STATISTICAL METHODS FOR RECURRENT MARKER PROCESS IN THE PRESENCE OF TERMINAL EVENTS

by

Yifei Sun

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

July, 2015

Abstract

Benefit-risk assessment is a crucial step in the medical decision process. In many biomedical studies, both longitudinal marker measurements and time to a terminal event serve as important endpoints for benefit-risk assessment. The effect of an intervention or a treatment on the longitudinal marker process, however, can be in conflict with its effect on the time to the terminal event. Thus questions arise on how to evaluate treatment effects based on the two endpoints, for the purpose of deciding on which treatment is most likely to benefit the patients. In this dissertation, we present a unified framework for benefit-risk assessment using the observed longitudinal markers and time to event data. We propose a cumulative weighted marker process to synthesize information from the two endpoints, and use its mean function at a pre-specified time point as a benefit-risk summary measure. We consider nonparametric estimation of the summary measure under two scenarios: (i) the longitudinal marker is measured intermittently during the study period, and (ii) the value of the longitudinal marker is observed throughout the entire follow-up period. The large-sample properties of the estimators are derived and compared. Simulation studies and the application to an AIDS clinical trial exhibit that the proposed methods are easy to implement and reliable for practical use.

In many follow-up or surveillance studies, marker data are collected conditioning on the occurrence of recurrent events. In contrast with the above situation that the marker measurements exists at any time before the terminal event, sometimes marker measurements are triggered by the occurrence of recurrent events. Examples include the medical cost for inpatient or outpatient cares, length-of-stay for hospitalizations, and prognostic or quality-of-life measurement repeatedly measured at multiple infections related to a certain disease. A recurrent marker process, defined between a pre-specified time origin and a terminal event, is composed of recurrent events and repeatedly measured marker measurements. We consider nonparametric estimation of the mean recurrent marker process in the situation when the occurrence of terminal event is subject to competing risks. Statistical methods and inference are developed to address a variety of questions and applications, for the purposes of estimating and comparing the integrated risk in relation to recurrent events, marker measurements and time to the terminal event for different competing risk groups. A SEER-Medicare linked database is used to illustrate the proposed approaches.

Acknowledgements

First and foremost I want to thank my advisor, Dr. Mei-Cheng Wang. It has been a great honor to be one of her students. With her exceptional guidance and precious personality, she has taught me how good research work is done. I sincerely appreciate all her support and care to walk me through tough times and complete my Ph.D. study. I also give my wholehearted thanks to my co-advisor, Dr. Chiung-Yu Huang, who has generously shared with me a lot of insights of research, and kept encouraging me like a sister. The enthusiasm they have for research was contagious and motivational, and I am thankful for the excellent examples they provided as successful women professors in biostatistics.

I would like to thank Dr. Xiaobin Wang for being the chair of my thesis committee and for her support during my work as a research assistant. I'm very grateful to Dr. Ciprian Crainiceanu for serving in my thesis committee and being my academic advisor. Also, I would like to thank Dr. Elizabeth Ogburn and Dr. Lawrence Moulton for being my alternate examiners.

I am very thankful to our wonderful faculty and staff members of Department of Biostatistics and School of Public health at Hopkins for their support throughout the past five years. I also thank Dr. Xiaobin Wang, Dr. Xiumei Hong and Dr. Christine Ladd-Acosta for the opportunities of working on a variety of interesting epidemiologic research. The collaboration experience with them has broadened my perspective in the research on public health.

I would like to thank all the students in this department for their friendship and support over these years. I also thank all the SLAM group members for their inspiring talks and valuable discussions. My special thank goes to Dr.

Yingying Wei, for numerous discussions in our summer study group and her generous help in research and job search.

Finally, I want to give my deepest appreciation to my parents, Qiang Sun and Yan Li, and my fiancé, Yiyang Pan. Their unconditional love are the motivation for me to be a better person. My gratitude and love for them are far beyond words.

Table of Contents

Table of Contents **vi**

List of Tables **x**

List of Figures **xii**

1 Introduction **1**

1.1 Overview of Statistical Problems 1

1.1.1 Benefit-risk assessment using marker process and time-to-event data 1

1.1.2 Recurrent marker process in the presence of competing terminal events 2

1.2 Motivating Examples 3

1.2.1 CPCRA ddI/ddC data 3

1.2.2 SEER-Medicare linked database for breast cancer 4

1.3 Organization 4

2 Literature Review **7**

2.1 Statistical Methods for Longitudinal marker and Time-to-Event Data 7

2.2 Statistical Methods for Cost/Utility Data 9

3 Benefit-risk Assessment Using Marker Process in the Presence of a Terminal Event 10

3.1 Cumulative Weighted Marker Process and Benefit-Risk Assessment 10

3.2 Nonparametric Estimation of $\mu(t)$ When Marker is Intermittently Observed 13

3.2.1 A kernel smoothing approach 14

3.2.2 A computationally more efficient approach 17

3.3 Nonparametric Estimation of $\mu(t)$ When Marker is Continuously Observed 19

3.4 Two-Sample Test 21

3.5 Simulation Studies 23

3.5.1 Simulation when $Y(\cdot)$ is intermittently observed 23

3.5.2 Simulation when $Y(\cdot)$ is continuously observed 27

3.6 Proofs 29

3.6.1 Proof of Theorem 3.2.1 and 3.2.2 29

3.6.2 Proof of Theorem 3.3.1 and comparison of asymptotic variance of $\hat{\mu}_A(t)$ and $\tilde{\mu}(t)$ 33

3.6.3 Proof of Lemma 3.6.1 35

3.6.4 Proof of Lemma 3.6.2 36

4 Recurrent Marker Process in the Presence of Competing Terminal Events 41

4.1 Recurrent Marker Processes 41

4.2 Nonparametric Estimation in Non-Competing Risks Model . . . 44

4.3 Nonparametric Estimation in Competing Risks Model 46

4.4 Improved Estimation in Competing Risks Model 49

4.5 Simulation Studies 52

4.6 Proofs 53

4.6.1 Proof of Theorem 4.2.1 55

4.6.2 Proof of Theorem 4.3.1 57

4.6.3 Proof of Theorem 4.4.1 59

4.6.4 Proof of Corollary 4.4.2 60

5 Analysis of Quality-of-life outcomes and Medical Costs Data: Application to AIDS and Cancer Studies 62

5.1 Analysis of Quality of life and Survival: CPCRA ddI/ddC Trial 62

5.2 Analysis of Censored Medical Cost Data: SEER-Medicare Linked Database 65

6 Discussion 71

List of Tables

3.1 Simulation summary statistics for $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$ 26

3.2 Empirical power of two-sample test when $Y(\cdot)$ is intermittently observed 27

3.3 Empirical power of two-sample test where $Y(\cdot)$ is a function of recurrent event process 29

4.1 Simulation summary statistics for $\hat{\Phi}_j^{imp}$, $\hat{\Phi}_j$ and $\tilde{\Phi}_j$ 54

5.1 Analysis of ddI/ddC trial of CPCRA 65

5.2 Analysis of SEER-Medicare Data 68

List of Figures

3.1 $M(t)$ (area of shaded region) when terminal event occurs before t (left panel) and terminal event occurs after t (right panel), in the special case where $w(\cdot) = 1$. 11

5.1 Estimated cumulative mean functions $\mu(t)$ using Karnofsky score and time to death (left), survival functions (middle) and mean Karnofsky score of survivors $E\{Y(t) \mid D \geq t\}$ (right) for ddI (solid line) and ddC (dashed line) treatment groups. 64

5.2 Estimated total mean cost over time since first diagnosis of breast cancer and point-wise 95% confidence intervals 68

5.3 Cost analysis for breast cancer mortality and cardiovascular disease mortality 69

5.4 Cost analysis for other causes mortality and long-term survivors 70

Chapter 1

Introduction

1.1 Overview of Statistical Problems

1.1.1 Benefit-risk assessment using marker process and time-to-event data

Assessing benefits and risks is a crucial step in the medical decision making process. The purpose of benefit-risk assessment is to determine whether the benefits of an intervention or treatment outweigh its risks based on a given measure. In many clinical trials or biomedical studies, a conventional way of risk assessment is to analyze the time to an event of interest. Statistical methods such as the log-rank test and Cox's proportional hazards model are widely used for risk assessment based on the event time. On the other hand, longitudinally measured patient-centered outcomes or biomarkers are also frequently collected, because they characterize patients' health status and quality of life over time. For example, in the Didanosine/Zalcitabine trial conducted by Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA) (Abrams et al., 1994), time to AIDS progression or death is the primary endpoint; moreover, the Karnofsky score, which quantifies patients' general well-being and physical quality of life, is assigned by study investigators at each follow-up visit. The longitudinal marker measurements, such as quality of life score, offer insights into patients' experience and perceptions and serve as important endpoints in evaluating a treatment. Thus question arises on how to assess benefits and risks based on both time to event and longitudinal marker, for the purpose of deciding on which treatment is most likely to benefit the patients.

1.1.2 Recurrent marker process in the presence of competing terminal events

In biomedical or prospective follow-up studies, longitudinal data are typically collected or observed with pre-specified or random sampling times where sampling times do not have specific biological or medical implications. In contrast, recurrent marker data are a type of repeated measurements, where the sampling times are recurrent event times, and a marker measurement is collected or observed conditioning on the occurrence of a recurrent event. In many situations, marker measurement does not even exist unless a recurrent event takes place. Examples include multiple medical cost for inpatient or outpatient cares, prognostic or quality-of-life measurement repeatedly measured at incidences of infections, and length-of-stay measurement for recurrent hospitalizations. In reality, the recurrent marker process could be terminated by a failure event such as death, and competing risks arise when subjects are exposed to several causes of terminal event (Kalbfleisch and Prentice, 2002). Besides analyzing recurrent marker data over time without discriminating the types of terminal event, investigators are also interested in the performance of recurrent marker process for subjects with a specific type of terminal event. This article presents a framework for nonparametric estimation of recurrent marker process in the presence of competing terminal events.

1.2 Motivating Examples

1.2.1 CPCRA ddI/ddC data

The Community Programs for the Clinical Research on AIDS (CPCRA) was established to study the effectiveness of various treatments for HIV. Comprised of 17 research units that represent a significant diversity of ethnicity, geography and risk group, the CPCRA provides opportunity of clinical research on patients underrepresented in traditional, university-based HIV studies. The CPCRA ddI/ddC study was designed to address the important clinical question of which one of the currently available nucleoside analogues should be given to a patient who can no longer tolerate or has failed ZDV therapy.

The CPCRA ddI/ddC study opened in December 1990 and enrolled 467 patients by September 20, 1991. Among the 467 subjects, 230 were randomized to receive ddI and 237 to receive ddC. All patients were followed for at least one year after the last patient was enrolled. As mentioned above, Karnofsky score is measured at each follow-up visit. Moreover, occurrence of all the opportunistic infections, which indicates deterioration in patients' health, is also recorded. At the end of study, 88 patients from the ddC group and 100 patients from the ddl group died; and the death terminates the existence of Karnofsky score as well as the occurrence of opportunistic infections. In this study, prolonged survival time is desired, but is not the only goal of the treatments. In chapter 3, we develop statistical methods for deciding which treatment is better based on (i)

Karnofsky score and time to death, (ii) Opportunistic infections and time to death.

1.2.2 SEER-Medicare linked database for breast cancer

The linked Surveillance, Epidemiology, and End Results (SEER)-Medicare data are a large population-based source of information for cancer-related health services research in the United States. The SEER-Medicare data for breast cancer patients provide detailed information about Medicare beneficiaries with breast cancer, including clinical, demographic, cause-of-death information and the Medicare claims for covered health care services from the time of a person's Medicare eligibility until death. Specifically, each Medicare claim includes the date of service, diagnostic codes and amounts for charges. The medical cost accumulation process is an example for recurrent marker process, since charges was recorded when each recurrent health service occurred.

There are different possible types of death for a breast cancer patient, including death from a toxic reaction to the therapy, an isolated local recurrence, development of a second type of cancer and so on. Thus, in addition to study the total medical cost, information on medical costs attributed to different types of failure may also be useful. Our work is the first attempt to develop statistical methods to study medical cost accumulation process when competing risk is present.

1.3 Organization

In this dissertation, we consider two types of marker process with terminal events, and develop statistical methods and inference to address a variety of questions and applications related with the two types of data.

For the first type of data, marker exist at any time before the terminal event, for example, the marker can be quality of life or biomarkers. In the context of synthesizing information from both the terminal event process and the longitudinal marker process, we develop a unified framework for benefit-risk assessment to to facilitate decision-making. A summary measure integrating the two outcomes is proposed, including the expected quality-adjusted survival time as a special case. We consider different nonparametric approaches for the summary measure under two scenarios: (i) the longitudinal marker is measured intermittently until terminal event or loss to follow-up, and (ii) the value of the longitudinal marker is observed throughout the entire follow-up period. The contents of this topic are organized as follows: In Chapter 3.1, we define cumulative weighted marker process and use its mean function at a pre-specified time point as a summary measure for benefit-risk assessment. In Chapter 3.2, we consider nonparametric estimation of the mean function of the cumulative weighted marker process when the longitudinal marker is intermittently observed. In Chapter 3.3, a nonparametric estimator of the mean function is proposed when the longitudinal marker is continuously observed. In Chapter 3.4, two-sample tests based on the proposed summary measure is developed. In Chapter 3.6, we report the results of some simulation studies.

For the second type of data, the marker measurement exists conditioning on the occurrence of a recurrent event, for example, the marker can be medical cost or length of stay associated with each hospitalization. In Chapter 4.1, a point process for recurrent events is generalized to a recurrent marker process by accounting for additional information from markers as well as risk types. A mean function is defined with or without competing risks specification. In Chapter 4.2, we consider nonparametric estimation for the mean function without competing risks specification. In Chapter 4.3, a different approach is proposed for the mean function with competing risks specification. In Chapter 4.4, an improved estimator of the mean function under competing risks model is proposed for an efficiency gain. In Chapter 4.5, simulation studies are presented.

In Chapter 5, we present the statistical analysis of two sets of data. First, the CPCRA ddI/ddC data serve as an example of how to conduct benefit-risk assessment based on quality of life and survival time. Second, analysis of the SEER-Medicare breast cancer data is presented to illustrate the proposed methodology on recurrent marker process, with a special focus on competing risks model. The medical costs for sujbects with different causes of death are carefully studied to understand the cost accumulation patterns.

Finally, discussion and directions for future research are provided in Chapter 6.

Chapter 2

Literature Review

2.1 Statistical Methods for Longitudinal marker and Time-to-Event Data

Because the longitudinal measurements and time to the event are often correlated in nature, the occurrence of the terminal event can induce informative drop-out to the collection of longitudinal markers. Thus a conventional longitudinal data analysis which fails to account for the correlated terminal event can result in biased estimation. In the literature, many authors have proposed to employ a joint model of the longitudinal marker process and the terminal event time process to make valid inference. For example, Wu and Carroll (1988), Tsiatis et al. (1995) and Hogan and Laird (1997) linked the two outcome processes via subject-specific random effects, while Henderson et al. (2000), Wang and Taylor (2001) and Xu and Zeger (2001) considered using a time-varying latent process to link the two processes. Although the joint modeling approach is appropriate for describing treatment effects on the longitudinal marker and the time to the terminal event separately, it may be inadequate for decision-making. If a treatment has favorable effects on both endpoints, the decision is straightforward; however, if one treatment shows an advantage on survivorship but a disadvantage on longitudinal marker, then the decision is more difficult to make. In the latter scenario, a summary measure that integrates information from event time and longitudinal marker is desired, and decision can be made by comparing the summary measure across different treatments.

Quality-adjusted survival analysis (Gelber et al., 1989; Glasziou et al., 1990) is a useful tool that incorporates survival time and quality of life into a summary measure. By weighting the durations of different health states by their respective utility values, a single endpoint is constructed to summarize the duration of survival and the quality of life. Nonparametric estimation of the expected quality-adjusted survival time has been studied by many authors, including Huang and Louis (1999), Shen et al. (1999), Zhao and Tsiatis (1999) and Murray and Cole (2000). When the transitions between health states are unclear or if they do not adequately reflect variations in quality of life, Hwang et al. (1996) and Glasziou et al. (1998) considered using quality of life measures over time as the utility weight, instead of assigning a fixed weight to a specific health state. In Hwang et al. (1996), in addition to a cohort study from which the survival function of the time to the terminal event is readily estimable, another cross-sectional survey needs to be conducted in order to estimate the quality-of-life weight. The validity of the estimator then relies on the assumption that the subjects in the cross-sectional survey must be a random sample from the original cohort study population. To ensure an accurate decision-making process, it is desirable to develop standardized and validated methodologies for studying quality-adjusted survival.

2.2 Statistical Methods for Cost/Utility Data

In the absence of competing risks, numerous nonparametric methods have been developed for estimating the total utility or cost at a pre-specifed time horizon. Huang and Louis (1998) studied nonparametric estimation of the joint distribution of a survival time and a mark variable or vector, where an important example of mark variable is the lifetime utility or cost. Lin et al. (1997), Bang and Tsiatis (2000), Strawderman (2000), Zhao and Tian (2001) proposed nonparametric estimators of the mean of the mark variable. Moreover, Huang (2002) and Sun et al. (2009) developed semiparametric models for inference on mark variable and survival outcome. In the literature, estimation of recurrent marker processes under competing risks models has never been considered. In the presence of competing risks, because the failure type is typically unknown for censored subjects, existing methods in the aforementioned papers are not directly extendable to estimate the mean utility or cost attributed to a specific failure type. In this dissertation, we analyze the recurrent marker data over time, with a special focus on competing risks model. The proposed methods are also relevant to the quality-of-life research. Specifically, when the marker is a quality-of-life measurement, the methodology can be extended to quality-adjusted survival analysis (Glasziou et al., 1990) when the terminal event occurs with competing risks.

Chapter 3

Benefit-risk Assessment Using Marker Process in the Presence of a Terminal Event

3.1 Cumulative Weighted Marker Process and Benefit-Risk Assessment

Let $\{Y(t), t \geq 0\}$ be a longitudinal marker process, where $Y(t)$ is a nonnegative marker measurement at time t. Denote the time to the terminal event of interest by D, where D is possibly correlated with marker process $Y(\cdot)$. Here we consider benefit-risk assessment based on the time to the terminal event and the longitudinal marker process before the terminal event, that is, $\{Y(t), D; 0 \leq t \leq D\}$, as the value of $Y(\cdot)$ after D is either not defined or not of interest. For ease of discussion, we assume that a larger marker value indicates a more favorable result throughout this paper. Define the *cumulative weighted marker process*

$$M(t) = \int_0^t w(u)Y(u)I(D \geq u)du,$$

where $w(u)$ is a pre-specified weight function and $Y(u)I(D > u)$ is a marker process that takes the value 0 after the terminal event. In the special case where $w(\cdot) = 1$, $M(t)$ is the area under the marker trajectory before the time point t or the terminal event, whichever occurs first, as shown in Figure 3.1. Note that $M(t)$ can be viewed as an endpoint that integrates information from both the longitudinal marker process and the survival time. An ideal treatment or intervention should prolong survival while maintaining higher marker values over time, thus leading to a large value of $M(t)$ at any time point.

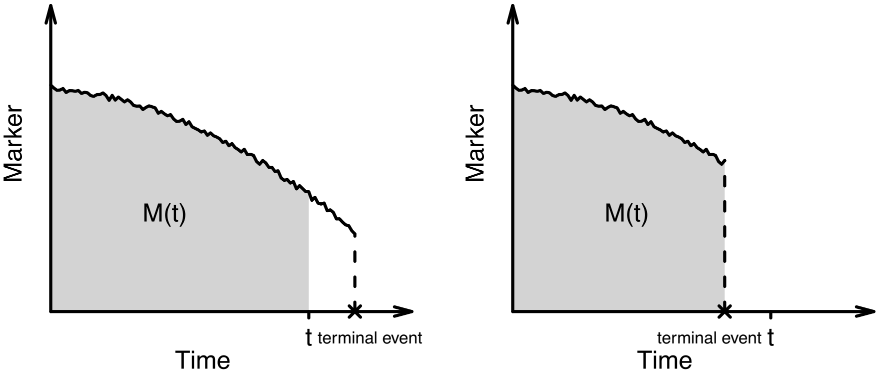


Figure 3.1: $M(t)$ (area of shaded region) when terminal event occurs before t (left panel) and terminal event occurs after t (right panel), in the special case where $w(\cdot) = 1$.

Taking expectation of $M(t)$, we define the cumulative mean function

$$\mu(t) = E\{M(t)\} = \int_0^t w(u)E\{Y(u)I(D \geq u)\}du.$$

This gives the area under curve for the weighted mean function $w(u)E\{Y(u)$ $I(D \geq u)\}$. In the special case where $w(\cdot) = 1$ and $Y(\cdot) = 1$, $\mu(t)$ reduces to the restricted mean survival time up to t (Irwin, 1949); moreover, in the absence of the terminal event, $\mu(t) = \int_0^t E\{Y(u)\}du$ is the area under the expected marker trajectory up to t (Sun and Wu, 2003). The weight function $w(\cdot)$ can be set to reflect the clinical importance of a marker at different time points. For example, if achieving a high marker value at earlier time points is more important than that at later time points, then $w(\cdot)$ can be set as a nonincreasing function of time. We propose to use $\mu(\tau)$, the cumulative mean function at a pre-specified time point τ, as a benefit-risk summary measure, where a treatment with higher value of $\mu(\tau)$ is preferred. In what follows, we consider two scenarios to illustrate the use of the proposed summary measure.

