}^{t_U}\Xi(\tau,t_0)\hat{\boldsymbol{\beta}}^{(2)}(\tau,t_0)dt_0-\hat{\Omega}_\tau\}$. Following the same line for proving Theorem 2.2.2, we can show that the limiting distribution of Γ_τ under H_{02} is normal with mean 0. A consistent variance estimate for Γ_τ may be given by $\hat{\sigma}^2_{\Gamma_\tau}$, which is the (2,2) element of

$$n^{-1}\sum_{i=1}^n\left[\int_{t_L}^{t_U}\{\Xi(\tau,t_0)-\frac{1}{t_U-t_L}\}\sqrt{n}\boldsymbol{D}_n(\tau,t_0)\boldsymbol{E}_n^{-1}(\tau,t_0)\hat{\boldsymbol{\zeta}}_i(\tau,t_0)dt_0\right]^{\otimes 2}.$$

A Wald-type test for H_{02} is then given by $T_{\Gamma_\tau}=\Gamma_\tau/\hat{\sigma}_{\Gamma_\tau}$. A similar testing procedure can be developed for testing the constancy over $t_0\in[t_L,t_U]$, $H_{04}:\ LCQRR(\tau;t_0)=C_{t_0},\ \tau\in[\tau_L,\tau_U]$.

2.3 An Extension to Adjusting for Covariates

Exploiting population heterogeneity in semi-competing risks dependence is often scientifically meaningful, and for example, can help uncover uncommon disease mechanisms in subgroups. To this end, we propose an extension, which adjusts for covariates (captured by $\tilde{\boldsymbol{Z}} \in R^p$) in the assessment of the dependence between the nonterminal event and the terminal event.

First, we define the covariate-adjusted log cross quantile residual ratio as

$$LCQRR(\tau; t_0|\tilde{\boldsymbol{Z}}) = \log\left[\frac{Q_\tau(T_2 - t_0|T_2 > t_0, T_1 > t_0, \tilde{\boldsymbol{Z}})}{Q_\tau(T_2 - t_0|T_2 > t_0, T_1 \leq t_0, \tilde{\boldsymbol{Z}})}\right].$$

When all covariates of interest are discrete, one may conduct stratified analyses based on the methods in Section 2.2 to estimate and make inference on $LCQRR(\tau; t_0|\tilde{\boldsymbol{Z}})$.

In many practical settings, covariates of interest can be continuous. Thus we investigate a general scenario where $\tilde{\boldsymbol{Z}}$ can include both continuous and discrete covariates. Specifically, we are interested in formulating linear covariate effects on $LCQRR$, which may be expressed as

$$LCQRR(\tau; t_0|\tilde{\boldsymbol{Z}}) = \check{\boldsymbol{Z}}^T \boldsymbol{\alpha}_0(\tau, t_0), \tag{2.4}$$

where $\check{\boldsymbol{Z}} = (1, \tilde{\boldsymbol{Z}}^T)^T$. The non-intercept coefficients in $\boldsymbol{\alpha}_0(\tau, t_0)$ depict how $LCQRR$ changes per unit change in the corresponding covariate.

To address the interest in the linear effects of covariates on $LCQRR$, we consider the following quantile residual lifetime model:

$$\begin{aligned} Q_\tau(T_2 - t_0|T_2 > t_0, I(T_1 > t_0), \tilde{\boldsymbol{Z}}) &= \exp\{\boldsymbol{Z}^T(t_0)\boldsymbol{\gamma}_0(\tau, t_0)\} \\ &\equiv \exp[\boldsymbol{\gamma}_0^{(1)}(\tau, t_0) + I(T_1 > t_0)\boldsymbol{\gamma}_0^{(2)}(\tau, t_0) + \tilde{\boldsymbol{Z}}^T \boldsymbol{\gamma}_0^{3:(2+p)}(\tau, t_0) \\ &\quad + \tilde{\boldsymbol{Z}}^T I(T_1 > t_0)\boldsymbol{\gamma}_0^{(3+p):(2+2p)}(\tau, t_0)], \end{aligned} \tag{2.5}$$

where $\boldsymbol{Z}(t_0) = (1, I(T_1 > t_0), \tilde{\boldsymbol{Z}}^T, \tilde{\boldsymbol{Z}}^T I(T_1 > t_0))^T$, and $\boldsymbol{v}^{a:b}$ denotes the vector that includes the ath to bth components of vector $\boldsymbol{v}$. It is important to note that (2.5) implies

$$LCQRR(\tau; t_0|\tilde{\boldsymbol{Z}}) = \boldsymbol{\gamma}_0^{(2)}(\tau, t_0) + \tilde{\boldsymbol{Z}}^T \boldsymbol{\gamma}_0^{(3+p):(2+2p)}(\tau, t_0).$$

When there are only discrete covariates, model (2.5) and model (2.4) can be equivalent. These suggest that under slightly stronger assumptions regarding the effects of continuous covariates, model (2.5) defines the same linear relationship between covariates and $LCQRR$ as does model (2.4). Compared to model (2.4), model (2.5) is more convenient to tackle. This is because model (2.5) takes the same form as the working quantile residual lifetime model (2.1) considered for the one-sample case. As shown below, this fact greatly facilitates an extension to the general case with covariates. By these considerations, we adopt model (2.5) as the vehicle to explore the linear covariate effects on $LCQRR$.

Suppose the observed data include n i.i.d. replicates, $(X_i^*, Y_i^*, \delta_i^*, \eta_i^*, L_i^*, \tilde{\boldsymbol{Z}}_i^*)_{i=1}^n$, where $\tilde{\boldsymbol{Z}}_i^*$ is the truncated counterpart of $\tilde{\boldsymbol{Z}}_i$ following the conditional distribution of $\tilde{\boldsymbol{Z}}$ given $Y > L$. We assume that D is independent of $(T_1, T_2, L, \tilde{\boldsymbol{Z}})$ and L is independent of T_2 given $(T_1, \tilde{\boldsymbol{Z}})$. Define $\boldsymbol{K}^*(t_0) = (1, I(X_i^* > t_0), \tilde{\boldsymbol{Z}}_i^{*T}, \tilde{\boldsymbol{Z}}_i^{*T} I(X_i^* > t_0))^T$. Adapting the idea presented for the one-sample case, we propose to estimate $\boldsymbol{\gamma}_0(\tau, t_0)$ by solving the following estimating equation for $\boldsymbol{r} \in R^{2+2p}$:

$$\boldsymbol{S}_n(\boldsymbol{r}, \tau, t_0) = \boldsymbol{0},$$

where

$$\boldsymbol{S}_n(\boldsymbol{r}, \tau, t_0) = n^{-1/2} \sum_{i=1}^{n} \frac{I(L_i^* \le t_0) I(Y_i^* > t_0) \eta_i^*}{\hat{G}(Y_i^* - L_i^*)} \boldsymbol{K}_i^*(t_0) \{ I[\log(Y_i^* - t_0) \le \boldsymbol{K}_i^{*T}(t_0) \boldsymbol{r}] - \tau \}.$$

The resulting estimator is denoted by $\hat{\boldsymbol{\gamma}}(\tau, t_0)$. It is easy to see that the subvector,

$\hat{\boldsymbol{\gamma}}^{(3+p):(2+2p)}(\tau, t_0)$, can be used to describe the linear effect of $\tilde{\boldsymbol{Z}}$ on $LCQRR$. With an additional assumption that $\tilde{\boldsymbol{Z}}$ is uniformly bounded (i.e. $\sup_i ||\tilde{\boldsymbol{Z}}_i|| \leq M_1 < \infty$), we can established the same asymptotic properties and inference procedures for $\hat{\boldsymbol{\gamma}}(\tau, t_0)$ as those presented in Section 2.2.

2.4 Simulation Studies

Simulation studies are conducted to examine the finite-sample performance of the proposed methods in the left-truncated semi-competing risks setting. Specifically, we generate (T_1, T_2) from a gamma frailty model,

$$P(T_1 > x, T_2 > y) = [P(T_1 > x)^{1-\theta} + P(T_2 > y)^{1-\theta} - 1]^{1/(1-\theta)},$$

in which T_i follows a Weibull(α_i, λ_i) distribution and $P(T_i > x) = exp(-\lambda_i x^{\alpha_i})$, $i = 1, 2$. The truncation time $L = r \times L_0$, where r is a random variable following Bernoulli distribution with probability p, and L_0 is a positive random variable that is independent of r. Such a truncation scenario mimics the Denmark diabetes registry study, where the distribution of L has a point mass at 0. We generate the censoring time C as $L + D$, where D is a positive-valued random variable independent of L.

The simulations are conducted under two scenarios,

Scenario 1: $T_1 \sim$ Weibull$(1.4, 0.6)$, $T_2 \sim$ Weibull$(3.5, 0.5)$, L_0 and D following uniform distributions.

Scenario 2: $T_1 \sim$ Weibull$(3, 0.85)$, $T_2 \sim$ Weibull$(3, 0.4)$, L_0 and D following Weibull distributions.

For Scenario 1, there is a low truncation level with $P(Y < L) = 0.3$, and a high dependent censoring rate with $P(\delta^* = 0, \eta^* = 1)$ close to 0.4. For Scenario 2, there is a high truncation level of 0.5 and a low dependent censoring rate around 0.15. In each scenario, we consider three different θ values, 1, 2 and 3, corresponding to inde-

pendence, moderate positive association, and high positive association respectively. The choice of p, detailed marginal distributions of L_0 and D as well as censoring and truncation proportions are shown in Table 2.1.

Table 2.1: Summary of simulation setups: the choices of $\{p, L_0, D\}$ as well as the resulting truncation and censoring proportions, where $p_1 = P(Y < L)$, $p_2 = P(\delta^* = 0)$, $p_3 = P(\eta^* = 0)$ and $p_4 = P(\delta^* = 0, \eta^* = 1)$.

θ	p	L_0	D_0	p_1	p_2	p_3	p_4
Scenario 1: $T_1 \sim$ Weibull$(1.4, 0.6)$, $T_2 \sim$ Weibull$(3.5, 0.5)$							
1	0.86	Unif(0,1.67)	Unif(0.05,3.2)	0.30	0.52	0.21	0.39
2	0.86	Unif(0,1.67)	Unif(0.17,2.6)	0.30	0.59	0.22	0.42
3	0.86	Unif(0,1.67)	Unif(0.15,2.55)	0.30	0.63	0.23	0.44
Scenario 2: $T_1 \sim$ Weibull$(3, 0.85)$, $T_2 \sim$ Weibull$(3, 0.4)$							
1	0.90	Wei(2.6,0.35)	Wei(1.1,0.38)	0.50	0.26	0.20	0.16
2	0.90	Wei(1.2,0.49)	Wei(1.3,0.3)	0.50	0.27	0.20	0.15
3	0.90	Wei(0.5,0.55)	Wei(1.5,0.22)	0.50	0.27	0.20	0.15

We perform the proposed methods on 1000 simulated datasets with sample size $n = 200$ or 400 for each simulation setup, where M is set as 10^7. For Scenario 1, Figure 2.1 presents the empirical bias (EmpBias), empirical standard error (EmpSE) and average estimated standard error (EstSE) for the proposed estimator of $LCQRR(\tau; t_0)$ under different combinations of (θ, τ, t_0), where $\tau = 0.25, 0.5, 0.75$, $t_0 = 0.55, 0.84, 1.1$ and circles denote corresponding values. It is observed that the proposed estimator of $LCQRR(\tau; t_0)$, performs well with moderate sample size. The point estimates have small biases. The corresponding standard error estimates agree well with empirical standard errors, and the agreement generally improves as sample size increases. We have very similar observations from Figure 2.2, which presents the simulation results for Scenario 2.

We also examine the proposed second-stage inferences. With fixed τ, we evaluate the average of $LCQRR$ over $t \in [t_L, t_U]$, and test whether $LCQRR(\tau; t_0)$ equals 0 for $t \in [t_L, t_U]$ and whether $LCQRR(\tau; t_0)$ is constant over $t \in [t_L, t_U]$. We consider

three τ values, 0.25, 0.5, and 0.75. For Scenario 1, we set $t_L = 0.42$ and $t_U = 1.20$. For Scenario 2, we set $t_L = 0.68$ and $t_U = 1.28$. We compute integrals using left Riemann sums on intervals of equal length 0.001 and choose the weight function $\Xi(\tau, t_0) = 2I[t_0 \leq (t_L + t_U)/2]/(t_U - t_L)$. In Table 2.2, we summarize the EmpBias, EmpSE and EstSE of $\hat{\Omega}_\tau$, and the empirical rejection rates (EmpRR) for the proposed Wald tests for H_{01} and H_{02}. Note that for both H_{01} and H_{02}, the EmpRR gives empirical sizes when $\theta = 1$ and empirical power when $\theta = 2, 3$. Table 2.2 shows that for both scenarios, the empirical biases of $\hat{\Omega}_\tau$ are small and the estimated standard errors match the empirical standard errors very well. The test for either H_{01} or H_{02} appear to have empirical sizes close to the nominal levels. The power for testing H_{01} is good, while the constancy tests appear to be conservative. The empirical power increases considerably as sample size and θ value increase for both tests.

With fixed t_0, we assess the second-stage inferences over $[\tau_L, \tau_U]$. For Scenario 1, we consider $t_0 = 0.55, 0.84, 1.10$ and set $[\tau_L, \tau_U] = [0.1, 0.87]$. For Scenario 2, we consider $t_0 = 0.85, 1.00, 1.20$ and set $[\tau_L, \tau_U] = [0.1, 0.9]$. In both scenarios, $\Xi(\tau, t_0) = 2I[\tau \leq (\tau_L + \tau_U)/2]/(\tau_U - \tau_L)$. Table 2.3 presents the EmpBias, EmpSE and EstSE of $\hat{\Omega}_{t_0}$ and the EmpRR for the proposed tests. Similarly, we observe small empirical biases, well-matched estimated and empirical standard errors, and pretty accurate empirical sizes. The power for the constancy tests is not high but increases as sample size increases.

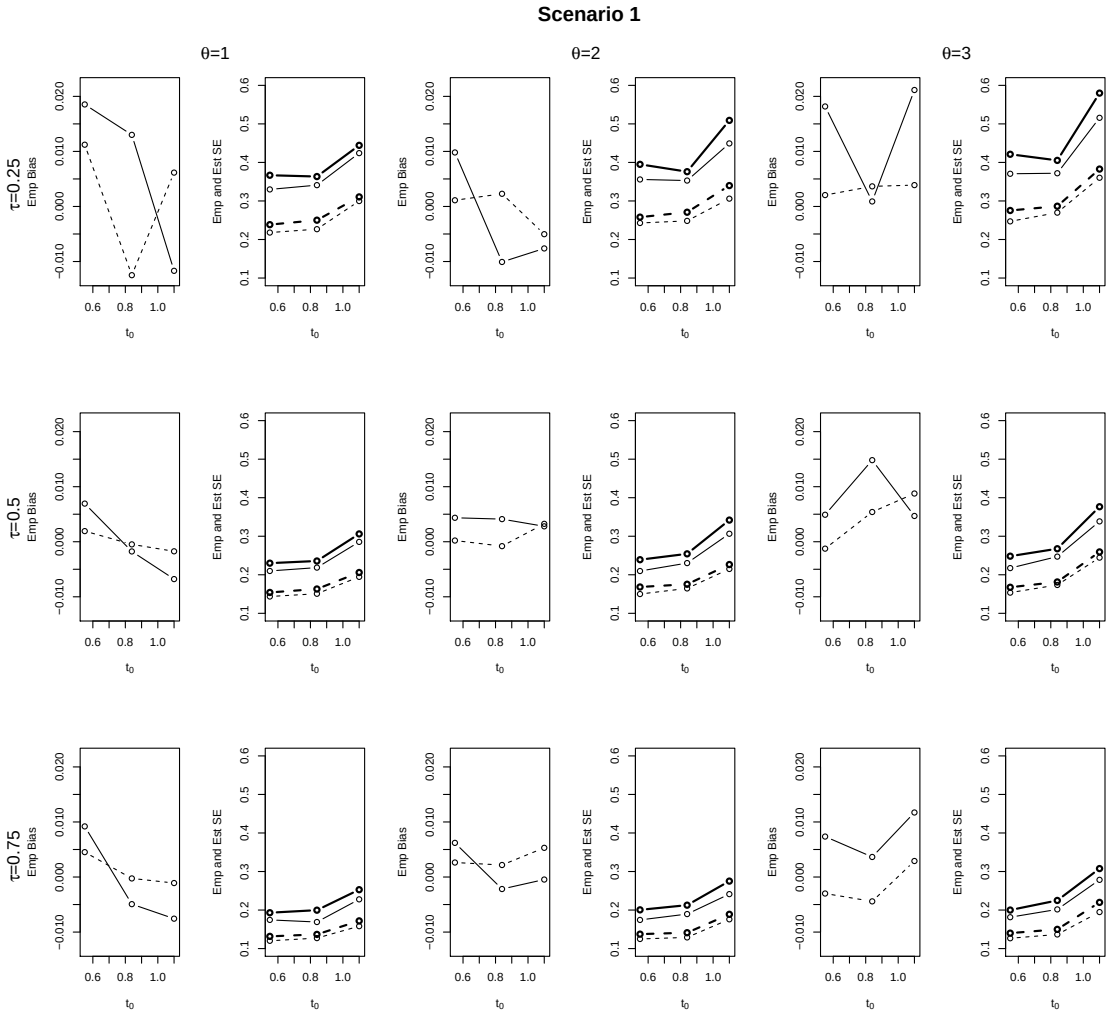

Figure 2.1: Simulation results for Scenario 1: Empirical bias (EmpBias), empirical standard error (EmpSE) and average estimated standard error (EstSE) of the proposed estimator of $LCQRR(\tau; t_0)$. EmpBias for $n = 200$ and EmpBias for $n = 400$ are plotted in solid lines and dotted lines respectively. EmpSE and EstSE for $n = 200$ are plotted in solid lines and bold solid lines respectively. EmpSE and EstSE for $n = 400$ are plotted in dotted lines and bold dashed lines respectively.

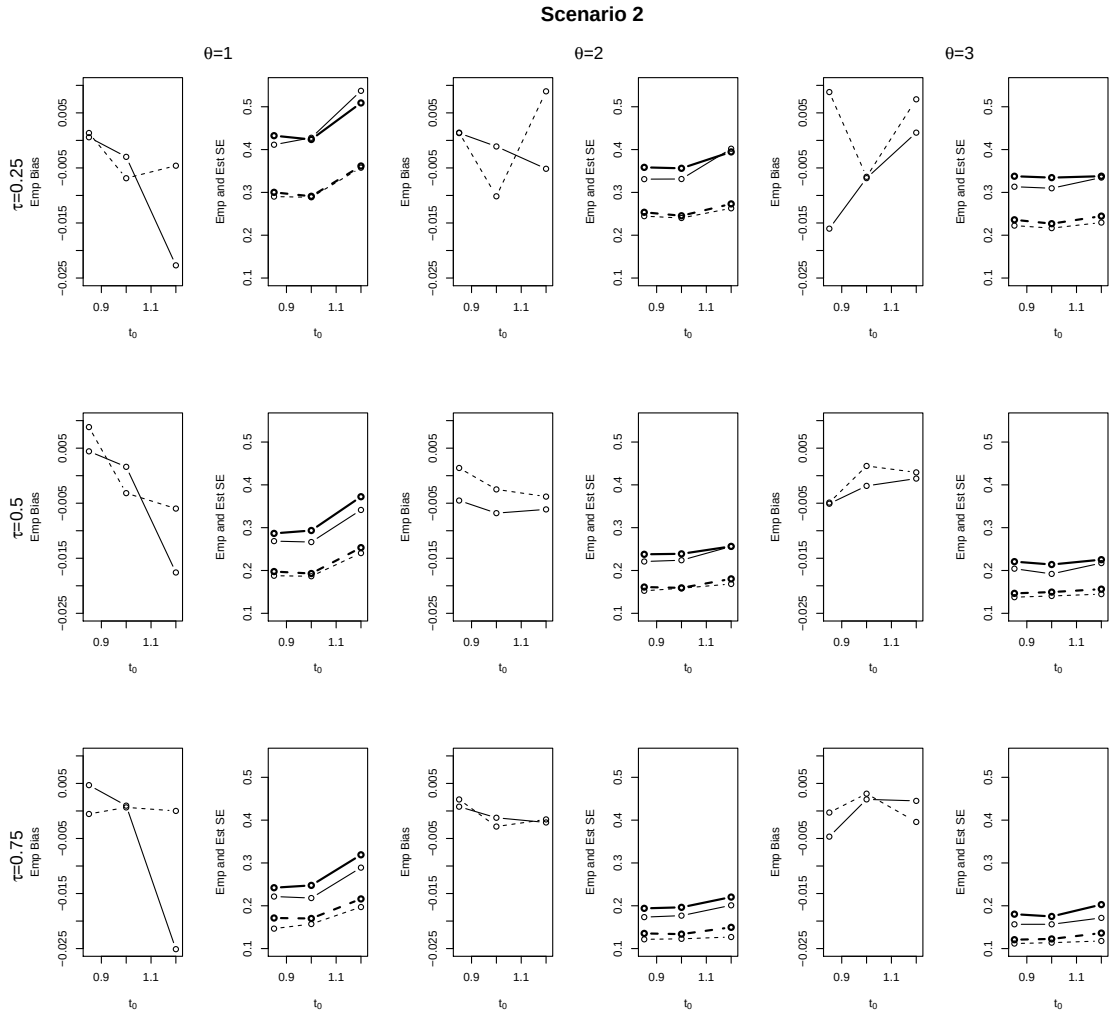

Figure 2.2: Simulation results for Scenario 2: Empirical bias (EmpBias), empirical standard error (EmpSE) and average estimated standard error (EstSE) of the proposed estimator of $LCQRR(\tau; t_0)$. EmpBias for $n = 200$ and EmpBias for $n = 400$ are plotted in solid lines and dotted lines respectively. EmpSE and EstSE for $n = 200$ are plotted in solid lines and bold solid lines respectively. EmpSE and EstSE for $n = 400$ are plotted in dotted lines and bold dashed lines respectively.

Table 2.2: EmpBias, EmpSE and EstSE of $\hat{\Omega}_\tau$ and empirical rejection rates for H_{01} and H_{02}.

				$\hat{\Omega}_\tau$		H_{01}	H_{02}
θ	τ	n	EmpBias	EmpSE	EstSE	EmpRR	EmpRR
				Scenario 1			
				$t_0 \in [0.42, 1.20]$			
1	0.25	200	0.008	0.181	0.191	0.041	0.051
		400	0.003	0.128	0.132	0.053	0.046
	0.50	200	0.006	0.153	0.160	0.056	0.043
		400	0.005	0.112	0.110	0.052	0.046
	0.75	200	0.008	0.143	0.149	0.065	0.037
		400	0.001	0.101	0.104	0.054	0.048
2	0.25	200	0.006	0.172	0.188	0.927	0.102
		400	0.003	0.119	0.127	1.000	0.149
	0.50	200	0.007	0.140	0.153	0.928	0.142
		400	0.003	0.102	0.107	0.995	0.215
	0.75	200	0.003	0.137	0.148	0.857	0.160
		400	0.005	0.099	0.104	0.988	0.234
3	0.25	200	0.014	0.166	0.182	0.999	0.118
		400	0.001	0.112	0.122	1.000	0.213
	0.50	200	0.011	0.143	0.149	0.999	0.184
		400	0.003	0.097	0.103	1.000	0.313
	0.75	200	0.010	0.143	0.147	0.981	0.200
		400	0.005	0.097	0.103	1.000	0.347
				Scenario 2			
				$t_0 \in [0.68, 1.28]$			
1	0.25	200	-0.004	0.243	0.237	0.066	0.066
		400	0.001	0.168	0.164	0.053	0.059
	0.50	200	-0.004	0.193	0.198	0.065	0.049
		400	0.002	0.133	0.138	0.048	0.050
	0.75	200	-0.001	0.174	0.179	0.080	0.047
		400	0.003	0.119	0.127	0.056	0.039
2	0.25	200	0.003	0.172	0.176	0.984	0.117
		400	0.001	0.124	0.120	1.000	0.211
	0.50	200	-0.004	0.140	0.141	0.984	0.128
		400	-0.001	0.098	0.097	1.000	0.194
	0.75	200	-0.001	0.128	0.133	0.965	0.096
		400	0.000	0.087	0.093	1.000	0.144
3	0.25	200	-0.007	0.133	0.137	1.000	0.161
		400	0.002	0.087	0.093	1.000	0.259
	0.50	200	-0.002	0.113	0.116	1.000	0.125
		400	0.000	0.078	0.080	1.000	0.161
	0.75	200	0.000	0.108	0.117	1.000	0.100
		400	0.001	0.075	0.081	1.000	0.118

Table 2.3: EmpBias, EmpSE and EstSE of $\hat{\Omega}_{t_0}$ and empirical rejection rates for H_{03} and H_{04}.

