Nonparametric and Semiparametric Methods for Recurrent Gap Time Data

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Chi Hyun Lee

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Xianghua Luo, Ph.D.

August, 2015

Acknowledgements

I owe the greatest debt of gratitude to my advisor, Dr. Xianghua Luo, for her guidance and support on this study. She has spent endless hours reading through my premature drafts and provided stimulating discussions. Without her patience and encouragement this work would have been impossible. Dr. Haitao Chu deserves a special note for being a valuable co-advisor and for giving thought-provoking questions and insightful suggestions. Thanks go to my other committee members, as well- Dr. Jim Hodges for his support throughout and suggestion to approach the problem in a different perspective, Dr. Gongjun Xu for his constructive comments on our model, and Dr. Daniel Weisdorf for his advice on real data analysis and agreement to join the committee late in the game. I am also grateful to Dr. Chiung-Yu Huang at Johns Hopkins University, who have contributed substantially in developing the idea and establishing our methods. Special thanks to Dr. Claudio Brunstein and Todd DeFor for providing assistance with the data. I also appreciate Minnesota Supercomputing Institute for providing computational resources.

Dedication

To God who gave me amazing love and unbelievable comfort & to my parents, Chongho Lee and Hyeisoon Ko, and my sister, Chi In Lee, who were always there for me.

Abstract

In many biomedical studies, recurrent events are frequently encountered where subjects experience an event repeatedly over time such as multiple infections after bone marrow transplant. The gap times between these recurrent events are often the natural outcome of interest. A number of studies have been conducted to describe the gap time distribution and examine the relationship between gap times and covariates. Despite rich literature, existing methods for recurrent gap time data commonly assume that all events are of the same type and thus that the gap times are identically distributed. In various cases, however, it is inappropriate to naively adopt this assumption. In this dissertation, we study two motivating datasets, the post-transplant infection data and the South Verona psychiatric case register (PCR) data, to which existing methods cannot be directly applied. We develop nonparametric and semiparametric methods to properly analyze these recurrent gap time data.

In the post-transplant infection data, enrollment to the study is triggered by transplant, which is a different event than the recurrent infections. Applying existing methods to this data leads to incorrect inferential results because the time from transplant to the first infection and the gap times between successive infections may have different distributions. We propose a nonparametric estimator of the joint distribution of the first event time and the gap times between consecutive infections. Then, a semiparametric regression method is proposed to identify risk factors for infections, accounting for the potentially different distributions of the two types of time.

Often times, recurrent events consist of two alternating states. The South Verona PCR dataset is a typical bivariate alternating recurrent event dataset, in which health care and break periods alternate during follow-up. Existing methods can analyze the gap times between consecutive contacts with health care services, but the break time is ignored, which wastes useful information carried by the duration of each state. We propose a semiparametric regression method for estimating simultaneously the relationship between covariates and the durations of the two alternating states.

Contents

List of Tables

List of Figures

Chapter 1

Background

In many biomedical studies, patients experience a clinical event repeatedly over time during follow-up. This type of data is referred to as recurrent event data. Examples of recurrent event data include repeated infections after bone marrow transplant, tumor recurrences among cancer patients, and readmissions to psychiatric hospitals. Recurrent events imply important information such as patients' underlying health condition, the severity of a disease, or quality of life. Statistical methods for recurrent event data have received much attention in the past few decades, and are still actively studied. Recurrent event methods can be either based on time from the initial event to each episode of event (i.e., time-to-event) or on time between recurrent events (i.e., gap time or interoccurrence time). In situations where time-to-event is of interest, the recurrent events are usually considered as a realization of a counting process. Statistical methods for time-to-event data are commonly constructed based on the intensity function or rate function of the recurrent event process.When modeling the time (or transformed time) variables directly, time-to-event could be of less interest than gap times between events because a delayed first episode, for example, will certainly prolong the total time to any subsequent episode. The focus of this thesis is on modeling recurrent gap time data.

1.1 Features of recurrent gap time data

We confront many challenges when analyzing recurrent gap time data due to the distinctive structure of recurrent events. One prominent characteristic is the presence of

correlations among recurrent events within a subject. Unlike univariate survival analysis, in which a single time-to-event is studied for each subject, in recurrent event data multiple events can be observed for a single subject during the follow-up period. Gap times between recurrent events of the same subject are generally correlated, and hence the correlation must be taken into account in statistical modeling and estimation. To deal with the correlation, one may consider applying multivariate clustered survival data methods by regarding a single subject as a cluster. However, this does not properly account for the ordinal nature of recurrent events. This distinctive sequential structure imposes difficulties in recurrent gap time data analysis.

First of all, the second or higher order gap times beyond the first event are subject to "induced dependent censoring". In general, recurrent events are collected until the end of the follow-up, and thus the event process is subject to right censoring. Even when the overall censoring time is assumed to be independent to the recurrent event process, the censoring times for the second or higher order gap times are dependent on their corresponding gap times due to the correlation between gap times. Identifiability is another important issue (Lin et al., 1999). The marginal distribution of the second or higher order gap times is only identifiable when the maximum support of the previous gap times is less than that of the censoring time. Finally, the last censored gap time in recurrent gap time data tends to be longer than the previous, completely observed gap times. This is due to "intercept sampling", where longer gap times are more likely to be censored than shorter ones (Wang and Chang, 1999). Thus, including the last censored gap time in an estimation procedure without accounting for this possible bias may lead to incorrect results. The key issues discussed in this section must be addressed appropriately in developing statistical methods for recurrent gap time data.

1.2 Overview of recurrent gap time data analysis

For univariate survival data such as time to death, the well-known Kaplan-Meier method (Kaplan and Meier, 1958) provides a useful tool for describing the distribution of the length of time from the onset of the follow-up (e.g., the time when a patient receives a treatment) to the occurrence of an event of interest. Corresponding one-sample methods are available for recurrent gap time data. For recurrent events of the same type (i.e.,

univariate recurrent events), Pena et al. (2001) established the generalized Nelson-Aalen and Kaplan-Meier estimators assuming that gap times are independent and identically distributed (i.i.d.). Wang and Chang (1999) relaxed the independence assumption on gap times and proposed moment-type estimators. Oftentimes, we observe multiple events of different nature in longitudinal studies such as the progression of AIDS: HIV infection, AIDS onset, and death, which are called multistate event data. For bivariate serial event data, nonparametric estimators were studied by Visser (1996), Wang and Wells (1998), and Lin et al. (1999). Nonparametric estimation methods for marked survival data developed by Huang and Louis (1998) can also be readily applied to bivariate serial gap time data by considering the vector of bivariate gap times as mark variables of the time to the second event. When the event process consists of two alternating events such as hospital admission and discharge, a nonparametric method for estimating the joint distribution of two alternating gap times has been proposed by Huang and Wang (2005).

Regression methods would be more attractive to researchers who are interested in identifying which risk factors are related to gap times. Extensive studies have been conducted on regression methods for recurrent gap time data. In the literature, covariate effects are assessed by modeling either the hazard functions of gap times (Huang and Chen, 2003; Sun et al., 2006) or the (transformed) gap times directly (Chang, 2004; Lu, 2005; Strawderman, 2005). More recently, quantile regression methods have been proposed for recurrent gap time data to account for data heteroscedasticity (Luo et al., 2013). These methods are designed for single-type recurrent event data. For gap times from different types of events, i.e., multistate gap time data, Huang (2002) proposed a semiparametric regression model where the number of states needs to be prefixed. Chang (2004) studied semiparametric regression methods for a bivariate alternating recurrent event process.