Example 3.1.1 (Quality of life and survival)**.** With advances in treatment and supportive care, treatment decision-making for patients with advanced cancer are increasingly complex. Because cure is elusive for thse patients, it has been recognized that prolonging survival is not the only goal of treatment and that maintaining quality of life is also an important outcome, as patients may be unwilling to accept worse quality of life to achieve longer survival. To integrate quality of life and survival into clinical decision analysis, we let D be the time to death and $Y(t)$ be the quality of life measurement at t. In the special case where $w(\cdot) = 1$, $\mu(\tau)$ is the mean quality-adjusted survival time restricted to time τ. Comparison based on $\mu(\tau)$ can assist investigators to evaluate trade-offs between survival and quality of life.

Example 3.1.2 (Multiple events and survival)**.** In many longitudinal studies, the occurrence of multiple events are commonly encountered and serve as important endpoints. For the Beta-Blocker Evaluation of Survival Trial (The Beta Blocker Evaluation of Survival Trial Investigators, 2001), an advanced chronic heart failure clinical trial, in addition to overall survival, which is the primary endpoint of the study, clinical outcomes such as hospitalization, myocardial infarction, and heart transplantation are also of interest. We denote by T_1, T_2

and T_3, the time of the three secondary endpoints, and by D the time to death. To incorporate information from the multiple event process, one can define $Y(t) = \sum_{i=1}^{3} I(T_i \geq t) + 1$. Then the stochastic process $Y(\cdot)$, which decreases by 1 when any one of the three non-fatal events occurs, can be viewed as a score that reflects patients' disease burden and health condition over time. By setting $w(\cdot) = 1$, the summary measure $\mu(\tau) = E\{\sum_{i=1}^{3} \min(T_i, D, \tau) + \min(D, \tau)\}$ is the expected sum of four types of event-free survival times up to τ (Claggett et al., 2014 - manuscript in preparation).

Note that in the first scenario the longitudinal maker process $Y(\cdot)$ is usually measured at intermittent time points, while in the second scenario $Y(\cdot)$ is completely observed throughout the follow-up period. We then develop different estimating procedures corresponding to the two types of observed data.

3.2 Nonparametric Estimation of $\mu(t)$ When Marker is Intermittently Observed

In this section, we consider nonparametric estimation of the cumulative mean function $\mu(t)$ $(0 \leq t \leq \tau)$ in the case where the longitudinal marker process $Y(\cdot)$ is measured intermittently. In practice, the survival time D is subject to right censoring due to study end or premature dropout. We denote the censoring time by C and assume that C is independent with $\{Y(\cdot), D\}$. Define $X = \min(D, C)$ and $\Delta = I(D \leq C)$. Let $N^*(\cdot)$ be the counting process for the potential data collecting times of the marker $Y(\cdot)$, where the rate function of $N^*(\cdot)$ is $\lambda^*(t)$, that is, $E\{dN^*(t)\} = \lambda^*(t)dt,\ t \geq 0$. Then the counting process $N(t) = I(X \geq t)N^*(t)$ gives the number of observations of the marker before time t, that is, $Y(\cdot)$ is observed only at the time points where $N(\cdot)$ jumps. We further assume that $N^*(\cdot)$ is independent with $\{D, C, Y(\cdot)\}$, then the rate function of the observation time process $N(t)$ is $\lambda(t) = S_X(t)\lambda^*(t)$ with $S_X(t) = \Pr(X \geq t)$. In other words, $\lambda(t)$ gives the instantaneous "risk" of the marker being measured at time t. The observations $\{X_i, \Delta_i, Y_i(t)dN_i(t), 0 \leq t \leq \tau, i = 1, \ldots, n\}$ are assumed to be independent replicates of $\{X, \Delta, Y(t)dN(t), 0 \leq t \leq \tau\}$.

Two major challenges lie in the estimation of $\mu(t) = E\{M(t)\}$. First, because $Y(\cdot)$ is observed at discrete time points during the course of follow-up, the cumulative weighted marker process $M(t)$ is not evaluable. Second, even in the ideal case that $Y(\cdot)$ is completely observed up to X, the induced informative censoring hampers the development of statistical methods. Although it is usually reasonable to assume that the terminal event time D and the censoring time C are independent, $M(D)$ and $M(C)$ are usually positive correlated. For example, a healthier subject may maintain a higher marker value over time, hence having larger $M(C)$ as well as $M(D)$. The naive method of treating $\{M_i(X_i), \Delta_i : i = 1, \ldots, n\}$ as right censored data and estimating the distribution of $M(D)$ using the Kaplan-Meier estimator can result in substantial bias. In what follows, we propose two consistent estimators for $\mu(t)$ and study their large-sample properties.

3.2.1 A kernel smoothing approach

To construct a nonparametric estimator for $\mu(t) = \int_0^t w(u)E\{Y(u)I(D \geq u)\}du$, we first note that the function $\mu(t)$ can be decomposed as

$$\mu(t) = \int_0^t w(u)S_D(u)E\{Y(u) \mid D \geq u\}du, \tag{3.2.1}$$

where $S_D(u) = \Pr(D \geq u)$ is the survival function of D and $r(u) = E\{Y(u) \mid D \geq u\}$ is the expected marker value of survivors at time u. Under independent censoring, subjects in the risk set at time u are a representative sample of event-free individuals at time u in the target population. As a result, it can be shown that $E\{Y(u) \mid D \geq u\} = E\{Y(u) \mid X \geq u\}$. We propose to estimate $r(u)$ with

$$\hat{r}_h(u) = \frac{\sum_{i=1}^n \int_0^\tau K_h(u-s)Y_i(s)I(X_i \geq s)dN_i^*(s)}{\sum_{i=1}^n \int_0^\tau K_h(u-s)I(X_i \geq s)dN_i^*(s)}, \ u \in [h, \tau - h], \tag{3.2.2}$$

where $K_h(x) = h^{-1}K(x/h)$ is a kernel function with bandwidth h, and $K(\cdot)$ satisfies $\int_{-1}^1 K(x)dx = 1$ and $\int_{-1}^1 xK(x)dx = 0$. It is easy to see that $\hat{r}_h$ is a locally weighted average of nearby marker values and is a natural extension of the Nadaraya-Watson estimator. If the uniform kernel is employed, that is, $K(x) = I(|x| < 1)/2$, the denominator of $\hat{r}_h(u)$ is the total number of observations in the time interval $[u-h, u+h]$, while the numerator is the sum of all the observed marker value in $[u-h, u+h]$. To avoid biased estimates in the boundary region $[0, h)$ and $(\tau - h, \tau]$, we set $\hat{r}_h(u) = \hat{r}_h(h)$ for $u \in [0, h)$, and $\hat{r}_h(u) = \hat{r}_h(\tau - h)$ for $u \in (\tau - h, \tau]$. It is shown in Appendix 3.6.3 that $\hat{r}_h(\cdot)$ is uniformly consistent on $[0, \tau]$.

Replacing $E\{Y(u) \mid D \geq u\}$ with $\hat{r}_h(u)$ and $S_D(u)$ with the Kaplan-Meier estimator $\hat{S}_D(u)$ in (3.2.1), we propose to estimate $\mu(t)$ by

$$\hat{\mu}_A(t) = \int_0^t w(u)\hat{S}_D(u)\hat{r}_h(u)du. \tag{3.2.3}$$

Theorem 3.2.1 summarizes the large-sample properties of $\hat{\mu}_A(t)$. Define $M_i^D(t) = N_i^D(t) - \int_0^t I(X_i \geq u)d\Lambda^D(u)$, where $\Lambda^D(t)$ is the cumulative hazard function of D and $N_i^D(t) = I(D_i \leq t, \Delta_i = 1)$.

Theorem 3.2.1. *Under Assumptions (A1)-(A5) in Appendix 3.6.1, the stochastic process* $n^{1/2}\{\hat{\mu}_A(t) - \mu(t)\}$ $(0 \le t \le \tau)$ *has an asymptotically i.i.d. representation* $n^{1/2}\{\hat{\mu}_A(t) - \mu(t)\} = n^{-1/2}\sum_{i=1}^{n} \Psi_i(t) + o_p(1)$*, where*

$$\Psi_i(t) = \int_0^t \frac{\mu(u)dM_i^D(u)}{S_X(u)} - \mu(t)\int_0^t \frac{dM_i^D(u)}{S_X(u)} + \int_0^t \frac{w(u)S_D(u)Y_i(u)I(X_i \ge u)dN_i^*(u)}{\lambda(u)} -$$

$$\int_0^t \frac{w(u)S_D(u)E\{Y(u)I(X \ge u)\}I(X_i \ge u)dN_i^*(u)}{\lambda(u)S_X(u)}.$$

Moreover, as $n \to \infty$, $\sqrt{n}\{\hat{\mu}_A(t) - \mu(t)\}$ $(0 \le t \le \tau)$ *converges weakly to a zero mean Gaussian process with the variance-covariance function* $E\{\Psi_1(s)\Psi_1(t)\}$.

It is worthwhile to point out that the main technical challenge in proving $\sqrt{n}$-consistency of $\hat{\mu}_A(t)$ is that the Kaplan-Meier estimator $\hat{S}_D(\cdot)$ is $\sqrt{n}$-consistent while the kernel-type estimator $\hat{r}_h(\cdot)$ is $\sqrt{nh}$-consistent, thus commonly used techniques such as functional delta method can not be directly applied. It is shown in Appendix 3.6.1 that, by under smoothing $r(u)$ using bandwidth $h = O(n^{-\nu})$ $(1/4 < \nu < 1/2)$, $\hat{\mu}_A(t)$ can achieve $\sqrt{n}$-consistency.

Remark 3.2.1. Although, as a common practice, marker values of survivors are summarized and analyzed for treatment comparison, caution should be paid when interpreting the function $r(u) = E\{Y(u) \mid D \ge u\}$, because the survivor population changes over time and may not be representative of the originally randomized population defined at time zero. To see this, suppose D and $Y(\cdot)$ are correlated through a frailty Z, where a larger value of Z inflates the risk of the terminal event and decreases the value of marker process simultaneously. If a treatment decreases the risk of the terminal event but does not affect $Y(\cdot)$, it can be shown that $E(Z \mid D \ge u)$ of the treatment group is larger than or equal to that of the control group at any time u. As a result, the survivors, based on which inference for $r(u)$ are drawn, are not comparable between treatment and controls as the terminal event occur along the time. In this case, $r(u)$ of treatment group may be lower than or equal to that of the control group. Hence comparisons based on $r(u)$ may yield incorrect conclusion about the treatment effects on the longitudinal marker process.

3.2.2 A computationally more efficient approach

In practice, numerical integration is employed to approximate the integral in (4.2.2). Thus the estimated curve $\hat{r}_h(u)$ needs to be evaluated at a large number of grid points. To reduce the computational burden, we consider an alternative estimator that does not require numerical integration in evaluating the estimator. Specifically, the second estimator is motivated by the equality

$$E\{Y(u)I(X \geq t)dN^*(u)\} = E\{Y(u) \mid D \geq u\}\lambda(u)du,$$

which holds under the assumption that $N^*(\cdot)$ is independent of $\{Y(\cdot), D, C\}$ and C is independent of $\{Y(\cdot), D\}$. Provided $\lambda(u) > 0$ for $u \in [0, t]$, we have

$$\mu(t) = \int_0^t w(u)S_D(u)\lambda(u)^{-1}E\{Y(u)I(X \geq u)dN^*(u)\}.$$

Note that $Y(u)I(X \geq u)dN^*(u)$ is a stochastic process that takes nonzero values only at the time when $dN^*(u) > 0$, so the stochastic process is completely observed and its mean function $E\{Y(u)I(X \geq u)dN^*(u)\}$ can be consistently estimated by its empirical average $n^{-1}\sum_{i=1}^n Y_i(u)I(X_i \geq t)dN_i^*(u)$. Then a nonparametric estimator for $\mu(t)$ is given by

$$\hat{\mu}_B(t) = \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{w(u)\hat{S}_D(u)Y_i(u)I(X_i \geq u)dN_i^*(u)}{\hat{\lambda}_h(u)}$$

where $\hat{\lambda}_h(u) = n^{-1}\sum_{i=1}^n \int_0^\tau K_h(u-s)dN_i(s)$ is a nonparametric smoothed estimator estimator for the rate function $\lambda(u)$. Note that $\hat{\lambda}_h(\cdot)$ can be viewed as an extension of kernel density estimator proposed by Wang and Chiang (2002). As before, we set $\hat{\lambda}_h(u) = \hat{\lambda}_h(h)$ for $u \in (0, h]$ and $\hat{\lambda}_h(u) = \hat{\lambda}_h(\tau - h)$ for $u \in [\tau - h, \tau]$ to avoid boundary effect of the kernel estimator. Theorem 3.2.2 summarizes the large sample properties of $\hat{\mu}_B(t)$, with proofs given in Appendix 3.6.1.

Theorem 3.2.2. *Under the assumptions in Theorem 3.2.1, the stochastic process* $n^{1/2}\{\hat{\mu}_B(t) - \mu(t)\}$ $(0 \le t \le \tau)$ *has an asymptotically i.i.d. representation* $n^{1/2}\{\hat{\mu}_B(t) - \mu(t)\} = n^{-1/2}\sum_{i=1}^n \Psi_i(t) + o_p(1)$. *Moreover, as* $n \to \infty$, $\sqrt{n}\{\hat{\mu}_B(t) - \mu(t)\}$ $(0 \le t \le \tau)$ *converges weakly to a zero mean Gaussian process with the variance-covariance function* $E\{\Psi_1(s)\Psi_1(t)\}$.

Interestingly, the two nonparametric estimators $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$ are asymptotically equivalent. Note that the latter evaluates the smoothed function $\hat{\lambda}_h(\cdot)$ only at the time when marker values are observed, while the former evaluates the smoothed function $\hat{r}_h(\cdot)$ on a much finer grid for numerical integration. Hence $\hat{\mu}_B(t)$ is computationally more convenient than $\hat{\mu}_A(t)$. The simulation study in Section 5 shows that the two estimators have similar performance with finite sample size, we then recommend the use of $\hat{\mu}_B(t)$ to estimate $\mu(t)$. For the standard error estimation, the variance-covariance function $E\{\Psi_1(s)\Psi_1(t)\}$ can be consistently estimated by $n^{-1}\sum_{i=1}^n \widehat{\Psi}_i(s)\widehat{\Psi}_i(t)$ under the assumptions in

Theorem 3.2.1, and

$$\widehat{\Psi}_i(t) = \int_0^t \frac{\hat{\mu}_B(u) d\hat{M}_i^D(u)}{\hat{S}_X(u)} - \hat{\mu}_B(t) \int_0^t \frac{d\hat{M}_i^D(u)}{\hat{S}_X(u)} + \int_0^t \frac{w(u)\hat{S}_D(u)Y_i(u)I(X_i \geq u)dN_i^*(u)}{\hat{\lambda}_h(u)} -$$

$$\int_0^t \frac{w(u)\hat{S}_D(u)\hat{r}_h(u)I(X_i \geq u)dN_i^*(u)}{\hat{\lambda}_h(u)},$$

where $\hat{\Lambda}^D(t) = n^{-1}\sum_{i=1}^n \int_0^t \hat{S}_X(u)^{-1} dN_i^D(u)$ is the Nelson-Aalen estimator of $\Lambda^D(t)$, and $\hat{M}_i^D(t) = N_i^D(t) - \int_0^t I(X_i \geq u) d\hat{\Lambda}_D(u)$.

3.3 Nonparametric Estimation of $\mu(t)$ When Marker is Continuously Observed

In this section, we consider estimation of $\mu(t)$ when the longitudinal marker process $Y(\cdot)$ is completely observed before the terminal event or censoring. The observed data $\{X_i, \Delta_i, I(X_i \geq t)Y_i(t) : 0 \leq t \leq \tau, i = 1, \ldots, n\}$ are assumed to independent replicates of $\{X, \Delta, I(X \geq t)Y(t) : 0 \leq t \leq \tau\}$. As in Section 3, the key step is to estimate the function $r(u) = E\{Y(u) \mid D \geq u\}$. Under the independent censoring assumption, for $u \in [0, \tau]$, we propose to estimate $r(u)$ by the moment type estimator

$$\tilde{r}(u) = \frac{\sum_{i=1}^n Y_i(u)I(X_i \geq u)}{\sum_{i=1}^n I(X_i \geq u)}.$$

Thus a straightforward estimator of $\mu(t)$ is

$$\tilde{\mu}(t) = \int_0^t w(u)\hat{S}_D(u)\tilde{r}(u)du.$$

Note that the moment-type estimator $\tilde{r}(u)$ is a $\sqrt{n}$-consistent estimator for $r(u)$, while the kernel-type estimator $\hat{r}_h(u)$ in (3.2.2) has a $\sqrt{nh}$ convergence rate. Interestingly, $\tilde{\mu}(t)$ can be shown to be more efficient than $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$.

Theorem 3.3.1 states the asymptotic properties of $\tilde{\mu}(t)$, with proof given in Appendix 3.6.2.

Theorem 3.3.1. *Under Assumptions (A1) and (A2) in Appendix 3.6.1, the stochastic process* $n^{1/2}\{\tilde{\mu}(t)-\mu(t)\}$ $(0 \le t \le \tau)$ *has an asymptotically i.i.d. representation* $n^{1/2}\{\tilde{\mu}(t)-\mu(t)\} = n^{-1/2}\sum_{i=1}^{n} U_i(t) + o_p(1)$*, where*

$$U_i(t) = \int_0^t \frac{\mu(u)dM_i^D(u)}{S_X(u)} - \mu(t)\int_0^t \frac{dM_i^D(u)}{S_X(u)} + \int_0^t \frac{w(u)S_D(u)Y_i(u)I(X_i \ge u)}{S_X(u)}du - \int_0^t \frac{w(u)S_D(u)E\{Y(u)I(X \ge u)\}I(X_i \ge u)}{S_X(u)^2}du.$$

Moreover, as $n \to \infty$, $\sqrt{n}\{\tilde{\mu}(t)-\mu(t)\}$ $(0 \le t \le \tau)$ *converges weakly to a zero mean Gaussian process with the variance-covariance function* $E\{U_1(s)U_1(t)\}$*. Moreover,* $E\{U_1(t)\}^2 \le E\{\Psi_1(t)\}^2$ *for all* $t \in [0,\tau]$.

An important application of the proposed methods is benefit-risk assessment that combines information from a multiple event process and a terminal event (see Example 3.1.2 in Section 3.1). Let $O(\cdot)$ denote the multiple event counting process that increase by one when a non-terminal event occurs. For ease of discussion, assume that a smaller value of $O(\cdot)$ at any time point is preferred. To perform benefit-risk assessment based on $\{X, \Delta, I(X \ge t)O(t) : 0 \le t \le \tau\}$, we set $Y(t)$ to be a function of $O(t)$, say, $Y(t) = f\{O(t)\}$, where f is a pre-specified non-increasing function with $f(\cdot) \ge 0$. In this case, $Y(t)$ can be viewed as a score that characterizes patient's disease burden and health condition, and a larger value of $Y(t)$ is desired. Without loss of generality, we set $w(\cdot) = 1$ since the weight function can be absorbed into f.

In practice, the function f can be determined by the investigators. We consider two choices of f for illustration. As suggested by Claggett et al.

(manuscript in preparation), a simple approach is to define a truncated reverse counting process with $f(x) = (K - x)I(K \geq x) + 1$, where K is a pre-specified integer. In this way, only the first K non-terminal events are of interest and $Y(\cdot)$ stays 1 after the Kth event until the terminal event occurs. Another approach is to define $f(x) = a^x$, where $0 < a < 1$. Then each subject starts with a score of 1, and the occurrence of a non-terminal event at time t discounts a patient's score $Y(t)$ by a factor of a. In contrast with the truncated reverse counting process approach, all the non-terminal events are of interest. We recommend the use of second approach when the number of event of interest that can be potentiallly observed is not fixed.

3.4 Two-Sample Test

In this section, we consider nonparametric tests for comparing the benefit-risk summary measure $\mu(\tau)$. Suppose there are two groups, say, group 1 and group 2. The notation used in this section is defined in a way similar to that in Section 3.2, with subscript j indicating the jth group. Assume that both groups can be potentially observed up to time τ. Let $\mu_j(\tau)$ be the mean function of the cumulative weighted marker process at time τ for group j, that is, $\mu_j(\tau) = \int_0^\tau w(u)E\{Y_{j1}(u)I(D_{j1} \geq u)\}du, j = 1, 2$. Consider the null hypothesis $H_0 : \mu_1(\tau) = \mu_2(\tau)$ for two-sample comparison. Let n_j be the number of subjects in the jth group, $n = n_1 + n_2$, and let $\pi_j = \lim_{n\to\infty} n_j/n$, $j = 1, 2$. In this section, $w(\cdot)$ is either known for can be consistently estimated from data by $\hat{w}(t)$. A possible choice of $\hat{w}(t)$ is the Gehan-type weight function $\hat{w}(t) = \{n\hat{S}_{X1}(t)\hat{S}_{X2}(t)\}/\{n_1\hat{S}_{X1}(t) + n_2\hat{S}_{X2}(t)\}$, where $\hat{S}_{Xj}(t)$ is the empirical survival function of $X_{j1}, j = 1, 2$.