				$\hat{\Omega}_{t_0}$		H_{03}	H_{04}
θ	t_0	n	EmpBias	EmpSE	EstSE	EmpRR	EmpRR
				Scenario 1			
				$\tau \in [0.1, 0.87]$			
1	0.55	200	0.006	0.193	0.195	0.060	0.041
		400	0.004	0.128	0.134	0.043	0.047
	0.84	200	0.006	0.197	0.201	0.052	0.036
		400	0.001	0.144	0.141	0.054	0.046
	1.10	200	0.005	0.262	0.247	0.061	0.050
		400	0.002	0.175	0.174	0.052	0.054
2	0.55	200	0.018	0.205	0.211	0.575	0.051
		400	0.007	0.139	0.144	0.893	0.069
	0.84	200	0.018	0.221	0.219	0.791	0.092
		400	0.006	0.153	0.152	0.982	0.139
	1.10	200	-0.005	0.269	0.274	0.697	0.053
		400	-0.004	0.191	0.193	0.955	0.091
3	0.55	200	0.017	0.220	0.216	0.918	0.072
		400	0.001	0.146	0.149	0.999	0.154
	0.84	200	-0.005	0.223	0.225	0.984	0.126
		400	0.006	0.156	0.157	1.000	0.292
	1.10	200	0.000	0.302	0.310	0.908	0.051
		400	-0.001	0.214	0.216	0.997	0.114
				Scenario 2			
				$\tau \in [0.1, 0.9]$			
1	0.85	200	0.003	0.244	0.236	0.066	0.045
		400	0.003	0.167	0.164	0.052	0.047
	1.00	200	0.004	0.242	0.232	0.061	0.045
		400	0.007	0.161	0.164	0.052	0.047
	1.20	200	-0.009	0.303	0.279	0.075	0.064
		400	-0.003	0.215	0.203	0.073	0.060
2	0.85	200	0.002	0.204	0.198	0.859	0.092
		400	0.005	0.138	0.139	0.992	0.180
	1.00	200	-0.012	0.188	0.194	0.929	0.157
		400	0.003	0.136	0.137	1.000	0.314
	1.20	200	-0.007	0.220	0.213	0.938	0.271
		400	0.004	0.156	0.151	0.997	0.426
3	0.85	200	0.010	0.185	0.182	0.998	0.236
		400	0.003	0.129	0.128	1.000	0.492
	1.00	200	0.003	0.178	0.179	1.000	0.363
		400	-0.002	0.126	0.126	1.000	0.664
	1.20	200	-0.007	0.187	0.190	1.000	0.492
		400	0.001	0.136	0.134	1.000	0.837

2.5 An Application to Denmark Diabetes Registry Data

We apply the proposed method to a dataset from the Denmark diabetes registry study (Andersen et al., 1993). The Denmark diabetes registry study is a prospective cohort study on insulin-dependent diabetes patients referred to the Steno Memorial Hospital in Greater Copenhagen. Diabetic nephropathy (DN), an indicator of kidney failure, is a significant complication among patients with diabetes. From 1933 to 1981, 2727 patients who were diagnosed with insulin-dependent diabetes mellitus prior to age 31 and between 1933 and 1972 were accrued. At entry, patients' age at diabetes diagnosis and the presence of DN were recorded. All patients were then followed until death, emigration or December 31, 1984. In our analysis, the time origin is the age at diabetes diagnosis, with event times recorded in years since diagnosis. It is seen that time to DN and time to death naturally formed a semi-competing risks structure because death terminated the observation on time to DN, but remained observable after the occurrence of DN. Administrative left truncation on mortality was also involved. That is, patients who had died before study enrollment were excluded. Out of 2727 patients, there were 731(26.8%) experiencing DN, 718(26.3%) dead in the end and 652(24%) with diabetic onset at entry. Summary statistics for the data are presented in Table 2.4.

Table 2.4: Summary statistics for diabetes registry data.

	$n(\%)$
$(\delta, \eta) = (0, 0)$	1729(63.4%)
$(\delta, \eta) = (0, 1)$	267(9.8%)
$(\delta, \eta) = (1, 0)$	280(10.3%)
$(\delta, \eta) = (1, 1)$	451(16.5%)
$L = 0$	652(24%)
$X < L$	116(4.25%)

Our focus is first to quantify the relationship between DN and death by using the proposed measure $LCQRR(\tau; t_0)$. We fit model (2.1) to the data and adopt $M = 10^7$ as in the simulations. We restrict t_0 to be within [6,40] to ensure reasonable sample sizes accumulated for strata defined by $I(X^* > t_0)$. In Figure 2.3, we display the results for $\tau = 0.25, 0.5, 0.75$ and t_0 values at an equally space grid on [6, 40] with step size=0.1. Estimated $LCQRR(\tau; t_0)$ are plotted in bold solid lines. The corresponding 95% pointwise confidence intervals are in dotted lines and the 95% pointwise Wald-type bootstrapping confidence intervals are in long-dashed lines. In Figure 2.3, we see that for all three τ values, the estimated $LCQRR(\tau; t_0)$ is generally positive; the lower bounds of confidence intervals are above 0 for t_0 less than 30, which is roughly the third quartile of X^*. This observation is consistent with the common belief that DN is positively associated with mortality. Our formal test for H_{01} yields p-values, < 0.001, 0.002, < 0.001, respectively, for $\tau = 0.25, 0.5, 0.75$, confirming that DN is a significant prognostic factor for mortality.

We note that the confidence intervals for $LCQRR(\tau; t_0)$ with $t_0 > 30$ become wider and mostly cover 0. This may be partly due to the reduced power/efficiency as t_0 approaches the upper tail of X, resulting in smaller effective sample sizes for the proposed estimator. The insignificant difference between $LCQRR(\tau; t_0)$ and 0 with $t_0 > 30$ may also have the implication that the occurrence of DN has diminished prognostic power for mortality among patients who had lived long since diabetes diagnosis. In addition, we observe that the estimated $LCQRR(\tau; t_0)$ appears rather constant for $\tau = 0.25$ and $\tau = 0.5$, but the decreasing trend in the estimated $LCQRR(\tau; t_0)$ with $\tau = 0.75$ is quite apparent. This observation is confirmed by the constancy tests for H_{02}, which yield p-values, 0.95, 0.23, and 0.01 for $\tau = 0.25, 0.5, 0.75$ respectively. The significant changing pattern of $LCQRR(\tau; t_0)$ may second the previously conjectured inhomogeneous prognostic ability of DN on mortality.

We also choose three t_0 values, $t_0 = 15, 21, 29$, which stand for the 25th, 50th

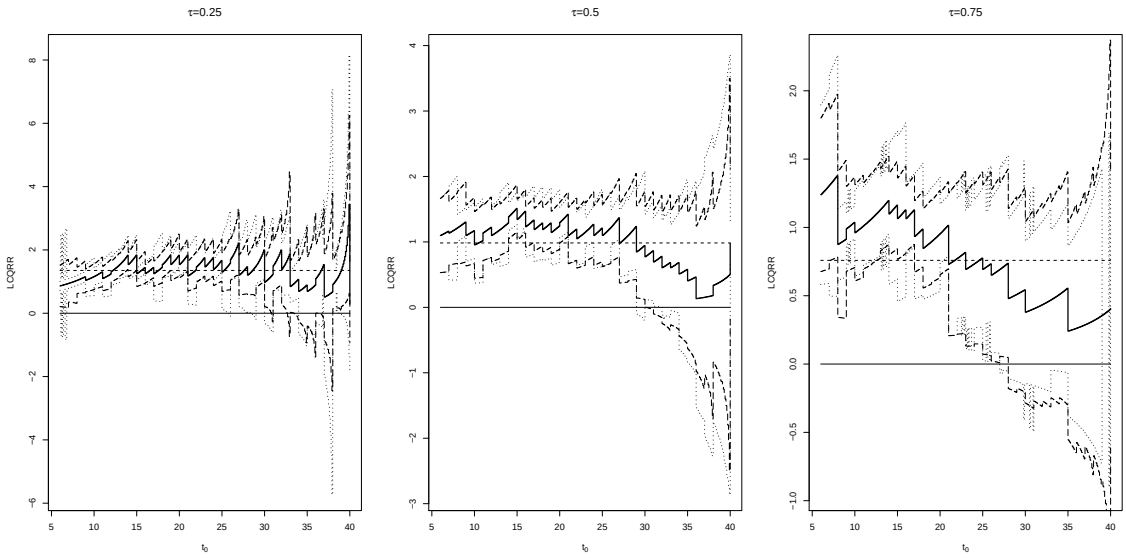

Figure 2.3: Denmark Diabetes Registry Study: Estimated $LCQRR(\tau; t_0)$ (bold solid lines), the corresponding 95% pointwise confidence intervals (dotted lines), 95% pointwise Wald-type bootstrapping confidence intervals (long-dashed lines), and the overall influence of DN across time (horizontal dashed lines).

and 75th quantile of X^*, respectively, to explore the patterns of $LCQRR(\tau; t_0)$ over $\tau \in [0.1, 0.82]$. Figure 2.4 displays estimated $\hat{R}_l(\tau, t_0)$ in bold solid lines at equally spaced τ-grids with step size 0.001, with the corresponding 95% pointwise confidence intervals in dotted lines and 95% pointwise Wald-type bootstrapping confidence intervals in long-dashed lines. We observe that $LCQRR(\tau; t_0)$ may be significantly different from 0 for all three t_0's. This is confirmed by tests for H_{03}, which give p-values, < 0.001, < 0.001, and 0.002, respectively. For $t_0 = 21$ and 29, we observe a clear decreasing trend in the estimated $LCQRR(\tau; t_0)$. Constancy tests for H_{04} yield p-values, 0.24, 0.004, 0.004, for $t_0 = 15, 21, 29$, respectively. The finding that $LCQRR(\tau; t_0)$ may decrease with τ aligns with previous results, manifesting a weak or negligible association between DN and mortality in long-term diabetes survivors.

Next, we study how diabetes onset age, a continuous covariate, affects the dependence between DN and mortality. We fit model (2.5) to the data and the coefficient $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ represent the change in $LCQRR$ per one year increase in diabetes onset

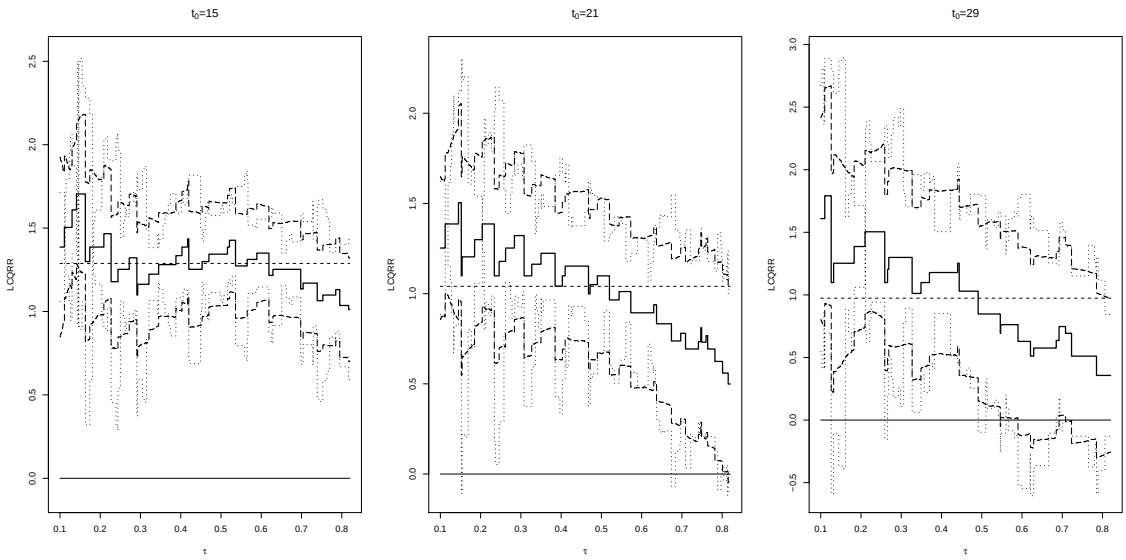

Figure 2.4: Denmark Diabetes Registry Study: Estimated $LCQRR(\tau; t_0)$ (bold solid lines), the corresponding 95% pointwise confidence intervals (dotted lines), 95% pointwise Wald-type bootstrapping confidence intervals (long-dashed lines), and overall influence of DN over τ (horizontal dashed lines)

age. For $\tau = 0.25, 0.5, 0.75$, we estimate $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ at an equally spaced grid on $[8, 36]$ with step size 0.1 for t_0. In Figure 2.5, we display the estimates for $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ along with their 95% pointwise confidence intervals. We see from Figure 2.5 that with all selected τ's, $\hat{\boldsymbol{\gamma}}_0^{(4)}(\tau, t_0)$ is generally significantly positive for t_0 belong to the first half of the time interval [8, 36], but loses significance from 0 for larger t_0. This suggests that for patients who were diagnosed with diabetes at older age, the occurrence of DN before t_0 may imply a bigger disadvantage in residual survival time. Such an effect of diabetes onset age may diminish for large t_0's, which point to the groups of patients who had survived for a long time since diagnosis. Tests for H_{01} over $t_0 \in [8, 22)$ confirm our observation from Figure 2.5, yielding three nearly zero p-values. Constancy tests for H_{02} gave p-values, 0.64,0.11,0.07, respectively, for $\tau = 0.25, 0.5, 0.75$. This provides some evidence for the observed diminishing effect of diabetes onset age over t_0.

We also evaluate $\hat{\boldsymbol{\gamma}}_0^{(4)}(\tau, t_0)$ over a τ-range $[0.1, 0.82]$ for fixed t_0 values, $15, 21, 29$.

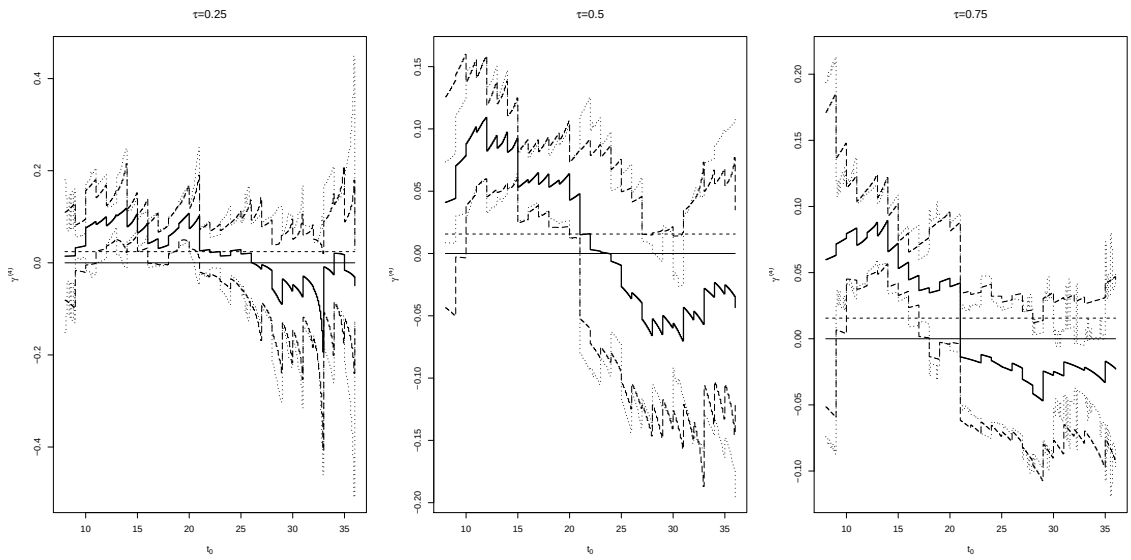

Figure 2.5: Denmark Diabetes Registry Study: Estimated $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ (bold solid lines), corresponding 95% pointwise confidence intervals (dotted lines), and 95% pointwise Wald-type bootstrapping confidence intervals (long-dashed lines).

Results displayed in Figure 2.6 suggest similar findings. That is, DN may have a bigger influence on subsequent mortality for patients with later diabetes diagnosis compared to those with earlier diagnosis. Such an effect of diagnosis age may varnish when t_0 is large.

2.6 Remarks

In this paper, we propose a robust measure to assess the dependence of the nonterminal event and the terminal event in a semi-competing risks setting. Evaluating this measure at multiple t_0 and τ allows us to perform a comprehensive and robust evaluation of semi-competing risks dependence. It also offers the flexibility to explore the dynamic pattern of the dependence structure. The developed estimation and inference procedures well utilize the semi-competing risks structure with left truncation, and can be extended to adjust for covariates. Simulation studies show that the proposed estimation procedure performs well in finite sample cases.

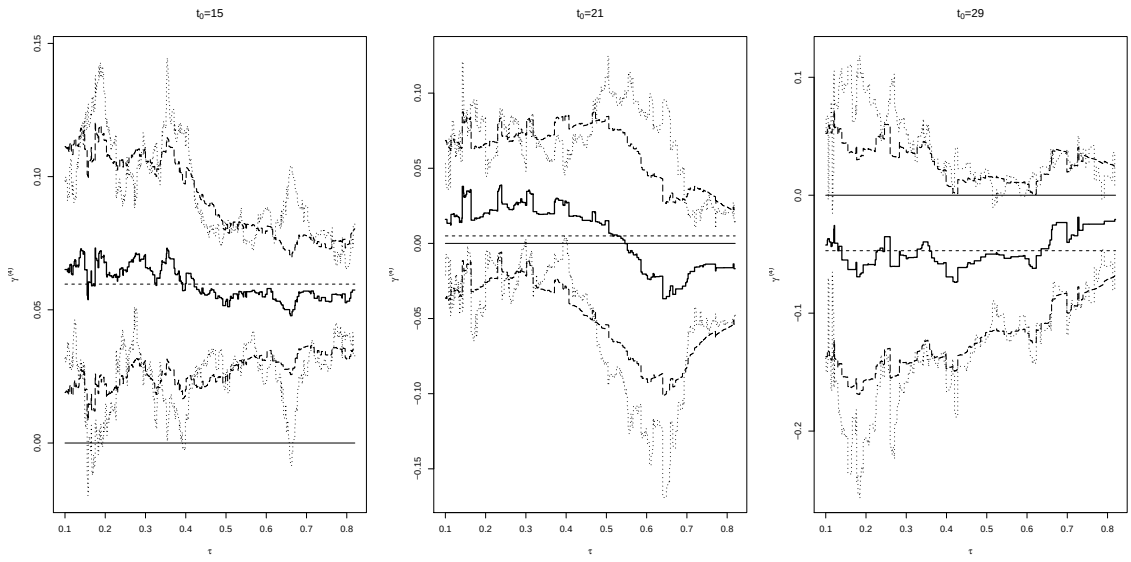

Figure 2.6: Denmark Diabetes Registry Study: Estimated $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ (bold solid lines), the corresponding 95% pointwise confidence intervals (dotted lines), and 95% pointwise Wald-type bootstrapping confidence intervals (long-dashed lines).

Other approaches to obtaining a nonparametric estimator of $LCQRR(\tau; t_0)$ are available. For example, in the standard semi-competing risks setting without left truncation, note that $T_1 \wedge T_2$ is only subject to independent censoring by C and thus the joint survival function of (T_1, T_2) on the upper wedge can be consistently estimated by using methods, such as Lin and Ying (1993). Then we can estimate the two conditional residual quantiles in $LCQRR(\tau; t_0)$ by reversing their corresponding conditional distribution estimates. Our preference of adopting a quantile residual lifetime regression framework is primarily because of the resulting simple extension to accommodate covariates in the consideration of $LCQRR(\tau; t_0)$. Our strategy of connecting $LCQRR$ with quantile residual lifetime regression models enables a unified approach to characterizing semi-competing risks dependence with or without covariates. Existing techniques for quantile regression can readily be applied to inferences and make our work neat.

In practice, the choices of τ and t_0 mainly depend on the interest of investigators. They may be adjusted according to the empirical observations of the data. For

example, the estimation efficacy may be unsatisfactory at small or large values of t_0. This is because the number of observations satisfying $X^* \leq t_0$ (or $X^* > t_0$) may be quite small when t_0 is small (or larger), making the estimate for $Q_\tau(T_2 - t_0|T_2 > t_0, T_1 \leq t_0)$ (or $Q_\tau(T_2 - t_0|T_2 > t_0, T_1 > t_0)$) inaccurate or unstable. Based on our numerical experiences, we find that our method works well for estimating both $LCQRR(\tau; t_0)$ and covariance matrix when $n_{t_0,1} \wedge n_{t_0,2} > 15$, where $n_{t_0,1} = \sum_{i=1}^n I(L_i^* \leq t_0, Y_i^* > t_0, X_i^* > t_0)\eta_i^*$ and $n_{t_0,2} = \sum_{i=1}^n I(L_i^* \leq t_0, Y_i^* > t_0, X_i^* \leq t_0)\eta_i^*$. For a larger τ, we may need $n_{t_0,1}$ and $n_{t_0,2}$ to be larger. These can serve as useful empirical rules to guide the selection of τ and t_0 in real data analysis.

2.7 Appendix

Define

$$\boldsymbol{S}_n(\boldsymbol{b},\tau,t_0) = n^{-1/2}\sum_{i=1}^{n}\frac{I(L_i^*\le t_0)I(Y_i^*>t_0)\eta_i^*}{\hat{G}(Y_i^*-L_i^*)}\boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^*-t_0)\le \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}]-\tau\},$$

$$\boldsymbol{S}_n^G(\boldsymbol{b},\tau,t_0) = n^{-1/2}\sum_{i=1}^{n}\frac{I(L_i^*\le t_0)I(Y_i^*>t_0)\eta_i^*}{G(Y_i^*-L_i^*)}\boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^*-t_0)\le \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}]-\tau\},$$

$$\boldsymbol{\mu}(\boldsymbol{b},\tau,t_0) = n^{-1/2}E\{\boldsymbol{S}_n^G(\boldsymbol{b},\tau,t_0)\}.$$

For brevity, we use $\sup_{\boldsymbol{b}}$, $\sup_{\tau}$ and $\sup_{t_0}$ to denote supremum taken over $\boldsymbol{b}\in R^2$, $\tau\in[\tau_L,\tau_U]$ and $t_0\in[t_L,t_U]$, respectively.

2.7.1 Proof of Theorem 2.2.1

By condition C1, we have $\sup_{t<\nu}|\hat{G}(t)-G(t)| = o(n^{-1/2+r})$, a.s., for every $r>0$. This implies that

$$\sup_{\boldsymbol{b},\tau,t_0}\|n^{-1/2}\boldsymbol{S}_n(\boldsymbol{b},\tau,t_0)-n^{-1/2}\boldsymbol{S}_n^G(\boldsymbol{b},\tau,t_0)\| = o(n^{-1/2+r}),\quad a.s.$$

Define $\mathcal{F} = \left\{\frac{I(L_i^*\le t_0)I(Y_i^*>t_0)\eta_i^*}{G(Y_i^*-L_i^*)}\boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^*-t_0)\le \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}]-\tau\}, \boldsymbol{b}\in R^2, \tau\in[\tau_L,\tau_U], t_0\in[t_L,t_U]\right\}$. The function class $\mathcal{F}$ is Donsker and thus Glivenko-Cantelli because the class indicator functions is Donsker and both $\boldsymbol{A}_i^*(t_0)$ and $G(Y_i^*-L_i^*)$ is uniformly bounded (Van der Vaart and Wellner, 1996). Then $\sup_{\boldsymbol{b},\tau,t_0}\|n^{-1/2}\boldsymbol{S}_n^G(\boldsymbol{b},\tau,t_0)-\boldsymbol{\mu}(\boldsymbol{b},\tau,t_0)\| = o(1), a.s.$ by the Glivenko-Cantelli Theorem and thus $\sup_{\boldsymbol{b},\tau,t_0}\|n^{-1/2}\boldsymbol{S}_n(\boldsymbol{b},\tau,t_0)-\boldsymbol{\mu}(\boldsymbol{b},\tau,t_0)\| = o(1), a.s.$. This, coupled with the fact that $\boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\} = 0$ and $n^{-1/2}\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}(\tau,t_0),\tau,t_0) = o(1), a.s.$, implies

that

$$\sup_{\tau,t_0} \|\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}(\tau,t_0),\tau,t_0\} - \boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}\| = o(1), \quad a.s.$$

Following the same line of Peng and Fine (2009), we can show that Condition C3 and the monotonicity of $\boldsymbol{\mu}(\boldsymbol{b},\tau,t_0)$ in $\boldsymbol{b}$ imply

$$\inf_{\boldsymbol{b}\notin B(\rho_0),\tau,t_0} \|\boldsymbol{\mu}\{\boldsymbol{b},\tau,t_0\} - \boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}\| \geq c_0\rho_0.$$

Consequently, $\{\hat{\boldsymbol{\beta}}(\tau,t_0) : \tau \in [\tau_L,\tau_U], t_0 \in [t_L,t_U]\} \subseteq B(\rho_0)$ for large enough n with probability 1. Applying Taylor expansion to $\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}(\tau,t_0),\tau,t_0\}$ around $\boldsymbol{\beta}_0(\tau,t_0)$ gives

$$\begin{aligned}
&\sup_{\tau,t_0}\|\hat{\boldsymbol{\beta}}(\tau,t_0) - \boldsymbol{\beta}_0(\tau,t_0)\| \\
&\quad = \sup_{\tau,t_0} \|H\{\breve{\boldsymbol{\beta}}(\tau,t_0),t_0\}^{-1}[\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}(\tau,t_0),\tau,t_0\} - \boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}]\| \\
&\quad \leq c_0^{-1} \sup_{\tau,t_0} \|\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}(\tau,t_0),\tau,t_0\} - \boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}\|
\end{aligned}$$

where $\breve{\boldsymbol{\beta}}(\tau,t_0)$ lies between $\hat{\boldsymbol{\beta}}(\tau,t_0)$ and $\boldsymbol{\beta}_0(\tau,t_0)$ and is therefore within $B(\rho_0)$ for large enough n. The uniform consistency of $\hat{\boldsymbol{\beta}}(\tau,t_0)$ to $\boldsymbol{\beta}_0(\tau,t_0)$ for $\tau \in [\tau_L,\tau_U], t_0 \in [t_L,t_U]$ then follows.

2.7.2 Proof of Theorem 2.2.2

From Pepe (1991), $\sup_{t\in[0,\nu)} \|n^{1/2}[\hat{G}(t)-G(t)] - n^{-1/2}\sum_{i=1}^n G(t)\int_0^t y(s)^{-1}dM_i^G(s)\| \to 0$. Using similar empirical process arguments for $\mathcal{F}$, we can show that $n^{-1}\sum_{i=1}^n \boldsymbol{A}_i^*(t_0)Y_i(t)I(L_i^* \leq t_0)I(Y_i^* > t_0)\eta_i^*\{I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}] - \tau\}G(Y_i^* - L_i^*)^{-1}$ converges to $\boldsymbol{w}(\boldsymbol{b},\tau,t_0,t)$ uniformly in $\boldsymbol{b},\tau,t_0$ and t.

Let $\approx$ denote asymptotic equivalence uniformly in $\tau \in [\tau_L,\tau_U]$ and $t_0 \in [t_L,t_U]$.