The remainder of this dissertation is organized as follows. In Section 1.3, we introduce the two datasets which motivated our study. In Chapter 2, nonparametric methods for recurrent gap time data are studied. We apply the method to the post-transplant recurrent infection data. Then, a semiparametric regression model for recurrent gap time data is developed in Chapter 3. In Chapter 4 we propose a semiparametric estimation method for bivariate alternating recurrent gap time data. An application to the

psychiatric case register data is provided for illustration. Finally, concluding remarks and future research are discussed in Chapter 5.

1.3 Motivating datasets

1.3.1 The post-hematopoietic stem cell transplantation infection data

In the post-hematopoietic stem cell transplantation (HSCT) data, infectious events were documented for a total of 1001 patients with malignant disease who received their first HSCT between 2000 and 2010 at the University of Minnesota. Demographic and treatment related variables for this population are summarized in Table 1.1.

Following Barker et al. (2005), an infectious episode was defined as any infection confirmed by culture, histology, polymerase chain reaction, or antigenemia for which treatment was initiated and clinically compatible time frames were used to define one infectious episode separated from a second episode with the same organism (see Barker et al., 2005, for details). The database includes various types of infections, while in this dissertation we restrict our analysis to the two most prevalent types of infections, bacterial and viral infections. Patients usually have the highest risk of infection prior to the engraftment of donor hematopoietic cells. Among them, neutrophils, a type of white blood cells important for fighting infections, could take as long as 42 days to engraft. If engraftment occurs after day 42, a patient would be considered an engraftment failure. Thus, in this dissertation, we focus on early infections occurring between day 0 and day 42 post HSCT, a short but critical period for infectious complication management.

Our goal is to describe the nature of early infection process in transplanted patients using nonparametric and semiparametric methods. We note that the time from transplant to the first infection has a different biological meaning than the gap times between consecutive recurrent infections. Hence, naively applying existing methods for single-type recurrent event data to the post-HSCT infection data is not appropriate. In Chapters 2 and 3 we propose a nonparametric method and a semiparametric regression method which account for the potentially different distributions of the two types of gap times.

Table 1.1: Summary of the patient- and transplant-related variables.

Variables	N (%) / Median (range)
Overall	1001
Age at Transplant	42.8 (0.5−74.2)
Gender	
Female	385 (38.5)
Male	616 (61.5)
Transplant Source	
UCB	519 (51.8)
Marrow	128 (12.8)
PBSC	346 (34.6)
Marrow + PBSC	8 (0.8)
Diagnosis	
Acute myeloid leukemia	354 (35.4)
Acute lymphoblastic leukemia	200 (20.0)
Non-Hodgkin's lymphoma	148 (14.8)
Myelodysplasia	99 (9.9)
Chronic myeloid leukemia	67 (6.7)
Myeloproliferative disease	19 (1.9)
Hodgkin's lymphoma	16 (1.6)
Multiple myeloma	15 (1.5)
Neuroblastoma	1 (0.1)
Other leukemia	56 (5.6)
Other malignancy	26 (2.6)
Conditioning Regimen	
Myeloablative	589 (58.8)
Non-myeloablative	412 (41.2)
CMV Serostatus[a]	
Positive	567 (56.7)
Negative	433 (43.3)

[a]One patient had missing CMV serostatus

1.3.2 The South Verona psychiatric case register data

A subsample of the South Verona psychiatric case register (PCR) data (Tansella, 1991) consist of 336 psychiatric patients who had contact with the register for the first time with a diagnosis of schizophrenia or related disorders between 1981 and 1995 in South Verona, Italy. Contacts with the South Verona Community Mental Health Service were documented prospectively, which includes day care, rehabilitation and home visits, a psychiatric unit in a general hospital, sheltered apartments, outpatient services, and 24-hour crisis intervention. Patients were in either periods of care (abbreviated as "care period") or break ("break period") during follow-up, and the two states alternated over time. This is a typical bivariate alternating recurrent event data example. According to the definition in Sturt et al. (1982) and Tansella et al. (1995), a break period is when no mental health service is used for over 90 days between consecutive mental health services, and a care period begins from the time a psychiatric contact is made until a break occurs. The South Verona PCR data have been analyzed in Huang and Wang (2005), and a detailed description of the data with a summary of the number of bivariate recurrence times can be found therein.

It is obvious that existing methods for single-type recurrent event data can only be applied to one of the two alternating states. For example, these methods can be used to analyze gap times between initial contacts of each care period, ignoring the break period. In Huang and Wang (2005), nonparametric estimators for the joint distribution of the two alternating recurrent gap times were proposed and applied to the PCR data. To researchers, regression methods can be more attractive. In this dissertation, we propose a new semiparametric regression method, competitive to the method proposed by Chang (2004), to evaluate the effects of the socio-demographic and economic factors on the alternating bivariate gap times in Chapter 4.

Chapter 2

Nonparametric methods for analyzing recurrent gap time data with application to infections after hematopoietic stem cell transplantation

2.1 Introduction

Serious infection is a major complication after hematopoietic stem cell transplantation (HSCT). It accounts for substantial morbidity and mortality among transplanted patients. Many patients experience infectious events repeatedly. Similar to the discussion in Lin et al. (1999), the time from transplant to the first episode of an infectious event and the gap time from one episode to the next episode of infection could both be of interest. This is because transplant and other adjuvant treatments may have different effects on time to the first infection and on gap times between consecutive infections. In this chapter, we are concerned with the gap time from HSCT to the first infection and the gap times between two consecutive infections. Specifically, we are interested in estimating the joint distribution of the time from transplant to the first episode of

infection and the gap times between consecutive infections.

In Section 1.2 we discussed existing methods for univariate recurrent gap time data such as Wang and Chang (1999, referred to as the "Wang-Chang estimator" hereafter) and Pena et al. (2001). These methods only consider the situation in which a patient's enrollment is triggered by an initial event that is of the same type as the recurrent events and assume that all gap times, including the first event time, are identically distributed. Applying them to the post-HSCT infection data by ignoring the fact that the initial event (i.e., transplant) is not of the same type as the recurrent events (i.e., infections) will inevitably lead to incorrect inferential results. This is because the time from transplant to the first infection has different clinical significance than gap times between recurrent infections after the first infection occurs: The immune response during immune reconstitution after transplant may have different profiles before and after the first infection.

Alternatively, one may analyze data after the first infection to make existing recurrent gap time methods applicable, but this introduces selection bias because only patients who have experienced at least one infection are included in the analysis. Other naive methods include applying univariate survival data methods on the time to the first infection only or using bivariate serial event data methods on the data up to the second infection, such as the estimators considered by Visser (1996), Huang and Louis (1998, referred to as the "Huang-Louis estimator" hereafter), Wang and Wells (1998), and Lin et al. (1999, referred to as the "Lin-Sun-Ying estimator" hereafter). Thus, all data beyond the first or the second infectious events is ignored. These approaches lead to either loss of information or failure to address appropriate scientific questions.

In this chapter, we develop a nonparametric estimator that efficiently uses data beyond the first and second infectious episodes for estimation of the joint distribution. In addition, we are interested in estimating the conditional distribution and conditional quantiles of the gap times between two consecutive infections given the time to the first infection after HSCT. It is also of interest to compare different patient subgroups defined by patient- or transplant-related characteristics. For example, patients who received myeloablative and non-myeloablative immunosuppressive regimen could have different risks of bacterial infections (van Burik and Weisdorf, 1999), while patients' cytomegalovirus (CMV) serostatus before transplant could affect CMV infection risk.