We consider testing H_0 in two scenarios: (i) When the marker process $Y(\cdot)$ is intermittently observed, corresponding to the two estimators in Section 3.2, two simple test statistics can be constructed as

$$W_A = \hat{\mu}_{A1}(\tau) - \hat{\mu}_{A2}(\tau) \text{ and } W_B = \hat{\mu}_{B1}(\tau) - \hat{\mu}_{B2}(\tau),$$

where $\tilde{\mu}_j(\tau) = \int_0^\tau \hat{w}(u)\hat{S}_{Dj}(u)\tilde{r}_{jh}(u)du$ and $\hat{\mu}_{Bj}(\tau) = \int_0^\tau \hat{w}(u)\hat{\lambda}_{jh}(u)^{-1}\hat{S}_{Dj}(u)\hat{E}\{Y_{j1}(u)dN_{j1}(u)\}$ for $j = 1, 2$. (ii) When $Y(\cdot)$ is continuously observed, a test statistics can be constructed as

$$\tilde{W} = \tilde{\mu}_1(\tau) - \tilde{\mu}_2(\tau),$$

where $\tilde{\mu}_j(\tau) = \int_0^\tau \hat{w}(u)\hat{S}_{Dj}(u)\tilde{r}_j(u)du$. Let Ψ_{ji} and $\widehat{\Psi}_{ji}$ be straightforward modifications of Ψ_i and $\widehat{\Psi}_i$ given in Theorem 3.2.1, and U_{ji} and $\widehat{U}_{ji}$ be straightforward modifications of U_i and $\widehat{U}_i$ given in Theorem 4.3.3. Theorem 3.4.1 states the asymptotic distribution of the test statistics W_A and W_B under the null hypothesis H_0.

Theorem 3.4.1. Let $\hat{\pi}_j = n_j/n$ be a consistent estimate of π_j, $j = 1, 2$.

Case with intermittently observed $Y(\cdot)$: With the regularity conditions in Theorem 3.2.1 being satisfied for each group, under $H_0 : \mu_1(\tau) = \mu_2(\tau)$, $(n_1 n_2/n)^{1/2} W_A$ and $(n_1 n_2/n)^{1/2} W_B$ converge in distribution to a zero-mean normal random variable with variance $\pi_2 E\left\{\Psi_{1i}(\tau)\right\}^2 + \pi_1 E\left\{\Psi_{2i}(\tau)\right\}^2$. The asymptotic variance can be consistently estimated by $\hat{\pi}_2 \cdot \sum_{i=1}^{n_1} \widehat{\Psi}_{1i}(\tau)^2/n_1 + \hat{\pi}_1 \cdot \sum_{i=1}^{n_2} \widehat{\Psi}_{2i}(\tau)^2/n_2$.

Case with continuously observed $Y(\cdot)$: With the regularity conditions in Theorem 4.3.3 being satisfied for each group, under $H_0 : \mu_1(\tau) = \mu_2(\tau)$,

$(n_1 n_2/n)^{1/2}\tilde{W}$ converges in distribution to a zero-mean normal random variable with variance $\pi_2 E\left\{U_{1i}(\tau)\right\}^2 + \pi_1 E\left\{U_{2i}(\tau)\right\}^2$. The asymptotic variance can be consistently estimated by $\hat{\pi}_2 \cdot \sum_{i=1}^{n_1} \widehat{U}_{1i}(\tau)^2/n_1 + \hat{\pi}_1 \cdot \sum_{i=1}^{n_2} \widehat{U}_{2i}(\tau)^2/n_2$.

Note that the test statistics W_A, W_B and $\tilde{W}$ estimate $\mu_1(\tau) - \mu_2(\tau)$, which can be interpreted as the weighted average difference in $\eta_j(t) = E\{Y_{j1}(t)I(D_{j1} \geq t)\}$ over the interval $[0, \tau]$, $j = 1, 2$. Consider the null hypothesis $H_0' : \eta_1(t) = \eta_2(t)$ for all $t \in [0, \tau]$ and the alternative that $\eta_1(t) \geq \eta_2(t)$ for all $t \in [0, \tau]$ and $\eta_1(\cdot) \neq \eta_2(\cdot)$. Then a natural statistics on which to base a test procedure would be W_A or $\tilde{W}$. Testing H_0' based on W_A or $\tilde{W}$ is a generalization of the distance test in Pepe and Fleming (1989). When setting $Y_{ji}(\cdot) = 1$ for $i = 1, \ldots, n_j$ and $j = 1, 2$, W_A and $\tilde{W}$ reduce to the weighted Kaplan-Meier statistic $WKM = \int_0^\tau \hat{w}(t)\{\hat{S}_{D1}(t) - \hat{S}_{D2}(t)\}dt$ for testing the null hypothesis $H_0'' : S_{D1}(t) = S_{D2}(t)$.

3.5 Simulation Studies

A series of simulation experiments are carried out to examine the finite-sample properties of the proposed methods. In Section 3.5.1, when $Y(\cdot)$ is intermittently observed, we examine the performance of $\hat{\mu}_A(t), \hat{\mu}_B(t)$ for one-sample estimation and W_A, W_B for two-sample test. In Section 3.5.2, when $Y(\cdot)$ is a function of a recurrent event process and is continuously observed, we consider the performance of test statistic $\tilde{W}$ with different choice of f.

3.5.1 Simulation when $Y(\cdot)$ is intermittently observed

In the following simulations, the association between D and $Y(\cdot)$ is induced by a shared subject-specific random effect Z, where Z is generated from a normal distribution with mean 0 and variance σ_1^2. Specifically, given Z, the terminal event time D is generated from exponential distribution with rate parameter $\lambda = a_0 + (Z + k\sigma_1)^2$. Moreover, the longitudinal marker process is generated from $Y(t) = g(t) + Z + \epsilon(t)$; where the error term $\epsilon(t)$ is a mean zero Gaussian process with independent increments and a time-invariant variance σ_2^2. Straightforward algebra gives $E\{Y(t) \mid D \geq t\} = g(t) - 2k\sigma_1^3 t/(1 + 2\sigma_1^2 t)$ and $P(D \geq t) = (1 + 2\sigma_1^2 t)^{-1/2} \exp\left(-a_0 t - k^2\sigma_1^2 t/(1 + 2\sigma_1^2 t)\right)$. Note that when $k = 0$, surviors' expected marker value $E\{Y(t) \mid D \geq t\}$ is $g(t)$, which is the same as $E\{Y(t)\}$; moreover, the difference between $E\{Y(t) \mid D \geq t\}$ and $E\{Y(t)\}$ becomes larger as $|k|$ increases. The model implies that subjects with Z close to $-k\sigma_1$ have smaller rate parameter for the terminal event, and tend to have longer survival time.

In our simulations for one-sample estimation, we set $k = 1, a_0 = 0.1, \sigma_1 = 0.5, \sigma_2 = 0.1, w(t) = 1, g(t) = t + 1$. The censoring time is generated from the uniform distribution on $[0, 5]$. The observation times are generated from $I(X \geq t)dN^*(t)$ and $N^*(t)$ is a Poisson process with constant rate $\lambda^*(t) = 5$. We examine the performance of $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$ when using the Epanechnikov kernel with bandwidth $h = n^{-1/3}, n^{-2/5}$, as the regularity condition (A5) given in the appendix is $h = O(n^{-\nu}), 1/4 < \nu < 1/2$. We also consider leave-one-out cross validation for choosing bandwidth as an extension of that for Nadaraya-Watson estimator. The averaged square error (Härdle et al., 2004) is a commonly used criteria that measures how close the estimate $\hat{r}_h$ is to the true curve r, which can be defined as $n^{-1}\sum_{i=1}^n \int_0^t \{r(u) - \hat{r}_h(u)\}^2 dN_i(u)$ in our case. We choose h that minimizing $CV(h) = n^{-1}\sum_{i=1}^n \int_0^t \{Y_i(u) - \hat{r}_{h,-i}(u)\}^2 dN_i(u)$, where $\hat{r}_{h,-i}(u)$ is the estimate for $r(u)$ leaving out the ith observation, since minimizing

$CV(h)$ is on average equivalent to minimizing $n^{-1}\sum_{i=1}^{n}\int_0^t\{r(u)-\hat{r}_h(u)\}^2 dN_i(u)$. Table 4.1 presents the summary statistics for $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$ based on 2000 replications. The performances of the estimators are not sensitive to the choice of bandwidth. As expected, the variances of the estimators increase as t increases and decrease as sample size n increases. Our proposed procedure performs well in finite-sample studies.

For two sample testing, we first consider scenarios that Y is observed at intermittent time points. Suppose group 1 is the treatment group and group 2 is the control group, and the data of group 1 and 2 are generated from the one-sample model above with different a_{0j},k_j and $g_j(t)$ for jth group, $j = 1, 2$. We consider the following five scenarios: (I) $a_{01} = a_{02} = 0.1, k_1 = k_2 = 1, g_1(t) = g_2(t) = g(t)$. (II) $a_{01} = a_{02} = 0.1, k_1 = 0, k_2 = 1, g_j(t) = 2k_j\sigma_1^3 t(1 + 2\sigma_1^2 t)^{-1} + \sqrt{1 + 2\sigma_1^2 t}\exp\{k_j^2\sigma_1^2 t/(1 + 2\sigma_1^2 t)^{-1}\}$, $j = 1, 2$. (III) $a_{01} = 0.1, a_{02} = 0.2, k_1 = k_2 = 1, g_1(t) = g_2(t) = g(t)$. (IV) $a_{01} = a_{02} = 0.1, k_1 = k_2 = 1, g_1(t) = g_2(t) + 0.2, g_2(t) = g(t)$. (V) $a_{01} = 0.1, a_{02} = 0.2, k_1 = k_2 = 1, g_1(t) = g_2(t) + 0.2, g_B(t) = g(t)$. The empirical powers of the proposed tests based on W_A and W_B are summarized in Table 3.2, with nominal Type I error rate 0.05. We also list the empirical power of the test based on the integrated difference in weighted Kaplan-Meier estimators as a reference, where the test statistics is $WKM = \int_0^\tau w(t)\{\hat{S}_{D1}(t) - \hat{S}_{D2}(t)\}dt$. We set $w(\cdot) = 1$ in our simulations.

For Scenario I, there is no difference in the longitudinal marker process or survival between group 1 and 2, thus null hypothesis H_0 holds and the proposed tests maintain the nominal Type I error rate. For Scenario II, group 1 performs better in terms of survival, while the summary measure of the two groups are equal. Our proposed tests offer a criterion for decision-making and

Table 3.1: Simulation summary statistics for $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$

		$\hat{\mu}_A(t)$			$\hat{\mu}_B(t)$			
	$\mu(t)$	Bias	SE	CP	Bias	SE	CP	SEE
$n = 100, h = n^{-1/3}$								
t = 1	1.076	0.016	0.062	0.932	0.015	0.061	0.938	0.061
t = 2	2.274	0.011	0.155	0.946	0.014	0.155	0.946	0.157
t = 3	3.540	0.001	0.302	0.948	0.014	0.301	0.945	0.301
$n = 100, h = n^{-2/5}$								
t = 1	1.076	0.008	0.062	0.943	0.009	0.061	0.945	0.061
t = 2	2.274	0.002	0.155	0.948	0.009	0.155	0.947	0.156
t = 3	3.540	0.011	0.302	0.944	0.009	0.301	0.944	0.300
$n = 100$, data-adaptive bandwidth								
t = 1	1.076	0.008	0.061	0.938	0.007	0.060	0.948	0.060
t = 2	2.274	0.003	0.160	0.936	0.008	0.161	0.936	0.156
t = 3	3.540	0.004	0.301	0.946	0.010	0.301	0.944	0.301
$n = 200, h = n^{-1/3}$								
t = 1	1.076	0.010	0.044	0.934	0.010	0.043	0.936	0.043
t = 2	2.274	0.008	0.111	0.945	0.010	0.112	0.944	0.111
t = 3	3.540	0.001	0.220	0.946	0.007	0.220	0.949	0.215
$n = 200, h = n^{-2/5}$								
t = 1	1.076	0.005	0.044	0.938	0.006	0.043	0.946	0.043
t = 2	2.274	0.002	0.112	0.942	0.006	0.112	0.944	0.111
t = 3	3.540	0.007	0.220	0.945	0.004	0.220	0.948	0.214
$n = 200$, data-adaptive bandwidth								
t = 1	1.076	0.003	0.045	0.938	0.003	0.044	0.942	0.043
t = 2	2.274	0.001	0.116	0.942	0.003	0.116	0.941	0.111
t = 3	3.540	0.006	0.222	0.939	0.004	0.222	0.940	0.215

Note: Bias is the empirical bias; SE is the empirical standard error; SEE is the empirical mean of the standard error estimates, and the two estimators have the same SEE; CP is the empirical coverage probability of the 95% confidence interval.

still maintain the nominal Type I error rate. For Scenario III, IV and V, there is no conflict of treatment effects on longitudinal marker process and time to terminal event, and group 1 performs better than group 2. Tests based on W_A or W_B can be viewed as testing the existence of treatment effect using synthesized information from both longitudinal marker process and survival outcome, while test based on WKM can be viewed as testing the existence of treatment effect using only survival outcome. For Scenario III, there is only a difference in survival outcome between two groups, the proposed tests and the test based on comparing integrated difference in survival function have similar powers in detecting treatment effect. For scenario IV and V, taking into account the longitudinal marker process helps to increase the power of detecting treatment effect and make correct decisions.

Table 3.2: Empirical power of two-sample test when $Y(\cdot)$ is intermittently observed

	$n_1 = n_2 = 100$			$n_1 = n_2 = 200$		
Scenario	W_A	W_B	WKM	W_A	W_B	WKM
I	0.04	0.04	0.05	0.05	0.05	0.05
II	0.05	0.05	0.45	0.05	0.04	0.75
III	0.18	0.19	0.18	0.32	0.32	0.30
IV	0.22	0.23	0.05	0.35	0.34	0.05
V	0.53	0.55	0.17	0.83	0.84	0.28

3.5.2 Simulation when $Y(\cdot)$ is continuously observed

We also consider the scenario where we have a recurrent event process with a terminal event. Suppose $Z \sim Gamma(\alpha, \alpha)$ is the subject-specific random effect. For the jth group, the recurrent event process is generated from a Poisson process with rate parameter Zc_j, and the terminal event time follows exponential distribution with rate parameter Zd_j. Note that larger α indicates that the two event processes are less correlated, and when $\alpha = \infty$, the two event processes are independent. We consider the following five scenarios: (VI) $\alpha = 4$, $d_1 = d_2 = 0.2$, $c_1 = c_2 = 2$. (VII) $\alpha = \infty$, $d_1 = d_2 = 0.2$, $c_1 = c_2 = 2$. (VIII) $\alpha = 4$, $d_1 = 0.2$, $d_2 = 0.3$, $c_1 = c_2 = 2$. (IX) $\alpha = 4$, $d_1 = d_2 = 0.2$, $c_1 = 2$, $c_2 = 3$. (X) $\alpha = 4$, $d_1 = 0.2$, $d_2 = 0.3$, $c_1 = 2$, $c_2 = 3$. Table 3.3 presents the empirical power of the proposed tests in Section 3.5 using different pre-specified functions f: (a) $f_1(x) = (3-x)I(x \leq 3)+1$, (b) $f_2(x) = 0.8^x$, (c) $f_3(x) = I(x \leq 1)$. Note that and f_3 corresponds to the composite endpoint approach (Meinert, 2012). The empirical powers are summarized in Table 3.3. For Scenario VI and VII, group 1 and group 2 have equal summary measure, and the tests maintain the nominal Type I error rate 0.05. For Scenario VIII, IX and X, group 1 performs better than group 2, and we are interested in the power of the three tests in detecting treatment effect. Tests using f_1 and f_2 have similar powers. However, the test using f_3 does not perform well in Scenario IX, since the composite event is very likely to be the first recurrent event and the recurrent event processes of group 1 and 2 follow the same distribution. Moreover, the composite endpoint approach is not as powerful as the other two test in Scenario X.

Table 3.3: Empirical power of two-sample test where $Y(\cdot)$ is a function of recurrent event process

	$n_1 = n_2 = 100$			$n_1 = n_2 = 200$		
Scenario	f_1	f_2	f_3	f_1	f_2	f_3
VI	0.04	0.05	0.05	0.06	0.05	0.05
VII	0.05	0.04	0.06	0.05	0.05	0.05
VIII	0.45	0.52	0.46	0.76	0.83	0.75
IX	0.14	0.16	0.05	0.27	0.30	0.05
X	0.75	0.80	0.52	0.96	0.98	0.81

3.6 Proofs

3.6.1 Proof of Theorem 3.2.1 and 3.2.2

Assumptions (A1)-(A5) are the regularity conditions in Theorem 3.2.1:

(A1) The censoring time C_i is independent of $\{D_i, N_i^*(\cdot), Y_i(\cdot)\}$ and $P(X_i \geq \tau) > 0$.

(A2) The marker process $Y_i(t)$ is bounded.

(A3) The counting process $N_i^*(\cdot)$ is independent of $\{D_i, C_i, Y_i(\cdot)\}$. The observation time process $I(X_i \geq t)N_i^*(t)$ is bounded and the second derivative of its rate function $\lambda(t)$ is bounded. Moreover, $\lambda(t) > 0$ on $[0, \tau]$.

(A4) Define $\xi(t)$ such that $\xi(t)dt = E[Y(t)I(X \geq t)dN^*(t)]$, the second derivative of $\xi(t)$ is bounded on $[0, \tau]$.

(A5) $K(\cdot)$ is a symmetric kernel function with bounded support and bounded variation, and $h = O(n^{-\nu}), 1/4 < \nu < 1/2$.

We first present two technical lemmas used in the proof. Lemma 3.6.1 states the uniform consistency of the proposed kernel-type estimators, and Lemma 3.6.2 is used when deriving the i.i.d. representation of $\hat{\mu}_A$ and $\hat{\mu}_B$. The two lemmas are proved later in Section 3.6.3 and 3.6.4.