Simple algebraic manipulations show that

$$
\begin{aligned}
&\boldsymbol{S}_n\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}\\
&=\boldsymbol{S}_n^G\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}+[\boldsymbol{S}_n\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}-\boldsymbol{S}_n^G\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}]\\
&=n^{-1/2}\sum_{i=1}^n\boldsymbol{\xi}_{1,i}(\tau,t_0)-n^{-1/2}\sum_{i=1}^n\boldsymbol{A}_i^*(t_0)\frac{\hat{G}(Y_i^*-L_i^*)-G(Y_i^*-L_i^*)}{\hat{G}(Y_i^*-L_i^*)G(Y_i^*-L_i^*)}I(L_i^*\le t_0)I(Y_i^*>t_0)\eta_i^*\\
&\quad\times\{I[\log(Y_i^*-t_0)\le\boldsymbol{A}_i^{*T}(t_0)\boldsymbol{\beta}_0(\tau,t_0)]-\tau\}\\
&\approx n^{-1/2}\sum_{i=1}^n\boldsymbol{\xi}_{1,i}(\tau,t_0)-n^{-1}\sum_{i=1}^n\boldsymbol{A}_i^*(t_0)\frac{n^{-1/2}\sum_{j=1}^n\int_0^\infty Y_i(s)y(s)^{-1}dM_j^G(s)}{G(Y_i^*-L_i^*)}I(L_i^*\le t_0)I(Y_i^*>t_0)\eta_i^*\\
&\quad\times\{I[\log(Y_i^*-t_0)\le\boldsymbol{A}_i^{*T}(t_0)\boldsymbol{\beta}_0(\tau,t_0)]-\tau\}\\
&=n^{-1/2}\sum_{i=1}^n\boldsymbol{\xi}_{1,i}(\tau,t_0)\\
&\quad-n^{-1/2}\sum_{i=1}^n\int_0^\infty\left\{\sum_{j=1}^n\frac{\boldsymbol{A}_j^*(t_0)Y_j(s)I(L_j^*\le t_0)I(Y_j^*>t_0)\eta_j^*\{I[\log(Y_j^*-t_0)\le\boldsymbol{A}_j^{*T}(t_0)\boldsymbol{\beta}_0(\tau,t_0)]-\tau\}}{nG(Y_j^*-L_j^*)}\right\}\\
&\quad\times\frac{dM_i^G(s)}{y(s)}\\
&\approx n^{-1/2}\sum_{i=1}^n\boldsymbol{\xi}_{1,i}(\tau,t_0)-n^{-1/2}\sum_{i=1}^n\int_0^\infty\boldsymbol{w}(\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0,s)\frac{dM_i^G(s)}{y(s)}\\
&=n^{-1/2}\sum_{i=1}^n\{\boldsymbol{\xi}_{1,i}(\tau,t_0)-\boldsymbol{\xi}_{1,2}(\tau,t_0)\}.
\end{aligned}
$$

We claim that $\mathcal{F}^*=\{\boldsymbol{\xi}_{1,i}(\tau,t_0),\tau\in[\tau_L,\tau_U],t_0\in[t_L,t_U]\}$ and $\mathcal{F}^{**}=\{\boldsymbol{\xi}_{2,i}(\tau,t_0),\tau\in[\tau_L,\tau_U],t_0\in[t_L,t_U]\}$ are Donsker classess by using similar arguments of Peng and Fine (2009). As a result of the Donsker theorem, $\boldsymbol{S}_n\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}$ converges weakly to a mean zero Gaussian process with covariance matrix $\boldsymbol{\Sigma}(\tau',t_0',\tau,t_0)=E\{\boldsymbol{\zeta}(\tau',t_0')\boldsymbol{\zeta}(\tau,t_0)^T\}$, where $\boldsymbol{\zeta}(\tau,t_0)=\boldsymbol{\xi}_1(\tau,t_0)-\boldsymbol{\xi}_2(\tau,t_0)$.

Next, we establish the asymptotic linearity of $\boldsymbol{S}_n^G(\boldsymbol{b},\tau,t_0)$ in the vicinity of $\boldsymbol{b}=$

$\boldsymbol{\beta}_0(\tau, t_0)$; that is, for any positive sequence of $\{d_n\}_{n=1}^{\infty}$ such that $d_n \to 0$,

$$\sup_{\boldsymbol{b},\boldsymbol{b}' \in B(\rho_0), \|\boldsymbol{b}-\boldsymbol{b}'\| \le d_n} \|\{\boldsymbol{S}_n^G(\boldsymbol{b}, \tau, t_0) - \boldsymbol{S}_n^G(\boldsymbol{b}', \tau, t_0)\} - n^{1/2}\{\boldsymbol{\mu}(\boldsymbol{b}, \tau, t_0) - \boldsymbol{\mu}(\boldsymbol{b}', \tau, t_0)\}\| = o(1), a.s. \tag{2.6}$$

Its proof greatly resembles the lines of Alexander (1984) and Lai and Ying (1988). The key is to show

$$\begin{aligned} &Var(I(L_i^* \le t_0) I(Y_i^* > t_0) \eta_i^* G(Y_i^* - L_i^*)^{-1} \boldsymbol{A}_i^*(t_0) \{I[\log(Y_i^* - t_0) \le \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}] \\ &\quad - I[\log(Y_i^* - t_0) \le \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}']\}) \le G_0 \|\boldsymbol{b} - \boldsymbol{b}'\|. \end{aligned}$$

This follows from the uniform boundedness of $f(t|\tilde{\boldsymbol{A}}(t_0))$ and boundedness of $B(\rho_0)$ and $G(t)$.

It follows from (2.6) that

$$\begin{aligned} &\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}(\tau, t_0), \tau, t_0) - \boldsymbol{S}_n(\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0) \\ &= n^{-1/2} \sum_{i=1}^{n} I(L_i^* \le t_0) I(Y_i^* > t_0) \eta_i^* G(Y_i^* - L_i^*)^{-1} \boldsymbol{A}_i^*(t_0) \{I[\log(Y_i^* - t_0) \le \boldsymbol{A}_i^{*T}(t_0)\hat{\boldsymbol{\beta}}(\tau, t_0)] \\ &\quad - I[\log(Y_i^* - t_0) \le \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{\beta}_0(\tau, t_0)]\} \\ &\quad + n^{-1/2} \sum_{i=1}^{n} I(L_i^* \le t_0) I(Y_i^* > t_0) \eta_i^* \boldsymbol{A}_i^*(t_0) \{I[\log(Y_i^* - t_0) \le \boldsymbol{A}_i^{*T}(t_0)\hat{\boldsymbol{\beta}}(\tau, t_0)] \\ &\quad - I[\log(Y_i^* - t_0) \le \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{\beta}_0(\tau, t_0)]\} \{\hat{G}(Y_i^* - L_i^*)^{-1} - G(Y_i^* - L_i^*)^{-1}\} \\ &\approx n^{1/2}[\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}(\tau, t_0), \tau, t_0\} - \boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0\}]. \end{aligned}$$

Taylor expansion of $\boldsymbol{\mu}(\boldsymbol{b})$ around $\boldsymbol{b} = \boldsymbol{\beta}_0(\tau, t_0)$, along with the fact that $\hat{\boldsymbol{\beta}}_0(\tau, t_0)$ uniformly converges to $\boldsymbol{\beta}_0(\tau, t_0)$, gives that

$$\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}(\tau, t_0), \tau, t_0) - \boldsymbol{S}_n(\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0) \approx \boldsymbol{H}\{\boldsymbol{\beta}_0(\tau, t_0), t_0\} n^{1/2} \{\hat{\boldsymbol{\beta}}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)\}.$$

This implies

$$n^{1/2}\{\hat{\boldsymbol{\beta}}(\tau,t_0)-\boldsymbol{\beta}_0(\tau,t_0)\}\approx -\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau,t_0),t_0\}^{-1}\boldsymbol{S}_n(\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0)$$

and then $n^{1/2}\{\hat{\boldsymbol{\beta}}(\tau,t_0)-\boldsymbol{\beta}_0(\tau,t_0)\}$ converges weakly to a mean zero Gaussian process with covariance matrix

$$\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau',t_0'),t_0'\}^{-1}E\{\boldsymbol{\zeta}(\tau',t_0')\boldsymbol{\zeta}(\tau,t_0)^T\}\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau,t_0),t_0\}^{-T}.$$

2.7.3 Justification for the proposed covariance estimate

Denote $\boldsymbol{b}_{n,j}(\tau,t_0)=\boldsymbol{S}_n^{-1}\{\boldsymbol{e}_{n,j}(\tau,t_0),\tau,t_0\}, j=1,2$. It is implied from the proof of Theorem 2.2.1 that $\{\boldsymbol{b}_{n,j}(\tau,t_0),\tau\in[\tau_L,\tau_U],t_0\in[t_L,t_U]\}$ is within $B(\rho_0)$ with probability 1 for large enough n, and thus $\sup_{\tau,t_0}\|\boldsymbol{b}_{n,j}(\tau,t_0)-\boldsymbol{\beta}_0(\tau,t_0)\|\to 0$, a.s., $j=1,2$. Using arguments similar to proof of weak convergence, we can show that

$$\boldsymbol{S}_n(\boldsymbol{b}_{n,j}(\tau,t_0),\tau,t_0)-\boldsymbol{S}_n(\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0)\approx \boldsymbol{H}\{\boldsymbol{\beta}_0(\tau,t_0),t_0\}n^{1/2}\{\boldsymbol{b}_{n,j}(\tau,t_0)-\boldsymbol{\beta}_0(\tau,t_0)\}.$$

The definitions of $\boldsymbol{D}_n(\tau,t_0)$ and $\boldsymbol{E}_n(\tau,t_0)$ imply $\boldsymbol{H}^{-1}\{\boldsymbol{\beta}_0(\tau,t_0),t_0\}\approx\sqrt{n}\boldsymbol{D}_n(\tau,t_0)\boldsymbol{E}_n^{-1}(\tau,t_0)$. It follows immediately that

$$n\boldsymbol{D}_n(\tau',t_0')\boldsymbol{E}_n^{-1}(\tau',t_0')\hat{\boldsymbol{\Sigma}}(\tau',t_0',\tau,t_0)\boldsymbol{E}_n^{-1}(\tau,t_0)\boldsymbol{D}_n^T(\tau,t_0)$$

is a consistent estimate for $\boldsymbol{\Phi}(\tau',t_0',\tau,t_0)=\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau',t_0'),t_0'\}^{-1}\boldsymbol{\Sigma}(\tau',t_0',\tau,t_0)\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau,t_0),t_0\}^{-T}$, which is the asymptotic covariance matrix of $\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau,t_0)-\boldsymbol{\beta}_0(\tau,t_0)\}$.

Chapter 3

Estimation of the New Dependence Measure for Semi-competing Risks Data under the General Truncation Scheme

In Chapter 2, we have proposed a dependence measure $LCQRR(\tau; t_0)$ that can capture a dynamic relationship between the nonterminal and the terminal event in the semi-competing risks scenario without requiring distributional assumptions. We have also developed an estimating approach for $LCQRR(\tau; t_0)$ that can address the semi-competing risks data subject to left truncation. One crucial assumption taken in Chapter 2 was that the gap time between truncation and censoring (i.e. $C - L$) was independent of the truncation time (L) itself. Such an assumption, however, can be restrictive in practice. For example, for some cohort studies that end up at a fixed calendar time, subjects who enter the study earlier intuitively are more likely to have longer follow-up time. This would make the independence of $C - L$ and L be inappropriate. In this chapter, we consider a more general scenario where L is allowed to depend on $C - L$. Our numerical studies will show that the proposed estimator in Chapter 2 is considerably biased under this general scenario. We propose a new estimator for $LCQRR(\tau; t_0)$ that can handle the left truncation, without requiring the independent assumption of $C - L$ and L.

3.1 Estimation and Inference Procedures

We first consider one-sample case. Here, we assume that (L, C) is independent of (T_1, T_2) and C has a continuous distribution.

3.1.1 The proposed estimator

To estimate $LCQRR(\tau; t_0)$ in the general truncation-censoring scheme that allows $C-L$ to depend on L, our idea is similar to that in Chapter 2. That is, we appropriately weigh the contributions of subjects who have complete observations on T_2 and also live beyond t_0, consisting of those satisfying conditions that $Y_i^* > t_0$, $L_i^* \leq t_0$ and $\eta_i^* = 1$. Define $D(s,t) = \frac{1}{\alpha} P(L \leq s, C > t)$, where $\alpha = P(Y > L)$. Note a fact

that $f_{(L^*,C^*,T_1^*,T_2^*)}(l,c,t_1,t_2) = \frac{1}{\alpha} f_{(L,C,T_1,T_2)}(l,c,t_1,t_2)$ in the region of $\{(l,c,t_1,t_2) : l \leq t_0, t_2 > t_0, t_2 < c, t_1 > t_0\}$, where $f_{(L,C,T_1,T_2)}(l,c,t_1,t_2)$ denote the joint distribution functions of (L,C,T_1,T_2), and $f_{(L^*,C^*,T_1^*,T_2^*)}(l,c,t_1,t_2)$ denote the joint distribution function of (L^*,C^*,T_1^*,T_2^*). We can show that

$$
\begin{aligned}
&E\left\{ \frac{I(L^* \leq t_0) I(Y^* > t_0) \eta^*}{D(t_0, Y^*)} \boldsymbol{A}^*(t_0) \{ I[\log(Y^* - t_0) \leq \boldsymbol{A}^{*T}(t_0) \boldsymbol{\beta}_0(\tau, t_0)] - \tau \} \right\} \\
&\quad = E\left\{ \frac{I(L^* \leq t_0) I(T_2^* > t_0, T_2^* < C^*)}{D(t_0, T_2^*)} \tilde{\boldsymbol{A}}^*(t_0) \{ I[\log(T_2^* - t_0) \leq \tilde{\boldsymbol{A}}^{*^T}(t_0) \boldsymbol{\beta}_0(\tau, t_0)] - \tau \} \right\} \\
&\quad = E\left\{ \frac{I(L \leq t_0) I(T_2 > t_0, T_2 < C)}{\alpha D(t_0, T_2)} \tilde{\boldsymbol{A}}(t_0) \{ I[\log(T_2 - t_0) \leq \tilde{\boldsymbol{A}}^T(t_0) \boldsymbol{\beta}_0(\tau, t_0)] - \tau \} \right\} \\
&\quad = E\left\{ \frac{I(T_2 > t_0) \tilde{\boldsymbol{A}}(t_0) \{ I[\log(T_2 - t_0) \leq \tilde{\boldsymbol{A}}^T(t_0) \boldsymbol{\beta}_0(\tau, t_0)] - \tau \}}{\alpha D(t_0, T_2)} E[I(L \leq t_0, C > T_2) | T_1, T_2] \right\} \\
&\quad = E\left\{ I(T_2 > t_0) \tilde{\boldsymbol{A}}(t_0) \{ I[\log(T_2 - t_0) \leq \tilde{\boldsymbol{A}}^T(t_0) \boldsymbol{\beta}_0(\tau, t_0)] - \tau \} \right\} \\
&\quad = 0.
\end{aligned}
$$

These suggest that $D(t_0, Y^*)$ sever as a weight for bias correction in the estimation of $\boldsymbol{\beta}_0(\tau, t_0)$. In light of the equations above, we propose to estimate $\boldsymbol{\beta}_0(\tau, t_0)$ by solving the following estimating equation for $\boldsymbol{b}$:

$$
\boldsymbol{S}_n(\boldsymbol{b}, \tau, t_0) = \boldsymbol{0}, \tag{3.1}
$$

where

$$
\boldsymbol{S}_n(\boldsymbol{b}, \tau, t_0) = n^{-1/2} \sum_{i=1}^{n} \frac{I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^*}{\hat{D}(t_0, Y_i^*)} \boldsymbol{A}_i^*(t_0) \{ I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0) \boldsymbol{b}] - \tau \}.
$$

We denote the resulting estimator by $\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0)$. To distinguish, we denote the estimator proposed in Chapter 2 as $\hat{\boldsymbol{\beta}}_{SP}(\tau, t_0)$. Here, $\hat{D}(s,t)$ is a reasonable estimate for $D(s,t)$.

To obtain an estimator for $D(t_0, Y_i^*)$, note that estimating equation (3.1) is based

on the subsample satisfying $Y_i^* > t_0$. It thus suffices to estimate the bivariate function $D(s,t)$ for $s < t$ rather than over the whole domain. Let $G(s) = P(L \leq s < C)$, $S(t|s) = P(C > t|L \leq s < C)$ and $S_{T_2}(s) = P(T_2 > s)$. Based on facts that $\alpha D(s,t) = G(s)S(t|s)$ for $s < t$ and $G(s) = \alpha P(L^* \leq s < Y^*)/S_{T_2}(s)$, we have

$$D(s,t) = P(L^* \leq s < Y^*)S(t|s)/S_{T_2}(s), \quad \text{for} \quad s < t.$$

Therefore, we propose to estimate $D(s,t)$ by substituting each of its elements with corresponding estimates, as

$$\hat{D}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(L_i^* \leq s < Y_i^*)\hat{S}(t|s)}{\hat{S}_{T_2}(s)}, \quad s < t.$$

$\hat{S}_{T_2}(s)$ can be a Kaplan-Meier type of estimator

$$\hat{S}_{T_2}(s) = \prod_{Y_j^* \leq s} \left\{ 1 - \frac{\sum_{i=1}^{n} I(Y_i^* = Y_j^*, \eta_i^* = 1)}{\sum_{i=1}^{n} I(L_i^* \leq Y_j^* \leq Y_i^*)} \right\}.$$

To construct an estimator for $S(t|s)$, one may base on the Nelson-Aalen type of estimator for $\Lambda(t|s)$, where $\Lambda(t|s) = \int_s^t \lambda(u|s)du$ and $\lambda(t|s) = \lim_{h\to 0} P(t \leq C < t + h|L \leq s < C, C \geq t)/h$. Define $W_s(u) = P(L^* \leq s < Y^* \leq u, \eta^* = 0)$ and $C_s(u) = P(L^* \leq s < u \leq Y^*)$. It is easy to show

$$\Lambda(t|s) = \int_s^t \frac{W_s(du)}{C_s(u)}, \quad s < t.$$

Thus, an estimator for $\Lambda(t|s)$ can take the form as

$$\Lambda_n(t|s) \equiv \int_s^t \frac{W_{n,s}(du)}{C_{n,s}(u)} = \sum_{s < Y_j^* \leq t} \frac{\sum_{i=1}^{n} I(L_i^* \leq s, Y_i^* = Y_j^*, \eta_i^* = 0)}{\sum_{i=1}^{n} I(L_i^* \leq s, Y_i^* \geq Y_j^*)},$$

where $W_{n,s}(u) = \frac{1}{n}\sum_{i=1}^{n} I(L_i^* \leq s < Y_i^* \leq u, \eta_i^* = 0)$ and $C_{n,s}(u) = \frac{1}{n}\sum_{i=1}^{n} I(L_i^* \leq$

$s < u \leq Y_i^*$). In light of the fact that $S(t|s) = \exp\{-\Lambda(t|s)\}$, one may estimate $S(t|s)$ as

$$\hat{S}(t|s) = \exp\{-\Lambda_n(t|s)\}.$$

Under some regularity conditions, we can show the uniform consistency of $\hat{D}(s,t)$, for $s < t$ (see Appendix).

When a proper $\hat{D}(s,t)$, $s < t$, is available, equation (3.1) can be easily solved by following the same lines of Chapter 2. Given the monotonicity of equation (3.1), we can transform its solution finding to locating the minimizer of the convex function $U_n(\boldsymbol{b}, \tau, t_0)$ given by

$$\begin{aligned} U_n(\boldsymbol{b}, \tau, t_0) = & \sum_{i=1}^{n} I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* \left| \frac{\log(Y_i^* - t_0)}{\hat{D}(t_0, Y_i^*)} - \boldsymbol{b}^T \frac{\boldsymbol{A}_i^*(t_0)}{\hat{D}(t_0, Y_i^*)} \right| \\ & + \left| M - (2\tau - 1)\boldsymbol{b}^T \sum_{i=1}^{n} I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* \frac{\boldsymbol{A}_i^*(t_0)}{\hat{D}(t_0, Y_i^*)} \right|, \end{aligned}$$

where M is a sufficiently large positive number that can bound $\left|(2\tau - 1)\boldsymbol{b}^T \sum_{i=1}^{n} I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* \frac{\boldsymbol{A}_i^*(t_0)}{\hat{D}(t_0, Y_i^*)}\right|$. Minimization of the L_1-type function $U_n(\boldsymbol{b}, \tau, t_0)$ can be solved by using standard software, like the $rq()$ function in the contributed R package *quantreg*.

3.1.2 Asymptotic results

For a non-negative random variable K, define $a_K = \inf\{k : P(K \leq k) > 0\}$ and $b_K = \sup\{k : P(K \leq k) < 1\}$. We assume the following regularity conditions:

C1. $a_L < a_Y$, $b_L \leq b_Y$.

C2. (i) $0 < \tau_L \leq \tau_U \leq 1$; (ii) $\inf_{s \in (a_{L^*}, b_{Y^*}], u \in (s, b_{Y^*}]} P(L^* \leq s < u \leq Y^*) > 0$; (iii) t_L and t_U are interior points of the support of X^*.

C3. (i) $\boldsymbol{\beta}_0(\tau, t_0)$ is Lipschitz continuous for $\tau \in [\tau_L, \tau_U]$ and $t_0 \in [t_L, t_U]$; (ii) $f(t|\tilde{\boldsymbol{A}}(t_0))$ is continuous and bounded above uniformly in t, t_0 and $\tilde{\boldsymbol{A}}(t_0)$, where $f(t|\tilde{\boldsymbol{A}}(t_0)) = dF(t|\tilde{\boldsymbol{A}}(t_0))/dt$ and $F(t|\tilde{\boldsymbol{A}}(t_0)) = E\{I(T_2 \leq t)|\tilde{\boldsymbol{A}}(t_0)\}$.

C4. For some $\rho_0 > 0$ and $c_0 > 0$, $\inf_{\boldsymbol{b} \in B(\rho_0), t_0 \in [t_L, t_U]} \text{eigmin} \boldsymbol{H}(\boldsymbol{b}, t_0) \geq c_0$, where $B(\rho) = \{\boldsymbol{b} \in R^2 : \inf_{\tau \in [\tau_L, \tau_U], t_0 \in [t_L, t_U]} ||\boldsymbol{b} - \boldsymbol{\beta}_0(\tau, t_0)|| \leq \rho\}$ and $\boldsymbol{H}(\boldsymbol{b}, t_0) = E[\tilde{\boldsymbol{A}}(t_0)^{\otimes 2} f(t_0 + \exp(\tilde{\boldsymbol{A}}^T(t_0)\boldsymbol{b})|\tilde{\boldsymbol{A}}^T(t_0)) \exp(\tilde{\boldsymbol{A}}^T(t_0)\boldsymbol{b})]$. Here $||\cdot||$ is the Euclidean norm and $\boldsymbol{u}^{\otimes 2} = \boldsymbol{u}\boldsymbol{u}^T$ for a vector $\boldsymbol{u}$.

We then have the following theorems:

Theorem 3.1.1. *Under conditions C1–C4,*

$$\lim_{n \to \infty} \sup_{\tau \in [\tau_L, \tau_U], t_0 \in [t_L, t_U]} ||\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)|| \to_p 0.$$

Theorem 3.1.2. *Under conditions C1–C4,* $\sqrt{n}\{\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)\}$ *weakly converge to a mean zero Gaussian process with covariance matrix given by*

$$\boldsymbol{\Phi}(\tau', t_0', \tau, t_0) = \boldsymbol{H}\{\boldsymbol{\beta}_0(\tau', t_0'), t_0'\}^{-1} E\{\boldsymbol{\iota_1}(\tau', t_0')\boldsymbol{\iota_1}(\tau, t_0)^T\}[\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau, t_0), t_0\}^{-1}]^T,$$

where the formal definition of $\boldsymbol{\iota_1}(\tau, t_0)$ *is provided in Appendix,* $\tau, \tau' \in [\tau_L, \tau_U]$ *and* $t_0, t_0' \in [t_L, t_U]$.

Detailed proof of Theorem 3.1.1 and 3.1.2 are provided in Section 3.6 Appendix.

3.1.3 Inference

For inference on $\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0)$, we use bootstrapping procedures, given the complexity in the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0)$ shown in the proof of Theorem 3.1.2. Denote $\boldsymbol{\beta}^*(\tau, t_0)$ as the bootstrap estimator. It can be shown that the distribution of

$n^{1/2}\{\boldsymbol{\beta}^*(\tau,t_0)-\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0)\}$ conditionally on the observed data and the unconditional distribution of $n^{1/2}\{\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0)-\boldsymbol{\beta}_0(\tau,t_0)\}$ have the same limiting distribution. By repeatedly resampling from the observed data $(X_i^*,Y_i^*,\delta_i^*,\eta_i^*,L_i^*)_{i=1}^n$, one may obtain a large number of realizations of $n^{1/2}\{\boldsymbol{\beta}^*(\tau,t_0)-\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0)\}$, the empirical distribution of which can be used to give the asymptotic covariance matrix estimate for $\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0)$ or the 95% Wald-type confidence interval for $\boldsymbol{\beta}_0(\tau,t_0)$.

Second-stage inferences can also be conducted in a similar fashion to that of Section 2.2.4. First, we can summarize the average of $LCQRR(\tau;t_0)$ over $t_0\in[t_L,t_U]$ by $\Omega_\tau=\frac{1}{t_U-t_L}\int_{t_L}^{t_U}\boldsymbol{\beta}_0^{(2)}(\tau,t_0)dt_0$. One natural estimate for Ω_τ may be obtained by simply replacing $\boldsymbol{\beta}_0^{(2)}(\tau,t_0)$ by $\hat{\boldsymbol{\beta}}_{GE}^{(2)}(\tau,t_0)$, that is, $\hat{\Omega}_\tau=\frac{1}{t_U-t_L}\int_{t_L}^{t_U}\hat{\boldsymbol{\beta}}_{GE}^{(2)}(\tau,t_0)dt_0$. We can show that $\hat{\Omega}_\tau$ is consistent and asymptotic normal. Bootstrapping-based inference on Ω_τ can be developed using realizations of $\Omega_\tau^*=\frac{1}{t_U-t_L}\int_{t_L}^{t_U}\boldsymbol{\beta}^{*(2)}(\tau,t_0)dt_0$, naturally rendering a Wald-type test for the null hypothesis $H_{01}:\ LCQRR(\tau;t_0)=0,t_0\in[t_L,t_U]$. Similar results can be obtained for the overall summary and testing of $LCQRR(\tau;t_0)$ over $\tau\in[\tau_L,\tau_U]$, corresponding to $\Omega_{t_0}=\frac{1}{\tau_U-\tau_L}\int_{\tau_L}^{\tau_U}\boldsymbol{\beta}_0^{(2)}(\tau,t_0)d\tau$, and $H_{03}:\ LCQRR(\tau;t_0)=0,\tau\in[\tau_L,\tau_U]$ respectively.