Knowledge obtained from subgroup analyses using one-sample estimation method can help to inform the regression analysis used to formally correlate risk factors to the risk of infections.

The remainder of this chapter is organized as follows. We introduce notation and describe the recurrent event process in Section 2.2. In Section 2.3, we first present the bivariate serial event method considered by Huang and Louis (1998). We then discuss how this method can be extended to recurrent event data in a way similar to the weighted risk-set method discussed by Luo and Huang (2011). Weak convergence of the proposed estimator is established by applying empirical processes theory in Section 2.4. In Section 2.5, we report results of simulation studies. In Section 2.6, we apply the proposed methods to the recurrent infection data of patients who received transplantation at the University of Minnesota. Some concluding remarks are provided in Section 2.7.

2.2 Notation and assumptions

Consider a study of post transplant infections, where patients are followed from transplant until a censoring time. For subject i, $i = 1, \ldots, n$, let W_{i0} denote the time from transplant to the first infection and W_{ij}, $j = 1, 2, \ldots$, the gap times between the following infections. Let $N_i = \{W_{ij}, j = 0, 1, \ldots\}$ denote the collection of all gap times since transplant for subject i. We allow the time from transplant to the first infection (i.e., the first event time) and subsequent gap times between successive infections to have different distributions. To highlight the difference between the two types of gap times, we denote the first event time by $X_i^0 \equiv W_{i0}$ and the recurrent gap times after the first infection by $Y_{ij}^0 \equiv W_{ij}$, $j = 1, 2, \ldots$. Let C_i be the censoring time from transplant, which has a survival function $G(t) = \Pr(C_i > t)$ with a fixed maximum support denoted by $\tau_C = \sup\{t : G(t) > 0\}$. Let m_i denote the number of completely observed infectious episodes for subject i. Figure 2.1 illustrates a typical patient's recurrent infection process, in which the first m_i infections are observed without censoring while the m_{i+1}^{th} infection is censored at time C_i (i.e., $\sum_{j=0}^{m_i-1} W_{ij} \leq C_i$ and $\sum_{j=0}^{m_i} W_{ij} > C_i$). As in existing recurrent gap time methods, such as the ones considered by Wang and Chang (1999) and many others, we assume there exists a subject-specific latent random variable or vector (i.e., frailty) γ_i characterizing the within-subject association among

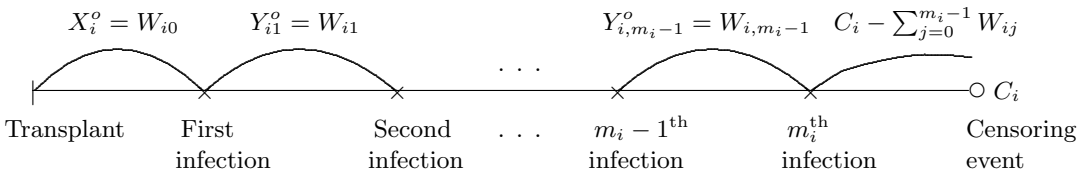

Figure 2.1: Illustration of time from transplant to the first infection and gap times between recurrent infections.

the gap times of the same subject, whose distribution is left unspecified. Then, we make the following assumptions:

Assumption 2.1 *Given γ_i, the gap times $(X_i^0, Y_{i1}^0, Y_{i2}^0, \ldots)$ are independent, and moreover, $Y_{ij}^0, j = 1, 2, \ldots,$ are identically distributed.*

It follows from Assumption 2.1 that unconditional on γ_i, the gap times $X_i^0, Y_{i1}^0, Y_{i2}^0, \ldots$ are correlated. Also, under Assumption 2.1, the gap times $\{Y_{ij}^0, j = 1, 2, \ldots\}$ of subject i are exchangeable and hence the gap time pairs $\{(X_i^0, Y_{ij}^0), j = 1, 2, \ldots\}$ are also exchangeable. Note that both the distribution of γ_i and the dependency between γ_i and the gap times $(X_i^0, Y_{i1}^0, Y_{i2}^0, \ldots)$ are left unspecified. Also note that the correlation between the first event time (X_i^0) and a subsequent gap time (Y_{ij}^0) is allowed to be different than that between two subsequent gap times (Y_{ij}^0 and $Y_{ij'}^0$). However, when correlation structure among gap times is more complex or changing over time, Assumption 2.1 could be inadequate.

Assumption 2.2 *The censoring time C_i is independent of N_i and γ_i.*

Under Assumption 2.2, the first event time X_i^0 is subject to independent censoring by C_i, whereas the subsequent gap times $(Y_{ij}^0, j = 1, 2, \ldots)$ are subject to dependent censoring by $C_i - X_i^0 - \cdots - Y_{ij-1}^0$, which is known as "induced dependent censoring" as discussed in Section 1.1.

2.3 Estimators

To estimate the joint distribution of the time from transplant to the first infection (X_i^0) and the gap times between following consecutive infections (Y_{ij}^0), a simple approach

would be to apply methods for bivariate serial event data, such as the Huang-Louis estimator, to the subset of the data that are composed of only the first two gap times. A brief review of the Huang-Louis estimator is as follows. We define $Z_{i1}^0 = X_i^0 + Y_{i1}^0$ and $\boldsymbol{V}_{i1}^0 = (X_i^0, Y_{i1}^0)$, which correspond to the survival time and the mark variables for the i^{th} individual in Huang and Louis (1998), respectively. For the post-transplant infection study, Z_{i1}^0 represents the time from transplant to the second infection and $\boldsymbol{V}_{i1}^0$ is the pair of the first two gap times. As discussed in Huang and Louis (1998) and Huang and Wang (2005), the equality,

$$F_{X^0,Y^0}(x,y) = F_{Z^0,\boldsymbol{V}^0}(x+y,(x,y)), \tag{2.1}$$

allows us to estimate the joint distribution of (X_i^0, Y_{i1}^0), $F_{X^0,Y^0}(x,y)$, through estimating $F_{Z^0,\boldsymbol{V}^0}(x+y,(x,y))$, where $F_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u}) = \Pr(Z_{i1}^0 \le t, X_i^0 \le u_1, Y_{i1}^0 \le u_2)$ and $\boldsymbol{u} = (u_1,u_2)$. The marginal survival function of the time to the second infection satisfies $S_{Z^0}(t) = 1 - F_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{\infty})$, where $\boldsymbol{\infty} = (\infty,\infty)$. Note that the time Z_{i1}^0 is subject to independent censoring by C_i. Let $Z_{i1} = \min(Z_{i1}^0, C_i)$ with a survival function denoted by $S_Z(\cdot)$, $\Delta_{i1} = I(Z_{i1}^0 \le C_i)$, $\boldsymbol{V}_{i1} = \boldsymbol{V}_{i1}^0\Delta_{i1}$, and $F_{Z,\boldsymbol{V}}(t,\boldsymbol{u}) = \Pr(Z_{i1}^0 \le t, \boldsymbol{V}_{i1}^0 \le \boldsymbol{u}, \Delta_{i1} = 1)$. Following the independent censoring assumption and the facts that $F_{Z,\boldsymbol{V}}(\mathrm{d}t,\boldsymbol{u}) = F_{Z^0,\boldsymbol{V}^0}(\mathrm{d}t,\boldsymbol{u})G(t-)$ and $S_Z(t-) = S_{Z^0}(t-)G(t-)$, we have

$$F_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u}) = \int_0^t S_{Z^0}(s-)\frac{F_{Z^0,\boldsymbol{V}^0}(\mathrm{d}s,\boldsymbol{u})}{S_{Z^0}(s-)} = \int_0^t S_{Z^0}(s-)\frac{F_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u})}{S_Z(s-)}, \tag{2.2}$$

which can be consistently estimated by $\hat{F}_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u}) = \int_0^t \hat{S}_{Z^0}(s-)\frac{\hat{F}_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u})}{\hat{S}_Z(s-)}$, where $\hat{F}_{Z,\boldsymbol{V}}(\cdot,\cdot)$ and $\hat{S}_Z(\cdot)$ are the corresponding empirical measures and $\hat{S}_{Z^0}(\cdot)$ is the Kaplan-Meier estimator.