Lemma 3.6.1. *Under Assumptions (A1)–(A5), let*

$$\hat{\xi}_h(t) = 1/n \sum_{i=1}^{n} \int K_h(t-u) Y_i(u) I(X_i \geq u) dN_i^*(u)$$

for $t \in [h, \tau - h]$, $\hat{\xi}_h(t) = \hat{\xi}_h(h)$ *for* $t \in [0, h)$ *and* $\hat{\xi}_h(t) = \hat{\xi}_h(\tau - h)$ *for* $t \in (\tau - h, \tau]$. *Then for* $t \in [0, \tau]$, $\hat{\xi}_h(t)$, $\hat{\lambda}_h(t)$ *and* $\hat{r}_h(t)$ *uniformly converge in probability to* $\xi(t)$, $\lambda(t)$ *and* $r(t)$, *respectively.*

Lemma 3.6.2. *Under Assumptions (A1)–(A5), for* $t \in [0, \tau]$,

$$\sqrt{n} \int_0^t \hat{\lambda}_h(u) du = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t I(X_i \geq u) dN_i^*(u) + o_p(1).$$

Similarly,

$$\sqrt{n} \int_0^t \hat{\xi}_h(u) du = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t Y_i(u) I(X_i \geq u) dN_i^*(u) + o_p(1).$$

We first prove the large-sample properties for $\hat{\mu}_A(t)$. For $\hat{\mu}_A(t)$,

$$\sqrt{n}\{\hat{\mu}_A(t) - \mu(t)\} = \int_0^t \sqrt{n}\{\hat{S}_D(u) - S_D(u)\} \frac{\hat{\xi}_h(u)}{\hat{\lambda}_h(u)} du + \sqrt{n} \int_0^t S_D(u) \left\{ \frac{\hat{\xi}_h(u)}{\hat{\lambda}_h(u)} - \frac{\xi(u)}{\hat{\lambda}_h(u)} \right\} du + \tag{3.6.1}$$

$$\sqrt{n} \int_0^t S_D(u) \left\{ \frac{\xi(u)}{\hat{\lambda}_h(u)} - \frac{\xi(u)}{\lambda(u)} \right\} du$$

Suppose Λ^D is the cumulative hazard function of D and

$$\hat{\Lambda}^D(t) = n^{-1} \sum_{i=1}^{n} \int_0^t \hat{S}_X(u)^{-1} dN_i^D(u)$$

is the Nelson-Aalen estimator, where $N_i^D(u) = I(D_i \leq u, \Delta_i = 1)$. By the uniform consistency of $\hat{\lambda}_h$ and $\hat{\xi}_h$, the first term in Equation (3.6.1) is equal to

$$\int_0^t \sqrt{n}\{\hat{S}_D(u) - S_D(u)\}\frac{\xi(u)}{\lambda(u)}du + o_p(1)$$

$$= -\int_0^t \sqrt{n}\{\hat{\Lambda}_D(u) - \Lambda_D(u)\}d\mu(u) + o_p(1).$$

The above equation holds because of the asymptotic equivalence of $\hat{S}_D$ and $e^{-\hat{\Lambda}_D}$. Moreover,

$$\sqrt{n}(\hat{\Lambda}_D(t) - \Lambda_D(t)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t S_X(u)^{-1}dM_i^D(u) + o_p(1),$$

and $M_i^D(t) = N_i^D(t) - \int_0^t I(X_i \geq u)d\Lambda_D(u)$. By the uniform consistency of $\hat{\lambda}_h$ and Lemma 2, and following similar steps as Mammen and Nielsen (2007), the second term in Equation (3.6.1) is equal to

$$\int_0^t \frac{S_D(u)}{\lambda(u)}\sqrt{n}\{\hat{\xi}_h(u) - \xi(u)\}du + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{S_D(u)}{\lambda(u)}Y_i(u)I(X_i \geq u)dN_i^*(u) - \sqrt{n}\mu(t) + o_p(1).$$

Similarly, the third term in Equation (3.6.1) is equal to

$$-\int_0^t \frac{S_D(u)\xi(u)}{\lambda(u)^2}\sqrt{n}\{\hat{\lambda}_h(u) - \lambda(u)\}du + o_p(1).$$

$$= -\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{S_D(u)\xi(u)}{\lambda(u)^2}I(X_i \geq u)dN_i^*(u) + \sqrt{n}\mu(t) + o_p(1).$$

Thus $\sqrt{n}\{\hat{\mu}_A(t)-\mu(t)\} = 1/\sqrt{n}\sum_{i=1}^n \Psi_i(t) + o_p(1)$. For $\hat{\mu}_B(t)$, we have

$$\sqrt{n}\{\hat{\mu}_B(t)-\mu(t)\} = \sqrt{n}\int_0^t \frac{\hat{S}_D(u)}{\hat{\lambda}_h(u)} d\{\hat{R}(u)-R(u)\} + \sqrt{n}\int_0^t \frac{\hat{S}_D(u)-S_D(u)}{\hat{\lambda}_h(u)} dR(u) + \tag{3.6.2}$$

$$\sqrt{n}\int_0^t S_D(u)\left\{\frac{1}{\hat{\lambda}_h(u)} - \frac{1}{\lambda(u)}\right\} dR(u)$$

Define $R(t) = \int_0^t E[Y(u)I(X \geq u)dN^*(u)]$ and $\hat{R}(t) = \int_0^t \hat{E}[Y(u)I(X \geq u)dN^*(u)]$, where $\hat{E}[Y(u)I(X \geq u)dN^*(u)] = 1/n\sum_{i=1}^n Y_i(u)I(X_i \geq u)dN_i^*(u)$. By the uniform consistency of $\hat{\lambda}_h$ and asymptotic equivalence of $\hat{S}_D$ and $e^{-\hat{\Lambda}^D}$, the first term in Equation (3.6.2) is equal to

$$\sqrt{n}\int_0^t \frac{S_D(u)}{\lambda(u)} d\{\hat{R}(u)-R(u)\} + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{S_D(u)}{\lambda(u)} Y_i(u)I(X_i \geq u)dN_i^*(u) - \sqrt{n}\mu(t) + o_p(1).$$

The second term is equal to $-\int_0^t \sqrt{n}\{\hat{\Lambda}_D(u)-\Lambda_D(u)\}d\mu(u) + o_p(1)$. And by the uniform consistency of $\hat{\lambda}_h$ and Lemma 2, the second term in Equation (3.6.2) is

$$-\int_0^t \frac{1}{\lambda(u)^2}\sqrt{n}\left\{\hat{\lambda}_h(u)-\lambda(u)\right\} dR(u) + o_p(1)$$

$$= -\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{r(u)}{\lambda(u)^2} I(X_i \geq u)dN_i^*(u) + \sqrt{n}\mu(t) + o_p(1).$$

Thus $\sqrt{n}\{\hat{\mu}_B(t)-\mu(t)\} = 1/\sqrt{n}\sum_{i=1}^n \Psi_i(t) + o_p(1)$. Since $\Psi_i(t)$ can be written as sums of monotone functions in t and are therefore manageable, then the weak convergence of $\sqrt{n}\{\hat{\mu}_B(t)-\mu(t)\}$ and $\sqrt{n}\{\hat{\mu}_A(t)-\mu(t)\}$ holds. The consistency of the variance estimates $1/n\sum_{i=1}^n \widehat{\Psi}_i(s)\widehat{\Psi}_i(t)$ follows from the arguments used in the proof of Lin et al. (1998).

3.6.2 Proof of Theorem 3.3.1 and comparison of asymptotic variance of $\hat{\mu}_A(t)$ and $\tilde{\mu}(t)$

The proof of Theorem 3.3.1 follows similar steps as Theorem 3.2.1. For $\tilde{\mu}(t)$, we have

$$\sqrt{n}\{\tilde{\mu}(t)-\mu(t)\} = \int_0^t \sqrt{n}\{\hat{S}_D(u)-S_D(u)\}\frac{n^{-1}\sum_{i=1}^n Y_i(u)I(X_i \geq u)}{n^{-1}\sum_{i=1}^n I(X_i \geq u)}du+$$

$$\sqrt{n}\int_0^t S_D(u)\left\{\frac{n^{-1}\sum_{i=1}^n Y_i(u)I(X_i \geq u)}{n^{-1}\sum_{i=1}^n I(X_i \geq u)} - \frac{E\{Y(u)I(X \geq u)\}}{n^{-1}\sum_{i=1}^n I(X_i \geq u)}\right\}du+$$

$$\sqrt{n}\int_0^t S_D(u)\left\{\frac{E\{Y(u)I(X \geq u)\}}{n^{-1}\sum_{i=1}^n I(X_i \geq u)} - \frac{E\{Y(u)I(X \geq u)\}}{S_X(u)}\right\}du$$

Similar as the proof for Theorem 3.2.1, we have

$$\sqrt{n}\{\tilde{\mu}(t)-\mu(t)\} = -\int_0^t \sqrt{n}\{\hat{\Lambda}_D(u)-\Lambda_D(u)\}d\mu(u) + \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{S_D(u)}{S_X(u)}Y_i(u)I(X_i \geq u)du-$$

$$\frac{1}{n}\sum_{i=1}^n \int_0^t \frac{S_D(u)E\{Y(u)I(X \geq u)\}}{S_X(u)^2}I(X_i \geq u)du + o_p(1)$$

$$= \frac{1}{n}\sum_{i=1}^n U_i(t) + o_p(1)$$

Moreover, by defining

$$\widehat{U}_i(t) = \int_0^t \frac{\tilde{\mu}(u)d\hat{M}_i^D(u)}{\hat{S}_X(u)} - \tilde{\mu}(t)\int_0^t \frac{d\hat{M}_i^D(u)}{\hat{S}_X(u)} + \int_0^t \frac{\hat{S}_D(u)Y_i(u)I(X_i \geq u)}{\hat{S}_X(u)}du-$$

$$\int_0^t \frac{\hat{S}_D(u)\hat{r}(u)I(X_i \geq u)}{\hat{S}_X(u)}du,$$

the covariance at (s,t) can also be consistently estimated by $1/n\sum_{i=1}^n \widehat{U}_i(s)\widehat{U}_i(t)$.

We now prove that $E\{\Psi_i(t)^2\} \geq E\{U_i(t)^2\}$. Taking $f_i(u) = S_D(u)[S_X(u)Y_i(u)I(X_i \geq u) - E\{Y(u)I(X \geq u)\}I(X_i \geq u)]/S_X(u)^2$, then $\Psi_i(t) = A_i(t) + \int_0^t f_i(u)/\lambda^*(u)dN_i^*(u)$

and $U_i(t) = A_i(t) + \int_0^t f_i(u)du$, where $A_i(t) = \int_0^t S_X(u)^{-1}\mu(u)dM_i^D(u) - \mu(t)\int_0^t S_X(u)^{-1}dM_i^D(u)$. Thus,

$$
\begin{aligned}
&E\{\Psi_i(t)^2\} - E\{U_i(t)^2\} \\
&= 2E\left[A_i(t)\left\{\int_0^t f_i(u)/\lambda^*(u)dN_i^*(u) - \int_0^t f_i(u)du\right\}\right] + E\left\{\int_0^t f_i(u)/\lambda^*(u)dN_i^*(u)\right\}^2 - \\
&\qquad E\left\{\int_0^t f_i(u)du\right\}^2 \\
&= 2E_{X,Y}\left[A_i(t)E_{N^*|X,Y}\left\{\int_0^t f_i(u)/\lambda^*(u)dN_i^*(u) - \int_0^t f_i(u)du\right\}\right] + \\
&\qquad E_{X,Y}E_{N^*|X,Y}\left\{\int_0^t f_i(u)/\lambda^*(u)dN_i^*(u)\right\}^2 - E_{X,Y}\left\{\int_0^t f_i(u)du\right\}^2 \\
&= E_{X,Y}E_{N^*|X,Y}\left\{\int_0^t f_i(u)/\lambda^*(u)dN_i^*(u)\right\}^2 - E_{X,Y}\left\{\int_0^t f_i(u)du\right\}^2 \\
&\geq E_{X,Y}\left\{E_{N^*|X,Y}\int_0^t f_i(u)/\lambda^*(u)dN_i^*(u)\right\}^2 - E_{X,Y}\left\{\int_0^t f_i(u)du\right\}^2 \\
&\geq E_{X,Y}\left\{\int_0^t f_i(u)du\right\}^2 - E_{X,Y}\left\{\int_0^t f_i(u)du\right\}^2 \\
&= 0.
\end{aligned}
$$

Therefore we prove that the asymptotic variance of $\tilde{\mu}(t)$ is smaller than that of $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$.

3.6.3 Proof of Lemma 3.6.1

We prove the result for ξ and the proof for λ is similar. For $t \in [h, \tau - h]$, we have

$$\hat{\xi}_h(t) = \int K_h(t-u) d\hat{R}_2(u),$$

and by integration by part

$$\begin{aligned}\sup_{t\in[h,\tau-h]} \left| \hat{\xi}_h(t) - E\{\hat{\xi}_h(t)\} \right| &= \sup_{t\in[h,\tau-h]} \left| \int_{t-h}^{t+h} \{\hat{R}_2(u) - R_2(u)\} dK_h(t-u) \right| \\ &\leq h^{-1} \sup_{t\in[0,\tau]} |\hat{R}_2(t) - R_2(t)| \cdot V(K).\end{aligned}$$

where $V(K)$ is the variation of the kernel function K. The functions $R_{2i}(t) = \int_0^t Y_i(u) I(X_i > u) dN_i^*(u)$ are monotone and bounded, therefore have pseudodimension at most 1. From Pollard (1990) p.37, $\sup_{t\in[0,\tau]} \sqrt{n} \mid \hat{R}_2(t) - R_2(t) \mid$ has uniformly subgaussian tail, that is, there exists a constant C such that

$$P(\sup_{t\in[0,\tau]} \sqrt{n} \mid \hat{R}_2(t) - R_2(t) \mid > t) < e^{-Ct^2}.$$

Then for any $\epsilon > 0$

$$P(\sup_{t\in[0,\tau]} \frac{1}{h} \mid \hat{R}_2(t) - R_2(t) \mid > \epsilon) = P(\sup_{t\in[0,\tau]} \sqrt{n} \mid \hat{R}_2(t) - R_2(t) \mid > \sqrt{n} h \epsilon) < e^{-Cnh^2\epsilon^2}.$$

So $\sup_{t\in[h,\tau-h]} \left| \hat{\xi}_h(t) - E\{\hat{\xi}_h(t)\} \right|$ converge to 0 in probability when $nh^2 \to \infty$. Also,

$$\sup_{t\in[h,\tau-h]} \left| E\{\hat{\xi}_h(t)\} - \xi(t) \right| = O(h),$$

and $\sup_{t\in[0,h]} \mid \xi(t) - \xi(h) \mid = O(h)$, $\sup_{t\in[\tau-h,\tau]} \mid \xi(t) - \xi(\tau - h) \mid = O(h)$. Therefore, the uniform consistency holds. Given the uniform consistency for $\hat{\lambda}_h$ and $\hat{\xi}_h$, $\hat{r}_h$ uniformly converges in probability to r on $[0, \tau]$.

3.6.4 Proof of Lemma 3.6.2

Now we prove that for $s \in [0, \tau]$,

$$\sqrt{n} \int_0^s \hat{\lambda}_h(t) dt = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^s dN_i(u) + o_p(1).$$

Define $\lambda_i(t) = \int_0^\tau K_h(t-u) dN_i(u)$ for $t \in [h, \tau - h]$,$\lambda_i(t) = \lambda_i(h)$ for $t \in [0, h)$, and $\lambda_i(t) = \lambda_i(\tau - h)$ for $t \in (\tau - h, \tau]$. Then $\hat{\lambda}_h(t) = \sum_{i=1}^n \lambda_i(t)$. We the above equation under four scenarios (a) $0 < s \leq h$, (b) $h < s \leq 3h$, (c) $3h < s \leq \tau - h$ and (d) $\tau - h < s \leq \tau$.

(a) For $0 < s \leq h$, we have

$$\begin{aligned}
& E\left\{\int_0^s \lambda_i(t) dt - \int_0^s dN_i(u)\right\} \\
& = E\left\{\int_0^s \int_0^{2h} K_h(h-u) dN_i(u) dt - \int_0^s r(u) dN_i(u)\right\} \\
& = \int_0^s \int_0^{2h} K_h(h-u) \lambda(u) du \cdot dt - \int_0^s \lambda(u) du \\
& = \int_0^s \lambda(h) \cdot dt - \int_0^s \lambda(u) du + O(h^2) \\
& = O(h^2). \qquad (3.6.3)
\end{aligned}$$

Moreover, $\int_0^s \lambda_i(t) dt$ is a bounded monotone function in s, thus $1/\sqrt{n} \sum_{i=1}^n [\int_0^s \lambda_i(t) dt - \int_0^s dN_i(u) - E\{\int_0^s \lambda_i(t) dt - \int_0^s dN_i(u)\}]$ converge weakly to a Gaussian process with variance less than $M_1 h$, where M_1 is a constant. Thus $1/\sqrt{n} \sum_{i=1}^n \{\int_0^s \lambda_i(t) dt - \int_0^s dN_i(u)\} = O(\sqrt{n} h^2) + o_p(1)$.

(b) When $h < s < 3h$, we have

$$\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)$$

$$= \left\{\int_0^h \lambda_i(t)dt - \int_0^h dN_i(u)\right\} + \left\{\int_h^s \lambda_i(t)dt - \int_h^s dN_i(u)\right\}$$

$$= \left\{\int_0^h \lambda_i(t)dt - \int_0^h dN_i(u)\right\} + \left\{\int_0^{s-h}\int_h^{u+h} K_h(t-u)dtdN_i(u) + \right.$$

$$\left.\int_{s-h}^{2h}\int_h^s K_h(t-u)dtdN_i(u) + \int_{2h}^{s+h}\int_{u-h}^s K_h(t-u)dtdN_i(u) - \int_h^s dN_i(u)\right\}$$

$$= \left\{\int_0^h \lambda_i(t)dt - \int_0^h dN_i(u)\right\} + \left\{\int_0^{2h}\int_h^{u+h} K_h(t-u)dtdN_i(u) - \int_h^{2h} dN_i(u)\right\} +$$

$$\left\{\int_{s-h}^s\int_{u+h}^s K_h(t-u)dtdN_i(u) + \int_s^{s+h}\int_{u-h}^s K_h(t-u)dtdN_i(u)\right\}$$

$$\overset{def}{=} \Pi_1 + \Pi_2 + \Pi_3.$$

We then prove $E(\Pi_1 + \Pi_2 + \Pi_3) = O(h^2)$. By Equation (3.6.3), we have $E(\Pi_1) =$

$O(h^2)$. Suppose $\dot{\lambda}$ is the first derivative of λ. For Π_2, we have

$$
\begin{aligned}
E(\Pi_2) &= E\left\{\int_0^{2h}\int_h^{u+h} K_h(t-u)dtdN_i(u) - \int_h^{2h} dN_i(u)\right\}\\
&= \int_0^{2h}\int_{1-u/h}^{1} K(x)dx\lambda(u)du - \int_h^{2h} \lambda(u)du\\
&= \int_0^{h}\int_{1-u/h}^{1} K(x)dx\lambda(u)du + \int_h^{2h}\int_{1-u/h}^{-1} K(x)dx\lambda(u)du\\
&= \int_0^{h}\int_{1-u/h}^{1} K(x)dx\{\lambda(u)-\lambda(0)\}du - \int_h^{2h}\int_{-1}^{1-u/h} K(x)dx\{\lambda(u)-\lambda(0)\}du\\
&= \int_0^{h}\int_{1-u/h}^{1} K(x)dx\dot{\lambda}(0)udu - \int_h^{2h}\int_{-1}^{1-u/h} K(x)dx\dot{\lambda}(0)udu + O(h^2)\\
&= O(h^2) + O(h^2) = O(h^2).
\end{aligned}
$$

For Π_3, we have

$$
\begin{aligned}
E(\Pi_3) &= E\left\{\int_{s-h}^{s}\int_{u+h}^{s} K_h(t-u)dtdN_i(u) + \int_s^{s+h}\int_{u-h}^{s} K_h(t-u)dtdN_i(u)\right\}\\
&= -\int_{s-h}^{s}\int_{(s-u)/h}^{1} K(x)dx\lambda(u)du + \int_s^{s+h}\int_{-1}^{(s-u)/h} K(x)dx\lambda(u)du\\
&= -\int_{s-h}^{s}\int_{(s-u)/h}^{1} K(x)dx\{\lambda(u)-\lambda(0)\}du + \int_s^{s+h}\int_{-1}^{(s-u)/h} K(x)dx\{\lambda(u)-\lambda(0)\}du\\
&= O(h^2).
\end{aligned}
$$

Thus $E\{\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)\} = O(h^2)$. Again, we have $E\left\{\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)\right\} = O(h^2)$ and $E\left\{\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)\right\}^2 \leq M_2 h$, where M_2 is a constant. Then we have the equation $n^{-1/2}\sum_{i=1}^n\{\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)\} = O(\sqrt{n}h^2) + o_p(1)$ hold.

(c) When $3h < s \leq \tau - h$,

$$\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)$$

$$= \left\{ \int_0^h \lambda_i(t)dt - \int_0^h dN_i(u) \right\} + \left\{ \int_0^{2h} \int_h^{u+h} K_h(t-u)dtdN_i(u) - \int_h^{2h} dN_i(u) \right\} +$$

$$\left\{ \int_{s-h}^{s+h} \int_{u-h}^{s} K_h(t-u)dtdN_i(u) - \int_{s-h}^{s} dN_i(u) \right\}$$

$$=\Pi_1 + \Pi_2 + \Pi_3.$$

Similarly as the arguments for $h < s \leq 3h$, we have $n^{-1/2}\sum_{i=1}^n \{\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)\} = O(\sqrt{n}h^2) + o_p(1)$.

(d) When $\tau - h < s \leq \tau$, we have

$$\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)$$

$$= \left\{ \int_0^{\tau-h} \lambda_i(t)dt - \int_0^{\tau-h} dN_i(u) \right\} + \left\{ \int_{\tau-h}^{s} \lambda_i(\tau-h)dt - \int_{\tau-h}^{s} dN_i(u) \right\}$$

$$= O(h^2) + o_p(n^{-1/2}) + \left\{ \int_{\tau-h}^{s} \lambda_i(\tau-h)dt - \int_{\tau-h}^{s} dN_i(u) \right\}.$$

And

$$E\left\{ \int_{\tau-h}^{s} \lambda_i(\tau-h)dt - \int_{\tau-h}^{s} dN_i(u) \right\}$$

$$= \int_{\tau-h}^{s} \int_{\tau-2h}^{\tau} K_h(\tau-h-u)\lambda(u)du \cdot dt - \int_{\tau-h}^{s} \lambda(u)du$$

$$= \int_{\tau-h}^{s} \lambda(\tau-h)dt - \int_{\tau-h}^{s} \lambda(u)du + O(h^2)$$

$$= O(h^2) + O(h^2) = O(h^2).$$

Moreover, we have $\int_{\tau-h}^{s} \lambda_i(\tau - h)dt - \int_{\tau-h}^{s} dN_i(u) = O(h^2) + o_p(n^{-1/2})$. Thus the equation $n^{-1/2}\sum_{i=1}^{n}\{\int_0^s \lambda_i(t)dt - \int_0^s dN_i(u)\} = O(\sqrt{n}h^2) + o_p(1)$ also holds. So if we take $h = n^{-\alpha}$, where $1/4 < \alpha < 1/2$, we have

$$\sqrt{n}\int_0^s \hat{\lambda}_h(t)dt - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^s dN_i(u) = o_p(1).$$

Along the same line as above, the equation $\sqrt{n}\int_0^s \hat{\xi}_h(t)dt = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^s Y_i(u)I(X_i > u)dN_i^*(u) + o_p(1)$ is proved by changing $dN_i(\cdot)$ to $Y_i(\cdot)dN_i(\cdot)$.