Another second-stage hypothesis of interest is given by $H_{02}:LCQRR(\tau;t_0)=C_\tau$, $t_0\in[t_L,t_U]$, where C_τ is an unspecified constant and may change with τ. To test H_{02}, one may adopt the test statistic $\Gamma_\tau=\sqrt{n}\{\int_{t_L}^{t_U}\Xi(\tau,t_0)\hat{\boldsymbol{\beta}}_{GE}^{(2)}(\tau,t_0)dt_0-\hat{\Omega}_\tau\}$, where $\Xi(\tau,t_0)$ is a non-constant weight function satisfying $\int_{t_L}^{t_U}\Xi(\tau,t_0)dt_0=1$. We can show that the limit distribution of Γ_τ under H_{02} is normal with mean 0. A consistent variance estimate for Γ_τ can be obtained through bootstrapping. This would render a Wald-type test for H_{02}. A similar testing procedure can also be developed for testing the constancy over $\tau\in[\tau_L,\tau_U]$, $H_{04}:\ LCQRR(\tau;t_0)=C_{t_0},\ \tau\in[\tau_L,\tau_U]$.

3.2 An Extension to Covariates Adjustment

As Chapter 2, we are interested in the linear effects of covariates on $LCQRR(\tau; t_0)$, expressed as model (2.4). Our strategy is also to employ model (2.5) as a working model.

To estimate $\boldsymbol{\gamma}_0(\tau, t_0)$ in model (2.5), we adapt the methods presented for the one-sample case based on the observed data $(X_i^*, Y_i^*, \delta_i^*, \eta_i^*, L_i^*, \tilde{\boldsymbol{Z}}_i^*)_{i=1}^n$. Under the assumption that (L, C) is independent of $(T_1, T_2, \tilde{\boldsymbol{Z}})$, we propose to estimate $\boldsymbol{\gamma}_0(\tau, t_0)$ by solving the following estimating equation for $\boldsymbol{r} \in R^{2+2p}$:

$$\boldsymbol{S}_n(\boldsymbol{r}, \tau, t_0) = \boldsymbol{0},$$

where

$$\boldsymbol{S}_n(\boldsymbol{r}, \tau, t_0) = n^{-1/2} \sum_{i=1}^{n} \frac{I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^*}{\hat{D}(t_0, Y_i^*)} \boldsymbol{K}_i^*(t_0) \{ I[\log(Y_i^* - t_0) \leq \boldsymbol{K}_i^{*T}(t_0) \boldsymbol{r}] - \tau \}.$$

With an additional assumption that $\tilde{\boldsymbol{Z}}$ is uniformly bounded (i.e. $\sup_i ||\tilde{\boldsymbol{Z}}_i|| \leq M_1 < \infty$), we can establish the same asymptotic properties and inference procedures for the estimator for $\boldsymbol{\gamma}_0(\tau, t_0)$, denoted by $\hat{\boldsymbol{\gamma}}(\tau, t_0)$, as those presented in Section 3.1.

3.3 Simulation Studies

Extensive simulation studies are conducted to evaluate the finite sample performances of the proposed estimators in the left-truncated semi-competing risks setting. Specifically, we generate (T_1, T_2) from a gamma frailty model,

$$P(T_1 > x, T_2 > y) = [P(T_1 > x)^{1-\theta} + P(T_2 > y)^{1-\theta} - 1]^{1/(1-\theta)},$$

in which T_i follows a Weibull(α_i, λ_i) distribution and $P(T_i > x) = exp(-\lambda_i x^{\alpha_i})$, $i = 1, 2$. The truncation time $L = r \times L_0$, where r is a random variable following Bernoulli distribution with probability p, and L_0 is a positive random variable that is independent of r. Such a truncation scenario mimics the Denmark diabetes registry study, where the distribution of L has a point mass at 0. We generate the censoring time C as $L + D_0$, where D_0 is a positive-valued random variable dependent on L.

The simulations are conducted under two scenarios and each scenario consists of three setups correspondent to three θ values, 1, 2 and 3, which separately reflects independence, moderate positive association and high positive association. In Scenario 1, let $T_1 \sim$ Weibull$(2, 0.5)$ and $T_2 \sim$ Weibull$(3.2, 0.35)$; the truncation rate $P(Y < L)$ is set low to moderate as 0.3, and the dependent censoring rate $P(\delta^* = 0, \eta^* = 1)$ is set moderate to high, close to 0.36. Scenario 2 involves a moderate to high level of truncation as 0.45 and a low to moderate level of dependent censoring around 0.2, with $T_1 \sim$ Weibull$(2.5, 0.75)$ and $T_2 \sim$ Weibull$(3, 0.4)$. In both scenarios, the proportion of zero truncation time $P(L^* = 0)$ is set as 0.2. Details about the choice of $\{p, L_0, D_0\}$ in each setup as well as corresponding censoring proportions are given in Table 3.1.

Table 3.1: Summary of simulation setups: the choice of $\{p, L_0, D_0\}$ and the resulting truncation and censoring proportions, where $p_1 = P(Y < L)$, $p_2 = P(\delta^* = 0)$, $p_3 = P(\eta^* = 0)$ and $p_4 = P(\delta^* = 0, \eta^* = 1)$.

θ	p	L_0	D_0	p_1	p_2	p_3	p_4
			Scenario 1: $T_1 \sim$ Weibull$(2, 0.5), T_2 \sim$ Weibull$(3.2, 0.35)$				
1	0.86	Unif(0, 1.9)	$Beta(1.4, 1) \times (1.95 - 0.8L)^2$	0.30	0.47	0.24	0.35
2	0.86	Unif(0, 1.9)	$Beta(1.4, 1) \times (1.95 - 0.8L)^2$	0.30	0.54	0.24	0.36
3	0.86	Unif(0, 1.9)	$Beta(1.4, 1) \times (1.95 - 0.8L)^2$	0.30	0.57	0.24	0.38
			Scenario 2: $T_1 \sim$ Weibull$(2.5, 0.75), T_2 \sim$ Weibull$(3, 0.4)$				
1	0.89	min{Weibull(1.4, 0.52), 1.9}	$Beta(1.3, 1) \times (2 - 0.8L)^2$	0.45	0.33	0.24	0.24
2	0.89	min{Weibull(1.2, 0.54), 1.9}	$Beta(1.4, 1) \times (2.2 - 1.1L)^2$	0.45	0.32	0.24	0.20
3	0.89	min{Weibull(1.0, 0.55), 1.9}	$Beta(1.6, 1) \times (2.2 - 1.2L)^2$	0.45	0.28	0.24	0.17

Under each setup, we implement the proposed methods on 1000 simulated datasets with sample size $n = 400$, where M is set as 10^7. For bootstrapping-based inference, we set the resampling size as 250. The empirical bias (EmpBias), empirical standard error (EmpSE) and average estimated standard error based on bootstrapping (EstSE) as well as the empirical coverage rate of the 95% Wald-type confidence intervals for the proposed estimator of $LCQRR(\tau; t_0)$, are reported under different combinations of (θ, τ, t_0), where $\tau = 0.25, 0.5, 0.75$, and t_0 are at grid over an interval $[t_L, t_U]$ with grid size 0.005. In Scenario 1, we set $t_L = 0.65$ and $t_U = 1.30$. In Scenario 2, we set $t_L = 0.62$ and $t_U = 1.20$. Figure 3.1 presents the results for Scenario 1, showing that the proposed estimator of $LCQRR(\tau; t_0)$ performs well with moderate sample size. The point estimates have small biases. The bootstrapping-based standard error estimates closely match their empirical counterparts, and the Wald-type confidence intervals based on normal approximation provide satisfactory empirical coverage. For comparison purpose, we also provide the empirical bias of the estimator proposed in Chapter 2 (i.e., $LCQRR_{SP}(\tau; t_0)$), which requires the independence of L and $C - L$. It is shown that, under our scenario that $C - L$ depends on L, $LCQRR_{SP}(\tau; t_0)$ can lead to substantial biases, in particular in cases with larger θ and t_0 values. We have similar findings in Figure 3.2, which presents the simulation results for Scenario 2.

We also evaluate the performance of the proposed average estimator of $LCQRR$ over $t_0 \in [t_L, t_U]$ for fixed τ, as well as the Wald tests for H_{01} and H_{02} in second-stage inferences. Still consider three τ values, 0.25, 0.5, 0.75, and the interval $[t_L, t_U]$ as previously mentioned. We compute integrals using left Riemann sums on intervals of equal length 0.005 and choose the weight function $\Xi(\tau, t_0) = 2I[t_0 \leq (t_L + t_U)/2]/(t_U - t_L)$. Table 3.2 summaries the EmpBias, EmpSE and EstSE of $\hat{\Omega}_\tau$, and the empirical rejection rates (EmpRR) for the two Wald tests. It shows that for both scenarios, the empirical biases of $\hat{\Omega}_\tau$ are small and the estimated standard errors agree very well with corresponding empirical standard errors. The test for either H_{01} or H_{02} appear

to have empirical sizes (when $\theta = 1$) close to the nominal levels. The power for testing H_{01} (when $\theta = 2$ and 3) is good, while the constancy tests seems conservative.

Table 3.2: Summary of simulation study: EmpBias, EmpSE and EstSE of $\hat{\Omega}_\tau$ and empirical rejection rates for H_{01} and H_{02}.

			$\hat{\Omega}_\tau$		H_{01}	H_{02}
θ	τ	EmpBias	EmpSE	EstSE	EmpRR	EmpRR
			Scenario 1			
			$t_0 \in [0.65, 1.30]$			
1	0.25	0.000	0.143	0.148	0.048	0.050
	0.50	0.001	0.125	0.125	0.052	0.047
	0.75	-0.002	0.113	0.114	0.055	0.045
2	0.25	-0.001	0.133	0.134	1.000	0.090
	0.50	-0.003	0.111	0.112	0.999	0.125
	0.75	-0.001	0.103	0.106	0.989	0.142
3	0.25	0.000	0.118	0.122	1.000	0.112
	0.50	0.000	0.102	0.106	1.000	0.171
	0.75	-0.00	0.103	0.103	1.000	0.202
			Scenario 2			
			$t_0 \in [0.62, 1.20]$			
1	0.25	0.000	0.150	0.153	0.051	0.053
	0.50	0.001	0.124	0.124	0.048	0.056
	0.75	-0.002	0.108	0.109	0.045	0.056
2	0.25	0.001	0.125	0.127	1.000	0.124
	0.50	0.001	0.098	0.102	1.000	0.163
	0.75	-0.004	0.087	0.093	0.999	0.167
3	0.25	0.001	0.107	0.108	1.000	0.162
	0.50	-0.003	0.087	0.091	1.000	0.173
	0.75	-0.000	0.103	0.103	1.000	0.202

For fixed t_0, we examine the second-stage inferences over $[\tau_L, \tau_U]$. We set $t_0 = 0.7, 1.0, 1.2$ for Scenario 1, and $t_0 = 0.7, 0.9, 1.1$ for Scenario 2. Let $[\tau_L, \tau_U] = [0.15, 0.85]$ with grid size 0.001 and $\Xi(\tau, t_0) = 2I[\tau \leq (\tau_L + \tau_U)/2]/(\tau_U - \tau_L)$. Table 3.3 presents the EmpBias, EmpSE and EstSE of $\hat{\Omega}_{t_0}$ and the EmpRR for the proposed tests. Similarly, we observe small empirical biases, well-matched estimated and empirical standard errors and empirical sizes close to nominal levels. The power for H_{03} is good, while for the constancy tests is not high.

Table 3.3: Summary of simulation study: EmpBias, EmpSE and EstSE of $\hat{\Omega}_{t_0}$ and empirical rejection rates for H_{03} and H_{04}.

			$\hat{\Omega}_{t_0}$		H_{03}	H_{04}
θ	t_0	EmpBias	EmpSE	EstSE	EmpRR	EmpRR
			Scenario 1			
1	0.7	0.004	0.139	0.145	0.045	0.045
	1.0	0.002	0.150	0.149	0.055	0.045
	1.2	0.002	0.179	0.175	0.055	0.046
2	0.7	0.003	0.154	0.155	0.901	0.072
	1.0	-0.002	0.155	0.153	0.979	0.118
	1.2	0.003	0.174	0.174	0.976	0.124
3	0.7	0.003	0.154	0.157	1.000	0.124
	1.0	0.000	0.159	0.160	1.000	0.257
	1.2	0.001	0.178	0.183	1.000	0.175
			Scenario 2			
1	0.7	-0.002	0.142	0.145	0.048	0.054
	0.9	-0.003	0.145	0.143	0.048	0.046
	1.1	0.001	0.173	0.170	0.058	0.052
2	0.7	0.001	0.143	0.142	0.965	0.083
	0.9	-0.001	0.132	0.135	0.995	0.167
	1.1	-0.004	0.143	0.146	0.998	0.237
3	0.7	0.002	0.140	0.142	1.000	0.186
	0.9	-0.003	0.129	0.131	1.000	0.386
	1.1	0.003	0.142	0.138	1.000	0.515

3.4 Denmark Diabetes Registry Data Analysis

In this section, we apply the proposed method to the Denmark diabetes registry study (Andersen et al., 1993), with the same objective as in Chapter 2. That is, we quantify the relationship between DN and death. Here, we focus on a subcohort of patients who had diabetes onset age greater than 19 in order to reduce the population heterogeneity as the disease mechanism of childhood diabetes may be different from that of adult diabetes. Among the 854 patients in this subcohort, 181(21%) subjects experienced DN and 239(28%) died during the study. Approximately 28% patients had diabetic onset at the study entry.

We fit model (2.1) to the dataset and adopt $M = 10^7$ as in the simulations. To en-

sure reasonable sample sizes accumulated for strata defined by $I(X^* > t_0)$, we restrict t_0 to be within $[12, 28)$. Table 3.4 shows results for estimated $LCQRR(\tau; t_0)$, corresponding 95% pointwise Wald-type bootstrapping confidence interval with resampling size of 500 and p-value for testing $LCQRR(\tau; t_0) = 0$ at several combinations of (τ, t_0), where $\tau = 0.25,\ 0.5$ and $t_0 = 12,\ 15,\ 18,\ 21,\ 25,\ 27$. Here, $t_0 = 15,\ 21$ separately represents the 25th, 50th quantile for time to DN. We see that for either fixed τ, the estimated $LCQRR(\tau; t_0)$ is generally significantly positive. This observation is consistent with the common belief that DN is a diabetes progression landmark, positively associated with mortality. Such belief is also confirmed by our formal test for H_{01}, in which t_0s are equally-spaced on $[12, 28)$ with step size=0.1, yielding p-values < 0.001. We note that the difference between $LCQRR(\tau; t_0)$ and 0 becomes insignificant when t_0 is generally beyond 25. This may imply that the occurrence of DN would lose its prognostic power for mortality among patients who has lived long since diabetes diagnosis. For fixed τ, the estimated $LCQRR(\tau; t_0)$ appears to have a decreasing trend over t_0, possibly indicating DN's prognostic power for mortality get weaker as time goes by.

In Table 3.4, we also provide results for the estimated $LCQRR(\tau; t_0)$ based on $\hat{\boldsymbol{\beta}}^{(2)}_{SP}(\tau, t_0)$. In contrast to the estimated $LCQRR(\tau; t_0)$ based on $\hat{\boldsymbol{\beta}}^{(2)}_{GE}(\tau, t_0)$, the approach in Chapter 2 would result in some quite different estimated $LCQRR(\tau; t_0)$ values. For example, at $(\tau, t_0) = (0.5, 21)$, $LC\hat{Q}RR_{SP}(\tau; t_0)$ (i.e., $\hat{\boldsymbol{\beta}}^{(2)}_{SP}(\tau, t_0)$) is 0.29 with p-value=0.62, showing a non-significant difference between 0; while our proposed $LC\hat{Q}RR_{GE}(\tau; t_0)$ (i.e., $\hat{\boldsymbol{\beta}}^{(2)}_{GE}(\tau, t_0)$) is 1.15 with p-value=0.004, indicating a significantly positive association between DN and mortality there. Such big discrepancy may be caused by the potential dependence of underlying $C - L$ and L for this real dataset. Recall our findings in simulation studies, a violation of the independent assumption of $C - L$ and L would lead to severely biased estimation in $LCQRR(\tau; t_0)$.

We next study whether diabetes onset age, a continuous covariate, affects the

Table 3.4: Denmark Diabetes Registry Study: estimated $LC\hat{Q}RR_{GE}(\tau;t_0)$ and $LC\hat{Q}RR_{SP}(\tau;t_0)$, 95% pointwise Wald-type bootstrapping confidence interval and corresponding p-value.

t_0	$LC\hat{Q}RR_{GE}(\tau;t_0)$	95% CI	P-value	$LC\hat{Q}RR_{SP}(\tau;t_0)$	95% CI	P-value
			$\tau = 0.25$			
12	1.95	$(1.40, 2.50)$	< 0.001	2.08	$(1.53, 2.62)$	< 0.001
15	1.87	$(1.23, 2.52)$	< 0.001	1.95	$(1.27, 2.62)$	< 0.001
18	1.87	$(1.41, 2.34)$	< 0.001	1.54	$(1.05, 2.03)$	< 0.001
21	1.30	$(0.78, 1.82)$	< 0.001	1.10	$(0.37, 1.83)$	0.003
25	0.98	$(0.21, 1.76)$	0.013	0.69	$(-0.66, 2.04)$	0.31
27	0.85	$(-0.32, 2.01)$	0.15	0.47	$(-0.97, 1.91)$	0.52
			$\tau = 0.5$			
12	1.75	$(1.20, 2.50)$	< 0.001	1.87	$(1.35, 2.40)$	< 0.001
15	1.44	$(1.10, 1.77)$	< 0.001	1.61	$(1.22, 2.00)$	< 0.001
18	1.48	$(0.89, 2.08)$	< 0.001	1.53	$(0.75, 2.31)$	< 0.001
21	1.15	$(0.37, 1.93)$	0.004	0.29	$(-0.87, 1.44)$	0.62
25	0.76	$(-0.36, 1.89)$	0.18	0.06	$(-1.02, 1.14)$	0.91
27	0.44	$(-0.73, 1.62)$	0.46	0.07	$(-0.85, 0.99)$	0.88

dependence between DN and mortality. We fit model (2.5) to the data and the coefficient $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ represents the change in $LCQRR(\tau;t_0)$ per one year increase in diabetes onset age. For $\tau = 0.25$ and 0.5, we estimate $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ at several t_0 values mentioned above. In Table 3.5, we display the estimated $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ along with corresponding 95% pointwise Wald-type bootstrapping confidence interval and p-value for testing $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0) = 0$. We see that there is no significant difference between $\boldsymbol{\gamma}_0^{(4)}(\tau, t_0)$ and 0. Tests for H_{01} confirm our observations, yielding p-values around 0.37. These suggest diabetes onset age may not affect the dependence between DN and mortality in groups of young adults.

3.5 Remarks

We propose a new robust estimator for $LCQRR(\tau;t_0)$ to address the general semi-competing risks scenario with left truncation. The proposed approach does not require

Table 3.5: Denmark Diabetes Registry Study: change rate of $LCQRR(\tau; t_0)$ by diabetes onset age ($i.e.$, $\hat{\gamma}_0^{(4)}(\tau, t_0)$), 95% pointwise Wald-type bootstrapping confidence interval and corresponding p-value.

t_0	change rate of $LCQRR(\tau; t_0)$	95% CI	P-value
	$\tau = 0.25$		
12	-0.04	$(-0.19, 0.10)$	0.55
15	-0.07	$(-0.24, 0.10)$	0.41
18	-0.01	$(-0.14, 0.11)$	0.84
21	-0.02	$(-0.27, 0.23)$	0.89
25	0.12	$(-0.25, 0.48)$	0.54
27	0.18	$(-0.28, 0.64)$	0.43
	$\tau = 0.5$		
12	0.05	$(-0.07, 0.17)$	0.42
15	-0.05	$(-0.18, 0.08)$	0.43
18	0.03	$(-0.16, 0.22)$	0.77
21	0.14	$(-0.10, 0.38)$	0.24
25	0.24	$(-0.12, 0.59)$	0.19
27	0.19	$(-0.23, 0.62)$	0.39

additional assumptions other than the typical independence of (L, C) and (T_1, T_2). This makes the proposed estimator be able to handel problems in a more general sampling schemes, while in comparison to the one proposed in Chapter 2. The developed estimation and inference procedures can be easily extended to adjust for covariates. Simulation studies show that the proposed estimation procedure performs well in finite sample cases.

To use $LCQRR(\tau; t_0)$ in practice, we recommend specifying τ and t_0 beforehand according to scientific interests. For example, common choices of τ are 0.25, 0.5 and 0.75, reflecting below average, average, and above average progression to the terminal event. The choice of t_0 may be at time points that landmark the development of the nonterminal event. Specifying τ and t_0 may also be adjusted according to the empirical observations of the data. For example, the estimation efficacy may be unsatisfactory at small values of t_0 because of a small weight $\hat{D}(t_0, Y_i^*)$ or a small number of observations satisfying $X^* \leq t_0$. Based on our numerical experiences,

we find that our method generally works well for estimating both $LCQRR(\tau; t_0)$ when $\min(n_{t_0,1}, n_{t_0,2}) > 30$, where $n_{t_0,1} = \sum_{i=1}^{n} I(L_i^* \leq t_0, Y_i^* > t_0, X_i^* > t_0)\eta_i^*$ and $n_{t_0,2} = \sum_{i=1}^{n} I(L_i^* \leq t_0, Y_i^* > t_0, X_i^* \leq t_0)\eta_i^*$. For a larger τ, we may need $n_{t_0,1}$ and $n_{t_0,2}$ to be larger. These can serve as useful empirical rules to guide the selection of τ and t_0 in real data analysis.

3.6 Appendix

Define

$$\boldsymbol{S}_n(\boldsymbol{b}, \tau, t_0) = n^{-1/2} \sum_{i=1}^{n} \frac{I(L_i^* \leq t_0) I(Y_i^* > t_0)\eta_i^*}{\hat{D}(t_0, Y_i^*)} \boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}] - \tau\},$$

$$\boldsymbol{S}_n^D(\boldsymbol{b}, \tau, t_0) = n^{-1/2} \sum_{i=1}^{n} \frac{I(L_i^* \leq t_0) I(Y_i^* > t_0)\eta_i^*}{D(t_0, Y_i^*)} \boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}] - \tau\},$$

$$\boldsymbol{\mu}(\boldsymbol{b}, \tau, t_0) = n^{-1/2} E\{\boldsymbol{S}_n^D(\boldsymbol{b}, \tau, t_0)\}.$$

For brevity, we use $\sup_{\boldsymbol{b}}$, $\sup_{\tau}$ and $\sup_{t_0}$ to denote supremum taken over $\boldsymbol{b} \in R^2$, $\tau \in [\tau_L, \tau_U]$ and $t_0 \in [t_L, t_U]$, respectively. In the sequel, $o_p^S(n^{-1/2})$ means root n convergence to 0 in probability uniformly on set S.

3.6.1 Proof of Theorem 3.1.1

The first step is to sort out the asymptotic properties of $\hat{D}(s,t)$, for $s < t$. To this end, we need to look at each specific element of this plug-in weight.

Define

$$\xi_s(L_i^*, Y_i^*, \eta_i^*, t) = \frac{I(L_i^* \le s < Y_i^* \le t, \eta_i^* = 0)}{C_s(Y_i^*)} - \int_s^t \frac{I(L_i^* \le s < u \le Y_i^*)}{C_s^2(u)} W_s(du)$$

$$R_{1n,s}(t) = \int_s^t \frac{[C_s(u) - C_{n,s}(u)]^2}{C_{n,s}(u) C_s^2(u)} W_s(du)$$

$$R_{2n,s}(t) = \int_s^t \left\{ \frac{1}{C_{n,s}(u)} - \frac{1}{C_s(u)} \right\} \Big[W_{n,s}(du) - W_s(du) \Big]$$

$$R_{n,s}(t) = R_{1n,s}(t) + R_{2n,s}(t)$$

Then

$$\Lambda_n(t|s) - \Lambda(t|s) = \frac{1}{n} \sum_{i=1}^{n} \xi_s(L_i^*, Y_i^*, \eta_i^*, t) + R_{n,s}(t)$$

with $E[\xi_s(L_i^*, Y_i^*, \eta_i^*, t)] = 0$. We can show

$$\sup_{(s,t) \in \mathcal{S} \times \mathcal{T}} |R_{n,s}(t)| = o_p(n^{-1/2}),$$

where $\mathcal{S} \times \mathcal{T} = \{(s,t) : s \in (a_{L^*}, b_{Y^*}], u \in (s, b_{Y^*}]\}$. This is because

$$|R_{1n,s}(t)| \le \sup_{s < u \le t} |C_s(u) - C_{n,s}(u)|^2 / \inf_{s < u \le t} [C_{n,s}(u) C_s^2(u)].$$

By the LIL for empirical distribution functions and almost surely uniformly zero-away boundness of both $C_{n,s}(u)$ and $C_s^2(u)$ on $\mathcal{S} \times \mathcal{T}$, we have

$$\sup_{(s,t) \in \mathcal{S} \times \mathcal{T}} |R_{1n,s}(t)| = O(n^{-1} \log \log n), a.s.$$

Similarly, we can show $\sup_{(s,t) \in \mathcal{S} \times \mathcal{T}} |R_{2n,s}(t)| = O(n^{-1} \log \log n), a.s.$ and thus

$$\sup_{(s,t) \in \mathcal{S} \times \mathcal{T}} |R_{n,s}(t)| = o_p(n^{-1/2}).$$

By Taylor expansion, we have

$$\hat{S}(t|s) - S(t|s) = \frac{1}{n}\sum_{i=1}^{n}\{-S(t|s)\xi_s(L_i^*, Y_i^*, \eta_i^*, t)\} + o_p^{\mathcal{S}\times\mathcal{T}}(n^{-1/2})$$
$$\equiv \frac{1}{n}\sum_{i=1}^{n} l_i(s,t) + o_p^{\mathcal{S}\times\mathcal{T}}(n^{-1/2}), \tag{3.2}$$

where $l_i(s,t) = -S(t|s)\xi_s(L_i^*, Y_i^*, \eta_i^*, t)$ and $El_i(s,t) = 0$.