To make better use of data beyond the second infection, we define $Z_{ij}^0 = X_i^0 + Y_{ij}^0$ and $\boldsymbol{V}_{ij}^0 = (X_i^0, Y_{ij}^0), j = 1,2,\ldots$, where Z_{ij}^0 can be thought of as the time from transplant to the artificial second infection time with the true second gap time Y_{i1}^0 in Z_{i1}^0 being replaced by Y_{ij}^0. Note that $Z_{ij}^0, j = 1,2,\ldots$ are identically (but not independently) distributed.

For ease of discussion, we let $m_i^* = m_i - 1$ denote the number of completely observed gap time pairs when $m_i \ge 2$, otherwise $m_i^* = 1$. Let $Z_{ij} = \min(Z_{ij}^0, C_i), j = 1,\ldots,m_i^*$.

For the reconstructed data $\{Z_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i^*\}$, we can estimate $S_{Z^0}(\cdot)$, in the same spirit as Wang and Chang (1999) and Huang and Wang (2005), by

$$\hat{S}^*_{Z^0}(t) = \prod_{t_k^* \le t} \left(1 - \frac{\hat{H}(t_k^*, \boldsymbol{\infty})}{\hat{R}(t_k^*)}\right),$$

where $t_1^*, t_2^*, \ldots$ are distinct and uncensored recurrence times from $\{Z_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i^*\}$ and $\hat{H}(\cdot)$ and $\hat{R}(\cdot)$ are, respectively, defined as $\hat{H}(t, \boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(m_i \ge 2)}{m_i^*} \sum_{j=1}^{m_i^*} I(Z_{ij} = t, X_i^0 \le u_1, Y_{ij}^0 \le u_2)$ and $\hat{R}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} I(Z_{ij} \ge t)$. Note that when $j \le m_i^*$ and $m_i \ge 2$, the variables X_i^0 and Y_{ij}^0 in $\hat{H}$ are observed quantities. Let $\hat{F}^*_{Z,\boldsymbol{V}}(t, \boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(m_i \ge 2)}{m_i^*} \sum_{j=1}^{m_i^*} I(Z_{ij} \le t, X_i^0 \le u_1, Y_{ij}^0 \le u_2)$. It is obvious that $\hat{F}^*_{Z,\boldsymbol{V}}(\mathrm{d}t, \boldsymbol{u}) = \hat{H}(t, \boldsymbol{u})$.

Now, motivated by Equation (2.2), we propose to estimate $F_{Z^0,\boldsymbol{V}^0}(t, \boldsymbol{u})$ by

$$\hat{F}^*_{Z^0,\boldsymbol{V}^0}(t, \boldsymbol{u}) = \sum_{t_k^* \le t} \left[\prod_{j<k} \left(1 - \frac{\hat{H}(t_j^*, \boldsymbol{\infty})}{\hat{R}(t_j^*)}\right)\right] \frac{\hat{H}(t_k^*, \boldsymbol{u})}{\hat{R}(t_k^*)}. \tag{2.3}$$

Note that our focus is to estimate the joint distribution, $F_{X^0,Y^0}(x, y)$, for the time from transplant to the first infection (X^0) and the between-infection gap times after the first infection (Y^0). It follows directly from (2.1) that $F_{X^0,Y^0}(x, y)$ can be estimated by $\hat{F}_{X^0,Y^0}(x, y) = \hat{F}^*_{Z^0,\boldsymbol{V}^0}(x + y, (x, y))$, which is identifiable for $x + y \le \tau_C$.

The conditional distribution of the recurrent gap times given the time to the first infection, denoted by $F_{Y^0|X^0}(y \mid x) = \Pr(Y^0 \le y \mid X^0 \le x)$, may be a scientifically interesting quantity. It can be estimated by

$$\hat{F}_{Y^0|X^0}(y \mid x) = \frac{\hat{F}_{X^0,Y^0}(x, y)}{1 - \hat{S}_{X^0}(x)}, x + y \le \tau_C, \tag{2.4}$$

where $\hat{S}_{X^0}(\cdot)$ is the Kaplan-Meier estimator for the time to the first infection. Note that the marginal distribution of gap times between consecutive infections, $F_{Y^0}(y)$ is not generally identifiable for $y > 0$ unless the support for X^0 is less than τ_C as noticed in Section 1.1. Similar discussion for bivariate serial gap time data can be found in Lin et al. (1999), Lin and Ying (2001), Schaubel and Cai (2004), Cook and Lawless (2007), and Lawless and Yilmaz (2011). As noted by Lin and Ying (2001), when x is large

enough, the conditional distribution $F_{Y^0|X^0}(y \mid x)$ may be a good approximation for the marginal distribution $F_{Y^0}(y)$.

The proposed conditional distribution estimator can be used to estimate quantiles of interest. The pth conditional quantile associated with the conditional distribution function $F_{Y^0|X^0}(y \mid x)$ is defined as $y_p(x) = \inf\{y : 1 - F_{Y^0|X^0}(y \mid x) \leq 1 - p\}$. When $p = 0.5$, $y_p(x)$ is the conditional median of the gap times beyond the first infection given that the time from transplant to the first infection is no longer than x. We can estimate $y_p(x)$ by $\hat{y}_p(x) = \inf\{y : 1 - \hat{F}_{Y^0|X^0}(y \mid x) \leq 1 - p\}$. In practice, the $100(1-\alpha)\%$ confidence interval (CI) for $\hat{y}_p(x)$ can be constructed using the bootstrap method without the need to estimate the variance of $\hat{F}_{Y^0|X^0}(y \mid x)$. Otherwise, one can obtain the linear Wald-type CI, $\{y : -Z_{1-\alpha/2} \leq [\hat{F}_{Y^0|X^0}(y \mid x) - p]/\widehat{\mathrm{Var}}^{1/2}[\hat{F}_{Y^0|X^0}(y \mid x)] \leq Z_{1-\alpha/2}\}$ with the variance estimate $\widehat{\mathrm{Var}}[\hat{F}_{Y^0|X^0}(y \mid x)]$. Other forms such as the log-log transformed and arcsine-square root transformed CIs can also be used (see Klein and Moeschberger, 2003, p. 120).