Chapter 4

Recurrent Marker Process in the Presence of Competing Terminal Events

4.1 Recurrent Marker Processes

Let $\{N(t), t \geq 0\}$ be a recurrent event process with $N(t)$ representing the total number of recurrent events occurring at or prior to t. Suppose the occurrence of a recurrent event at t, i.e., $dN(t) = 1$, is marked by a measurement $Y(t)$, and for ease of discussion, we assume that $Y(t)$ is nonnegative. Then, $Y(t)$ can be considered as a marker measurement for the recurrent event occurring at t, and we represent the marked recurrent event process by $\{N(t), Y(t)|_{dN(t)=1}; t \geq 0\}$. Consider the case where a terminal event is present, and the marked recurrent event process $\{N(t), Y(t)|_{dN(t)=1}; t \geq 0\}$ vanishes after the terminal event. Specifically, in situations where the marker $Y(\cdot)$ is a utility measurement such as medical cost or length of stay in hospital, the main interest would naturally be the utility consumed by survivors in the population. Let the time to a terminal event be represented by D, which is possibly correlated with $\{N(t), Y(t)|_{dN(t)=1}; t \geq 0\}$, then the stochastic process $\{Y(t)dN(t), 0 \leq t \leq D\}$

is frequently of interest.

In the presence of a terminal event without competing risks, the recurrent marker process is defined as

$$M^{total}(t) = \int_0^t Y(u)I(D > u)dN(u), \tag{4.1.1}$$

and we use the superscript "total" to distinguish (4.1.1) from recurrent marker processes under competing risks model, which is defined later in this section. The mean function (MF) is defined as

$$\Phi(t) = \mathrm{E}\{M^{total}(t)\} = \mathrm{E}\left\{\int_0^t Y(u)I(D \geq u)dN(u)\right\} \ .$$

When $Y(\cdot)$ is a utility measure, the MF $\Phi(t)$ corresponds to the average of cumulative utility consumed before the terminal event during time interval $[0, t]$, which is the utility of real life in case death is the terminal event. Also note that by setting $Y(\cdot) = 1$, the recurrent marker process reduces to recurrent event process and $M^{total}(t) = N(\min(D, t))$. Define $\phi(t)$ as the derivative of $\Phi(t)$, that is, $d\Phi(t) = \mathrm{E}\{Y(t)I(D \geq t)dN(t)\} = \phi(t)dt$. The function $\phi(t)$ is the rate of change of the MF at time t, which can be regarded as a counterpart of the rate function for a recurrent event process.

Now further consider recurrent marker process in the presence of a terminal event with competing risks. Suppose the occurrence of the terminal event is caused by one of J different types of risks, where the risk-type indicator is denoted by $\Pi \in \{1, \ldots, J\}$. Let τ be a pre-specified constant; for example, τ could be the maximum length of follow-up time or the length of time where investigators in a research project wish to study the recurrent marker process. For practical consideration, due to limited follow-up time, subjects with $D < \tau$

are classified according to the original risk-type of terminal event $\Pi = j$ ($j = 1, \ldots, J$), and subjects with $D \geq \tau$ are classified into the last category of risk-type, $J + 1$. In applications, subjects of risk-type $J + 1$ can be thought of as those "long-term survivors"or "cured cases".

For $0 \leq t \leq \tau$ and $j = 1, 2, \ldots, J$, the recurrent marker process with type-j terminal event is defined as

$$M_j(t) \quad = \quad \int_0^t Y(u)I(\Pi = j, u \leq D < \tau)dN(u),$$

and the recurrent marker process with type-$(J+1)$ terminal event is defined as

$$M_{J+1}(t) \quad = \quad \int_0^t Y(u)I(D \geq \tau)dN(u).$$

Taking expectation of $M_j(t)$, the type-j MF at time t is

$$\Phi_j(t) = \mathrm{E}\{M_j(t)\}, \quad j = 1, 2, \ldots, J + 1.$$

For $j = 1, 2, \ldots, J+1$, the MF $\Phi_j(t)$ is the expectation of cumulative recurrent marker prior to time t attributed to type-j risk. To connect the function $\Phi(t)$ with $\Phi_j(t)$, it is clear that for each $t \in [0, \tau]$, we have

$$\Phi(t) = \sum_{j=1}^{J+1} \Phi_j(t). \tag{4.1.2}$$

In practice, another quantity of interest is the conditional mean function of recurrent marker process given risk-type j, that is,

$$\Phi_j^c(t) = \begin{cases} E\{M^{total}(t) \mid \Pi = j, D \leq \tau\} & j = 1, \ldots, J, \\ E\{M^{total}(t) \mid D \geq \tau\} & j = J + 1. \end{cases}$$

Note that $\Phi_j^c(t)$ is the expectation of cumulative recurrent marker up to time t of subpopulation with type-j risk. We shall consider nonparametric estimation of $\Phi(t)$, $\Phi_j(t)$ and $\Phi_j^c(t)$ ($j = 1, \ldots, J + 1$) in later sections.

4.2 Nonparametric Estimation in Non-Competing Risks Model

We first consider nonparametric estimation of $\Phi(t)$ without the complication of competing risks. In practice, the terminal event time D is subject to right censoring due to study end or premature dropout, and the recurrent marker process cannot be observed after censoring. We denote the censoring time by C and assume that C is independent of $\{D, Y(t)dN(t); 0 \leq t \leq \tau\}$. The observed terminal event time is $X = \min(D, C)$ with a censoring indicator $\Delta = I(D < C)$. The observed data $\{X_i, \Delta_i, Y_i(t)I(X_i \geq t)dN_i(t), I(X_i \geq t)N_i(t) : 0 \leq t \leq \tau,\ i = 1, \ldots, n\}$ are assumed to be independent replicates of $\{X, \Delta, Y(t)I(X \geq t)dN(t), I(X \geq t)N(t) : 0 \leq t \leq \tau\}$.

Let $S_D(\cdot)$ be survival function of the terminal event time D. Under the independent censoring assumption, subjects in the risk set at time u are a representative sample of event-free individuals at time u in the target population, and we note that

$$\phi(u)du = S_D(u) \cdot \mathrm{E}\{Y(u)dN(u)|\ D \geq u\} = S_D(u) \cdot \mathrm{E}\{Y(u)dN(u)|\ X \geq u\}.$$

Thus, for $0 < t \leq \tau$, we have

$$\Phi(t) = \int_0^t S_D(u) \cdot \mathrm{E}\{Y(u)dN(u) \mid X \geq u\}. \tag{4.2.1}$$

To estimate $\Phi(t)$, one can use the moment estimator of $\mathrm{E}\{Y(u)dN(u) \mid X \geq u\}$ based on subjects in the risk set $\{i : X_i \geq u\}$ and estimate $S_D(u)$ by the Kaplan-Meier estimate $\hat{S}_D(u)$, then a nonparametric estimator of $\Phi(t)$ can be constructed as

$$\hat{\Phi}(t) = \int_0^t \hat{S}_D(u) \cdot \frac{\sum_{i=1}^n Y_i(u)I(X_i \geq u)dN_i(u)}{\sum_{i=1}^n I(X_i \geq u)}. \tag{4.2.2}$$

The estimator in (4.2.2) can be viewed as an extension of the nonparametric estimator of mean frequency function in Ghosh and Lin (2000), where the special case $Y(\cdot)|_{dN(\cdot)=1} = 1$ was considered. To study large sample properties of $\hat{\Phi}(t)$, the following notations are introduced. Let S_C and Λ_C be the survival function and cumulative hazard function of the censoring time C, respectively; and let S_X denote the survival function of the observed failure time X. We then define $\mathcal{N}_i^C(t) = I(C_i \leq t, \Delta_i = 0)$ and $\mathcal{M}_i^C(t) = \mathcal{N}_i^C(t) - \int_0^t I(X_i \geq u) d\Lambda_C(u)$. Theorem 4.2.1 summarizes the asymptotic property of $\hat{\Phi}(t)$, with proof given in Appendix 4.6.1.

Theorem 4.2.1. Under Assumption (A1) and (A2) in Appendix, for $t \in (0, \tau]$, the stochastic process $n^{1/2}\{\hat{\Phi}(t) - \Phi(t)\}$ has an asymptotically i.i.d. representation $n^{1/2}\{\hat{\Phi}(t) - \Phi(t)\} = n^{-1/2}\sum_{i=1}^n a_i(t) + o_p(1)$, where

$$a_i(t) = \int_0^t S_C(u)^{-1} Y_i(u) I(X_i \geq u) dN_i(u) - \Phi(t) + \Phi(t)\int_0^t S_X(u)^{-1} d\mathcal{M}_i^C(u)$$

$$- \int_0^t \Phi(u) S_X(u)^{-1} d\mathcal{M}_i^C(u).$$

Moreover, as $n \to \infty$, $n^{1/2}\{\hat{\Phi}(t) - \Phi(t)\}$ converges weakly to a zero-mean tight Gaussian process whose covariance function at (t_1, t_2) can be consistently estimated by $n^{-1}\sum_{i=1}^n \hat{a}_i(t_1)\hat{a}_i(t_2)$ for $t_1, t_2 \in (0, \tau]$, with $\hat{a}_i(t)$ defined in Appendix 4.6.1.

Remark. To estimate the rate function $\phi(t)$ for $t \in (0, \tau]$, one can use the following kernel estimate,

$$\hat{\phi}_h(t) = \frac{1}{n\hat{S}_C(t)} \sum_{i=1}^n \int_0^\tau K_h(t-u) Y_i(u) I(X_i \geq u) dN_i(u),$$

where $K_h(x) = h^{-1}K(x/h)$ is a kernel function with bandwidth h, with $K(\cdot)$ satisfing $\int_{-1}^{1} K(x)dx = 1$ and $\int_{-1}^{1} xK(x)dx = 0$. Note that $\hat{\phi}_h(t)$ can be viewed as an extension of the kernel-type estimator proposed by Wang and Chiang (2002).

4.3 Nonparametric Estimation in Competing Risks Model

When the terminal event occurs with competing risks, nonparametric estimation of the MF will need to take into account the data structure that the risk-type indicator, Π, is available only when the terminal event is observed. Under competing risks model, the censoring time C is assumed to be independent of $\{D, \Pi, Y(t)dN(t); 0 \le t \le \tau\}$. The observed data $\{X_i, \Delta_i, \Pi_i, Y_i(t)I(X_i \ge t)dN_i(t), I(X_i \ge t)N_i(t) : 0 \le t \le \tau, i = 1, \ldots, n\}$ are assumed to be independent replicates of $\{X, \Delta, \Pi, Y(t)I(X \ge t)dN(t), I(X \ge t)N(t) : 0 \le t \le \tau\}$.

Similar to the formula in (4.2.1), for $j = 1, \ldots, J$ and $0 \le t \le \tau$, one derives

$$\Phi_j(t) = \int_0^t E\{Y(u)I(\Pi = j, u \le D < \tau)dN(u)\}$$

$$= \int_0^t S_D(u) \ \mathrm{E}\{I(\Pi = j)Y(u)dN(u) \mid X \ge u\} - S_D(\tau)E\{I(\Pi = j)M^{total}(t) \mid X \ge \tau\}.$$

Therefore, for $j = 1, \ldots, J$, along the same line of the estimator $\hat{\Phi}(t)$ in (4.2.2), a hypothetical estimator of $\Phi_j(t)$ that utilizes the recurrent marker history data from all the subjects can be obtained as,

$$\hat{\Phi}_j^H(t) = \int_0^t \hat{S}_D(u) \cdot \frac{\sum_{i=1}^n I(X_i \ge u, \Pi_i = j)Y_i(u)dN_i(u)}{\sum_{i=1}^n I(X_i \ge u)}$$

$$- \hat{S}_D(\tau)\frac{\sum_{i=1}^n I(X_i \ge \tau, \Pi_i = j)M_i^{total}(t)}{\sum_{i=1}^n I(X_i \ge \tau)}. \qquad (4.3.1)$$

And for risk type $J+1$, an estimator of $\Phi_{J+1}(t) = S_D(\tau)E\{M^{total}(t) \mid X \geq \tau\}$ can be constructed as

$$\hat{\Phi}^H_{J+1}(t) = \hat{S}_D(\tau)\frac{\sum_{i=1}^n I(X_i \geq \tau)M_i^{total}(t)}{\sum_{i=1}^n I(X_i \geq \tau)}.$$

Clearly, the estimator $\hat{\Phi}^H_j(t)$ $(j = 1, \ldots, J)$ depends on data information of Π from all subjects in the risk set at the observed recurrent event times. In reality, however, knowledge of Π is rarely available from subjects whose terminal event are censored $(\Delta = 0)$, therefore the hypothetical estimator in (4.3.1) fails to serve as a proper estimator for most of the applications. We next propose an estimation approach which is useful for commonly encountered recurrent marker data with competing terminal events.

For different risk types $j = 1, \ldots, J$, we define $H_j(t, m, u) = \Pr(\Pi = j, M^{total}(t) \leq m, D \leq u)$. With straightforward algebra, the mean function for type-j risk is

$$\Phi_j(t) = \int_0^\infty mH_j(t, dm, \tau).$$

Note that for $u \in [0, \tau]$, $H_j(t, \infty, u) = \Pr(\Pi = j, D \leq u)$ is the cumulative incidence function in standard competing risks model (Prentice et al., 1978; Gray, 1988; Fine and Gray, 1999). Define $U_j(t, m, u) = \Pr(\Pi = j, M^{total}(t) \leq m, \Delta = 1, X \leq u)$. Under the assumption that C is independent of $\{D, \Pi, Y(t)dN(t); 0 \leq t \leq \tau\}$, one derives

$$H_j(t, m, u) = \int_0^u S_D(v)\frac{H_j(t, m, dv)}{S_D(v)} = \int_0^u S_D(v)\frac{U_j(t, m, dv)}{S_X(v)},$$

where S_X is the survival function of the observed failure time X. Note that when a terminal event is uncensored, the risk type is observed and therefore

$U_j(t,m,u)$ can be estimated by its empirical average, that is, $\hat{U}_j(t,m,u) = n^{-1}\sum_{i=1}^n I(\Pi_i = j, M_i^{total}(t) \le m, \Delta_i = 1, X_i \le u)$. By plugging into the Kaplan-Meier estimate $\hat{S}_D$ and the empirical estimates ($\hat{S}_X$, $\hat{U}_j$), $H_j(t,m,u)$ can be nonparametrically estimated by

$$\hat{H}_j(t,m,u) = \int_0^u \hat{S}_D(v)\frac{\hat{U}_j(t,m,dv)}{\hat{S}_X(v)}. \tag{4.3.2}$$

Thus, an estimator of $\Phi_j(t)$ $(j = 1,\ldots,J)$ can be constructed as

$$\hat{\Phi}_j(t) = \int_0^\infty m\hat{H}_j(t,dm,\tau). \tag{4.3.3}$$

and $\Phi_{J+1}(t)$ is still estimated by

$$\hat{\Phi}_{J+1}(t) \equiv \hat{\Phi}_{J+1}^H(t) = \hat{S}_D(\tau)\frac{\sum_{i=1}^n I(X_i \ge \tau)M_i^{total}(t)}{\sum_{i=1}^n I(X_i \ge \tau)}.$$

Theorem 4.3.1 summarizes the large-sample properties of $\hat{\Phi}_j(t)$, $j = 1,\ldots,J+1$, with proof given in Appendix 4.6.2.

Theorem 4.3.1. Under Assumption (A1') and (A2) in Appendix, for $t \in (0,\tau]$ and $j = 1,\ldots,J+1$, the stochastice process $n^{1/2}\{\hat{\Phi}_j(t) - \Phi_j(t)\}$ has an asymptotically i.i.d representation $n^{1/2}\{\hat{\Phi}_j(t) - \Phi_j(t)\} = n^{-1/2}\sum_{i=1}^n b_{ji}(t) + o_p(1)$, where for $j = 1,\ldots,J$,

$$\begin{aligned} b_{ji}(t) =& M_i^{total}(t)I(\Pi_i = j, D_i \le \tau)\Delta_i S_C(D_i)^{-1} - \Phi_j(t) + \Phi_j(t)\int_0^\tau S_X(u)^{-1}d\mathcal{M}_i^C(u) \\ &- \int_0^\tau E\{I(\Pi = j, D \le u)M^{total}(t)\}S_X(u)^{-1}d\mathcal{M}_i^C(u), \end{aligned}$$

and for $j = J+1$,

$$b_{ji}(t) = I(X_i \ge \tau)M_i^{total}(t)S_C(\tau)^{-1} - \Phi_{J+1}(t) + \Phi_{J+1}(t)\int_0^\tau S_X(u)^{-1}d\mathcal{M}_i^C(u).$$

Moreover, as $n \to \infty$, $n^{1/2}\{\hat{\Phi}_j(t) - \Phi_j(t)\}$ converges weakly to a zero-mean tight Gaussian process with covariance function $E\{b_{j1}(t_1)b_{j1}(t_2)\}$ for $t_1, t_2 \in (0, \tau]$, and the covariance can be consistently estimated by $n^{-1}\sum_{i=1}^{n} \hat{b}_{ji}(t_1)\hat{b}_{ji}(t_2)$, with $\hat{b}_{ji}(t)$ defined in Appendix 4.6.2.

4.4 Improved Estimation in Competing Risks Model

As the estimator $\hat{\Phi}_j(t)$ utilizes marker information only from uncensored subjects, the estimation may be inefficient when censoring is heavy; in contrast, for the estimation of $\Phi(t)$, the estimator $\hat{\Phi}(t)$ in Section 3.1 utilizes marker history data from both censored and uncensored subjects. Thus, it is of no surprise that $\hat{\Phi}(t)$ is more efficient than $\sum_{j=1}^{J+1} \hat{\Phi}_j(t)$ for estimating $\Phi(t)$. When formulating estimators for type-j MF $\Phi_j(t)$, a question arises as to whether it is possible to borrow information from $\hat{\Phi}(t)$ to improve the estimation of $\Phi_j(t)$.

Note that, based on (4.1.2), we can construct an alternative estimator of $\Phi_j(t)$ as

$$\tilde{\Phi}_j(t) = \hat{\Phi}(t) - \sum_{k \neq j} \hat{\Phi}_k(t).$$

In general, $\tilde{\Phi}_j(t)$ may or may not be more efficient than $\hat{\Phi}_j(t)$, even though the former estimator involves marker information from both censored and uncensored subjects and the latter only uses marker information from uncensored ones. In what follows, we propose an estimator that is more efficient than $\tilde{\Phi}_j(t)$ and $\hat{\Phi}_j(t)$. We consider a class of linearly combined estimators $\mathcal{W}_{jt} = \{w_{jt}\hat{\Phi}_j(t) + (1 - w_{jt})\tilde{\Phi}_j(t) : w_{jt} \in \mathbb{R}\}$ and propose the use of the most efficient estimator from this class. Let $\bar{\Phi}_j(t) = w_{jt}\hat{\Phi}_j(t) + (1 - w_{jt})\tilde{\Phi}_j(t)$ be a weighted average of $\hat{\Phi}_j(t)$ and $\tilde{\Phi}_j(t)$. Clearly, $\hat{\Phi}_j(t)$ and $\tilde{\Phi}_j(t)$ both belong to $\mathcal{W}_{jt}$. Using results from Theorem 4.2.1 and 4.3.1, estimators in $\mathcal{W}_{jt}$ are consistent and asymptotically normal. Thus, a question of interest is to identify the estimator in $\mathcal{W}_{jt}$ which has the minimal asymptotic variance.