By results in Gijbels and Wang (1993), we have

$$\hat{S}_{T_2}(s) - S_{T_2}(s) = \frac{1}{n}\sum_{i=1}^{n} a_i(s) + o_p^{[0,\tilde{b}]}(n^{-1/2}), \quad \tilde{b} < b_Y,$$

where

$$a_i(s) = -S_{T_2}(s)\left\{\frac{I(L_i^* \le Y_i^* \le s, \eta_i^* = 1)}{\tilde{R}(Y_i^*)} - \int_0^s \frac{I(L_i^* \le u \le Y_i^*)}{\tilde{R}^2(u)} d\tilde{W}(u)\right\},$$

$\tilde{W}(u) = P(L^* \le Y^* \le u, \eta^* = 1)$, $\tilde{R}(u) = P(L^* \le u \le Y^*)$ and $Ea_i(s) = 0$. Taylor expansion indicates

$$\frac{\hat{S}(t|s)}{\hat{S}_{T_2}(s)} - \frac{S(t|s)}{S_{T_2}(s)} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{l_i(s,t)}{S_{T_2}(s)} - \frac{S(t|s)a_i(s)}{S_{T_2}^2(s)}\right\} + o_p^{\mathcal{S}\times\mathcal{T}}(n^{-1/2})$$
$$\equiv \frac{1}{n}\sum_{i=1}^{n} k_i(t,s) + o_p^{\mathcal{S}\times\mathcal{T}}(n^{-1/2})$$

with $Ek_i(s,t) = 0$. Further,

$$\hat{D}(s,t) - D(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[I(L_i^* \le s < Y_i^*)\frac{S(t|s)}{S_{T_2}(s)} - D(s,t)\right] + \frac{1}{n}\sum_{j=1}^{n} k_i(s,t)I(L_j^* \le s < Y_j^*)\right\}$$
$$+ o_p^{\mathcal{S}\times\mathcal{T}}(n^{-1/2})$$
$$\equiv \frac{1}{n}\sum_{i=1}^{n} d_i(s,t) + o_p^{\mathcal{S}\times\mathcal{T}}(n^{-1/2}) \tag{3.3}$$

with $Ed_i(s,t) = 0$. If we can claim that the functional class $\{d_i(s,t), s \in \mathcal{S}, t \in \mathcal{T}\}$ is

Donsker, then thus Glivenko-Cantelli and $\hat{D}(s,t)$ is uniformly consistent for $D(s,t)$ on $\mathcal{S}\times\mathcal{T}$. By the functional law of the iterated logarithm (Goodman et al., 1981), equation (3.3) implies $\sup_{(s,t)\in\mathcal{S}\times\mathcal{T}}|\hat{D}(s,t)-D(s,t)|=o(n^{-1/2+r})$ for $0<r<\frac{1}{2}$ and consequently

$$\sup_{\boldsymbol{b},\tau,t_0}\|n^{-1/2}\boldsymbol{S}_n(\boldsymbol{b},\tau,t_0)-n^{-1/2}\boldsymbol{S}_n^D(\boldsymbol{b},\tau,t_0)\|=o(n^{-1/2+r}),\quad a.s. \tag{3.4}$$

To show $\{d_i(s,t),s\in\mathcal{S},t\in\mathcal{T}\}$ is Donsker, we first prove that $\{l_i(s,t),s\in\mathcal{S},t\in\mathcal{T}\}$ forms a Donsker class. This can be shown because the class of indicator functions is Donsker, $C_s(u)$ and $C_s(Y_i^*)$ are uniformly bounded away from 0 on $\mathcal{S}\times\mathcal{T}$ and the map $\pi: x(s,u)\to\int_s^t x(s,u)W_s(du)$ is a linear continuous map. Similar arguments and the boundness of $S_{T_2}^{-1}(s)$ on $\mathcal{S}$ lead to show $\{d_i(s,t),s\in\mathcal{S},t\in\mathcal{T}\}$ is Donsker.

Now we can finish the proof by following the same line of Section 2.7.1. Define $\mathcal{F}=\left\{\frac{I(L_i^*\le t_0)I(Y_i^*>t_0)\eta_i^*}{D(t_0,Y_i^*)}\boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^*-t_0)\le\boldsymbol{A}_i^{*T}(t_0)\boldsymbol{b}]-\tau\},\boldsymbol{b}\in R^2,\tau\in[\tau_L,\tau_U],t_0\in[t_L,t_U]\right\}$. The function class $\mathcal{F}$ is Donsker. Then $\sup_{\boldsymbol{b},\tau,t_0}\|n^{-1/2}\boldsymbol{S}_n^D(\boldsymbol{b},\tau,t_0)-\boldsymbol{\mu}(\boldsymbol{b},\tau,t_0)\|=o(1),a.s.$ by the Glivenko-Cantelli Theorem and thus $\sup_{\boldsymbol{b},\tau,t_0}\|n^{-1/2}\boldsymbol{S}_n(\boldsymbol{b},\tau,t_0)-\boldsymbol{\mu}(\boldsymbol{b},\tau,t_0)\|=o(1),a.s.$. This, coupled with the fact that $\boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}=0$ and $n^{-1/2}\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0),\tau,t_0)=o(1),a.s.$, implies that

$$\sup_{\tau,t_0}\|\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0),\tau,t_0\}-\boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}\|=o(1),\quad a.s.$$

Following the same line of Peng and Fine (2009), we can show that Condition C3 and the monotonicity of $\boldsymbol{\mu}(\boldsymbol{b},\tau,t_0)$ in $\boldsymbol{b}$ imply

$$\inf_{\boldsymbol{b}\notin B(\rho_0),\tau,t_0}\|\boldsymbol{\mu}\{\boldsymbol{b},\tau,t_0\}-\boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}\|\ge c_0\rho_0.$$

Consequently, $\{\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0):\tau\in[\tau_L,\tau_U],t_0\in[t_L,t_U]\}\subseteq B(\rho_0)$ for large enough n with

probability 1. Applying Taylor expansion to $\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0),\tau,t_0\}$ around $\boldsymbol{\beta}_0(\tau,t_0)$ gives

$$\begin{aligned}
&\sup_{\tau,t_0}\|\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0)-\boldsymbol{\beta}_0(\tau,t_0)\| \\
&\quad=\sup_{\tau,t_0}\|H\{\breve{\boldsymbol{\beta}}(\tau,t_0),t_0\}^{-1}[\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0),\tau,t_0\}-\boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}]\| \\
&\quad\le c_0^{-1}\sup_{\tau,t_0}\|\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0),\tau,t_0\}-\boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}\|
\end{aligned}$$

where $\breve{\boldsymbol{\beta}}(\tau,t_0)$ lies between $\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0)$ and $\boldsymbol{\beta}_0(\tau,t_0)$ and is therefore within $B(\rho_0)$ for large enough n. The uniform consistency of $\hat{\boldsymbol{\beta}}_{GE}(\tau,t_0)$ to $\boldsymbol{\beta}_0(\tau,t_0)$ for $\tau\in[\tau_L,\tau_U],t_0\in[t_L,t_U]$ then follows.

3.6.2 Proof of Theorem 3.1.2

Let $\approx$ denote asymptotic equivalence uniformly in $\tau\in[\tau_L,\tau_U]$ and $t_0\in[t_L,t_U]$. First, by equation (3.3), simple algebraic manipulations show that

$$\begin{aligned}
&\boldsymbol{S}_n\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\} \\
&=\boldsymbol{S}_n^D\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}+[\boldsymbol{S}_n\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}-\boldsymbol{S}_n^D\{\boldsymbol{\beta}_0(\tau,t_0),\tau,t_0\}] \\
&\approx n^{-1/2}\sum_{i=1}^n\boldsymbol{\xi}_{1,i}(\tau,t_0)-n^{-1/2}\sum_{i=1}^n\boldsymbol{A}_i^*(t_0)\frac{n^{-1}\sum_{j=1}^n d_j(t_0,Y_i^*)}{D^2(t_0,Y_i^*)}I(L_i^*\le t_0)I(Y_i^*>t_0)\eta_i^* \\
&\times\{I[\log(Y_i^*-t_0)\le\boldsymbol{A}_i^{*T}(t_0)\boldsymbol{\beta}_0(\tau,t_0)]-\tau\} \\
&=n^{-1/2}\sum_{i=1}^n\boldsymbol{\xi}_{1,i}(\tau,t_0) \\
&-n^{-1/2}\sum_{i=1}^n\left\{\sum_{j=1}^n\frac{\boldsymbol{A}_j^*(t_0)I(L_j^*\le t_0)I(Y_j^*>t_0)\eta_j^*\{I[\log(Y_j^*-t_0)\le\boldsymbol{A}_j^{*T}(t_0)\boldsymbol{\beta}_0(\tau,t_0)]-\tau\}}{nD^2(t_0,Y_j^*)}d_i(t_0,Y_j^*)\right\} \\
&\approx n^{-1/2}\sum_{i=1}^n\{\boldsymbol{\xi}_{1,i}(\tau,t_0)-\boldsymbol{\xi}_{2,i}(\tau,t_0)\}.
\end{aligned}$$

where $\boldsymbol{\xi}_{1,i}(\tau,t_0)=I(L_i^*\le t_0)I(Y_i^*>t_0)\eta_i^*\boldsymbol{A}_i^*(t_0)\{I[\log(Y_i^*-t_0)\le\boldsymbol{A}_i^{*T}(t_0)\boldsymbol{\beta}_0(\tau,t_0)]-\tau\}D(t_0,Y_i^*)^{-1}$ and $\boldsymbol{\xi}_{2,i}(\tau,t_0)=E_{\boldsymbol{\omega}_j^*}[\boldsymbol{\xi}_{1,j}(\tau,t_0)D(t_0,Y_j^*)^{-1}d_i(t_0,Y_j^*)]$ with $\boldsymbol{\omega}_i^*$ de-

noting $(X_i^*, Y_i^*, \delta_i^*, \eta_i^*, L_i^*)$ and $E_{\boldsymbol{\omega}_j^*}$ representing the expectation over $\boldsymbol{\omega}_j^*$, $j = 1, ..., n$. Following similar arguments in the proof of Theorem 3.1.1, we claim that $\{\boldsymbol{\xi}_{1,i}(\tau, t_0) - \boldsymbol{\xi}_{2,i}(\tau, t_0), \tau \in [\tau_L, \tau_U], t_0 \in [t_L, t_U]\}$ is also a Donsker class. Thus, $\boldsymbol{S}_n\{\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0\}$ converges weakly to a mean zero Gaussian process with covariance matrix $\boldsymbol{\Sigma}(\tau', t_0', \tau, t_0) = E\{\boldsymbol{\iota_1}(\tau', t_0')\boldsymbol{\iota_1}(\tau, t_0)^T\}$, where $\boldsymbol{\iota}_i(\tau, t_0) = \boldsymbol{\xi}_{1,i}(\tau, t_0) - \boldsymbol{\xi}_{2,i}(\tau, t_0)$.

Next, we establish the asymptotic linearity of $\boldsymbol{S}_n^D(\boldsymbol{b}, \tau, t_0)$ in the vicinity of $\boldsymbol{b} = \boldsymbol{\beta}_0(\tau, t_0)$; that is, for any positive sequence of $\{d_n\}_{n=1}^{\infty}$ such that $d_n \to 0$,

$$\sup_{\boldsymbol{b}, \boldsymbol{b}' \in B(\rho_0), \|\boldsymbol{b} - \boldsymbol{b}'\| \leq d_n} \|\{\boldsymbol{S}_n^D(\boldsymbol{b}, \tau, t_0) - \boldsymbol{S}_n^D(\boldsymbol{b}', \tau, t_0)\} - n^{1/2}\{\boldsymbol{\mu}(\boldsymbol{b}, \tau, t_0) - \boldsymbol{\mu}(\boldsymbol{b}', \tau, t_0)\}\| = o(1), a.s. \tag{3.5}$$

Its proof greatly resembles the lines of Alexander (1984) and Lai and Ying (1988).

It follows from (3.5) that

$$\begin{aligned}
&\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0), \tau, t_0) - \boldsymbol{S}_n(\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0) \\
&\quad = n^{-1/2} \sum_{i=1}^{n} I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* D(t_0, Y_i^*)^{-1} \boldsymbol{A}_i^*(t_0) \{I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0) \hat{\boldsymbol{\beta}}_{GE}(\tau, t_0)] \\
&\qquad - I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0) \boldsymbol{\beta}_0(\tau, t_0)]\} \\
&\qquad + n^{-1/2} \sum_{i=1}^{n} I(L_i^* \leq t_0) I(Y_i^* > t_0) \eta_i^* \boldsymbol{A}_i^*(t_0) \{I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0) \hat{\boldsymbol{\beta}}_{GE}(\tau, t_0)] \\
&\qquad - I[\log(Y_i^* - t_0) \leq \boldsymbol{A}_i^{*T}(t_0) \boldsymbol{\beta}_0(\tau, t_0)]\} \{\hat{D}(t_0, Y_i^*)^{-1} - D(t_0, Y_i^*)^{-1}\} \\
&\quad \approx n^{1/2}[\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0), \tau, t_0\} - \boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0\}].
\end{aligned}$$

Taylor expansion of $\boldsymbol{\mu}(\boldsymbol{b})$ around $\boldsymbol{b} = \boldsymbol{\beta}_0(\tau, t_0)$, along with the fact that $\hat{\boldsymbol{\beta}}_0(\tau, t_0)$ uniformly converges to $\boldsymbol{\beta}_0(\tau, t_0)$, gives that

$$\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0), \tau, t_0) - \boldsymbol{S}_n(\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0) \approx \boldsymbol{H}\{\boldsymbol{\beta}_0(\tau, t_0), t_0\} n^{1/2} \{\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)\}.$$

This implies

$$n^{1/2}\{\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)\} \approx -\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau, t_0), t_0\}^{-1}\boldsymbol{S}_n(\boldsymbol{\beta}_0(\tau, t_0), \tau, t_0)$$

and then $n^{1/2}\{\hat{\boldsymbol{\beta}}_{GE}(\tau, t_0) - \boldsymbol{\beta}_0(\tau, t_0)\}$ converges weakly to a mean zero Gaussian process with covariance matrix

$$\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau', t_0'), t_0'\}^{-1}E\{\boldsymbol{\iota_1}(\tau', t_0')\boldsymbol{\iota_1}(\tau, t_0)^T\}\boldsymbol{H}\{\boldsymbol{\beta}_0(\tau, t_0), t_0\}^{-T}.$$

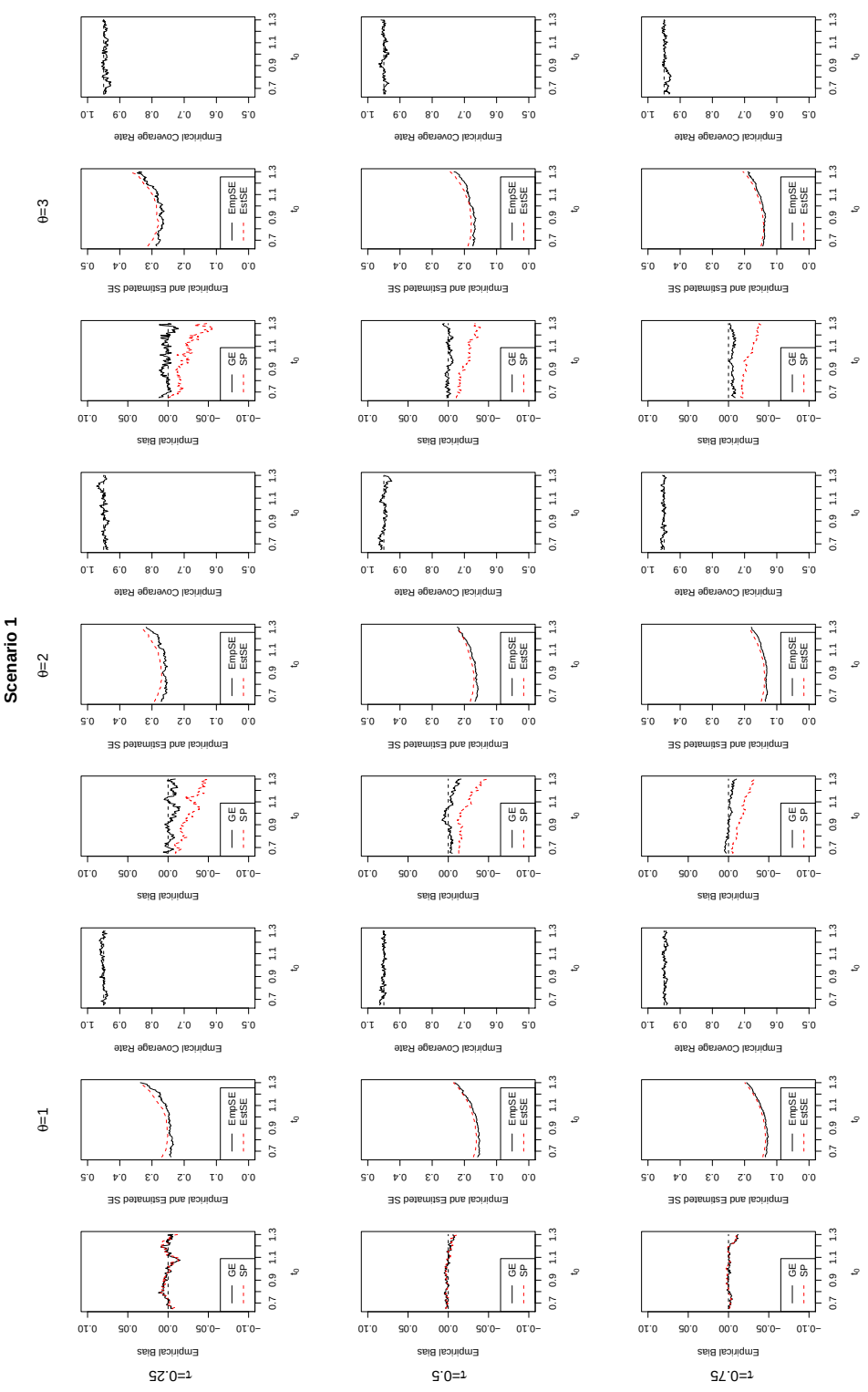

Figure 3.1: Scenario 1: EmpBias of the estimator proposed in Chapter 2 for $LCQRR(\tau;t_0)$ (i.e., $LCQRR_{SP}(\tau;t_0)$); EmpBias, EmpSE and EstSE for the estimator proposed in this chapter (i.e., $LCQRR_{GE}(\tau;t_0)$).

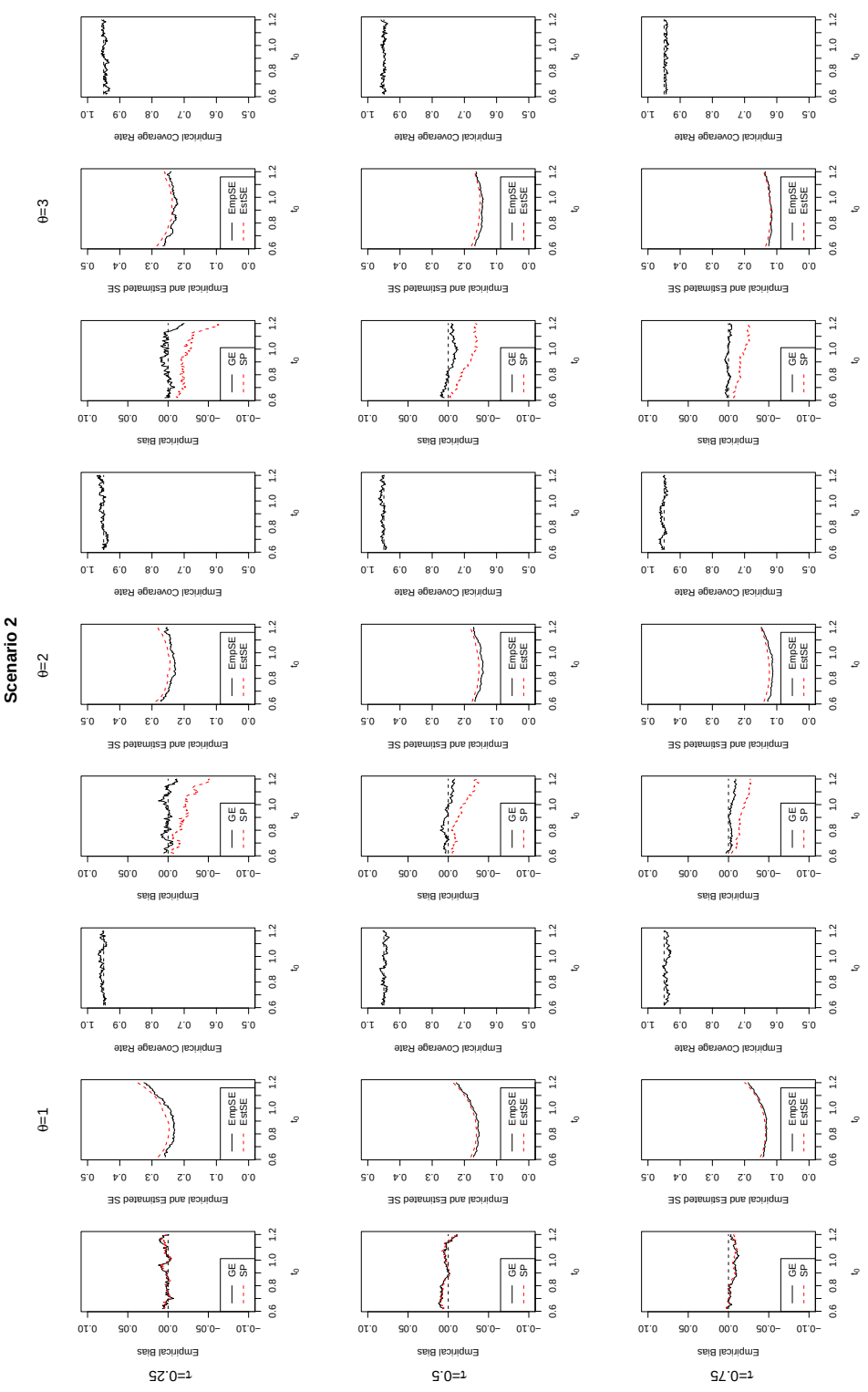

Figure 3.2: Scenario 2: EmpBias of the estimator proposed in Chapter 2 for $LCQRR(\tau;t_0)$ (i.e., $LCQRR_{SP}(\tau;t_0)$); EmpBias, EmpSE and EstSE for the estimator proposed in this chapter (i.e., $LCQRR_{GE}(\tau;t_0)$).

Chapter 4

Semiparametric Regression Procedures for the Association between Bivariate Recurrent Processes

4.1 Association Measure and Model

4.1.1 Data and notation

Let $\boldsymbol{T}_k = \{T_k^{(1)}, T_k^{(2)}, ...\}$ be the recurrence times of type-k events and $(L_k, R_k]$ be an observation window to which $\boldsymbol{T}_k$ is subject, where $k = 1, 2$. For type-k events, denote $N_k(t) = \sum_{j=1}^{\infty} I(T_k^{(j)} \leq t)$ as the underlying process of recurrent events, $\tilde{N}_k(t) = \sum_{j=1}^{\infty} I(L_k \leq T_k^{(j)} \leq t \wedge R_k)$ as the counting process of observed recurrent events and $Y_k(t) = I(L_k \leq t \leq R_k)$ as the at risk process. Let $\boldsymbol{Z} = (1, Z_1, ..., Z_p)$ be the associated $(p+1) \times 1$ vector of covariates. Define $\mu_{\boldsymbol{Z},k}(t) = E[N_k(t)|\boldsymbol{Z}]$, which represents the expected frequency of type-k events by time t given covariates $\boldsymbol{Z}$. Define $\tau_{\boldsymbol{Z},k}(u) = \inf\{t \geq 0 : \mu_{\boldsymbol{Z},k}(t) \geq u\}$, representing time to expected recurrence frequency u of type-k events (Huang and Peng, 2009).

In this work, without loss of generality, we assume that type-1 and type-2 recurrent events share a common observation window $(L, R]$, i.e., $L_1 = L_2 = L$ and $R_1 = R_2 = R$, and also assume that L and R are independent of $N_k(\cdot)$ given $\boldsymbol{Z}$, $k = 1, 2$. The observed data consists of n i.i.d. replicates of $\{N_1(\cdot), N_2(\cdot), \boldsymbol{Z}, L, R\}$, denoted by $\{N_{i1}(\cdot), N_{i2}(\cdot), \boldsymbol{Z}_i, L_i, R_i\}_{i=1}^n$.

Note that when the event of interest can occur only once, u is restricted to $[0, 1]$ and $\tau_{\boldsymbol{Z},k}(u)$ becomes the conditional quantile of $T_k^{(1)}$ given $\boldsymbol{Z}$.