2.4 Asymptotic properties

Set region $\Omega = \{(t, (u_1, u_2)) : 0 \leq u_1 + u_2 \leq t \leq L\}$ where L is any number smaller than τ_C. In this region, the proposed estimator can obviously be identified. We assume that $G(t)$ and $F_{Z^0,\boldsymbol{V}^0}(t, \boldsymbol{u})$ are absolutely continuous on $[0, L]$ and Ω, respectively. Define $\Lambda(t, \boldsymbol{u}) = \int_0^t F_{Z,\boldsymbol{V}}(\mathrm{d}s, \boldsymbol{u})/S_Z(s)$. Then, function $S_{Z^0}(t)$ can be re-expressed as a function of $\Lambda(t, \boldsymbol{u})$ as follows,

$$S_{Z^0}(t) = \prod_{[0,t]} \{1 - \Lambda(\mathrm{d}s, \boldsymbol{\infty})\}. \tag{2.5}$$

Let $\mathcal{S}(\Omega)$ denote the space of bivariate right-continuous functions on Ω with left-hand limits. Combining (2.2) and (2.5), both of which lie in $\mathcal{S}(\Omega)$, we can define a mapping $\Phi : \Lambda \to F_{Z^0,\boldsymbol{V}^0}$ as follows

$$F_{Z^0,\boldsymbol{V}^0}(t, \boldsymbol{u}) = \Phi(\Lambda)(t, \boldsymbol{u}) = \int_0^t \prod_{[0,s)} \{1 - \Lambda(\mathrm{d}s, \boldsymbol{\infty})\} \Lambda(\mathrm{d}s, \boldsymbol{u}). \tag{2.6}$$

Let the estimator of Λ be

$$\hat{\Lambda}(t,\boldsymbol{u}) = \int_0^t \frac{\hat{F}^*_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u})}{\hat{R}(s)} \tag{2.7}$$

by replacing $F_{Z,\boldsymbol{V}}$ and S_Z with $\hat{F}^*_{Z,\boldsymbol{V}}$ and $\hat{R}$, respectively. Plugging (2.7) into (2.6), we can derive the proposed estimator of $F_{Z^0,\boldsymbol{V}^0}$, that is, $\hat{F}^*_{Z^0,\mathbf{V}^0} = \Phi(\hat{\Lambda})$. It can be easily verified that this is equivalent to (2.3).

By the exchangeability of the pairs $\{(X_i^0, Y_{i1}^0), (X_i^0, Y_{i2}^0), \ldots\}$, we have $\mathrm{E}\left[\frac{I(m_i\geq 2)}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij} \leq t, \mathbf{V}_{ij}^0 \leq \boldsymbol{u})\right] = \mathrm{E}[I(Z_{i1}^0 \leq t, \mathbf{V}_{i1}^0 \leq \boldsymbol{u}, \Delta_{i1} = 1)] = F_{Z,\mathbf{V}}(t,\boldsymbol{u})$ and $\mathrm{E}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij} \geq t)\right] = \mathrm{E}[I(Z_{ij} \geq t)] = S_Z(t)$. Hence, the moment type estimators $\hat{F}^*_{Z,\mathbf{V}}$ and $\hat{R}$ involved in $\hat{\Lambda}$ both converge weakly to the same limit as their counterparts $\hat{F}_{Z,\mathbf{V}}$ and $\hat{S}_Z$ in the Huang-Louis estimator. The mapping Φ is compactly differentiable at each point of $\mathcal{S}(\Omega)$ with the derivative

$$\begin{aligned}\{\mathrm{d}\Phi(\Lambda)\cdot h\}(t,\boldsymbol{u}) = &\int_0^t \{F_{Z^0,\boldsymbol{V}^0}(s,\boldsymbol{u}) - F_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u})\}h(\mathrm{d}s,\boldsymbol{\infty}) \\ &+ \int_0^t \{1 - F_{Z^0,\boldsymbol{V}^0}(s,\boldsymbol{\infty})\}h(\mathrm{d}s,\boldsymbol{u}),\end{aligned}$$

where $h \in \mathcal{S}(\Omega)$ (Andersen et al., 1993, Proposition 2.8.7). Because the mapping is differentiable, it is sufficient to study the asymptotic properties of $\hat{\Lambda}$ to derive the large samples properties of the proposed method. It is straightforward that the moment type estimators $\hat{F}^*_{Z,\boldsymbol{V}}$ and $\hat{R}$ included in $\hat{\Lambda}$ both satisfy the weak convergence theorem. The proof is provided in Appendix A.1. By applying the functional delta method, the large sample properties of $\hat{\Lambda}$ can be derived. Define the function $\psi_i(t,\boldsymbol{u}) = \frac{I(m_i\geq 2)}{m_i^*}\sum_{j=1}^{m_i^*}\frac{I(Z_{ij}\leq t, \boldsymbol{V}_{ij}\leq \boldsymbol{u})}{S_Z(Z_{ij})} - \int_{[0,t]}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij}\geq s)\frac{F_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u})}{S_Z^2(s)}$. Then, $\hat{\Lambda}$ has the following property.

Theorem 2.1 *For any $L < \tau_C$ and $(t,\boldsymbol{u}) \in \Omega$, the stochastic process $n^{1/2}\{\hat{\Lambda}(t,\boldsymbol{u}) - \Lambda(t,\boldsymbol{u})\}$ has an asymptotically i.i.d. representation*

$$\sqrt{n}\{\hat{\Lambda}(t,\boldsymbol{u}) - \Lambda(t,\boldsymbol{u})\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(t,\boldsymbol{u}) + o_p(1),$$

which converges weakly to a Gaussian process with mean zero and variance-covariance function $E[\psi_1(t_1,\boldsymbol{u}_1)\psi_1(t_2,\boldsymbol{u}_2)]$, where $(t_j,\boldsymbol{u}_j) \in \Omega, j = 1,2$.

The variance-covariance function of the limiting distribution can be consistently estimated by $n^{-1}\sum_{i=1}^{n}\hat{\psi}_i(t_1,\boldsymbol{u}_1)\hat{\psi}_i(t_2,\boldsymbol{u}_2)$, where $\hat{\psi}_i$ is the estimator of ψ_i derived by replacing $F_{Z,\boldsymbol{V}}$ and S_Z with $\hat{F}^*_{Z,\boldsymbol{V}}$ and $\hat{R}$, respectively.

Define the function $\phi_i(t,\boldsymbol{u}) = \int_{[0,t]}F_{Z^0,\boldsymbol{V}^0}(s,\boldsymbol{u})\psi_i(\mathrm{d}s,\boldsymbol{\infty}) + \int_{[0,t]}S_{Z^0}(s)\psi_i(\mathrm{d}s,\boldsymbol{u}) - F_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u})\psi_i(t,\boldsymbol{\infty})$. The asymptotic properties of the proposed estimator follow in Theorem 2.2.

Theorem 2.2 *For any $L < \tau_C$ and $(t,\boldsymbol{u}) \in \Omega$, the stochastic process $n^{1/2}\{\hat{F}^*_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u}) - F_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u})\}$ has an asymptotically i.i.d. representation*

$$\sqrt{n}\{\hat{F}^*_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u}) - F_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u})\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_i(t,\boldsymbol{u}) + o_p(1),$$

which converges weakly to a Gaussian process with mean zero and variance-covariance function $E[\phi_1(t_1,\boldsymbol{u}_1)\phi_1(t_2,\boldsymbol{u}_2)]$, where $(t_j,\boldsymbol{u}_j) \in \Omega, j = 1,2$.

The variance-covariance function of the limiting distribution can be consistently estimated by $n^{-1}\sum_{i=1}^{n}\hat{\phi}_i(t_1,\boldsymbol{u}_1)\hat{\phi}_i(t_2,\boldsymbol{u}_2)$, where $\hat{\phi}_i$ is the estimator of ϕ_i derived by replacing $F_{Z^0,\boldsymbol{V}^0}$, S_{Z^0}, and ψ_i with $\hat{F}^*_{Z^0,\boldsymbol{V}^0}$, $\hat{S}^*_{Z^0}$, and $\hat{\psi}_i$, respectively. Therefore, for $0 \le x+y \le L$, $n^{1/2}\{\hat{F}_{X^0,Y^0}(x,y) - F_{X^0,Y^0}(x,y)\}$ converges weakly to a Gaussian process with mean zero and variance-covariance function $\mathrm{E}[\phi_1(x_1+y_1,(x_1,y_1))\phi_1(x_2+y_2,(x_2,y_2))]$, where $0 \le x_j + y_j \le L, j = 1,2$.