To study the asymptotic variance of the estimators in $\mathcal{W}_{jt}$, we define $\boldsymbol{\Sigma}_j(t_1, t_2) = \{\sigma_{jpq}(t_1, t_2)\}_{2\times 2}$ for $t_1, t_2 \in [0, \tau]$, where $\sigma_{j11}(t_1, t_2), \sigma_{j12}(t_1, t_2), \sigma_{j21}(t_1, t_2)$ and $\sigma_{j22}(t_1, t_2)$ are the asymptotic covariance of $[\sqrt{n}\{\hat{\Phi}_j(t_1) - \Phi_j(t_1)\}, \sqrt{n}\{\hat{\Phi}_j(t_2) - \Phi_j(t_2)\}]$, $[\sqrt{n}\{\hat{\Phi}_j(t_1) - \Phi_j(t_1)\}, \sqrt{n}\{\tilde{\Phi}_j(t_2) - \Phi_j(t_2)\}]$, $[\sqrt{n}\{\tilde{\Phi}_j(t_1) - \Phi_j(t_1)\}, \sqrt{n}\{\hat{\Phi}_j(t_2) - \Phi_j(t_2)\}]$ and $[\sqrt{n}\{\tilde{\Phi}_j(t_1) - \Phi_j(t_1)\}, \sqrt{n}\{\tilde{\Phi}_j(t_2) - \Phi_j(t_2)\}]$, respectively. We further assume that $\boldsymbol{\Sigma}_j(t, t)$ is a nonsingular matrix for $t \in (0, \tau]$. Define the vector $\boldsymbol{w}_{jt} = (w_{jt}, 1 - w_{jt})^{\mathsf{T}}$, then the asymptotic variance of $\sqrt{n}\{\bar{\Phi}_j(t) - \Phi_j(t)\}$ is $\boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t, t)\boldsymbol{w}_{jt}$. We consider the optimization of asymptotic variance to obtain the most efficient estimator in the class of $\mathcal{W}_{jt}$:

$$\underset{\boldsymbol{w}_{jt}}{\text{minimize}} \quad \boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t, t)\boldsymbol{w}_{jt}.$$

By the method of Lagrange multipliers, the optimal weight can be derived as

$$\boldsymbol{w}_{jt}^* = \frac{\boldsymbol{\Sigma}_j(t, t)^{-1}\boldsymbol{e}}{\boldsymbol{e}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t, t)^{-1}\boldsymbol{e}},$$

where we define $\boldsymbol{e} = (1, 1)^{\mathsf{T}}$. And using this optimal weight $\boldsymbol{w}_{jt}^*$, the asymptotic variance of $\sqrt{n}\{\bar{\Phi}_j(t) - \Phi_j(t)\}$ is $\{\boldsymbol{e}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t, t)^{-1}\boldsymbol{e}\}^{-1}$. By the Cauchy-Schwartz inequality, for any weight $\boldsymbol{w}_{jt}$, we have

$$\{\boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t, t)\boldsymbol{w}_{jt}\}\{\boldsymbol{e}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t, t)^{-1}\boldsymbol{e}\} \geq (\boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{e})^2 = 1,$$

or equivalently,

$$\boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t, t)\boldsymbol{w}_{jt} \geq \{\boldsymbol{e}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t, t)^{-1}\boldsymbol{e}\}^{-1}. \tag{4.4.1}$$

Thus, in the class $\mathcal{W}_{jt}$, using the weight $\boldsymbol{w}_{jt}^*$ results in the estimator with smallest asymptotic variance. In real applications, the optimal $\boldsymbol{w}_{jt}^*$ involves $\boldsymbol{\Sigma}_j(t,t)$, which is unknown and needs to be estimated from data. By applying the results of Theorem 4.2.1 and 4.3.1, we can consistently estimate $\boldsymbol{\Sigma}_j(t,t)$ by $\widehat{\boldsymbol{\Sigma}}_j(t,t)$, with details given in the Appendix 4.6.3. We propose the following improved estimator for $\Phi_j(t)$,

$$\hat{\Phi}_j^{imp}(t) = \hat{w}_{jt}^* \hat{\Phi}_j(t) + (1 - \hat{w}_{jt}^*)\tilde{\Phi}_j(t), \tag{4.4.2}$$

where $\hat{\boldsymbol{w}}_{jt}^* = (\hat{w}_{jt}^*, 1 - \hat{w}_{jt}^*)^\mathsf{T} = \widehat{\boldsymbol{\Sigma}}_j(t,t)^{-1}\boldsymbol{e}/\boldsymbol{e}^\mathsf{T}\widehat{\boldsymbol{\Sigma}}_j(t,t)^{-1}\boldsymbol{e}$. Here we indicate that the estimator using the estimated weight $\hat{\boldsymbol{w}}_{jt}^*$ possesses the same efficiency as the estimator with $\boldsymbol{w}_{jt}^*$ as the weight. Theorem 4.4.1 summarizes the large-sample properties of $\hat{\Phi}_j^{imp}(t)$ and the proof is given in Appendix 4.6.3.

Theorem 4.4.1. Under Assumption (A1'),(A2) and (A3) in Appendix, for $j = 1, \ldots, J+1$, $\sqrt{n}\{\hat{\Phi}_j^{imp}(t) - \Phi_j(t)\}(0 < t \le \tau)$ converges weakly to a zero-mean tight Gaussian process with

$$\frac{\boldsymbol{e}^\mathsf{T}\boldsymbol{\Sigma}_j(t_1,t_1)^{-1}\boldsymbol{\Sigma}_j(t_1,t_2)\boldsymbol{\Sigma}_j(t_2,t_2)^{-1}\boldsymbol{e}}{\boldsymbol{e}^\mathsf{T}\boldsymbol{\Sigma}_j(t_1,t_1)^{-1}\boldsymbol{e}\boldsymbol{e}^\mathsf{T}\boldsymbol{\Sigma}_j(t_2,t_2)^{-1}\boldsymbol{e}}$$

as the covariance function at (t_1, t_2) for $t_1, t_2 \in (0, \tau]$, and the covariance can be consistently estimated by $n^{-1}\sum_{i=1}^n \hat{f}_{ji}(t_1)\hat{f}_{ji}(t_2)$, with $\hat{f}_{ji}(t)$ defined in Appendix 4.6.3.

Clearly, the result of Theorem 4.1 implies that the asymptotic variance of $\sqrt{n}\{\hat{\Phi}_j^{imp}(t) - \Phi_j(t)\}$, $\{\boldsymbol{e}^\mathsf{T}\boldsymbol{\Sigma}_j(t,t)^{-1}\boldsymbol{e}\}^{-1}$, is smaller than or equal to the asymptotic variance of $\sqrt{n}\{\hat{\Phi}_j(t) - \Phi_j(t)\}$ and $\sqrt{n}\{\tilde{\Phi}_j(t) - \Phi_j(t)\}$.

In real data applications, it would also be of interest to estimate the mean recurrent marker process within each risk group, $\Phi_j^c(t)$. For $j = 1, \ldots, J$, since the equation $\Phi_j^c(t) = \Phi_j(t)/P(\Pi = j, D \leq \tau)$ holds, we propose the following estimator,

$$\hat{\Phi}_j^c(t) = \frac{\hat{\Phi}_j^{imp}(t)}{\hat{H}_j(0, \infty, \tau)},$$

where $\hat{H}_j(0, \infty, \tau)$ is defined in equation (4.3.2). For risk-type $j = J + 1$, since we have $\Phi_{J+1}^c(t) = \Phi_{J+1}(t)/P(D \geq \tau)$, we propose the estimator

$$\hat{\Phi}_{J+1}^c(t) = \frac{\hat{\Phi}_{J+1}^{imp}(t)}{\hat{S}_D(\tau)}.$$

Corollary 4.2 summarizes the large-sample properties of $\hat{\Phi}_j^c(t)$, with proof given in Appendix 4.6.4.

Corollary 4.4.2. Under Assumption (A1'),(A2) and (A3) in Appendix, for $j = 1, \ldots, J + 1$, the stochastic process $n^{1/2}\{\hat{\Phi}_j^c(t) - \Phi_j^c(t)\}(0 < t \leq \tau)$ converges weakly to a zero-mean tight Gaussian process with covariance function $E\{g_{j1}(t_1)g_{j1}(t_2)\}$ for $t_1, t_2 \in (0, \tau]$, and the covariance function can be consistently estimated by $n^{-1}\sum_{i=1}^n \hat{g}_{ji}(t_1)\hat{g}_{ji}(t_2)$, with $\hat{g}_{ji}(t)$ defined in Appendix 4.6.4.

4.5 Simulation Studies

A series of simulation experiments are carried out to examine finite-sample performance of the proposed methods. We simulate the data so that the association among the random variables $\{D, Y(\cdot)|_{dN(\cdot)=1}, N(\cdot), \Pi\}$ is induced by a subject-specific random effect Z, where Z is generated from a gamma distribution with shape parameter $\alpha = 2$ and rate parameter $\beta = 0.5$. Specifically, given Z, the terminal event time D is generated from Weibull distribution with rate parameter $.01 \times Z$ and shape parameter $\nu = 3$; the recurrent event process $N(\cdot)$ is a

Poisson process with rate function $\lambda(t) = I(Z > z_0) + 1$, where z_0 is the median of Z; and the marker process is generated from $Y(t)|_{dN(t)=1} = 1 + t + Z + \epsilon(t)$; where the error term $\epsilon(t)$ is a mean zero Gaussian process with independent increments and a time-invariant standard deviation $\sigma = 0.1$. We assume there are two types of terminal event, and that the cause of death is determined by Z: we set $\Pi = 1$ when $Z \leq z_0$ and set $\Pi = 2$ when $Z > z_0$. Moreover, we set $\tau = 5$, and subjects with $D > 5$ belong to the third risk type. The censoring time C is generated from Uniform$[0, 19]$ to produce a 25% censoring rate. We set the sample size $n = 200$ and $n = 400$. The simulation results are based on 2000 replications and are summarized in Table 1.

It can be seen that the proposed estimators $\hat{\Phi}_j$, $\tilde{\Phi}_j$ and $\hat{\Phi}_j^{imp}$ all perform well with moderate sample sizes. In our simulation, note that $\hat{\Phi}_j$ has smaller standard error than $\tilde{\Phi}_j$ for $j = 1, 3$, while the standard error of $\tilde{\Phi}_2$ outperforms that of $\hat{\Phi}_2$. In whichever case, the improved estimator $\hat{\Phi}_j^{imp}$ is either more efficient than or as efficient as $\hat{\Phi}_j$ and $\tilde{\Phi}_j$. And, as expected, the standard error of the proposed estimators increase with time and decrease with sample size.

4.6 Proofs

We first introduce a few regularity conditions used in the theorems:

(A1) The censoring time C is independent of $\{D, Y(t)dN(t); t \in [0, \tau]\}$ and $P(X \geq \tau) > 0$.

(A1') The censoring time C is independent of $\{D, Y(t)dN(t), \Pi; t \in [0, \tau]\}$ and $P(X \geq \tau) > 0$.

Table 4.1: Simulation summary statistics for $\hat{\Phi}_j^{imp}$, $\hat{\Phi}_j$ and $\tilde{\Phi}_j$

		$\hat{\Phi}_j^{imp}$				$\hat{\Phi}_j$		$\tilde{\Phi}_j$	
		Bias	SE	SEE	CP	Bias	SE	Bias	SE
$n = 200$									
$\Phi_1(t)$	$t = 1$	0.001	0.095	0.093	0.935	0.002	0.095	0.004	0.152
	$t = 2$	0.003	0.189	0.185	0.937	0.005	0.189	0.011	0.277
	$t = 3$	0.002	0.297	0.291	0.939	0.002	0.297	0.006	0.406
$\Phi_2(t)$	$t = 1$	0.014	0.307	0.298	0.936	0.004	0.324	0.002	0.308
	$t = 2$	0.020	0.589	0.587	0.946	0.008	0.613	0.002	0.590
	$t = 3$	0.021	0.893	0.889	0.941	0.012	0.924	0.008	0.895
$\Phi_3(t)$	$t = 1$	0.012	0.175	0.170	0.932	0.003	0.179	0.006	0.197
	$t = 2$	0.020	0.355	0.349	0.932	0.003	0.361	0.008	0.384
	$t = 3$	0.028	0.587	0.575	0.928	0.001	0.595	0.005	0.617
$n = 400$									
$\Phi_1(t)$	$t = 1$	0.001	0.067	0.066	0.947	0.001	0.067	0.000	0.109
	$t = 2$	0.001	0.131	0.132	0.947	0.001	0.131	0.003	0.199
	$t = 3$	0.001	0.206	0.207	0.948	0.002	0.207	0.005	0.291
$\Phi_2(t)$	$t = 1$	0.006	0.211	0.212	0.951	0.003	0.221	0.004	0.211
	$t = 2$	0.012	0.420	0.417	0.946	0.003	0.438	0.002	0.420
	$t = 3$	0.010	0.644	0.630	0.943	0.010	0.664	0.007	0.646
$\Phi_3(t)$	$t = 1$	0.003	0.121	0.122	0.949	0.001	0.123	0.002	0.138
	$t = 2$	0.006	0.249	0.250	0.945	0.003	0.253	0.001	0.272
	$t = 3$	0.014	0.411	0.410	0.945	0.002	0.418	0.002	0.433

Note: Bias is the empirical bias; SE is the empirical standard error; SEE is the empirical mean of the standard error estimates; CP is the empirical coverage probability of the 95% confidence interval.

(A2) The stochastic process $N(t)$ and $Y(t)dN(t)$ are bounded for $t \in [0, \tau]$.

(A3) The covariance matrix $\Sigma_j(t,t)$ is nonsingular for $t \in (0, \tau]$.

4.6.1 Proof of Theorem 4.2.1

Assumptions (A1) and (A2) are the regularity conditions for Theorem 4.2.1. Define $\mathcal{N}_i^C(t) = I(C_i \le t, \Delta_i = 0)$, we consider Nelson-Aalen estimator for the cumulative hazard function of censoring time, $\Lambda_C(\cdot)$,

$$\hat{\Lambda}_C(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{d\mathcal{N}_i^C(u)}{\hat{S}_X(u)},$$

where $\hat{S}_X(t) = n^{-1}\sum_{i=1}^n I(X_i \ge t)$ is the empirical estimator for $S_X(t)$. Note that $S_C(t)$ can be estimated by $e^{-\hat{\Lambda}_C(t)}$, which is asymptotically equivalent to the Kaplan-Meier estimator. We use $\hat{S}_C(t) = e^{-\hat{\Lambda}_C(t)}$ in what follows. For $0 < t \le \tau$,

$$\begin{aligned}\Phi(t) &= \int_0^t \frac{S_D(u)}{S_X(u)} E\{Y(u)I(X > u)dN(u)\}\\ &= \int_0^t \frac{E\{Y(u)I(X > u)dN(u)\}}{S_C(u)}\\ &= \int_0^t e^{\Lambda_C(u)}dB(u),\end{aligned}$$

where $B(t) = \int_0^t E\{Y(u)I(X > u)dN(u)\}$ and can be estimated by $\hat{B}(t) = n^{-1}\sum_{i=1}^n \int_0^t Y_i(u)I(X_i > u)dN_i(u)$; By the martingale central limit theorem, we have

$$\sqrt{n}\{\hat{\Lambda}_C(t) - \Lambda_C(t)\} = n^{-1/2}\sum_{i=1}^n \int_0^t S_X(u)^{-1}d\mathcal{M}_i^C(u) + o_p(1),$$

where $\mathcal{M}_i^C(t) = \mathcal{N}_i^C(t) - \int_0^t I(X_i \ge u)d\Lambda_C(u)$.

Given the estimator $\hat{\Phi}(t) = \int_0^t e^{\hat{\Lambda}_C(u)} d\hat{B}(u)$, by the functional delta method (van der Vaart, 2000; Theorem 20.8, Lemma 20.10), the functional $(F_1, F_2) \mapsto \int_0^t e^{F_1} dF_2$ is Hadamard-differentiable as a map into the set of cadlag functions $D[0, \tau]$, and the derivative is $(h_1, h_2) \mapsto \int_0^t e^{F_1} dh_2 + \int_0^t e^{F_1} h_1 dF_2$, thus $\sqrt{n}\{\hat{\Phi}(t) - \Phi(t)\}$ converge weakly to a Gaussian process. Moreover,

$$\sqrt{n}\{\hat{\Phi}(t) - \Phi(t)\}$$

$$= \sqrt{n} \int_0^t e^{\Lambda_C(u)} d\{\hat{B}(u) - B(u)\} + \sqrt{n} \int_0^t e^{\Lambda_C(u)} \{\hat{\Lambda}_C(u) - \Lambda_C(u)\} dB(u) + o_p(1)$$

$$= \sqrt{n} \int_0^t e^{\Lambda_C(u)} d\{\hat{B}(u) - B(u)\} + \sqrt{n} \int_0^t \{\hat{\Lambda}_C(u) - \Lambda_C(u)\} d\Phi(u) + o_p(1)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n a_i(t) + o_p(1),$$

where

$$a_i(t) = \int_0^t \frac{I(X_i \geq u) Y_i(u) dN_i(u)}{S_C(u)} - \Phi(t) + \Phi(t) \int_0^t \frac{d\mathcal{M}_i^C(u)}{S_X(u)} - \int_0^t \Phi(u) \frac{d\mathcal{M}_i^C(u)}{S_X(u)}.$$

The limiting covariance function at (t_1, t_2) is $E\{a_1(t_1) a_1(t_2)\}$, which can be consistently estimated by $n^{-1} \sum_{i=1}^n \hat{a}_i(t_1) \hat{a}_i(t_2)$ where $\hat{a}_i(t)$ is obtained by replacing all the unknown parameters in $a_i(t)$ with their respective empirical estimators. The consistency of the variance estimator can be proved using arguments similar to, for example, the proof of Theorem 3 of Lin et al. (1998).

4.6.2 Proof of Theorem 4.3.1

Assumptions (A1') and (A2) are the regularity conditions for Theorem 4.2.1. We first consider the large sample properties for $\hat{\Phi}_j(t), j = 1, \ldots, J$. By straightforward algebra, we derive

$$\begin{aligned}
\hat{\Phi}_j(t) &= \int_0^{\infty} m\hat{H}_j(t, dm, \tau) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i I(\Pi_i = j, D_i \leq \tau) M_i^{total}(t)}{\hat{S}_C(D_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_0^{\tau} \frac{I(\Pi_i = j) M_i^{total}(t)}{\hat{S}_C(u)} d\mathcal{N}_i^D(u) \\
&= \int_0^{\tau} \frac{\hat{G}_j(du, t)}{\hat{S}_C(u)},
\end{aligned}$$

where $\mathcal{N}^D(t) = I(D \leq t, \Delta = 1)$, and $\hat{G}_j(u, t) = n^{-1}\sum_{i=1}^{n} I(\Pi_i = j)M_i^{total}(t)\mathcal{N}_i^D(u)$ is an estimator of $G(u, t) = E\{I(\Pi = j)M^{total}(t)\mathcal{N}^D(u)\}$. Moreover, we have

$$\begin{aligned}
\int_0^{\tau} \frac{G_j(du, t)}{S_C(u)} &= \int_0^{\tau} \frac{E\{I(\Pi = j)M^{total}(t) d\mathcal{N}^D(u)\}}{S_C(u)} \\
&= E[I(\Pi = j)M^{total}(t)I(D \leq \tau)E\left\{\Delta S_C(D)^{-1} \mid D, M^{total}(t), \Pi\right\}] \\
&= E\{I(\Pi = j)M^{total}(t)I(D \leq \tau)\} \\
&= \Phi_j(t).
\end{aligned}$$

Note that $\hat{\Phi}_j(t)$ can be viewed as a functional $(F_3(u), F_4(u, t)) \mapsto \int_0^{\tau} e^{F_3(u)} F_4(du, t)$ from the domain $D[0, \tau] \times BV_M[0, \tau]^2$ to $D[0, \tau]$, where $BV_M[0, \tau]^2$ means the set of cadlag functions $F_4 : [0, \tau] \times [0, \tau] \mapsto [0, M]$ with $\sup_{t \in [0, \tau]} \int_0^{\tau} |F_4(du, t)| < M$. Since $M^{total}(t)$ is increasing with t and $\mathcal{N}_D(u)$ is increasing with u, the functional class $\{I(\Pi = j)M^{total}(t)\mathcal{N}^D(u),\ t, u \in [0, \tau]\}$ is Donsker by Lemma 4.1 and Corollary 9.32 in Kosorok (2007). Thus, the stochastic process $\sqrt{n}\{\hat{G}_j(u,t) - G_j(u,t)\}$ converges weakly to a Gaussian process. Applying techniques similar to Lemma 20.10 in van der Vaart (2000), the functional $(F_3(u), F_4(u,t)) \mapsto \int_0^\tau e^{F_3(u)}F_4(du,t)$ is Hadamard-differentiable and

$$\begin{aligned}
&\sqrt{n}\{\hat{\Phi}_j(t) - \Phi_j(t)\} \\
&= \sqrt{n}\int_0^\tau e^{\Lambda^C(u)}\{\hat{G}_j(du,t) - G_j(du,t)\} + \sqrt{n}\int_0^\tau e^{\Lambda^C(u)}\{\hat{\Lambda}_C(u) - \Lambda_C(u)\}G_j(du,t) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n b_{ji}(t) + o_p(1)
\end{aligned}$$

where

$$\begin{aligned}
b_{ji}(t) = M_i^{total}(t)I(\Pi_i = j)\int_0^\tau \frac{d\mathcal{N}_i^D(u)}{S_C(u)} - \Phi_j(t) + \Phi_j(t)\int_0^\tau \frac{d\mathcal{M}_i^C(u)}{S_X(u)} \\
- \int_0^\tau E\{I(\Pi = j)M^{total}(t)I(D \le u)\}\frac{d\mathcal{M}_i^C(u)}{S_X(u)}\ .
\end{aligned}$$

For $\hat{\Phi}_{J+1}(t)$, we have

$$\begin{aligned}
\sqrt{n}\{\hat{\Phi}_{J+1}(t) - \Phi_{J+1}(t)\} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\frac{I(X_i \ge \tau)M_i^{total}(t)}{S_C(\tau)} - \frac{E\{I(X \ge \tau)M^{total}(t)\}}{S_C(\tau)}\right] + \\
&E\{I(X \ge \tau)M^{total}(t)\}\sqrt{n}\left[\exp\{\hat{\Lambda}_C(t)\} - \exp\{\Lambda_C(t)\}\right] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\frac{I(X_i \ge \tau)M_i^{total}(t)}{S_C(\tau)} - \Phi_{J+1}(t) + \Phi_{J+1}(t)\int_0^\tau \frac{d\mathcal{M}_i^C(u)}{S_X(u)}\right\} + \\
&o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n b_{J+1,i}(t) + o_p(1)\ .
\end{aligned}$$

For $j = 1, \ldots, J+1$, note that the stochastic process $b_j(t)$ $(0 < t \leq \tau)$ has zero mean and can be written as the sum of monotone cadlag processes and is therefore Donsker (Lemma 4.1, Kosorok (2007)). We then have $n^{-1/2}\sum_{i=1}^n b_{ji}(t)$ $(0 < t \leq \tau)$ converges weakly to a tight and zero-mean Gaussian process. Again, the variance-covariance function can be consistently estimated by $n^{-1}\sum_{i=1}^n \hat{b}_{ji}(t_1)\hat{b}_{ji}(t_2)$, where $\hat{b}_{ji}(t)$ is obtained by replacing all the unknown parameters in $b_{ji}(t)$ with their respective empirical estimators. Specifically, for $j = 1, \ldots, J$, the estimator for $E\{I(\Pi = j)M(t)I(D \leq u)\}$ can be constructed as $\int_0^u \frac{\hat{G}_j(dv,t)}{\hat{S}_C(v)}$.