4.1.2 Proposed association measure for bivariate recurrent event data

We start with considering the conditional covariance of $N_1(s)$ and $N_2(t)$, defined as,

$$Cov_{\boldsymbol{Z}}(s, t) = E_{\boldsymbol{Z}}[N_1(s)N_2(t)] - \mu_{\boldsymbol{Z},1}(s)\mu_{\boldsymbol{Z},2}(t), \tag{4.1}$$

where $E_{\boldsymbol{Z}}(\cdot) = E(\cdot|\boldsymbol{Z})$. This measure inherits the general interpretation of covariance. It is easy to see that when $N_1(\cdot)$ and $N_2(\cdot)$ are two independent processes, $Cov_{\boldsymbol{Z}}(s,t) = 0$ for all $s,t > 0$. Since $Cov_{\boldsymbol{Z}}(s,t)$ is formulated based on the counting process notation of recurrent events data, it has the flexibility to capture a varying association structure between the two different types of recurrent events. However, it is important to note that both $E_{\boldsymbol{Z}}[N_1(s)N_2(t)]$ and $\mu_{\boldsymbol{Z},1}(s)\mu_{\boldsymbol{Z},2}(t)$ are increasing functions of s and t. This fact may confound the interpretation of a large value of $Cov_{\boldsymbol{Z}}(s,t)$. That is, an increase in $Cov_{\boldsymbol{Z}}(s,t)$ may result from a scale shift in $E_{\boldsymbol{Z}}[N_1(s)N_2(t)]$ and $\mu_{\boldsymbol{Z},1}(s)\mu_{\boldsymbol{Z},2}(t)$ as s and t increase, which is irrelevant to the association of interest.

A similar issue also exists in the simpler bivariate survival setting, where the outcome may be represented by (Y_1, Y_2). Li et al. (2014) proposed a novel solution to this problem by assessing the association based on quantiles. More specifically, they presented a quantile-specific probability ratio by comparing the joint probability that Y_1 and Y_2 were simultaneously less than their respective τ_1th and τ_2th quantiles to the expected probability under independence, namely,

$$qpr(\tau_1, \tau_2|\boldsymbol{Z}) = \frac{P\{Y_1 \leq Q_1(\tau_1|\boldsymbol{Z}), Y_2 \leq Q_2(\tau_2|\boldsymbol{Z})|\boldsymbol{Z}\}}{P\{Y_1 \leq Q_1(\tau_1|\boldsymbol{Z})|\boldsymbol{Z}\} \times P\{Y_2 \leq Q_2(\tau_1|\boldsymbol{Z})|\boldsymbol{Z}\}}, \tag{4.2}$$

where $Q_k(\tau_k|\boldsymbol{Z}) = \inf\{t : P(Y_k \leq t|\boldsymbol{Z}) \geq \tau_k\}, \tau_k \in (0,1)$, denotes the τ_kth quantile function of Y_k $(k = 1, 2)$. Such a quantile-specific measure has a nice property of scale invariance because the values of indicators $I[Y_1 \leq Q_1(\tau_1|\boldsymbol{Z})]$ and $I[Y_2 \leq Q_2(\tau_2|\boldsymbol{Z})]$ keep the same regardless of any scale change in Y_1 and Y_2. Note that these indicators are also invariant to any monotone transformations of Y_1 and Y_2. With different selections of (τ_1, τ_2), this measure can also provide a comprehensive view of the association between Y_1 and Y_2.

Motivated by the work of Li et al. (2014), we propose to assess the association between two types of recurrent events based on the frequency scale rather than the

time scale. To exploit this idea, we first rewrite (4.2) as

$$qpr(\tau_1, \tau_2|\boldsymbol{Z}) = \frac{E_{\boldsymbol{Z}}\{I[Y_1 \leq Q_1(\tau_1|\boldsymbol{Z})] \times I[Y_2 \leq Q_2(\tau_2|\boldsymbol{Z}]\}}{E_{\boldsymbol{Z}}\{I[Y_1 \leq Q_1(\tau_1|\boldsymbol{Z})]\} \times E_{\boldsymbol{Z}}\{I[Y_2 \leq Q_2(\tau_1|\boldsymbol{Z})]\}}. \tag{4.3}$$

The representation in equation (4.3) suggests a natural adaptation of $qpr(\tau_1, \tau_2|\boldsymbol{Z})$ to the bivariate recurrent event setting. That is, we replace $I(Y_1 \leq \cdot)$ and $I(Y_2 \leq \cdot)$ by $N_1(\cdot)$ and $N_2(\cdot)$, respectively. As investigated by Huang and Peng (2009), $\tau_{\boldsymbol{Z},k}(u)$ that represents time to expected recurrence frequency u for type-k events can be considered as an analogue to the quantile in the recurrent event setting. We thus substitute $Q_1(\tau_1|\boldsymbol{Z})$ and $Q_2(\tau_2|\boldsymbol{Z})$ with $\tau_{\boldsymbol{Z},1}(u)$ and $\tau_{\boldsymbol{Z},2}(v)$ and propose a frequency-specific association measure for bivariate recurrent event data, taking the form,

$$\rho_{\boldsymbol{Z}}(u, v) = \frac{E_{\boldsymbol{Z}}\{N_1[\tau_{\boldsymbol{Z},1}(u)]N_2[\tau_{\boldsymbol{Z},2}(v)]\}}{uv}, \quad u, v > 0. \tag{4.4}$$

Similar to the quantile-specific measure $qpr(\tau_1, \tau_2|\boldsymbol{Z})$, $\rho_{\boldsymbol{Z}}(u, v)$ is invariant to any scale change or monotone transformation of $\boldsymbol{T}_k$.

The definition of $\rho_{\boldsymbol{Z}}(u, v)$ also reflects calibrations of the two marginal mean functions by setting $\mu_{\boldsymbol{Z},1}(s) = u$ and $\mu_{\boldsymbol{Z},2}(t) = v$. By such calibrations, the two arguments in $\rho_{\boldsymbol{Z}}(\cdot, \cdot)$ are both on the frequency scale. This makes $\rho_{\boldsymbol{Z}}(u, v)$ more comparable among different covariate groups, and hence facilitates the interpretation of potential regression analysis for this new association measure. It is easy to see that

$$\rho_{\boldsymbol{Z}}(u, v) = 1 + \frac{Cov_{\boldsymbol{Z}}(\tau_{\boldsymbol{Z},1}(u), \tau_{\boldsymbol{Z},2}(v))}{uv}.$$

By this connection, we name $\rho_{\boldsymbol{Z}}(u, v)$ as frequency-specific adjusted covariance measure.

The proposed measure $\rho_{\boldsymbol{Z}}(u, v)$ is easy to interpret. For example, $\rho_{\boldsymbol{Z}}(u, v) > 1$ indicates that the counts of type-1 recurrent events at time $\tau_{\boldsymbol{Z}}(u)$ and the counts

of type-2 recurrent events at time $\tau_{\boldsymbol{Z}}(v)$ are positively associated. That is, greater (smaller) cumulative recurrences of type-1 events at time to expected frequency u tend to be associated with greater (smaller) cumulative recurrences of type-2 events at time to expected frequency v, conditionally on $\boldsymbol{Z}$. Similarly, $0 < \rho_{\boldsymbol{Z}}(u, v) < 1$ may suggest that greater (smaller) cumulative recurrences of type-1 events at time to expected frequency u tend to be associated with smaller (greater) cumulative recurrences of type-2 events at time to expected frequency v, conditionally on $\boldsymbol{Z}$. When $N_1(\cdot)$ and $N_2(\cdot)$ are independent, we have $\rho_{\boldsymbol{Z}}(u, v) = 1$ for all $u, v > 0$.

4.1.3 Proposed regression model for $\rho_{\boldsymbol{Z}}(u, v)$

Given the fact that $\rho_{\boldsymbol{Z}}(u, v)$ is always positive, we propose a regression model for $\rho_{\boldsymbol{Z}}(u, v)$ that takes the form,

$$\rho_{\boldsymbol{Z}}(u, v) = \exp\{\boldsymbol{Z}^T \boldsymbol{\alpha}_0(u, v)\}, \quad u, v > 0, \tag{4.5}$$

where $\boldsymbol{\alpha}_0(u, v)$ is a $(p+1) \times 1$ vector of regression coefficients and a function of u and v. The intercept term represents the the association between $N_1(\cdot)$ and $N_2(\cdot)$ in the baseline group (i.e., $Z_1 = ... = Z_p = 0$) and the remaining p coefficients depict the deviations from the baseline association resulted from one unit or category change in the corresponding covariates.

Note that $\rho_{\boldsymbol{Z}}(\cdot, \cdot)$ involves $\tau_{\boldsymbol{Z},1}(\cdot)$ and $\tau_{\boldsymbol{Z},2}(\cdot)$, the marginal time to expected frequency given $\boldsymbol{Z}$. To facilitate the estimation of $\boldsymbol{\alpha}_0(\cdot, \cdot)$, we propose to estimate $\tau_{\boldsymbol{Z},k}(\cdot)$ first, assuming accelerated recurrence time models for both types of recurrent events. More specifically, we adopt the model

$$\tau_{\boldsymbol{Z},k}(u) = \exp\{\boldsymbol{Z}^T \boldsymbol{\beta}_{k0}(u)\}, \quad u > 0, \tag{4.6}$$

for $k = 1, 2$, where $\boldsymbol{\beta}_{k0}(u)$ is a $(p + 1) \times 1$ vector of regression coefficients and a

function of u. The estimation of $\boldsymbol{\beta}_{k0}(u)$ in model (4.6) was studied by Sun et al. (2015) for recurrent events data subject to window observation.

4.2 Estimation Procedure

4.2.1 Estimation of $\boldsymbol{\beta}_{k0}(\cdot)$

We estimate $\boldsymbol{\beta}_{k0}(\cdot)$ by following the method of Sun et al. (2015). Specifically, we employ the estimating equation

$$n^{1/2}\boldsymbol{S}_{kn}(\boldsymbol{\beta}_k, u) = 0, \quad k = 1, 2,$$

where

$$\boldsymbol{S}_{kn}(\boldsymbol{\beta}_k, u) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ N_{ki}(e^{\boldsymbol{Z}_i^T \boldsymbol{\beta}_k(u)}) - \int_0^u Y_{ki}(e^{\boldsymbol{Z}_i^T \boldsymbol{\beta}_k(s)}) ds \right\}.$$

The estimator of $\boldsymbol{\beta}_{k0}(\cdot)$, denoted by $\hat{\boldsymbol{\beta}}_k(\cdot)$, is defined as a right-continuous piecewise constant function jumping only on pre-specified grids that were denoted by $\{0 = u_0 < u_1 < ... < u_{L(n)} = U\}$. Set $\exp\{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_k(0)\} = 0$ for all i because of $\tau_{\boldsymbol{Z},k}(0) = 0$. Then $\hat{\boldsymbol{\beta}}_k(u_l)$, $l = 1, ..., L(n)$, can be obtained by sequentially solving the estimating equation,

$$n^{-1/2}\sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ N_{ki}(e^{\boldsymbol{Z}_i^T \boldsymbol{\beta}_k(u_l)}) - \sum_{p=0}^{l-1} Y_{ki}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_k(u_p)})(u_{p+1} - u_p) \right\} = 0,$$

for $\boldsymbol{\beta}_k(u_l)$. Under certain conditions, Sun et al. (2015) established the uniform consistency and the root-n weak convergence of the resultant estimators.

4.2.2 Proposed estimation procedure for $\boldsymbol{\alpha}_0(\cdot,\cdot)$

To construct an estimating equation for $\boldsymbol{\alpha}_0(u,v)$, we use the fact that

$$E_{\boldsymbol{Z}}[\tilde{N}_1(s)\tilde{N}_2(t)] = E_{\boldsymbol{Z}}\bigg\{\int_0^s\int_0^t Y_1(x)Y_2(y)\frac{\partial^2}{\partial x\partial y}E_{\boldsymbol{Z}}[N_1(x)N_2(y)]dydx\bigg\}. \tag{4.7}$$

Equation (4.7) holds because under the assumption that L and R are independent of $N_k(\cdot)$ given $\boldsymbol{Z}$, $k=1,2$, we have

$$\begin{aligned}
&\frac{\partial^2}{\partial x\partial y}E_{\boldsymbol{Z}}[\tilde{N}_1(x)\tilde{N}_2(y)]\\
&= \lim_{\Delta x\to 0,\Delta y\to 0}\frac{1}{\Delta x\Delta y}E_{\boldsymbol{Z}}[\{\tilde{N}_1(x+\Delta x)-\tilde{N}_1(x)\}\{\tilde{N}_2(y+\Delta y)-\tilde{N}_2(y)\}]\\
&= \lim_{\Delta x\to 0,\Delta y\to 0}\frac{1}{\Delta x\Delta y}E_{\boldsymbol{Z}}[E(\{\tilde{N}_1(x+\Delta x)-\tilde{N}_1(x)\}\{\tilde{N}_2(y+\Delta y)-\tilde{N}_2(y)\}|L,R)]\\
&= \lim_{\Delta x\to 0,\Delta y\to 0}\frac{1}{\Delta x\Delta y}E_{\boldsymbol{Z}}[I(L\le x\le R)I(L\le y\le R)\{N_1(x+\Delta x)-N_1(x)\}\{N_2(y+\Delta y)-N_2(y)\}]\\
&= E_{\boldsymbol{Z}}[Y_1(x)Y_2(y)]\frac{\partial^2}{\partial x\partial y}E_{\boldsymbol{Z}}[N_1(x)N_2(y)].
\end{aligned}$$

To simplify the notation, we denote $\Phi_{\boldsymbol{Z}}(x,y) = E_{\boldsymbol{Z}}[N_1(x)N_2(y)]$. Then under models (4.5) and (4.6), we have $\Phi_{\boldsymbol{Z}}(e^{\boldsymbol{Z}^T\boldsymbol{\beta}_{10}(s)}, e^{\boldsymbol{Z}^T\boldsymbol{\beta}_{20}(t)}) = st\exp\{\boldsymbol{Z}^T\boldsymbol{\alpha}_0(s,t)\}$ for $s,t>0$. Equation (4.7) implies that

$$\begin{aligned}
E_{\boldsymbol{Z}}\bigg\{&\tilde{N}_1(e^{\boldsymbol{Z}^T\boldsymbol{\beta}_{10}(u)})\tilde{N}_2(e^{\boldsymbol{Z}^T\boldsymbol{\beta}_{20}(v)})\\
&-\int_0^u\int_0^v Y_1(e^{\boldsymbol{Z}^T\boldsymbol{\beta}_{10}(s)})Y_2(e^{\boldsymbol{Z}^T\boldsymbol{\beta}_{20}(t)})\frac{\partial^2}{\partial s\partial t}[st\exp\{\boldsymbol{Z}^T\boldsymbol{\alpha}_0(s,t)\}]dtds\bigg\} = 0.
\end{aligned} \tag{4.8}$$

Therefore, we propose the following estimating equation for $\boldsymbol{\alpha}_0(u,v)$:

$$n^{1/2}\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2,\boldsymbol{\alpha},u,v) = 0, \tag{4.9}$$

where

$$\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \boldsymbol{\alpha}, u, v) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i \Bigg\{ \tilde{N}_{i1}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_1(u)}) \tilde{N}_{i2}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_2(v)}) - \int_0^u \int_0^v Y_{i1}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_1(s)}) Y_{i2}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_2(t)}) \frac{\partial^2}{\partial s \partial t} [st \exp\{\boldsymbol{Z}_i^T \boldsymbol{\alpha}(s,t)\}] dt ds \Bigg\}.$$

We can show that equation (4.9) is asymptotically unbiased given equation (4.8) and the uniform consistency of $\hat{\boldsymbol{\beta}}_k(\cdot)$, $k = 1, 2$.

4.2.3 Algorithm to obtain the estimator of $\boldsymbol{\alpha}_0(\cdot, \cdot)$

The stochastic integral representation of $\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \boldsymbol{\alpha}, u, v)$ suggests a grid-based estimation procedure for $\boldsymbol{\alpha}_0(\cdot, \cdot)$. Define grids $S_{L_1(n)} = \{0 = u_0 < u_1 < ... < u_{L_1(n)} = U\}$ and $S_{L_2(n)} = \{0 = v_0 < v_1 < ... < v_{L_2(n)} = V\}$. The proposed estimator $\hat{\boldsymbol{\alpha}}(\cdot, \cdot)$ is defined as a block-wise constant function jumping only at grids $\{(u_p, v_q) : p = 1, ..., L_1(n), q = 1, ..., L_2(n)\}$. We let grids $S_{L_1(n)}$ and $S_{L_2(n)}$ satisfy the conditions of Sun et al. (2015) and follow their algorithm to obtain $\hat{\boldsymbol{\beta}}_1(\cdot)$ and $\hat{\boldsymbol{\beta}}_2(\cdot)$. Given $st \exp\{\boldsymbol{Z}^T \boldsymbol{\alpha}_0(s,t)\} = 0$ for $st = 0$, we set $st \exp\{\boldsymbol{Z}_i^T \hat{\boldsymbol{\alpha}}(s,t)\} = 0$ for all i.

Based on equation (4.9), we propose to obtain $\hat{\boldsymbol{\alpha}}(u_l, v_m)$, $l = 1, ..., L_1(n), m = 1, ..., L_2(n)$, by sequentially solving the following estimating equation for $\boldsymbol{\alpha}(u_l, v_m)$:

$$n^{-1/2} \sum_{i=1}^{n} \boldsymbol{Z}_i \Bigg\{ \tilde{N}_{i1}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_1(u_l)}) \tilde{N}_{i2}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_2(v_m)}) - Y_{i1}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_1(u_{l-1})}) Y_{i2}(e^{\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_2(v_{m-1})}) \Big[u_l v_m \exp\{\boldsymbol{Z}_i^T \boldsymbol{\alpha}(u_l, v_m)\}$$
$$- u_l v_{m-1} \exp\{\boldsymbol{Z}_i^T \hat{\boldsymbol{\alpha}}(u_l, v_{m-1})\} - u_{l-1} v_m \exp\{\boldsymbol{Z}_i^T \hat{\boldsymbol{\alpha}}(u_{l-1}, v_m)\} + u_{l-1} v_{m-1} \exp\{\boldsymbol{Z}_i^T \hat{\boldsymbol{\alpha}}(u_{l-1}, v_{m-1})\} \Big]$$
$$- S_i(l-2, m-2, \hat{\boldsymbol{\alpha}}) \Bigg\} = 0, \tag{4.10}$$

where

$$S_i(l,m,\boldsymbol{\alpha}) = \sum_{q=0}^{m}\sum_{p=0}^{l} Y_{i1}(e^{\boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}_1(u_p)})Y_{i2}(e^{\boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}_2(v_q)})\Big[u_{p+1}v_{q+1}\exp\{\boldsymbol{Z}_i^T\boldsymbol{\alpha}(u_{p+1},v_{q+1})\}$$
$$- u_p v_{q+1}\exp\{\boldsymbol{Z}_i^T\boldsymbol{\alpha}(u_p,v_{q+1})\} - u_{p+1}v_q\exp\{\boldsymbol{Z}_i^T\boldsymbol{\alpha}(u_{p+1},v_q)\} + u_p v_q\exp\{\boldsymbol{Z}_i^T\boldsymbol{\alpha}(u_p,v_q)\}\Big]$$

More specifically, we can follow the following algorithm to obtain $\hat{\boldsymbol{\alpha}}(u_l, v_m)$:

1. Let $\bar{p} = p = 1, \bar{q} = q = 1$. Obtain $\hat{\boldsymbol{\alpha}}(u_1, v_1)$ by solving equation (4.10).

2. Let $p = p + 1$. Obtain $\hat{\boldsymbol{\alpha}}(u_p, v_{\bar{q}})$ by solving equation (4.10). Repeat this step until $p = l$.

3. Let $q = q + 1$. Obtain $\hat{\boldsymbol{\alpha}}(u_{\bar{p}}, v_q)$ by solving equation (4.10). Repeat this step until $q = m$.

4. Let $\bar{p} = \bar{p} + 1$ and $\bar{q} = \bar{q} + 1$.

5. Go back to step 2 unless $\bar{p} = l$ or $\bar{q} = m$.

6. If $\bar{p} = l$, $\bar{q} = m$, then output $\hat{\boldsymbol{\alpha}}(u_{\bar{p}}, v_{\bar{q}})$. If $\bar{p} < l$, $\bar{q} = m$, then repeat step 2 until $p = l$ and output $\hat{\boldsymbol{\alpha}}(u_p, v_{\bar{q}})$. If $\bar{p} = l$, $\bar{q} < m$, then repeat step 3 until $q = m$ and output $\hat{\boldsymbol{\alpha}}(u_{\bar{p}}, v_q)$.

Note that, for each fixed (u_l, v_m), equation (4.10) is continuous and monotone in $\boldsymbol{\alpha}(u_l, v_m)$ and thus is not prone to the multiple solution issue. This fact facilitates the computation. In numerical studies, we adopt the *nleqslv*() function in R package *nleqslv*, which implements the algorithm of Dennis and Schnabel (1996).

4.2.4 Inference

To make inference on $\boldsymbol{\alpha}_0(u, v)$, bootstrapping procedures can be used. Denote $\hat{\boldsymbol{\alpha}}^*(u, v)$ as the bootstrap estimator. It can be shown that the distribution of

$n^{1/2}\{\hat{\boldsymbol{\alpha}}^*(u,v) - \hat{\boldsymbol{\alpha}}(u,v)\}$ conditionally on the observed data and the unconditional distribution of $n^{1/2}\{\hat{\boldsymbol{\alpha}}(u,v) - \boldsymbol{\alpha}_0(u,v)\}$ have the same limiting distribution. By repeatedly resampling from the observed data $\{N_{i1}(\cdot), N_{i2}(\cdot), \boldsymbol{Z}_i, L_i, R_i\}_{i=1}^n$, one may obtain a large number of realizations of $n^{1/2}\{\hat{\boldsymbol{\alpha}}^*(u,v) - \hat{\boldsymbol{\alpha}}(u,v)\}$, the empirical distribution of which can be used to give the covariance estimate for $\hat{\boldsymbol{\alpha}}(u,v)$ or the confidence interval for $\boldsymbol{\alpha}_0(u,v)$.

4.3 Simulation Studies

Simulation studies are conducted to examine the finite sample properties of the proposed methods. We consider a scenario, where the type-k recurrent events are generated based on a process which has the intensity function,

$$\lambda_k(t|Z,\gamma) = [2t + c_{k1}\gamma]I(Z=1) + c_{k2}\gamma I(Z=0), \quad k=1,2,$$

where $Z \sim Bernoulli(0.5)$, γ is a Gamma frailty with mean 1 and variance σ^2, c_{k1} and c_{k2} are some constants for $k=1,2$. It can be shown that

$$\tau_{Z,k}(u) = \exp\{\log(u/c_{k2}) + [\log(\frac{-c_{k1}+\sqrt{c_{k1}^2+4u}}{2}) - \log(u/c_{k2})]Z\}, \quad k=1,2,$$
$$\rho_Z(u,v) = \exp\{\log(\sigma^2+1) + [\log(\frac{4c_{11}c_{21}\sigma^2}{(c_{11}+\sqrt{c_{11}^2+4u})(c_{21}+\sqrt{c_{21}^2+4v})}+1) - \log(\sigma^2+1)]Z\}.$$

Covariate Z has varying-effects on $\tau_{Z,k}(u)$, $k=1,2$, and $\rho_Z(u,v)$.

We set $c_{11}=1.5, c_{12}=1.8, c_{21}=1, c_{22}=1.5$ and choose σ^2 to be 0 or 1. Note that $\sigma^2=0$ indicate intra-subject event times are independent. We generate L from $\omega \cdot Unif(0,1.5)$ and R from $Unif(L+1,3)$, where ω is a $Bernoulli(0.8)$ variate. Then the average numbers of observed recurrent events per subject for type-1 and type-2 events are about 4.5 and 5.3, respectively. Under each selection of σ^2, we generate 1000 simulated data sets of sample size $n=400$. For bootstrapping-based

inference, the resampling size of 200 is chosen. We adopt an equally-spaced grid on both $u \in (0, 3]$ and $v \in (0, 3]$ with grid size, 0.02.

Figure 4.1 presents the empirical bias (EmpBias) of the proposed estimator, as well as its empirical standard error (EmpSE), average estimated standard error based on bootstrapping (EstSE) and the ratio, EmpBias/EmpSE, for the setup with $\sigma^2 = 1$. The first row is for the intercept and the second row is for the covariate coefficient. The plot of EmpBias in the first row shows that the intercept estimate has small bias except for those corresponding to small u or v. The plot of EmpBias/EmpSE also indicates the magnitude of bias is mostly within 10% of the corresponding standard error. We have similar observations for the estimator of the covariate coefficient. The EmpBias/EmpSE ratios are smaller; most have a magnitude less than 6%. Bootstrapping-based standard error estimates for both intercept and covariate coefficient agree well with corresponding empirical standard errors except at small u or v.

Figure 4.2 presents the simulation results of $\hat{\boldsymbol{\alpha}}(u, v)$ for the setup with $\sigma^2 = 0$ and shows similar observations. Except at small u or v, the estimates have small bias. The absolute values of EmpBias/EmpSE for estimated intercept are mostly within 10% and those for estimated covariate coefficient are within 6%. Bootstrapping-based standard error estimates agree with the empirical standard errors well except at small u or v.