The proof of Theorems 2.1 and 2.2 follows closely the proof in Huang and Wang (2005) and is provided in the Appendix A.2 and A.3.

To establish the asymptotic properties for the conditional distribution $\hat{F}_{Y^0|X^0}(y \mid x)$, we need to introduce additional notation for the first event time. Let $X_i = \min(X_i^0, C_i)$ denote the observed first gap time and $S_X(\cdot)$ and $F_X(\cdot)$ denote its survival function and distribution function, respectively. The corresponding empirical functions for X are $\hat{S}_X(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \ge t)$ and $\hat{F}_X(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t)$. Define the function $\xi_i(t,\boldsymbol{u}) = \frac{1}{1-S_{X^0}(u_1)}\phi_i(t,\boldsymbol{u}) - \frac{F_{X^0,Y^0}(u_1,u_2)}{\{1-S_{X^0}(u_1)\}^2}\phi_i'(u_1)$, where $\phi_i'(u_1) = S_{X^0}(u_1)\left\{\frac{I(X_i\le C_i)I(X_i\le u_1)}{S_X(X_i)} - \int_0^{u_1}\frac{I(X_i\ge s)}{S_X^2(s)}\mathrm{d}F_X(s)\right\}$. The weak convergence of $\hat{F}_{Y^0|X^0}(y \mid x) = \hat{F}_{X^0,Y^0}(x,y)/\{1 - \hat{S}_{X^0}(x)\}$ follows naturally from the weak convergence of the proposed estimator $\hat{F}_{X^0,Y^0}(x,y)$ in the following theorem.

Theorem 2.3 *For any $L < \tau_c$ and $(t, \boldsymbol{u}) = (u_1+u_2, (u_1, u_2)) \in \Omega$, the stochastic process $n^{1/2}\{\hat{F}_{Y^0|X^0}(u_2 \mid u_1) - F_{Y^0|X^0}(u_2 \mid u_1)\}$ has an asymptotically i.i.d. representation*

$$\sqrt{n}\{\hat{F}_{Y^0|X^0}(u_2 \mid u_1) - F_{Y^0|X^0}(u_2 \mid u_1)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i(t, \boldsymbol{u}) + o_p(1),$$

which converges weakly to a Gaussian process with mean zero and variance-covariance function $E[\xi_1(t_1, \boldsymbol{u}_1)\xi_1(t_2, \boldsymbol{u}_2)]$, where $(t_j, \boldsymbol{u}_j) \in \Omega, j = 1, 2$.

The variance-covariance function of the limiting distribution can be consistently estimated by $n^{-1}\sum_{i=1}^{n} \hat{\xi}_i(t_1, \boldsymbol{u}_1)\hat{\xi}_i(t_2, \boldsymbol{u}_2)$, where $\hat{\xi}_i$ is the estimator of ξ_i derived by replacing S_{X^0}, S_X, F_X, F_{X^0,Y^0}, ϕ_i with $\hat{S}_{X^0}$, $\hat{S}_X$, $\hat{F}_X$, $\hat{F}_{X^0,Y^0}$, $\hat{\phi}_i$, respectively. The proof of Theorem 2.3 can also be found in Appendix A.4.

2.5 Simulation studies

We did a series of simulation studies to evaluate the performance of the proposed method, with 1000 simulated datasets and a sample size of $n = 500$ per dataset for each scenario. For all scenarios, we assume the censoring time C_i follows a uniform distribution $(0, U)$, where $U = 75$ and 150 for different censoring rates.

We consider two different scenarios. In the first scenario, we assume a common frailty shared by all gaps and equal error variance which leads to equal correlation between any two gap times from $\{X_i^0, Y_{i1}^0, Y_{i2}^0, \ldots\}$, including the first event time; whereas in the second scenario, we assume a bivariate frailty, which allows the correlation between the first event time (X_i^0) and a subsequent gap time (Y_{ij}^0) to be different than that between two subsequent gap times (Y_{ij}^0 and $Y_{ij'}^0$).

Simulation Scenario 1

We generate time from transplant to the occurrence of the first infection and gap times between consecutive recurrent infections from the following model:

$$\log(W_{ij}) = b_0 + b_1 I(j > 0) + \gamma_i + \epsilon_{ij}, i = 1, \ldots, n, j = 0, 1, \ldots,$$

where b_1 is a non-zero value allowing time to first infection to be distributed differently from the following gap times, γ_i is a subject-specific random variable following $N(0, \sigma^2)$, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is the measurement error term. Note that the larger the variance of γ_i is, the greater is the heterogeneity among subjects. We let $b_0 = 3$, $b_1 = -1$, $\sigma_\epsilon^2 = 0.1$, and consider $\sigma^2 = 0.1$ and 0.5 for different levels of within-subject correlation between gap times.

Simulation Scenario 2

In this scenario, we assume a bivariate frailty $(\gamma_{i0}, \gamma_{i1})$. The gap times are generated from the model:

$$\log(W_{ij}) = b_0 + b_1 I(j > 0) + \{\gamma_{i0} I(j = 0) + \gamma_{i1} I(j > 0)\} + \epsilon_{ij}, i = 1, \ldots, n, j = 0, 1, \ldots,$$

where the frailty $(\gamma_{i0}, \gamma_{i1})$ follows a bivariate normal distribution with mean zero and variance-covariance matrix $\begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}$. Under this setting, the covariance between the (transformed) first gap time ($\log(X_i^0)$) and a subsequent gap time ($\log(Y_{ij}^0)$) is σ_{01} and that between two transformed subsequent gap times, $\log(Y_{ij}^0)$ and $\log(Y_{ij'}^0)$, $j \neq j'$, is σ_1^2. We let $\sigma_0^2 = \sigma_1^2 = 0.5$ and consider two levels of σ_{01} ($\sigma_{01} = 0$ and 0.25) in the simulations. The other parameters are the same as in Simulation Scenario 1.

The performance of the proposed estimator of $F_{X^0,Y^0}(x, y)$ is compared with that of the Huang-Louis estimator. Tables 2.1 and 2.2 summarize the simulation results for the two simulation scenarios, respectively, at grid points (x, y), where $x = 15, 20, 30$ and $y = 5, 7, 15$. In each scenario, both the proposed method and the Huang-Louis estimator provide virtually unbiased estimates. The efficiency of the Huang-Louis estimator relative to the proposed estimator, measured by the squared quotient of the respective standard deviation estimates, is less than 1 in all scenarios. The efficiency gain of the proposed estimator is greater when more gap times are observed (U increases from 75 to 150), but diminishes with larger y.

We also conducted a simulation study, using the data from Simulation Scenario 1 with $U = 150$ to investigate the performance of the conditional distribution estimator $\hat{F}_{Y^0|X^0}(y \mid x)$ for a large x ($x = 100$) on approximating the marginal distribution of the gap times between consecutive infections, $F_{Y^0}(y)$. The bias and power for the proposed

Table 2.1: Summary of the results for Simulation Scenario 1 with true values of the bivariate joint distribution $F_{X^0,Y^0}(x,y)$ (True); relative bias $\times 10^3$ (Monte-Carlo SD $\times 10^3$ and average asymptotic SE $\times 10^3$) of Huang-Louis estimator (H-L) and the proposed estimator (Pro); and relative efficiency (re) of H-L vs. Proposed.