4.6.3 Proof of Theorem 4.4.1

Assumptions (A1'), (A2) and (A3) are the regularity conditions for Theorem 4.4.1. We first derive the optimal weight function. Consider the Lagrange function defined by

$$V(\boldsymbol{w}_{jt}, k) = \boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t,t)\boldsymbol{w}_{jt} + k(\boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{e} - 1).$$

Taking derivative of $V(\boldsymbol{w}_{jt})$ with respective to $\boldsymbol{w}_{jt}$, we have

$$\left\{\begin{array}{l} \frac{\partial V(\boldsymbol{w}_{jt},k)}{\partial \boldsymbol{w}_{jt}} = 2\boldsymbol{\Sigma}_j(t,t)\boldsymbol{w}_{jt} + k\boldsymbol{e} = 0 \\ \frac{\partial V(\boldsymbol{w}_{jt},k)}{k} = \boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{e} - 1 = 0 \end{array}\right.$$

And solving the equations gives us

$$\left\{\begin{array}{l} k = -\frac{2}{\boldsymbol{e}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t,t)^{-1}\boldsymbol{e}} \\ \boldsymbol{w}_{jt} = \frac{\Sigma_j(t,t)^{-1}\boldsymbol{e}}{\boldsymbol{e}^{\mathsf{T}}\Sigma_j(t,t)^{-1}\boldsymbol{e}} \end{array}\right.$$

Together with the inequality, we know that $\boldsymbol{w}_{jt}^* \equiv (w_{jt}^*, 1 - w_{jt}^*)^{\mathsf{T}} = \frac{\boldsymbol{\Sigma}_j(t,t)^{-1}\boldsymbol{e}}{\boldsymbol{e}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t,t)^{-1}\boldsymbol{e}}$ is the weight that minimize $\boldsymbol{w}_{jt}^{\mathsf{T}}\boldsymbol{\Sigma}_j(t,t)\boldsymbol{w}_{jt}$. We then use $\widehat{\boldsymbol{\Sigma}}_j(t_1,t_2)$, that is,

$$\frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \hat{b}_{ji}(t_1)\hat{b}_{ji}(t_1) & \sum_{i=1}^n \hat{b}_{ji}(t_1)\{\hat{a}_i(t_2) - \sum_{k\neq j}\hat{b}_{ki}(t_2)\} \\ \sum_{i=1}^n\{\hat{a}_i(t_1) - \sum_{k\neq j}\hat{b}_{ki}(t_1)\}\hat{b}_{ji}(t_2) & \sum_{i=1}^n\{\hat{a}_i(t_1) - \sum_{k\neq j}\hat{b}_{ki}(t_1)\}\{\hat{a}_i(t_2) - \sum_{k\neq j}\hat{b}_{ki}(t_2)\} \end{pmatrix},$$

to estimate $\boldsymbol{\Sigma}_j(t_1, t_2)$, and use $\hat{\boldsymbol{w}}^*_{jt} = \frac{\widehat{\boldsymbol{\Sigma}}_j(t,t)^{-1}\boldsymbol{e}}{\boldsymbol{e}^\intercal \widehat{\boldsymbol{\Sigma}}_j(t,t)^{-1}\boldsymbol{e}}$ to estimate the optimal weight $\boldsymbol{w}^*_{jt}$.

To obtain the i.i.d. representation of $\hat{\Phi}_j^{imp}(t)$, we have

$$\begin{aligned}\sqrt{n}\{\hat{\Phi}_j^{imp}(t) - \Phi_j(t)\} &= \sqrt{n}\hat{w}^*_{jt}\{\hat{\Phi}_j(t) - \Phi_j(t)\} + \sqrt{n}(1 - \hat{w}^*_{jt})\{\tilde{\Phi}_j(t) - \Phi_j(t)\} \\ &= n^{-1/2}\sum_{i=1}^{n}[w^*_{jt}b_{ji}(t) + (1 - w^*_{jt})\{a_i(t) - \sum_{k\neq j} b_{ki}(t)\}] + o_p(1) \\ &\equiv n^{-1/2}\sum_{i=1}^{n} f_{ji}(t) + o_p(1).\end{aligned}$$

It's easy to see that $\sqrt{n}\{\hat{\Phi}_j^{imp}(t) - \Phi_j(t)\}$ $(0 < t \leq \tau)$ converges weakly to a tight and zero-mean Gaussian process. The covariance at (t_1, t_2) can be consistently estimated by $n^{-1}\sum_{i=1}^{n} \hat{f}_{ji}(t_1)\hat{f}_{ji}(t_2)$, where $\hat{f}_{ji}(t) = \hat{w}^*_{jt}\hat{b}_{ji}(t) + (1 - \hat{w}^*_{jt})\{\hat{a}_i(t) - \sum_{k\neq j} \hat{b}_{ki}(t)\}$.

4.6.4 Proof of Corollary 4.4.2

We assume $H_j(0, \infty, \tau) > 0$. First, note that $\sqrt{n}\{\hat{H}_j(0, \infty, \tau) - H_j(0, \infty, \tau)\} = n^{-1/2}\sum_{i=1}^{n} d_{ji}(\tau) + o_p(1)$, where

$$\begin{aligned}d_{ji}(\tau) = &\int_0^\tau \frac{I(\Pi_i = j)d\mathcal{N}_i^D(u)}{S_C(u)} - H_j(0, \infty, \tau) + H_j(0, \infty, \tau)\int_0^\tau \frac{d\mathcal{M}_i^C(u)}{S_X(u)} - \\ &\int_0^\tau E\{I(\Pi = j, D \leq u)\}\frac{d\mathcal{M}_i^C(u)}{S_X(u)}.\end{aligned}$$

For $j = 1, \dots, J$, we have

$$
\begin{aligned}
&\sqrt{n}\{\hat{\Phi}_j^c(t) - \Phi_j^c(t)\} \\
&= \frac{1}{H_j(0,\infty,\tau)}\sqrt{n}\{\hat{\Phi}_j^{imp}(t) - \Phi_j^{imp}(t)\} - \frac{\Phi_j(t)}{H_j(0,\infty,\tau)^2}\{\hat{H}_j(0,\infty,\tau) - H_j(0,\infty,\tau)\} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{f_{ji}(t)}{H_j(0,\infty,\tau)} - \frac{\Phi_j(t)d_{ji}(\tau)}{H_j(0,\infty,\tau)^2}\right\} + o_p(1) \\
&\overset{def}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{ji}(t) + o_p(1).
\end{aligned}
$$

Also, note that $g_j(t)$ $(0 < t \leq \tau)$ can be written as sum of monotone cadlag processes and is Donsker, thus $\sqrt{n}\{\hat{\Phi}_j^c(t) - \Phi_j^c(t)\}$ $(0 < t \leq \tau)$ converges weakly to a tight and zero-mean Gaussian process, whose covariance function can be consistently estimated by $n^{-1}\sum_{i=1}^{n}\hat{g}_{ji}(t_1)\hat{g}_{ji}(t_2)$, where we define

$$
\hat{g}_{ji}(t) = \frac{\hat{f}_{ji}(t)}{\hat{H}_j(0,\infty,\tau)} - \frac{\hat{\Phi}_j^{imp}(t)\hat{d}_{ji}(\tau)}{\hat{H}_j(0,\infty,\tau)^2}.
$$

Chapter 5

Analysis of Quality-of-life outcomes and Medical Costs Data: Application to AIDS and Cancer Studies

5.1 Analysis of Quality of life and Survival: CPCRA ddI/ddC Trial

We illustrate the proposed methods by analyzing data from a clinical trial conducted by Terry Beirn Community Programs for Clinical Research on AIDS, a federally funded national network of community-based research groups. The study compared didanosine (ddI) and zalcitabine (ddC) as treatments for HIV-infected patients who were intolerant or had failed treatment with zidovudine. The trial randomized 230 patients to receive ddI treatment and 237 to receive ddC. The primary endpoint is time to disease progression or death. The secondary endpoints include changes in the Karnofsky performance score and opportunistic infections, where a reduction in the Karnofsky score and the occurrence of opportunistic disease indicate a deterioration in health. Both survival time and quality of life are regarded as important indexes for treatment success. The analysis in Abrams et al. (1994) suggested that ddC treatment may have provided a survival advantage over ddI treatment, with borderline significance based on a proportional hazards model. We investigated the treatment effects on the cumulative weighted marker process for a more comprehensive assessment of the benefits and risks of the treatments. In our analysis, death is the terminal event of interest, and Karnofsky score and incidence of opportunistic infections are used as measures for quality of life. The analysis with Karnofsky score illustrates the proposed methodology in the case whre the longitudinal marker is intermitently observed, while the analysis with incidence of opportunistic infections illustrates the situation where the longitudinal marker is completely observed throughout the follow-up period.

In the first set of analysis, we divide the Karnofsky score by 100 and transform it to a 0 to 1 scale and set $w(\cdot) = 1$. The results are summarized in Table 5.2. The mean of the cumulative weighted marker process at 500th day (≈ 1.37 year) is 0.876 for the ddI group and is 0.907 for the ddC group. The two-sided p-value deriived from the proposed two-sample test is 0.38. Our analysis suggests that ddC performs slightly better than ddI in terms of the proposed summary measure, though the difference is not statistically significant. Figure 5.1 displays the estimated cumulative mean function $\mu(t)$, survival function and mean Karnofsky score of survivors in the ddI and ddC treatment groups. The plots show that ddC performs better in terms of survival but worse in terms of survivors' physical quality of life, and the estimated summary measures for the two treatments are very close.

In the second set of analysis, we consider benefit-risk assessment based on

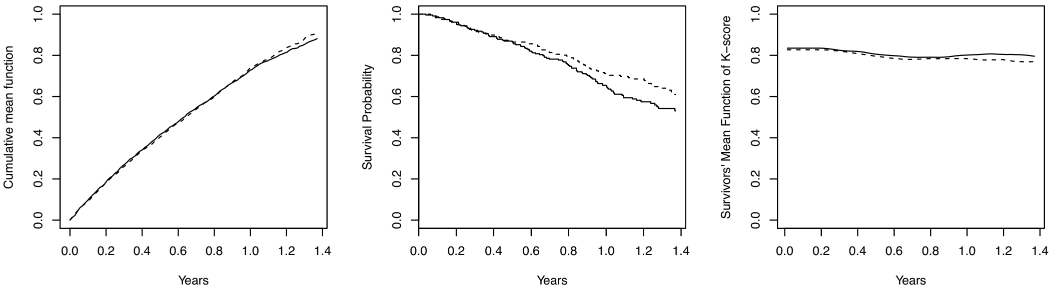


Figure 5.1: Estimated cumulative mean functions $\mu(t)$ using Karnofsky score and time to death (left), survival functions (middle) and mean Karnofsky score of survivors $E\{Y(t) \mid D \geq t\}$ (right) for ddI (solid line) and ddC (dashed line) treatment groups.

opportunistic infections and death. A total of 363 confirmed or probable opportunistic diseases indicating disease progression (Neaton et al., 1994) were reported. The number of opportunistic infections per subject ranges from 0 to 5, with median 1 and mean 0.78. Denote by $O(u)$ the total number opportunistic infections occurred at or before time u, and set $Y(\cdot) = 0.8^{O(\cdot)}$. Then the occurrence of opportunistic infection at time t discounts a patient's score $Y(t)$ by 0.8. Then the estimated summary measure is 0.998 for the ddI group and 1.028 for the ddC group. The p-value derived from the proposed two-sample test is 0.27. Our analysis again suggests that ddC outperforms ddI in terms of the proposed summary measure on survival and opportunistic disease, although the advantage is not statistically significant.

Table 5.1: Analysis of ddI/ddC trial of CPCRA

	ddI		ddC		
Marker	Estimate	95% CI	Estimate	95% CI	p-Value
Karnofsky score	0.876	(0.828, 0.925)	0.907	(0.858, 0.955)	0.38
OI	0.998	(0.932, 1.063)	1.028	(0.959, 1.097)	0.27

Note: Estimate is the estimated $\mu(\tau)$ ($\tau \approx 1.37$year) , 95% CI is the 95% confidence interval based on standard error estimate. OI stands for opportunistic infection.

5.2 Analysis of Censored Medical Cost Data: SEER-Medicare Linked Database

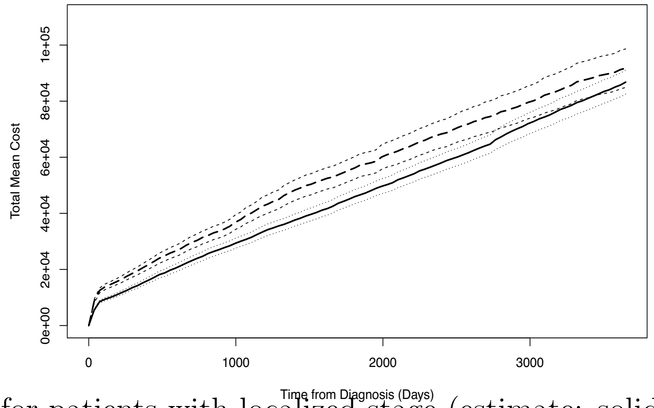
The proposed methods are applied to SEER–Medicare linked database; see Warren et al. (2002) for an overview of the data. For illustration, we assess the medical cost of breast cancer patients diagnosed at age 65+ in 1994 among Medicare enrollees. We take the time of first diagnosis of breast cancer to be the time origin, and D is the time from first diagnosis of breast cancer to death. $N(\cdot)$ is the counting process that characterizes the number of inpatient or outpatient cares, and upon the occurrence of inpatient or outpatient cares, $Y(t)$ is the cost charged for medical treatment. As a well known fact, cardiovascular disease competes with breast cancer as the leading cause of death for older females diagnosed with breast cancer (Patnaik et al., 2011). In particular, we are interested in three competing terminal events within ten years since diagnosis of breast cancer: (i) breast cancer mortality ($\Pi = 1$), (ii) death from a cardiovascular disease ($\Pi = 2$), (iii) mortality from other causes ($\Pi = 3$). The subjects are divided into two groups by the historic stage determined at diagnosis: 6156 subjects with localized stage and 2540 subjects with regional stages. In the following analysis, the cost accumulation process from first diagnosis to $\tau = 10$ years is of interest, and people who survives more than 10 years are classified to the fourth category (long-term survivors).

We begin with estimating the cumulative incidence in standard competing risks model. For patients diagnosed with localized stage, the 10-year cumulative incidence was 8.5% (SE : 0.5%) for breast cancer (BC) deaths, 12.4% (SE : 0.5%) for cardiovascular disease (CVD) death, and 25.7% (SE : 0.7%) for other cause mortality. For patients diagnosed with regional stage, the 10-year cumulative incidence was 30.8% (SE : 1.2%) for breast cancer deaths, 13.4% (SE : 0.9%) for cardiovascular disease death, and 24.6% (SE : 1.1%) for other cause mortality. Thus at the end of the tenth year, the regional stage group has larger proportion of patients with breast cancer death than the localized stage group, and the two groups has similar proportions of cardiovascular disease death and death due to other causes.

We then analyze the medical cost up to a time horizon $\tau = 10$ years with our proposed methods. The results are presented in Table 2. When we do not distinguish the three types of death, the estimator in Section 3.1 is employed to estimate the total medical cost. It can be seen that the regional stage group has higher average ten-year medical cost than the localized stage group. We further take into account the three competing risks, $\Pi = 1, 2, 3$, and the estimates are obtained by using the improved estimators in Chapter 4.4. For each of the localized and regional stage groups, the average cost of patients with CVD death ($\Phi_2^c(\tau)$) differs slightly from the average cost of patients with other mortality ($\Phi_3^c(\tau)$), but is much higher than the average cost of patients with BC mortality ($\Phi_1^c(\tau)$). In contrast, for the overall spending of medical cost, $\Phi_j(\tau)$, patients with other mortality spent the most when compared with BC and CVD mortality, which is largely explained by the large proportion of patients of with other cause of death at the tenth year.

For a better plot presentation, we consider the average medical cost over time from first diagnosis of breast cancer to the tenth year after diagnosis. The medical costs over time of localized and regional stage group are presented in Figure 1, and regional stage group has consistently higher medical cost over time. The estimated medical costs over time for competing risk types $\Pi = 1$ and 2 are shown in Figure 2. For the overall spending of medical cost $(\Phi_j(t), j = 1, 2)$, it can be seen that the expected cost attributed to BC mortality of regional stage patients is much higher than that of localized stage patients, which is mainly due to the higher 10-year cumulative incidence of BC mortality of regional stage patients. The average costs of patients with each cause of death $(\Phi_j^c(t), j = 1, 2)$ are similar between localized and regional group, though the cost over time for CVD mortality patients is consistently higher than the cost for BC mortality patients.

Figure 5.2: Estimated total mean cost over time since first diagnosis of breast cancer and point-wise 95% confidence intervals



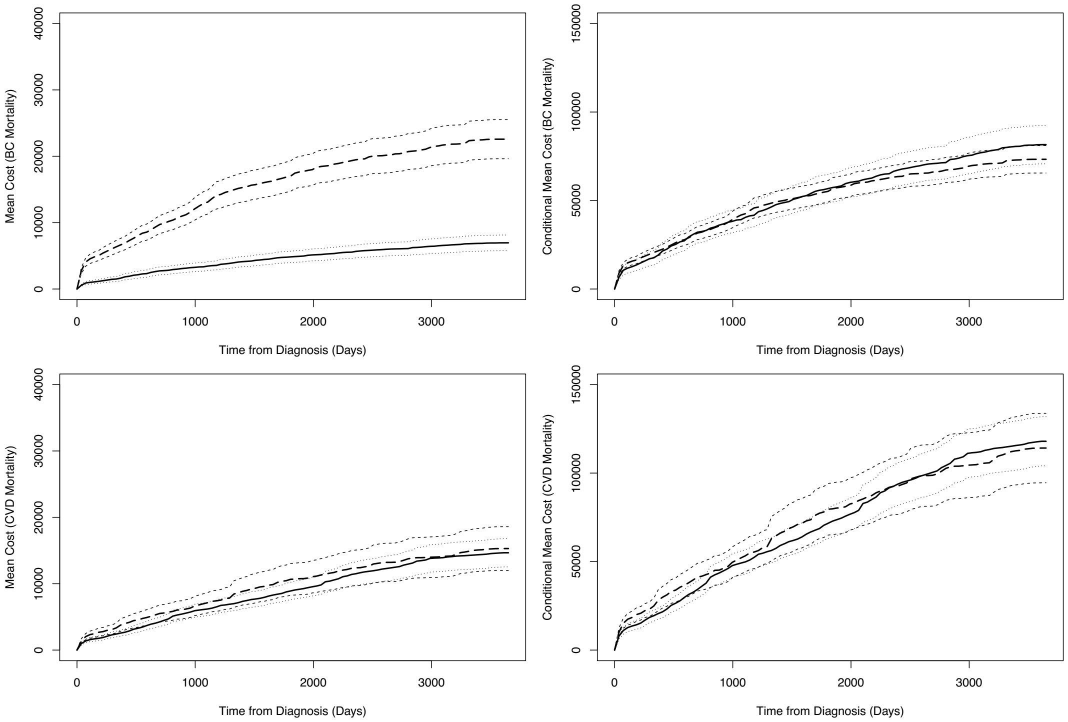
NOTE: Costs for patients with localized stage (estimate: solid line; CI: dotted line) and regional stage(estimate: long-dash line; CI: dashed line).