4.4 An Application to CFFPR Data

Cystic Fibrosis (CF) is a lethal autosomal disease without known cure yet that commonly affects Caucasians due to mutation of CFTR gene. Pseudomonas aeruginosa

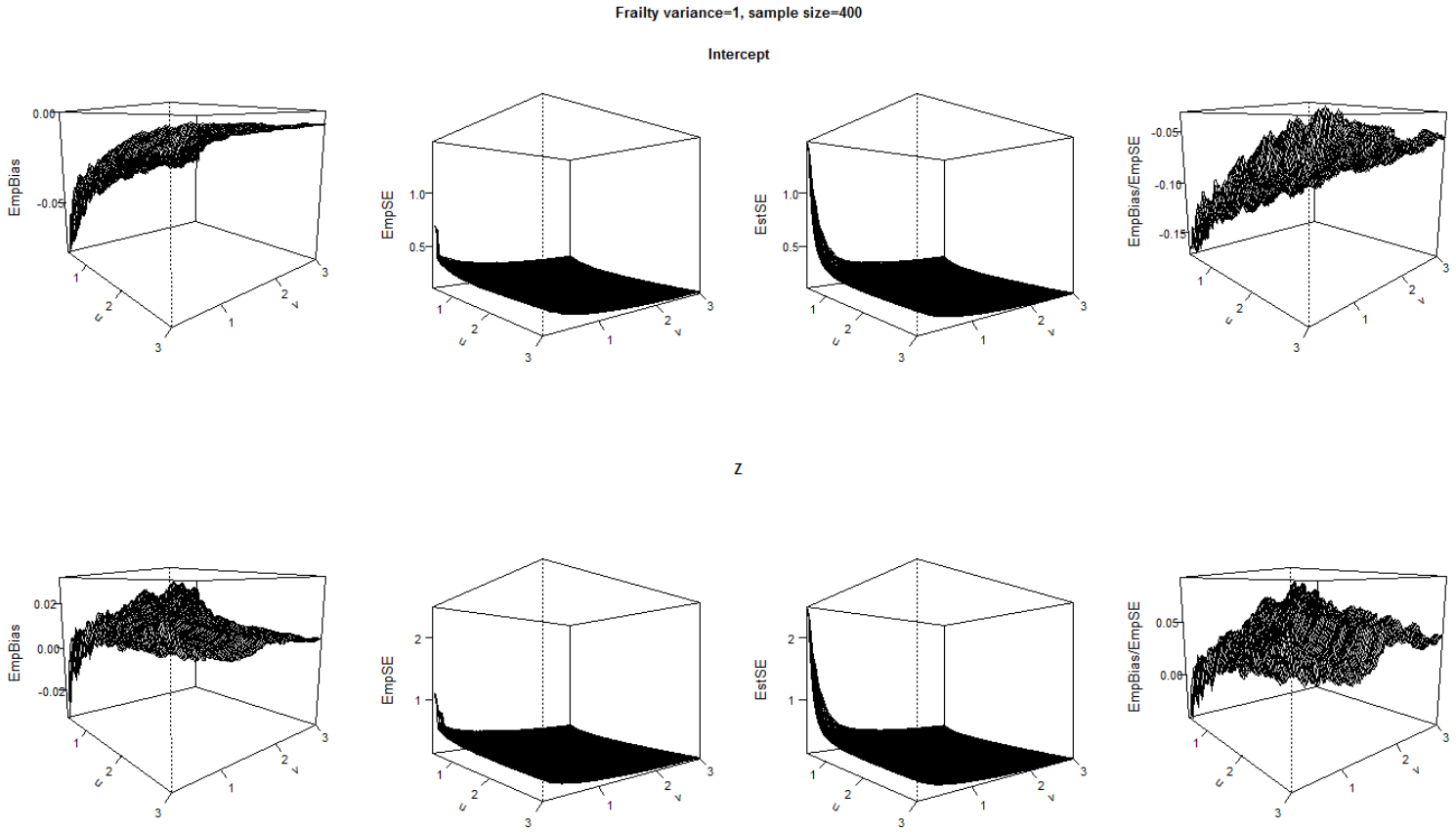

Figure 4.1: Simulation Results for sample size n=400 based on the setup with Gamma frailty of variance 1, $u \in (0.3, 3], v \in (0.3, 3]$.

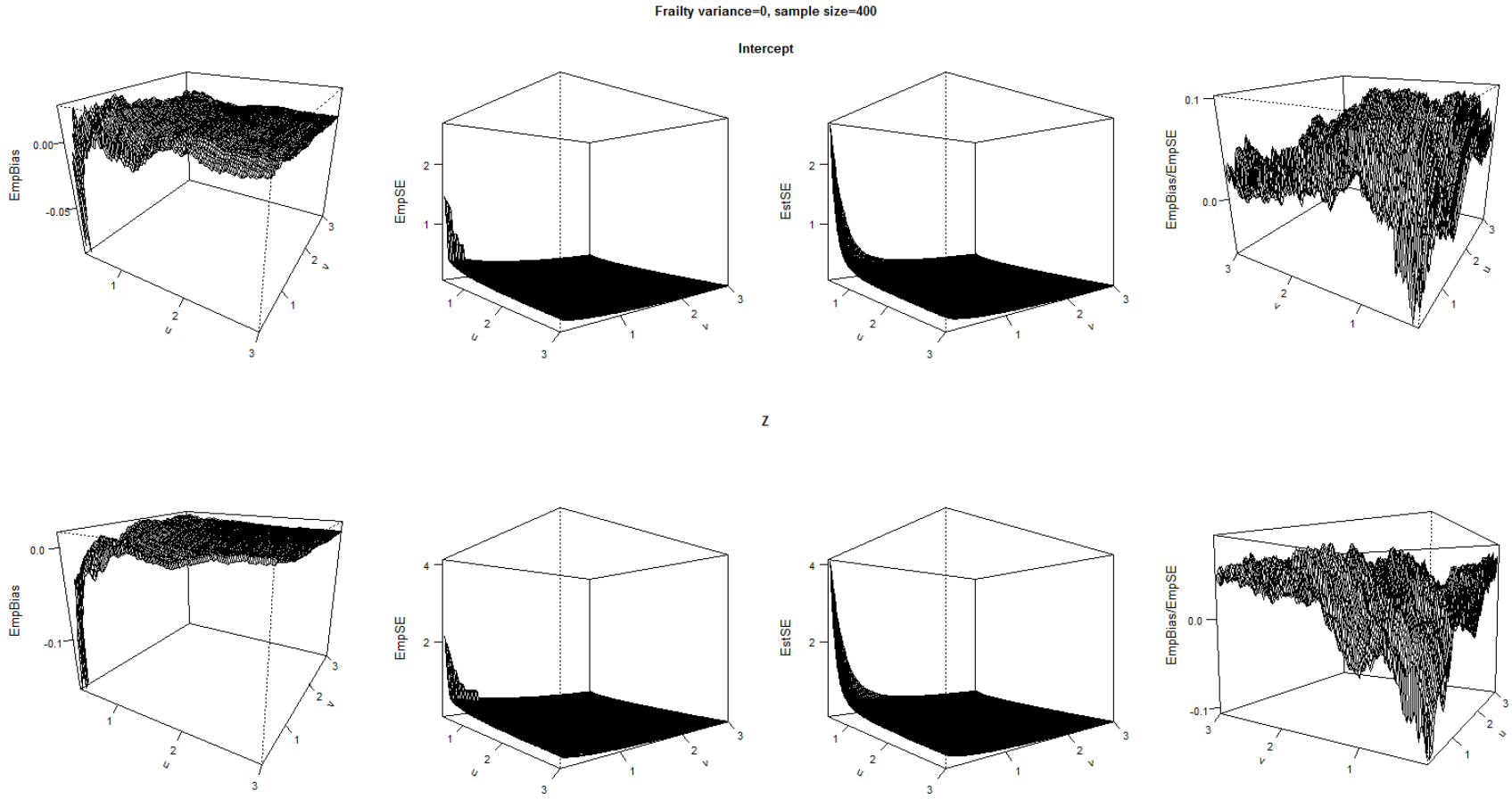

Figure 4.2: Simulation Results for sample size n=400 based on the setup with Gamma frailty of variance 0, $u \in (0.3, 3], v \in (0.3, 3]$.

(PA) and Staphylococcus aureus (SA) are two major pathogens of medical concerns for CF patients, but the interplay occurring between the two remains largely unknown. To address this question and also see how the interplay would be influenced by potential risk factors, we apply the proposed method to the data from 2799 children documented in 1986-2008 CFF Patient Registry (CFFPR), who were born in or after 1998 with at least one F508del mutation, insufficient pancreatic status (defined as ever on pancreatic enzymes) and at least 5 years of follow-up in the registry. The vector $\boldsymbol{Z}$ consists of three covariates, representing gender, patient's CFTR genotype (F508del homozygous/heterozygous) and meconium ileus (MI) status. We let $Z_1 = 1$ for girls and 0 otherwise; $Z_2 = 1$ for F508del heterozygous and 0 otherwise; $Z_3 = 1$ for MI and 0 otherwise.

In our analysis, time from birth to registry entry constitutes L and we treat L less than one month as zero. Out of 2799 children, there are 309 (11%) children with $L = 0$, 1403(50%) girls, 1047 (37%) with heterozygous F508del mutations and 779 (28%) with MI. Age at the first CFFPR visit ranges from 0 to 5.4 years with mean=0.7 years and median=0.4 years. Mean numbers of PA infections and SA infections per subject are 3.9 and 9.5, respectively. Corresponding median numbers are 2 and 8, respectively.

We first study the effects of covariates on the timing of PA infections by fitting marginal model (4.6) over $u \in (0, 2]$. The resulting coefficient estimates and 95% pointwise confidence intervals are displayed In Figure 4.3. It shows that gender and CFTR genotype have little impact on the timing of PA infection recurrences. The coefficient for MI is significantly smaller than zero when u is small but is close to zero when $u > 0.3$. This may suggest some disadvantage for CF patients with MI in the early occurrence of PA infections.

We make similar analysis on the timing of SA infections over $v \in (0, 7]$. Figure 4.4 depicts the estimated effects of covariates with the 95% pointwise confidence intervals.

Interestingly, we observe that MI demonstrates a strong positive effect on the timing of SA recurrence. This suggests that MI may have some protecting effect for CF children in terms of SA infection recurrence.

Next, we study the effects of covariates on the frequency-specific association between early recurrences of SA infection and recurrences of PA infection, $\rho_{\boldsymbol{Z}}(u, v)$, by fitting model (4.5). We pick three expected frequency v values of 0.5, 1, 1.5, for SA infection. In Figure 4.5, we plot the estimated coefficients $\hat{\boldsymbol{\alpha}}(u, v)$ with corresponding 95% pointwise Wald-type bootstrapping confidence intervals at these selected v, respectively. In this figure, negative estimates are found for the intercept and generally significant over $u \in (0, 1]$ at $v = 1$ and $u \in (0, 0.76]$ at $v = 1.5$. This suggests that for the reference group, which consists of CF boys with homozygous F508del mutations and no MI, early recurrences of SA infection would postpone early recurrences of PA infection. Gender and CFTR genotype seem to have no effects on on $\rho_{\boldsymbol{Z}}(u, v)$, but MI has positive significant effect over $u \in (0, 1.34]$ at $v = 1$ and $u \in (0, 2]$ at $v = 1.5$, respectively.

To garner a clearer picture about the association between recurrences of SA infection and recurrences of PA infection, we further depict the estimated $\hat{\rho}_{\boldsymbol{Z}}(u, v)$ in 8 subgroups that are defined by the eight possible combinations of covariate values. Figure 4.6 and Figure 4.7 plot the estimated $\hat{\rho}_{\boldsymbol{Z}}(u, v)$ with corresponding 95% pointwise bootstrapping confidence intervals at fixed expected frequency v values of 0.5, 1, 1.5, respectively. In Figure 4.6, the first column is for the reference group, CF boys with homozygous F508del mutations and no MI. Negative estimated $\hat{\rho}_{\boldsymbol{Z}}(u, v)$ generally over $u \in (0, 1]$ at $v = 1$ and $v = 1.5$ indicate early recurrences of SA infection being negatively associated with early recurrences of PA infection, and the same as

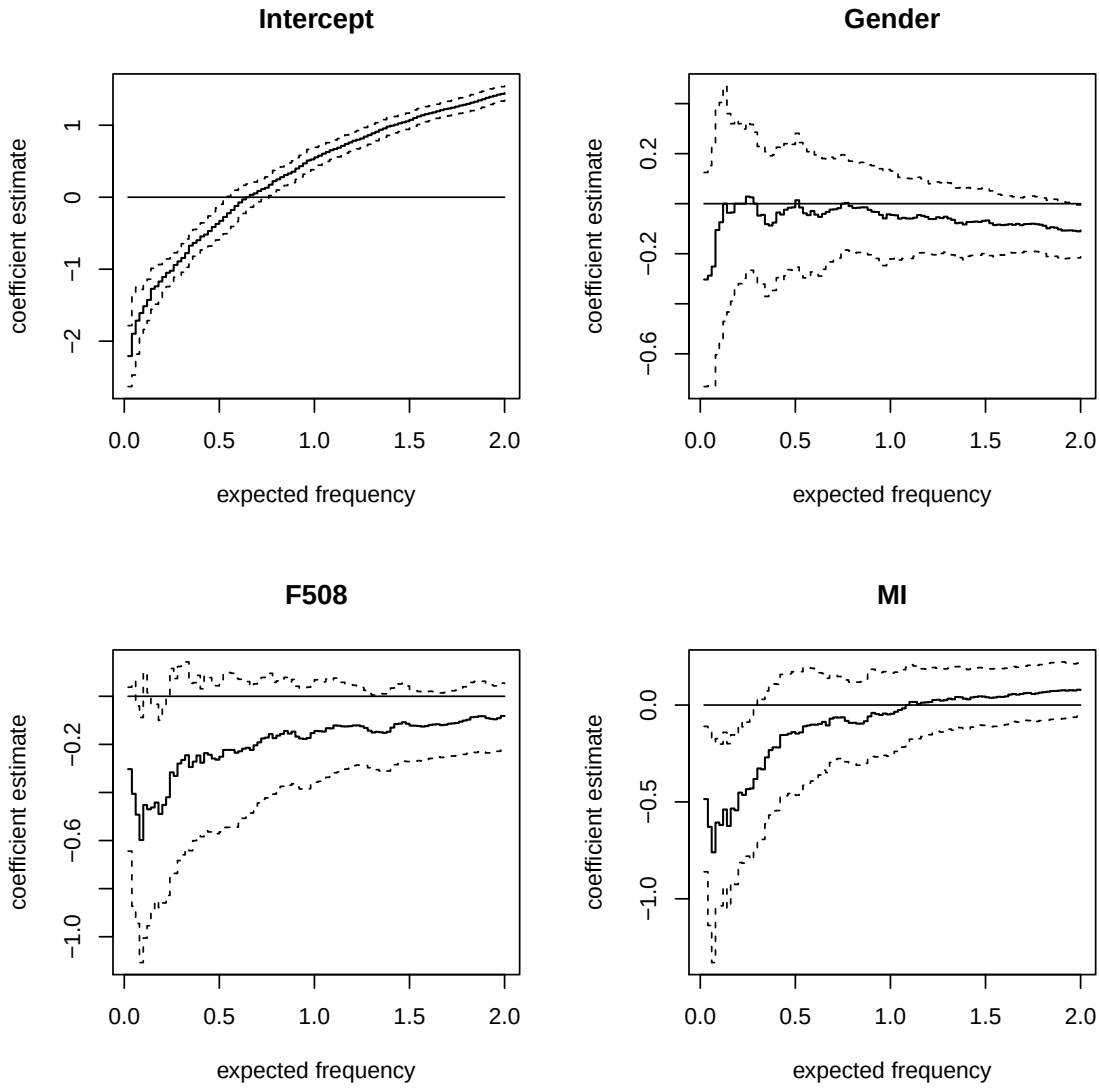

Figure 4.3: Analysis of CFFPR data: effects of covariates on the timing of PA infections; coefficient estimates (solid lines) and 95% pointwise confidence intervals (dashed lines).

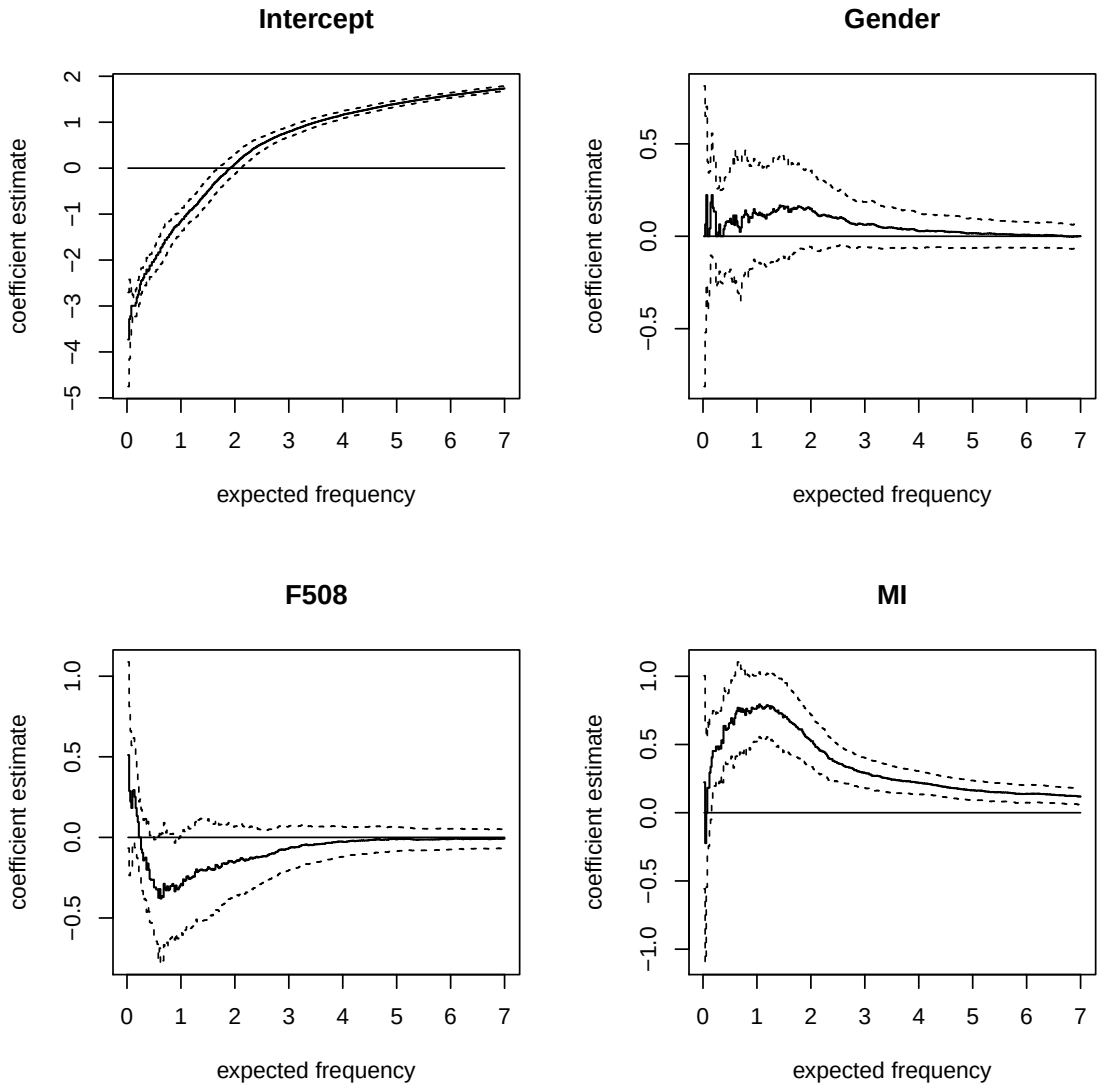

Figure 4.4: Analysis of CFFPR data: effects of covariates on the timing of SA infections; coefficient estimates (solid lines) and 95% pointwise confidence intervals (dashed lines).

suggested by Figure 4.5. The third column is for CF boys with heterozygous F508del mutations and no MI. Its significantly negative estimates generally over $u \in (0.6, 1.1)$ at $v = 1$ and $v = 1.5$ also imply that early recurrences of SA infection would postpone early recurrences of PA infection. In Figure 4.7, similar negative association patterns are found for CF girls without MI, regardless of their CFTR genotypes.

We follow the same procedures to study the effects of covariates on the association between early recurrences of PA infection and recurrences of SA infection, by choosing three expected frequency values of 0.5, 1, 1.5, for PA infection. Figure 4.8 plot the estimated coefficients $\hat{\boldsymbol{\alpha}}(u, v)$ with corresponding 95% pointwise Wald-type bootstrapping confidence intervals, showing that MI has significantly positive effect on $\rho_{\boldsymbol{Z}}(u, v)$ generally over $v \in (0, 4.5]$. Negative association are still found in the reference group, generally over $v \in (0, 2]$ at $u = 0.5$ and at $u = 1$, suggesting early recurrences of PA infection and early recurrences of PA infection are negatively associated. From Figure 4.9 and Figure 4.10, we can see that early recurrences of PA infection is negatively associated with early recurrences of SA infection in subgroups of CF girls without MI. For CF boys with heterogynous F508del mutations and MI, estimated $\hat{\rho}_{\boldsymbol{Z}}(u, v)$ is positively significant for $v > 2$. This may indicate early recurrences of PA infections do not influence early recurrence of SA infection, but would boost the latter's later-on recurrences.

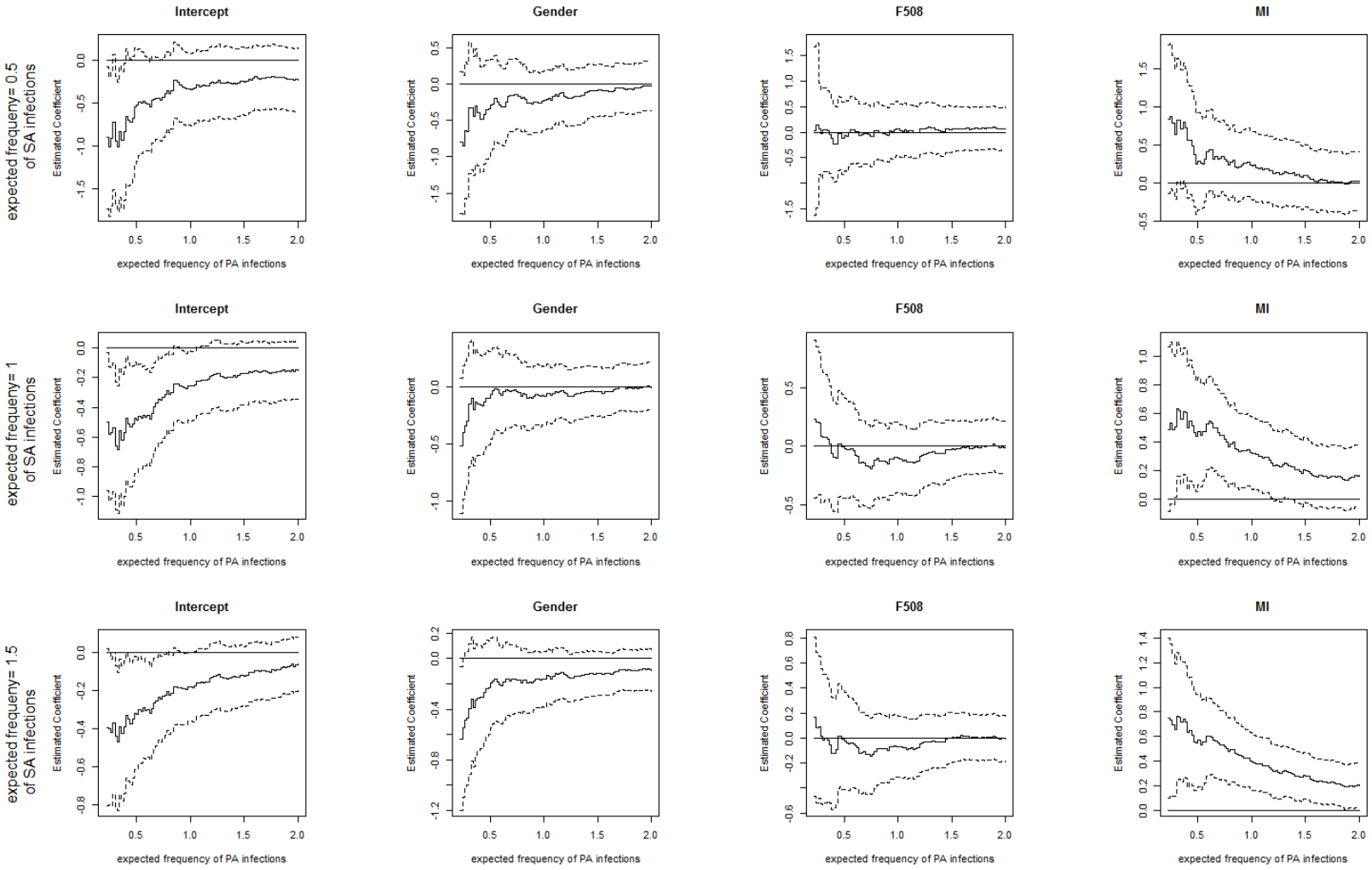

Figure 4.5: Analysis of CFFPR data: coefficient estimates (solid lines) and 95% pointwise confidence intervals (dashed lines) of $\hat{\boldsymbol{\alpha}}(u, v)$ at fixed expected frequency $v = 0.5, 1, 1.5$, of SA infections.