		$C_i \sim Unif(0,75)$			$C_i \sim Unif(0,150)$		
x	y =	5	7	15	5	7	15
$\sigma^2 = 0.1$		$\bar{m}^a = 3.20$; $cr^b = 0.30$			$\bar{m} = 8.14$; $cr = 0.15$		
15	True	0.1007	0.1827	0.2551	0.1007	0.1827	0.2551
	H-L	-2.7 (15, 15)	2.9 (20, 20)	-0.8 (23, 22)	2.4 (14, 14)	4.4 (19, 18)	2.1 (21, 21)
	Pro	-3.2 (12, 12)	0.5 (18, 18)	-0.5 (23, 22)	-0.3 (11, 11)	4.7 (16, 16)	2.0 (21, 21)
	re^c	0.6401	0.8200	0.9952	0.5711	0.7801	0.9936
20	True	0.1497	0.3066	0.4891	0.1497	0.3066	0.4891
	H-L	0.8 (19, 18)	2.5 (24, 24)	-0.2 (28, 27)	2.7 (17, 17)	2.4 (22, 22)	1.2 (24, 24)
	Pro	2.5 (14, 14)	1.1 (21, 21)	-0.1 (28, 27)	2.0 (12, 12)	3.0 (18, 18)	1.1 (24, 24)
	re	0.5813	0.7381	0.9869	0.4920	0.6713	0.9811
30	True	0.1847	0.4201	0.7898	0.1847	0.4201	0.7898
	H-L	0.3 (21, 20)	1.9 (27, 27)	0.1 (24, 24)	-1.1 (19, 19)	0.0 (24, 24)	0.3 (20, 20)
	Pro	1.4 (15, 15)	0.6 (21, 21)	0.1 (24, 23)	-0.9 (12, 12)	0.8 (17, 17)	0.1 (20, 20)
	re	0.5372	0.6393	0.9469	0.4304	0.5299	0.9089
$\sigma^2 = 0.5$		$\bar{m} = 4.39$; $cr = 0.34$			$\bar{m} = 10.39$; $cr = 0.18$		
15	True	0.2432	0.3093	0.3514	0.2432	0.3093	0.3514
	H-L	0.8 (21, 21)	2.4 (22, 23)	2.3 (23, 24)	0.0 (20, 20)	1.0 (22, 22)	0.1 (23, 23)
	Pro	2.1 (19, 19)	2.0 (21, 22)	2.2 (23, 24)	-1.0 (18, 18)	-0.1 (21, 21)	0.2 (23, 22)
	re	0.8212	0.9168	0.9974	0.7920	0.8991	0.9970
20	True	0.2823	0.3911	0.4916	0.2823	0.3911	0.4916
	H-L	-0.9 (22, 22)	1.5 (24, 25)	4.1 (25, 26)	-3.1 (21, 21)	-2.4 (23, 23)	0.1 (24, 24)
	Pro	1.0 (19, 20)	1.1 (22, 23)	4.1 (25, 26)	-2.7 (18, 18)	-2.3 (21, 21)	0.2 (24, 24)
	re	0.7665	0.8575	0.9911	0.7274	0.8257	0.9875
30	True	0.3035	0.4533	0.6712	0.3035	0.4533	0.6712
	H-L	-0.2 (23, 23)	1.8 (26, 26)	0.5 (26, 27)	-3.0 (22, 22)	-1.3 (24, 24)	-0.7 (24, 23)
	Pro	0.7 (19, 20)	0.9 (23, 23)	0.6 (26, 26)	-3.7 (18, 18)	-1.3 (21, 20)	-0.5 (23, 23)
	re	0.7285	0.7907	0.9665	0.6810	0.7358	0.9403

[a] Average number of observed infections per subject

[b] Average proportion of subjects without any infections

[c] Efficiency of H-L estimator relative to the proposed estimator measured by squared quotient of standard deviations

Table 2.2: Summary of the results for Simulation Scenario 2 with true values of the bivariate joint distribution $F_{X^0,Y^0}(x,y)$ (True); relative bias $\times 10^3$ (Monte-Carlo SD $\times 10^3$ and average asymptotic SE $\times 10^3$) of Huang-Louis estimator (H-L) and the proposed estimator (Pro); and relative efficiency (re) of H-L vs. Proposed.

		$C_i \sim Unif(0,75)$			$C_i \sim Unif(0,150)$		
x	y =	5	7	15	5	7	15
$\sigma_{01} = 0$		$\bar{m}^a$= 3.44; cr^b= 0.34			$\bar{m} = 9.10$; $cr = 0.18$		
15	True	0.1081	0.1671	0.2901	0.1081	0.1671	0.2901
	H-L	-2.9 (15, 15)	-5.3 (19, 18)	-1.9 (24, 23)	3.3 (15, 15)	0.3 (18, 18)	-1.5 (21, 22)
	Pro	-0.0 (14, 14)	-4.4 (17, 17)	-2.4 (23, 23)	3.0 (13, 13)	-0.9 (16, 16)	-1.8 (21, 21)
	re^c	0.7890	0.8489	0.9552	0.7492	0.8100	0.9301
20	True	0.1535	0.2368	0.4097	0.1535	0.2368	0.4097
	H-L	-9.0 (18, 18)	-9.5 (22, 22)	-1.8 (26, 26)	-5.1 (17, 17)	-6.7 (20, 20)	-3.2 (24, 24)
	Pro	-6.2 (16, 16)	-8.3 (20, 20)	-2.2 (25, 25)	-5.0 (15, 15)	-7.5 (18, 18)	-3.7 (23, 23)
	re	0.7808	0.8383	0.9485	0.7373	0.7943	0.9170
30	True	0.2147	0.3309	0.5730	0.2147	0.3309	0.5730
	H-L	-6.3 (22, 21)	-7.0 (26, 25)	-0.4 (28, 28)	2.1 (20, 20)	-1.1 (23, 23)	0.3 (25, 24)
	Pro	-5.1 (19, 19)	-6.4 (23, 23)	-0.6 (27, 27)	1.1 (17, 17)	-1.7 (20, 20)	-0.1 (23, 23)
	re	0.7691	0.8214	0.9357	0.7181	0.7678	0.8863
$\sigma_{01} = 0.25$		$\bar{m} = 3.94$; $cr = 0.34$			$\bar{m} = 9.85$; $cr = 0.18$		
15	True	0.1674	0.2313	0.3295	0.1674	0.2313	0.3295
	H-L	-11.4 (18, 18)	-5.3 (21, 21)	-9.2 (24, 24)	-7.7 (18, 17)	-4.6 (20, 20)	-9.1 (22, 22)
	Pro	-10.5 (17, 16)	-5.4 (20, 20)	-9.5 (24, 23)	-8.7 (16, 15)	-4.9 (18, 18)	-9.6 (22, 22)
	re	0.8166	0.8790	0.9752	0.7849	0.8515	0.9625
20	True	0.2136	0.3039	0.4544	0.2136	0.3039	0.4544
	H-L	-6.3 (21, 20)	-1.2 (24, 23)	-7.2 (26, 26)	-2.4 (19, 19)	0.4 (22, 22)	-5.1 (24, 24)
	Pro	-4.8 (18, 18)	-1.3 (22, 22)	-7.4 (26, 26)	-2.8 (17, 17)	-0.0 (20, 20)	-5.6 (23, 23)
	re	0.7938	0.8547	0.9654	0.7556	0.8193	0.9453
30	True	0.2640	0.3879	0.6165	0.2640	0.3879	0.6165
	H-L	-8.1 (23, 23)	-0.3 (26, 26)	-4.3 (27, 27)	-2.7 (21, 21)	1.7 (23, 23)	-4.2 (24, 24)
	Pro	-6.3 (20, 20)	-0.0 (23, 23)	-4.6 (27, 27)	-3.1 (18, 18)	0.7 (20, 21)	-4.1 (22, 23)
	re	0.7663	0.8215	0.9465	0.7194	0.7725	0.9056