Table 5.2: Analysis of SEER-Medicare Data

		Localized		Regional	
		Estimate	SE	Estimate	SE
Total	$\Phi(\tau)$	86794.9	2173.1	91761.8	3504.0
BC mortality	$\Phi_1(\tau)$	6958.5	603.33	22573.6	1502.6
	$\Phi_1^c(\tau)$	81580.1	5540.9	73279.5	3952.3
CVD mortality	$\Phi_2(\tau)$	14661.5	1088.7	15296.6	1681.2
	$\Phi_2^c(\tau)$	117963.1	7082.9	114158.2	10012.0
Other mortality	$\Phi_3(\tau)$	31374.9	1652.7	29170.1	2562.0
	$\Phi_3^c(\tau)$	122222.9	5452.9	118937.0	8907.5
Long-term survivors	$\Phi_4(\tau)$	63330.0	2343.5	24682.7	2198.8
	$\Phi_4^c(\tau)$	45911.9	1364.7	79116.7	6342.7

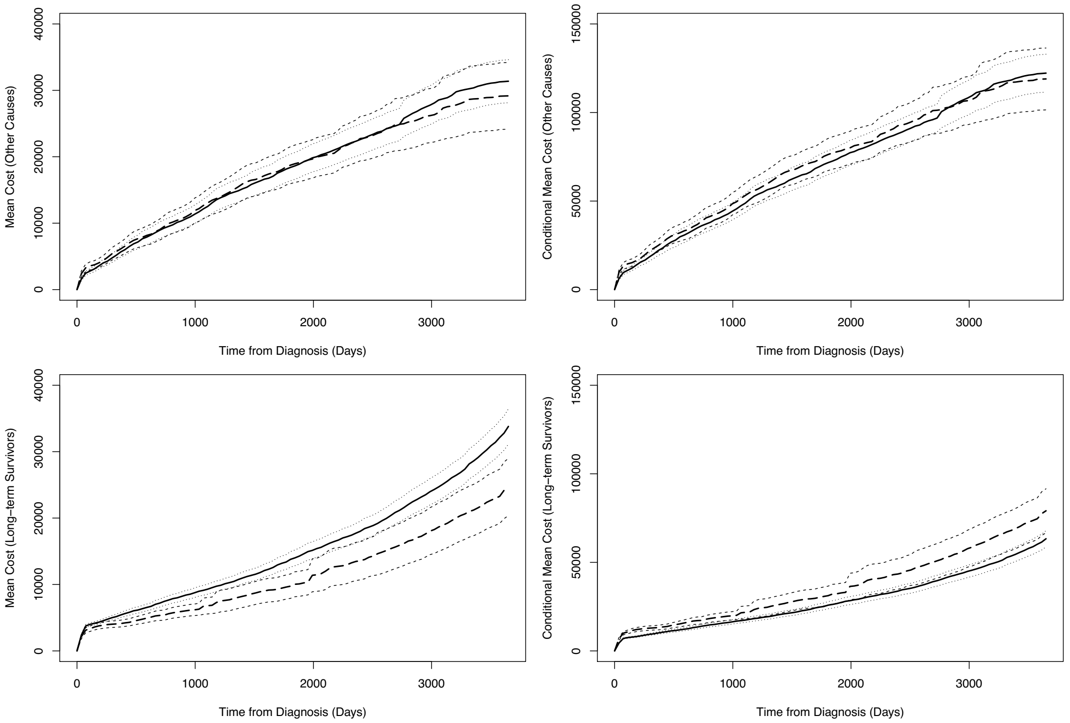
NOTE: Estimate is the cost in US Dollar at $\tau = 10$ year, SE is the standard error estimate. BC is for breast cancer and CVD is for cardiovascular disease.

Figure 5.3: Cost analysis for breast cancer mortality and cardiovascular disease mortality



NOTE: The left panels are estimated mean cost since first diagnosis of breast cancer for BC mortality $\Phi_1(t)$ (Upper) and CVD mortality $\Phi_2(t)$ (Lower) with point-wise 95% confidence intervals (CI). The right panels are estimated conditional mean cost of BC mortality $\Phi_1^c(t)$ (Upper) and CVD mortality $\Phi_2^c(t)$ (Lower). The solid lines are estimates for patients with localized stage (CI: dotted line) and the dashed line are estimates for patients with regional stage (CI: dashed line).

Figure 5.4: Cost analysis for other causes mortality and long-term survivors



NOTE: The left panels are estimated mean cost since first diagnosis of breast cancer for other-cause mortality $\Phi_3(t)$ (Upper) and long-term survivors $\Phi_4(t)$ (Lower) with point-wise 95% confidence intervals (CI). The right panels are estimated conditional mean cost of other-cause mortality $\Phi_3^c(t)$ (Upper) and long-term survivors $\Phi_4^c(t)$ (Lower). The solid lines are estimates for patients with localized stage (CI: dotted line) and the dashed line are estimates for patients with regional stage (CI: dashed line).

Chapter 6

Discussion

In this dissertation, we first consider benefit-risk assessment based on longitudinal marker measurements and time to event data. The proposed method is especially useful when conflict results about the treatment effects are reported for the two outcomes. Our estimation and testing procedures are more robust than the existing methods, such as Hwang et al. (1996), in the sense that the statistical procedures can be derived from one single data set. Statistical inference properties are established for point estimate and hypothesis testing, hence the proposed methodology is expected to be attractive for practitioners to facilitate accurate decision-making.

The proposed methodologies have a wide range of applicability in biomedical and publich health research. Besides the examples discussed in Section 3.1, the longitudinal measure $Y(\cdot)$ can also be the value of an surrogate biomarker for the survival outcome of interest; for example, CD4 cell count has been used as a surrogate for progression to AIDS or death in many AIDS studies. In the case where the follow-up duration is not long enough to accumulate adequate number of events for meaningful analysis, the clinical study may have insufficient power to detect treatment effects on the survival outcome. Compared with the conventional survival analysis, the proposed methods utilize additional information from the surrogate marker and possess the potential to increase power in detecting real treatment effects. Finally, instead of using a single marker process, a benefit-risk summary measure integrating multiple marker processes and time to event is under investigation.

In this work, we have focused on the one- and two- sample problems, and the proposed summary measure is estimated using kernel smoothing techniques. It would be interesting to consider extending the methodology to a regression setting.

For example, we may consider the following frailty model. Let V represents the covariates, then the summary measure adjusted for covariates can be defined as $\mu(\tau \mid V) = \int_0^\tau E\{Y(u)I(D \geq u) \mid V\}du$. We can estimate $\mu(\tau \mid V)$ based on a joint model of longitudinal and survival data. For example, we assume $E\{Y(t) \mid V, Z\} = g(t) + V\beta + Z$ and hazard function $h(t \mid V, Z) = Zh_0(t)e^{V\gamma}$, where the frailty random variable Z is independent of V and has a gamma distribution with unit mean and variance $1/\alpha$. When $\alpha \to \infty$, the correlation of $Y(\cdot)$ and D goes to 0. It can be further shown that $E\{Y(t)I(D \geq t) \mid V\} = \alpha^{\alpha+1}\{\alpha + H_0(t)e^{V\gamma}\}^{-\alpha-1} + \alpha^{\alpha}\{\alpha + H_0(t)e^{V\gamma}\}^{-\alpha}\{V\beta + g(t)\}$. Suppose V is the treatment indicator, coded 0 if control and coded 1 if treated, then $\mu(\tau \mid V = 1)$ and $\mu(\tau \mid V = 0)$ are deterministic functions of (β, γ). If $\beta > 0$ and $\gamma < 0$, both longitudinal and survival components for the treatment group would be better. If $\beta < 0$ and $\gamma < 0$, survival outcome for the treatment group is better but longitudinal outcome is worse, then it would be difficult to make decision based on the two separate components. Our proposed summary measure offers a way to summarize joint modeling results for a conclusive benefit-risk assessment, that is, we can compare $\mu(\tau \mid V = 1)$ and $\mu(\tau \mid V = 0)$ for decision-making.

The work can also be extend to the situation where the terminal event time is subject to left truncation. One can modify the risk-set indicator to get estimation under left truncation.

This dissertation also proposed nonparametric estimators of the mean recurrent marker process, with specific focus on competing risks model. In Section 4.2, we considered a nonparametric estimation approach which uses marker history information from both censored and uncensored subjects, but the estimator cannot be generalized to handle problems involving competing risks because the risk type information is unknown for those censored subjects. A consistent and asymptotically normally distributed estimator of type-j mean function is then constructed in Section 4.3 under competing risks model, where the proposed estimator only uses risk type information from uncensored subjects. Furthermore, using auxiliary information from the estimate of the non-competing risks mean function, an optimal estimator among a class of weighted estimators is proposed in Section 4.4 to improve the estimation efficiency over the estimator proposed in Section 4.3.

In this article, we mainly considered one-sample estimation, the authors are considering to extend the non-parametric estimation to regression setting. For example, we may consider the following marginal models. We assume the survival time in the target population follows the Cox proportional hazards model

$$h(t \mid V) = h_0(t) \exp(\alpha V),$$

where $\lambda(t)$ is an unspecified baseline hazard function and α is the coefficient. For the marked recurrent event process, we define the conditional rate function given $D = t$ and $V = v$, that is, $\eta(u, t \mid v)du = E\{Y(u)dN(u) \mid D = t, V = v\}$. Then we consider the following proportional rate model

$$\eta(u, t \mid v) = \eta_0(u, t) \exp(\beta v),\ 0 \le u \le t \tag{6.0.1}$$

where $h(u, t)$ is an unspecified baseline rate function at time u given $D = t$. Note that $h(u, t \mid v)$ is the rate of accumulation of cost or utility at u given that survival time is t and covariate is v. In many studies, survival time is typically the primary endpoint whereas lifetime medical cost is a secondary outcome, thus it is meaningful to compare the cost/utility accumulation process among subjects with the same survival time. Moreover, medical costs usually increase during the time period prior to death because of the intensive care for dying patients, and the patterns of medical costs is often closely linked to the death time (Liu et al., 2007). Therefore, it is natural to model cost trajectory conditional on terminal event time. Define $H(t \mid v) = \int_0^t \eta(u, t \mid v)du = E\{\int_0^D Y(u)dN(u) \mid D = t, V = v\}$, which is the expected lifetime cost or utility given survival time t and covariate v. Note that (6.0.1) implies

$$H(t \mid v) = H_0(t) \exp(\beta v), \tag{6.0.2}$$

where $H_0(t) = \int_0^t \eta_0(u, t)du$ is the baseline lifetime cost/utility given survival time is t. Compared to equation (6.0.1), equation (6.0.2) makes stronger assumption of the stochastic process $Y(\cdot)dN(\cdot)$ before the terminal event. Estimating equations can be constructed using similar methods in Chan (2009).

In conclusion, the nonparametric methods proposed in this dissertation for marker processes with a terminal event may initiate a variety of future works on both statistical methods and applications, which could facilitate a comprehensive understanding of the marker process and survival time.

Bibliography

Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., et al. (1994), "A Comparative Trial of Didanosine or Zalcitabine after Treatment with Zidovudine in Patients With Human Immunodeficiency Virus Infection," *New England Journal of Medicine*, 330, 657–662.

Bang, H. and Tsiatis, A. A. (2000), "Estimating Medical Costs With Censored Data," *Biometrika*, 87, 329–343.

Chan, K. C. G. (2009), *Recurrent Marker Process Before Failure Event: A Backward Process Approach*, ProQuest.

Fine, J. P. and Gray, R. J. (1999), "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, 94, 496–509.

Gelber, R., Gelman, R., and Goldhirsch, A. (1989), "A Quality-of-life-oriented Endpoint for Comparing Therapies," *Biometrics*, 45(3), 781–795.

Ghosh, D. and Lin, D. Y. (2000), "Nonparametric Analysis of Recurrent Events and Death," *Biometrics*, 56, 554–562.

Glasziou, P. P., Cole, B. F., Gelber, R. D., Hilden, J., and Simes, R. J. (1998), "Quality Adjusted Survival Analysis with Repeated Quality of Life Measures," *Statistics in Medicine*, 17, 1215–1229.

Glasziou, P. P., Simes, R. J., and Gelber, R. D. (1990), "Quality Adjusted Survival Analysis," *Statistics in Medicine*, 9, 1259–1276.

Gray, R. J. (1988), "A Class of K-sample Tests for Comparing the Cumulative Incidence of a Competing Risk," *The Annals of Statistics*, 1141–1154.

Härdle, W., Sperlich, S., Werwatz, A., and Müller, M. (2004), *Nonparametric and Semiparametric Models*, New York: Springer–Verlag.

Henderson, R., Diggle, P., and Dobson, A. (2000), "Joint Modelling of Longitudinal Measurements and Event Time Data," *Biostatistics*, 1, 465–480.

Hogan, J. W. and Laird, N. M. (1997), "Mixture Models for the Joint Distribution of Repeated Measures and Event Times," *Statistics in Medicine*, 16, 239–257.

Huang, Y. (2002), "Calibration Regression of Censored Lifetime Medical Cost," *Journal of the American Statistical Association*, 97, 318–327.

Huang, Y. and Louis, T. A. (1998), "Nonparametric Estimation of the Joint Distribution of Survival Time and Mark Variables," *Biometrika*, 85, 785–798.

— (1999), "Expressing Estimators of Expected Quality Adjusted Survival as Functions of Nelson-Aalen Estimators," *Lifetime Data Analysis*, 5, 199–212.

Hwang, J. S., Tsauo, J. Y., and Wang, J. D. (1996), "Estimation of Expected Quality Adjusted Survival by Cross-sectional Survey," *Statistics in Medicine*, 15, 93–102.

Irwin, J. (1949), "The standard error of an estimate of expectational life," *Journal of Hygiene*, 47, 188–189.

Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, New York: Willey, 2nd ed.

Kosorok, M. R. (2007), *Introduction to Empirical Processes and Semiparametric Inference*, Springer Science & Business Media.

Lin, D., Feuer, E. J., Etzioni, R., and Wax, Y. (1997), "Estimating Medical Costs from Incomplete Follow-up Data," *Biometrics*, 419–434.

Lin, D. Y., Wei, L. J., and Ying, Z. (1998), "Accelerated Failure Time Models for Counting Processes," *Biometrika*, 85, 605–618.

Liu, L., Wolfe, R. A., and Kalbfleisch, J. D. (2007), "A Shared Random Effects Model for Censored Medical Costs and Mortality," *Statistics in medicine*, 26, 139–155.

Mammen, E. and Nielsen, J. P. (2007), "A General Approach to the Predictability Issue in Survival Analysis with Applications," *Biometrika*, 94, 873–892.

Meinert, C. L. (2012), *Clinical Trials Dictionary: Terminology and Usage Recommendations*, Hoboken, NJ: John Wiley & Sons, Inc.

Murray, S. and Cole, B. (2000), "Variance and Sample Size Calculations in Quality-of-Life-Adjusted Survival Analysis (Q-TWiST)," *Biometrics*, 56, 173–182.

Neaton, J. D., Wentworth, D. N., Rhame, F., Hogan, C., Abrams, D. I., and Deyton, L. (1994), "Considerations in Choice of a Clinical Endpoint for AIDS Clinical Trials," *Statistics in Medicine*, 13, 2107–2125.

Patnaik, J. L., Byers, T., DiGuiseppi, C., Dabelea, D., and Denberg, T. D. (2011), "Cardiovascular Disease Competes with Breast Cancer as the Leading Cause of Death for Older Females Diagnosed with Breast Cancer: A Retrospective Cohort Study," *Breast Cancer Research*, 13, R64.

Pepe, M. S. and Fleming, T. R. (1989), "Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data," *Biometrics*, 45, 497–507.

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, Hayward, CA: Inst. Math. Statist.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr., A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978), "The Analysis of Failure Times in the Presence of Competing Risks," *Biometrics*, 34, 541–554.

Shen, L. Z., Pulkstenis, E., and Hoseyni, M. (1999), "Estimation of Mean Quality Adjusted Survival Time," *Statistics in Medicine*, 18, 1541–1554.

Strawderman, R. L. (2000), "Estimating the Mean of an Increasing Stochastic Process at a Censored Stopping Time," *Journal of the American Statistical Association*, 95, 1192–1208.

Sun, Y. Q., Gilbert, P. B., and McKeague, I. W. (2009), "Proportional Hazards Models with Continuous Marks," *Annals of Statistics*, 37, 394.

Sun, Y.-Q. and Wu, H.-L. (2003), "AUC-based Tests for Nonparametric Functions with Longitudinal Data," *Statistica Sinica*, 13, 593–612.

The Beta Blocker Evaluation of Survival Trial Investigators (2001), "A Trial of the Beta-blocker Bucindolol in Patients with Advanced Chronic Heart Failure," *The New England Journal of Medicine*, 344, 1659.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995), "Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS," *Journal of the American Statistical Association*, 90, 27–37.

van der Vaart, A. W. (2000), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

Wang, M.-C. and Chiang, C.-T. (2002), "Non-parametric Methods for Recurrent Event Data with Informative and Non-informative Censorings," *Statistics in Medicine*, 21, 445–456.

Wang, Y. and Taylor, J. M. G. (2001), "Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome," *Journal of the American Statistical Association*, 96, 895–905.

Warren, J. L., Klabunde, C. N., Schrag, D., Bach, P. B., and Riley, G. F. (2002), "Overview of the SEER-Medicare Data: Content, Research Applications, and

Generalizability to the United States Elderly Population," *Medical Care*, 40, IV–3.

Wu, M. C. and Carroll, R. J. (1988), "Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process," *Biometrics*, 44, 175–188.

Xu, J. and Zeger, S. L. (2001), "Joint Analysis of Longitudinal Data Comprising Repeated Measures and Times to Events," *Journal of the Royal Statistical Society: Series C*, 50, 375–387.

Zhao, H. and Tian, L. (2001), "On Estimating Medical Cost and Incremental Cost-Effectiveness Ratios with Censored Data," *Biometrics*, 57, 1002–1008.

Zhao, H. and Tsiatis, A. A. (1999), "Efficient Estimation of the Distribution of Quality-Adjusted Survival Time," *Biometrics*, 55, 1101–1107.

Yifei Sun

615 North Wolfe Street, E3005
Baltimore, MD 21205

410-868-4864
ysun26@jhu.edu

Research Interests

Survival analysis, longitudinal data analysis, recurrent event analysis, nonparametric and semiparametric methods

Education

Johns Hopkins University, Baltimore, MD

Ph.D., Biostatistics, *Expected:* August 2015

- Thesis Topic: *Statistical Methods for Recurrent Marker Process in the Presence of Terminal Events*
- Advisors: Dr. Mei-Cheng Wang and Dr. Chiung-Yu Huang

Zhejiang University, Hangzhou, China

B.S., Statistics, July 2010

Awards and Honors

The Best Paper Award, American Statistical Association Section on Risk Analysis Student/Young Researcher Paper Competition, 2015

The Jane and Steve Dykacz Award, Johns Hopkins University, 2015

- Honors outstanding work in medical statistics

The Glaxo SmithKline Award, Johns Hopkins University, 2011

- Honors outstanding achievement on the first-year comprehensive exam

The First Prize Scholarship and Excellent Thesis Award, Zhejiang University, 2007-2010

Research Experience

Research Assistant September 2014 - present

Supervisor: Dr. Gary Chan, Department of Biostatistics, University of Washington

Research in efficient estimation of accelerated failure time model with length-biased data

Research Assistant September 2012 - August 2014

Supervisors: Dr. Xiaobin Wang and Dr. Xiumei Hong, Department of Population, Family, and Reproductive Health, JHSPH

Research in Genome-wide association study and DNA methylation analysis on child-hood food allergy and preterm birth

Papers and Publications

1. Hong X, Hao K, Ladd-Acosta C, Hansen K D, Tsai H-J, Liu X, Xu X, Thornton T A, Caruso D, Keet C A , **Sun Y**, Wang G, Luo W, Kumar R, Fuleihan R, Singh A M, Kim J S, Story R E, Gupta R S, Gao P, Chen Z, Walker S O, Bartell T R, Beaty T H, Fallin M D, Schleimer R, Holt P G, Nadeau K C, Wood R A, Pongracic J A, Weeks D E and Wang X (2015). "Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children".*Nature Communications* 6. Article number: 6304.

2. **Sun Y**, Huang C-Y and Wang M-C (2015). "Nonparametric Benefit-risk Assessment Using Marker Process in the Presence of a Terminal Event."*In revision at Journal of the American Statistical Association.*
   *** An earlier version won the Best Paper Award, 2015 ASA Section on Risk Analysis Student Paper Competition**

3. **Sun Y** and Wang M-C (2015). "Recurrent Marker Process in the Presence of Competing Terminal Events."*In revision at Journal of the American Statistical Association.*

4. **Sun Y**, Chan G and Qin J (2015). "Fast Over-identified Rank Estimation for Right-censored Length-biased Data." *To be submitted.*

5. Chan G, **Sun Y**, Huang C-Y and Wang M-C (2015). "Semiparametric Joint Modeling of Recurrent Marker Process and a Terminal Event in the Presence of Left Truncation." *In Preparation.*

6. **Sun Y**, Huang C-Y and Qin J (2015). "Missing Information Principle for Left-truncated and Right-censored data with a Known Truncation Time Distribution."*In Preparation.*

7. Marr K A, Tsai H-L, **Sun Y**, Avery R K, Shoham S, Alp S, LaRue R, Ostrander D, Lu N, Jones R, Montgomery R and Huang C-Y. "Infections after Organ and Hematopoietic Stem Cell Transplantation: A Prospective Cohort Study." *To be submitted.*

Professional Service

Review: Life Time Data Analysis

Professional Development

Language Skills: English (Fluent); Chinese (Fluent)
Computer Skills: R, Matlab, Stata, SAS, C/C++