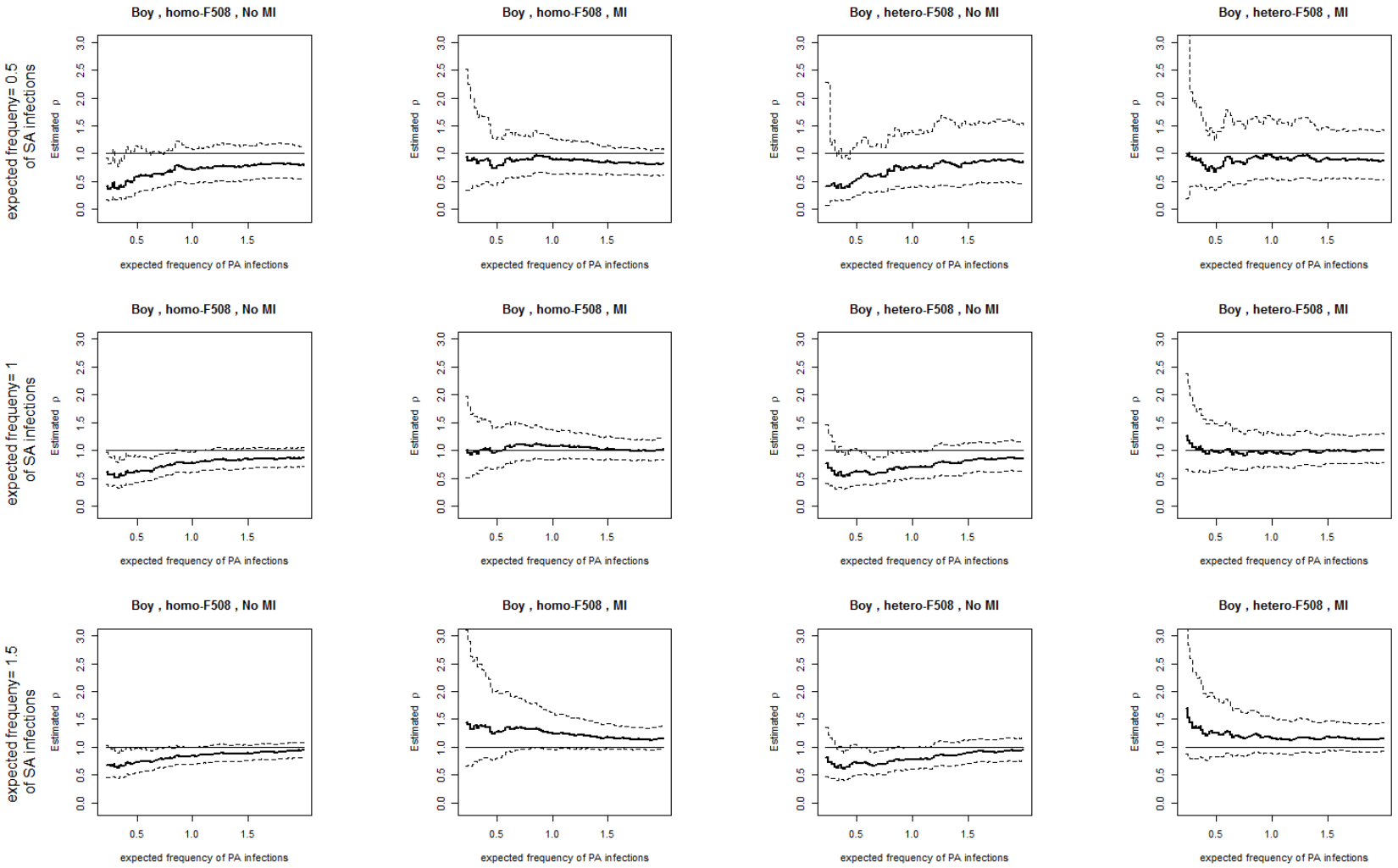

Figure 4.6: Analysis of CFFPR data: estimates of $\hat{\rho}_{\boldsymbol{Z}}(u, v)$ (solid lines) and 95% pointwise confidence intervals (dashed lines) at fixed expected frequency $v = 0.5, 1, 1.5$, for SA infection.

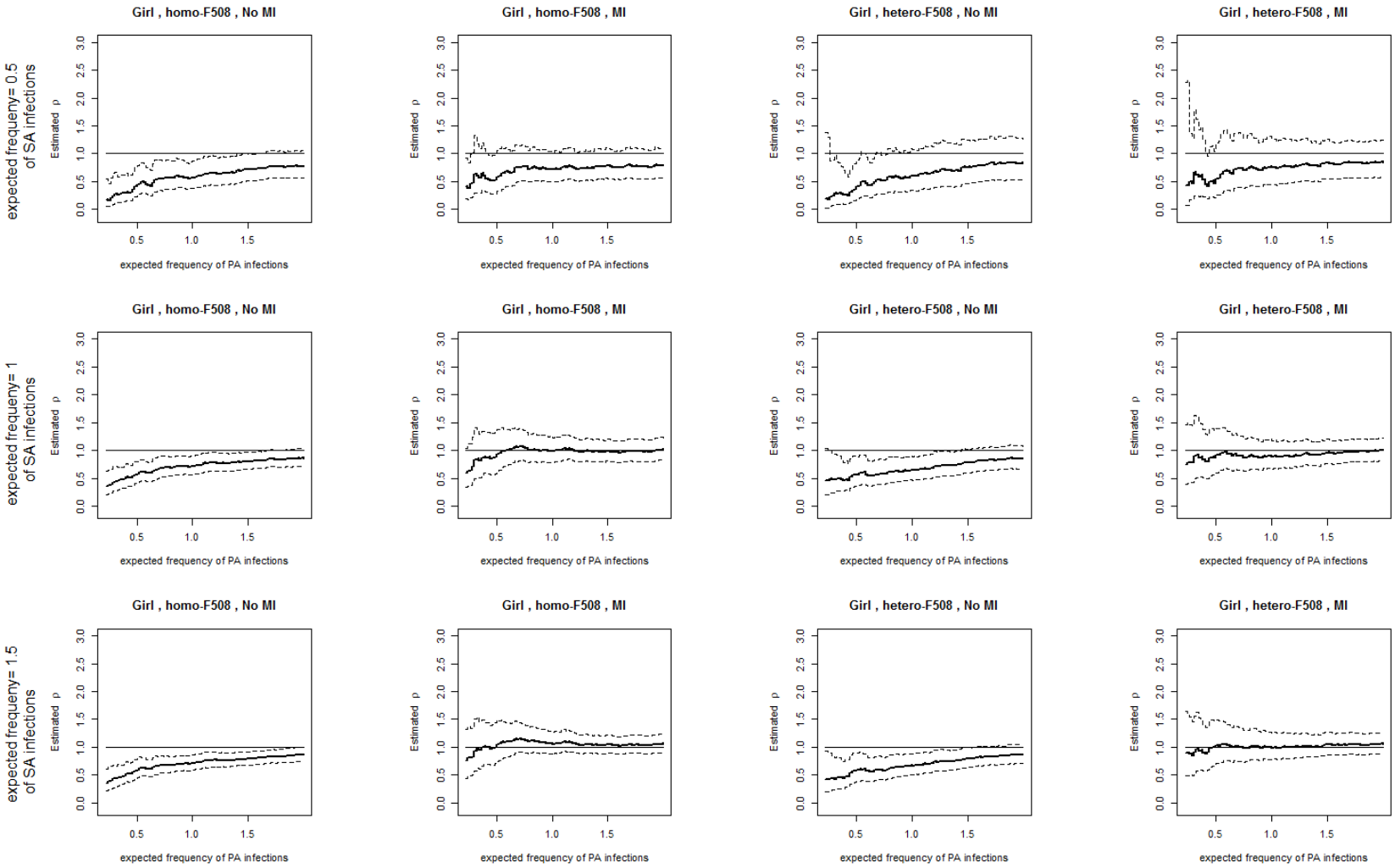

Figure 4.7: Analysis of CFFPR data: estimates of $\hat{\rho}_{\boldsymbol{Z}}(u, v)$ (solid lines) and 95% pointwise confidence intervals (dashed lines) at fixed expected frequency $u = 0.5, 1, 1.5$, for SA infection.

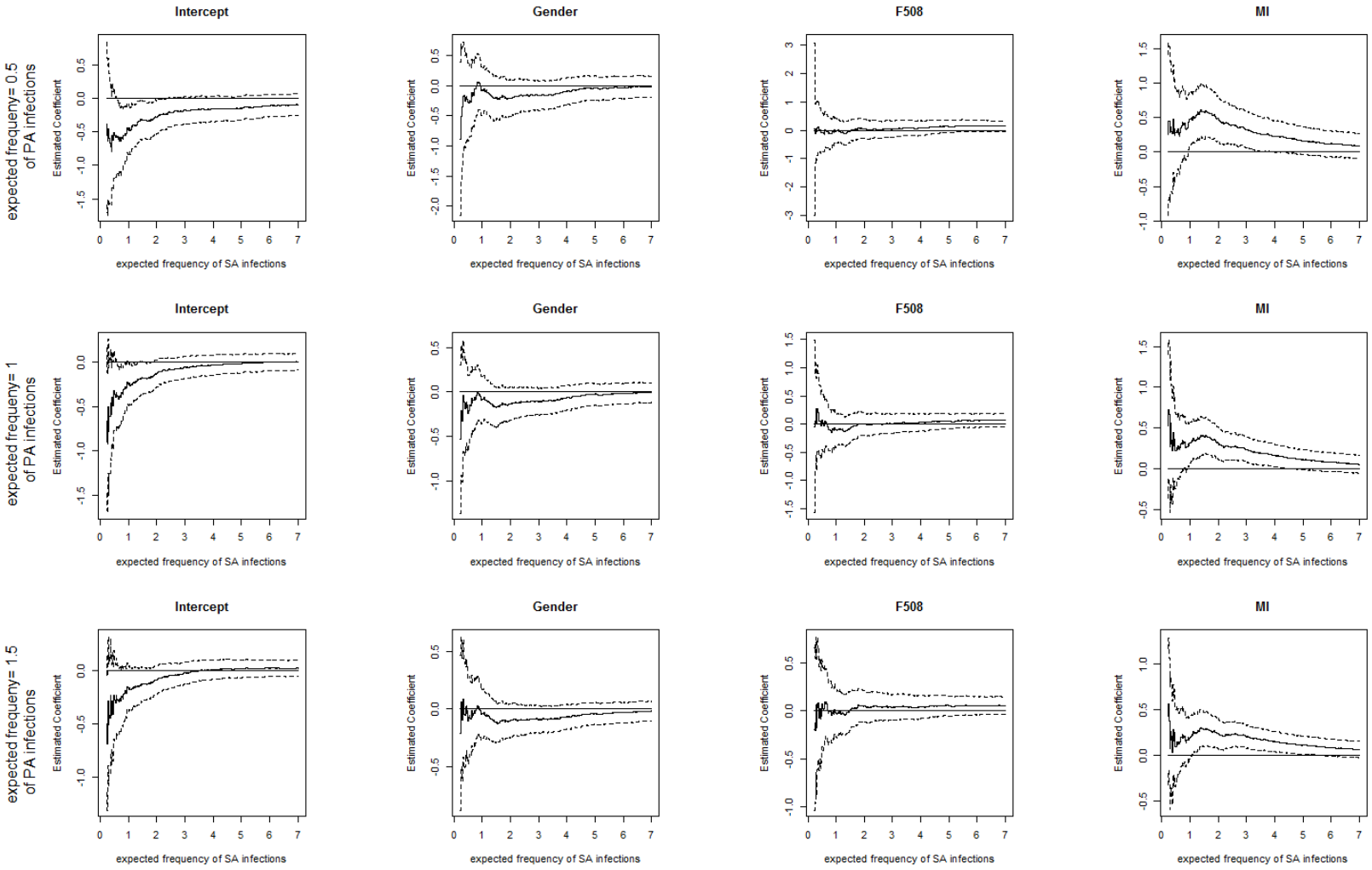

Figure 4.8: Analysis of CFFPR data: coefficient estimates (solid lines) and 95% pointwise confidence intervals (dashed lines) of $\hat{\boldsymbol{\alpha}}(u, v)$ at fixed expected frequency $u = 0.5, 1, 1.5$, for PA infection.

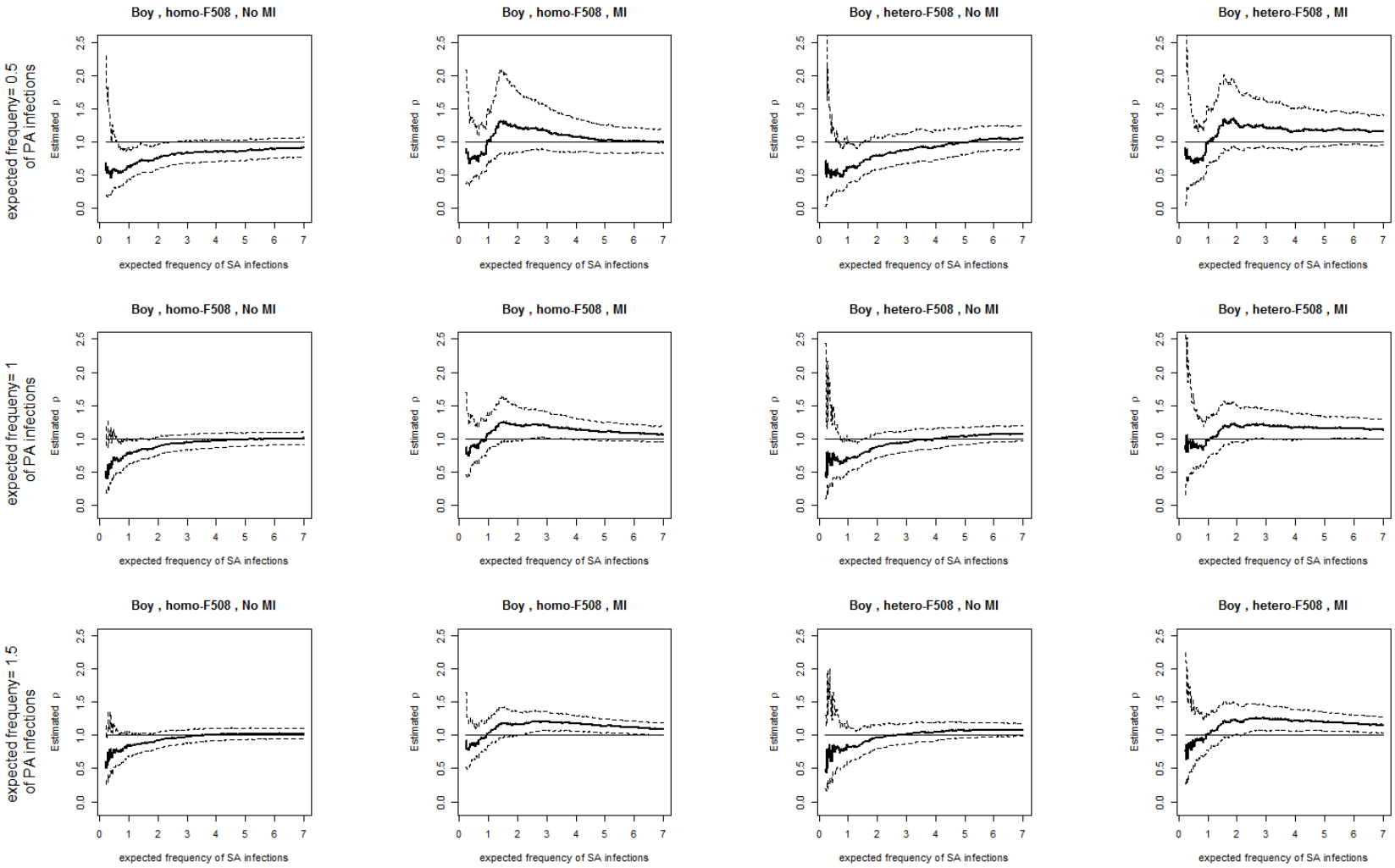

Figure 4.9: Analysis of CFFPR data: estimates of $\hat{\rho}_{\boldsymbol{Z}}(u, v)$ (solid lines) and 95% pointwise confidence intervals (dashed lines) at fixed expected frequency $u = 0.5, 1, 1.5$, for PA infection.

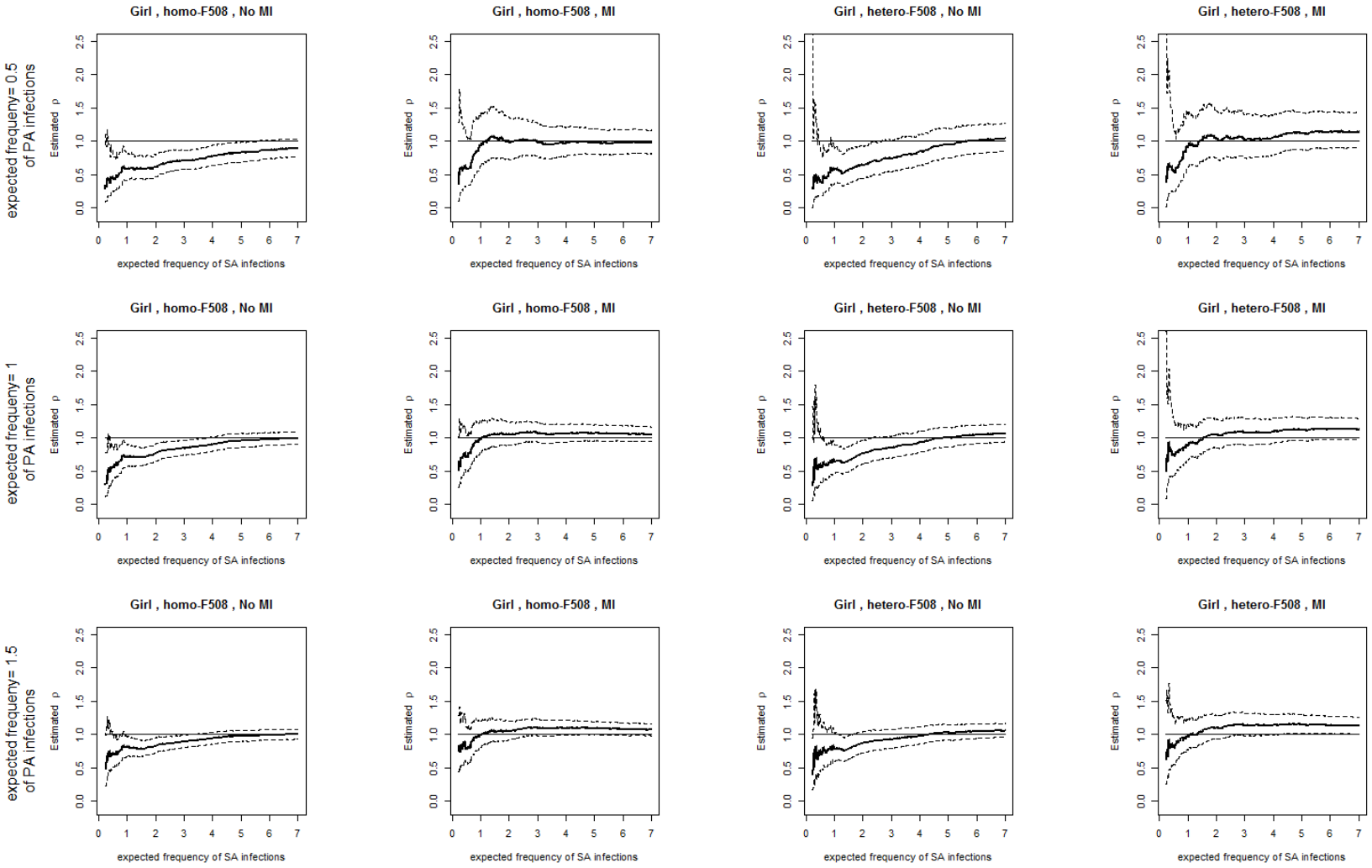

Figure 4.10: Analysis of CFFPR data: estimates of $\hat{\rho}_{\boldsymbol{Z}}(u, v)$ (solid lines) and 95% pointwise confidence intervals (dashed lines) at fixed expected frequency $u = 0.5, 1, 1.5$, of PA infections.

Chapter 5

Summary and Future Work

5.1 Summary

In this dissertation, we focus on two data scenarios that are often encountered in biomedical follow-up studies, semi-competing risks data and bivariate recurrent events data. We develop methods to study dynamic association patterns embedded in these data scenarios..

In the semi-competing risks scenario, we propose a robust measure that can flexibly capture the dynamic pattern of the dependence structure between the nonterminal event and the terminal event. We develop a simple nonparametric estimator that can account for left truncation, but require that the gap time between truncation and censoring is independent of the truncation time itself. The established asymptotic results as well as estimating and inference procedures can be extended to adjust for covariates. Simulation studies show satisfactory performance of our method with moderate sample size. An application to the Denmark diabetes registry data demonstrates practical utility of our proposal.

We further develop an estimator for the proposed semi-competing risks dependence measure which can accommodate a more general left truncation scenario. Simulation studies demonstrate that the new proposal performances well with moderate sample size, while the former approach can lead to considerably biased estimation due to the violation of the strong left truncation assumption. The new method is also illustrated by an application to the Denmark diabetes registry data.

For bivariate recurrent events data setting, we propose to explore the association between bivariate recurrent events processes under an observation window structure. We develop a regression framework for the proposed measure to allow for assessing how the association is in influenced by covariates. The estimating and inference procedure are proposed along with an efficient iterative algorithm. Simulation studies suggest the validity of our proposal. We analyze the CFFPR data by using this new method.

5.2 Future Work

We plan to complete the ongoing work on the bivariate recurrent event data in the near future. First, we will establish the asymptotic properties, including uniform consistency and weak convergence, of the proposed estimator. We will further conduct additional simulations, for example, for the case with a continuous covariate.

In what follows we describe some possible topics for future work. One direction is to explore the covariates effects on the dependence measure $LCQRR(\tau; t_0)$ of the nonterminal and terminal event in a more general scenario, where (L, C) is allowed to depend on covariates. It would also be very desirable to develop methods that can accommodate time-dependent covariates.

For the bivariate recurrent event data, we employ grid-based estimation procedures, for which sufficiently small grid size is warranted for nice asymptotic results. It may be interesting to develop a grid-free approach. This may also merit future research.

Bibliography

Abu-Libdeh, H., Turnbull, B. W., and Clark, L. C. (1990). Analysis of multi-type recurrent event in longitudinal studies: application to a skin cancer prevention trial. *Biometrics* **46,** 1017–1034.

Alexander, K. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *The Annals of Probability* **12,** 1041–1067.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

Asgharian, M., M'Lan, C., and Wolfson, D. (2002). Length-biased sampling with right censoring. *The Annals of Probability* **97,** 201–209.

Cai, J. and Schaubel, D. E. (2004). Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Analysis* **10,** 121–138.

Chen, X., Wang, Q., Cai, J., and Shankar, V. (2012). Semiparametric additive marginal regression models for multiple type recurrent events. *Lifetime Data Analysis* **18,** 504–527.

Chen, Y. H. (2012). Maximum likelihood analysis of semicompeting risks data with semiparametric regression models. *Lifetime Data Analysis* **18,** 36–57.

Clayton, D. G. (1978). A model for association in bivariate life tables and its appli-

cation in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65,** 141–151.

Cook, R. J., Lawless, J. F., and Lee, K. A. (2010). A copula-based mixed poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine* **29,** 694–707.

Dennis, J. J. and Schnabel, R. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Siam.

Doss, H. (1989). On estimating the dependence between two point processes. *The Annals of Statistics* **17,** 749–763.

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

Fygenson, M. and Ritov, Y. (1994). Monotone estimating equations for censored data. *Biometrika* **22,** 732–746.

Ghosh, D. (2006). Semiparametric inferences for association with semi-competing risks data. *Statistics in Medicine* **25,** 2059–2070.

Gijbels, I., Veraverbeke, N., and Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics and Data Analysis* **55,** 1919–1932.

Gijbels, I. and Wang, J. (1993). Strong representations of the survival function estimator for truncated and censored data with applications. *Journal of Multivariate Analysis* **47,** 210–229.

Gorfine, M., Zucker, D. M., and Hsu, L. (2006). Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Biometrika* **93,** 735–741.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.

Hsieh, J.-J., Wang, W., and Ding, A. A. (2008). Regression analysis based on semi-competing risks data. *Journal of the Royal Statistical Society Series B* **70,** 3–20.

Hsu, L., Gorfine, M., and Malone, K. (2007). On robustness of marginal regression coefficient estimates and hazard functions in multivariate survival analysis of family data when the frailty distribution is mis- specified. *Statistics in Medicine* **26,** 4657–4678.

Hsu, L. and Prentice, R. (1996). On assessing the strength of dependency between failure time variates. *Biometrika* **83,** 491–506.

Huang, Y. and Peng, L. (2009). Accelerated recurrence time models. *Scandinavian Journal of Statistics* **36,** 636–648.

Jin, Z., Ying, Z., and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88,** 381–390.

Lai, T. and Ying, Z. (1988). Stochastic integrals of empirical-type processes with applications to censored regression. *Journal of Multivariate Analysis* **27,** 334–358.

Lakhal, L., Rivest, L.-P., and Abdous, B. (2008). Estimating survival and association in a semicompeting risks model. *Biometrics* **64,** 180–188.

Li, R., Cheng, Y., and Fine, J. P. (2014). Quantile assciation regression models. *Journal of the American Statistical Association* **109,** 230–242.

Li, R. and Peng, L. (2011). Quantile regression for left-truncated semi-competing risks data. *Biometrics* **67,** 701–710.

Lin, D. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* **80,** 573–581.

Ma, Y. and Yin, G. (2010). Semiparametric median residual life model and inference. *The Canadian Journal of Statistics* **34,** 665–679.

Ning, J., Chen, Y., Cai, C., Huang, X., and Wang, M. (2015). On the dependence structure of bivariate recurrent event processes: Inference and estimation. *Biometrika* in press.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society Series B* **44,** 414–422.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84,** 487–493.

Parzen, M., Wei, L., and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81,** 341–350.

Peng, L. and Fine, J. P. (2007). Regression modeling of semi-competing risks data. *Biometrics* **63,** 96–108.

Peng, L. and Fine, J. P. (2009). Competing risks quantile regression. *Journal of the American Statistical Association* **104,** 1140–1453.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* **86,** 770–778.

Prentice, R. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79,** 495–512.

Ripley, B. D. (1976). The second-order analysis of stationary point process. *Journal of Applied Probability* **13,** 255–266.

Schaubel, D. E. and Cai, J. (2005). Semiparametric methods for clustered recurrent event data. *Lifetime Data Analysis* **11,** 405–425.

Shen, Y. and Thall, P. F. (1998). Parametric likelihoods for multiple non-fatal competing risks and death. *Statistics in Medicine* **17,** 999–1015.

Sun, L., Zhu, L., and Sun, J. (2009). Regression analysis of multivariate recurrent event data with time-varying covariate effects. *Journal of Multivariate Analysis* **100,** 2214–2223.

Sun, X., Peng, L., Huang, Y., and Lai, H. (2015). A generalized framework for censored quantile regression based on counting process. *Journal of the American Statistical Association* in press.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer, New York.

Ventura, V., Cai, C., and Kass, R. E. (2005). Statistical assessment of time-varying dependency between two neurons. *J Neurophysiol* **94,** 2940–7.

Veraverbeke, N., Omelka, M., and Gijbels, I. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics* **38,** 766–780.

Wang, M. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association* **86,** 130–143.

Wang, M. (2003). Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society Series B* **65,** 257–273.

Yan, J. and Fine, J. P. (2005). Functional association models for multivariate survival processes. *Journal of the American Statistical Association* **100,** 184–196.