[a] Average number of observed infections per subject

[b] Average proportion of subjects without any infections

[c] Efficiency of H-L estimator relative to the proposed estimator measured by squared quotient of standard deviations

Table 2.3: Summary of the simulation results for comparing marginal distribution estimators for gap times beyond the first infection.

y	True marginal probability $\Pr(Y^0 \leq y)$	Bias[a] and coverage probability[b] Proposed[c]	W-C[d]	K-M[e]
$\sigma^2 = 0.1$				
5	0.1914	-0.8, 0.944	51.0, 0.903	51.4, 0.915
7	0.4519	0.9, 0.947	36.2, 0.840	35.6, 0.905
15	0.9436	-0.6, 0.934	6.0, 0.838	6.1, 0.874
$\sigma^2 = 0.5$				
5	0.3070	-2.8, 0.950	129.3, 0.473	130.5, 0.600
7	0.4719	-1.0, 0.951	110.9, 0.305	110.9, 0.448
15	0.8175	1.9, 0.948	62.5, 0.104	62.0, 0.190

[a]Monte-Carlo relative bias $\times 10^3$
[b]Based on 95% CI
[c]$\hat{F}_{Y^0|X^0}(y|100)$ based on the proposed method
[d]Wang-Chang estimator for the 2nd and higher gaps
[e]Kaplan-Meier estimator for the 2nd gap time data only

method, together with two naive estimators, namely the Wang-Chang estimator and the Kaplan-Meier estimator, which use only data beyond the first infection, are presented in Table 2.3. The results show that the two naive methods have non-ignorable biases, while the conditional distribution based on the proposed method provides satisfactory estimates for the marginal distribution. We further evaluated the conditional median estimator for the between-infection gap times given different length of the time from transplant to the first infection with all simulated data. The detailed result is shown in Table 2.4. The conditional median estimates were virtually unbiased in all scenarios with the coverage probabilities of 95% CIs all close to 0.95.

2.6 Application

We apply the proposed method to the post-HSCT infection data introduced in Section 1.3.1 to make inference about the joint distribution of time from transplant to the first infection and the gap times between recurrent infections. The post-HSCT data consist of 1001 patients who received their first HSCT between 2000 and 2010 at the University of Minnesota. In this section, we focus on bacterial and viral infections which

Table 2.4: Summary of the results for true values of the conditional median of between-infection gap times given different length of time from transplant to the first infection (True); relative bias $\times 10^3$ (RE-Bias); and the coverage probability (CP) of 95% bootstrap confidence intervals.

		$C_i \sim Unif(0, 75)$		$C_i \sim Unif(0, 150)$	
x	True	RE-Bias	CP	RE-Bias	CP
Scenario 1					
$\sigma^2 = 0.1$		$\bar{m}^a$= 3.20; cr^b= 0.30		$\bar{m} = 8.14$; $cr = 0.15$	
15	5.599	-2.2	0.935	-1.0	0.935
20	6.193	-2.0	0.941	-1.7	0.924
30	6.884	-0.4	0.936	-0.1	0.944
$\sigma^2 = 0.5$		$\bar{m} = 4.39$; $cr = 0.34$		$\bar{m} = 10.39$; $cr = 0.18$	
15	3.836	1.5	0.941	3.7	0.934
20	4.526	4.1	0.933	5.4	0.918
30	5.561	-4.6	0.924	-1.3	0.934
Scenario 2					
$\sigma_{01} = 0$		$\bar{m} = 3.44$; $cr = 0.34$		$\bar{m} = 9.10$; $cr = 0.18$	
15	7.385	7.0	0.909	-4.3	0.942
20	7.369	5.7	0.929	0.5	0.936
30	7.375	3.4	0.922	-1.2	0.939
$\sigma_{01} = 0.25$		$\bar{m} = 3.94$; $cr = 0.34$		$\bar{m} = 9.85$; $cr = 0.18$	
15	5.273	2.2	0.929	2.4	0.937
20	5.723	-1.4	0.922	-0.5	0.922
30	6.311	0.2	0.938	-0.6	0.925

[a]Average number of observed infections per subject
[b]Average proportion of subjects without any infections

Table 2.5: Summary of the number of patients who experienced k infections within 42 days after transplant, $k = 0, 1, \cdots, 5, \geq 6$

Group	No. patients (%)	No. of infections observed for a patient						
		0	1	2	3	4	5	≥6
Bacterial infections								
Overall	1001	523	286	136	38	13	2	3
	(100.00)	(52.25)	(28.57)	(13.59)	(3.80)	(1.30)	(0.20)	(0.30)
Myeloablative	589	274	182	90	30	9	2	2
	(100.00)	(46.52)	(30.90)	(15.28)	(5.09)	(1.53)	(0.34)	(0.34)
Non-myeloablative	412	249	104	46	8	4	0	1
	(100.00)	(60.44)	(25.24)	(11.17)	(1.94)	(0.97)	(0.00)	(0.24)
Viral infections								
Overall[a]	1001	661	263	63	10	3	0	1
	(100.00)	(66.03)	(26.27)	(6.29)	(1.00)	(0.30)	(0.00)	(0.10)
CMV seropositive	567	332	171	53	7	3	0	1
	(100.00)	(58.55)	(30.16)	(9.35)	(1.23)	(0.53)	(0.00)	(0.18)
CMV seronegative	433	328	92	10	3	0	0	0
	(100.00)	(75.75)	(21.25)	(2.31)	(0.69)	(0.00)	(0.00)	(0.00)

[a]One patient had missing CMV serostatus

occurred within 42 days after HSCT. By day 42 after transplant, 61 (6%) out of 1001 patients were censored by death, 27 (3%) by relapse, 13 (1%) by a second transplant, and 900 (90%) were alive without a second transplant or relapse. Among the 61 patients who died, 17 (1.7% out of 1001) were found to be infection-related, of whom 4 (0.4% out of 1001) were due to bacterial infections and 4 (0.4% out of 1001) were due to viral infections. Hence, the noninformative censoring assumption was not expected to be violated in this dataset.

About 48% of the patients experienced at least one bacterial infection and 34% at least one viral infection. Overall, 752 bacterial and 437 viral infections were observed within 42 days after transplant. A summary of number of infections for different infection organisms can be found in Table 2.5.

Due to the short follow-up period, we do not expect the exchangeability assumption on the gap times after the first infection to be violated. Nevertheless, we carried out trend tests by Wang and Chen (2000) to assure that the gap times after the first infection are identically distributed. The test results (see p-values in Table 2.6) show that neither bacterial nor viral infections had a significant trend in gap times after the first infection.

Table 2.6 presents the joint distribution estimates of the time from transplant to the first infection and the gap times between two consecutive infections of the same type, for bacterial (upper panel) and viral infections (lower panel) separately. Based on the estimated joint probabilities, bacterial infections had a greater than 2-fold higher probability than viral infections at all presented time points.

Conditional probability distributions of the gap times between consecutive infections, namely $\Pr(Y^0 \leq y \mid x_1 \leq X^0 \leq x_2)$, $x_1 < x_2$, may also be scientifically interesting as we can identify which patients experienced more frequent recurrent infections after experiencing the first infection within a certain period of time. Similar to (2.4), this conditional probability can be estimated by $\{\hat{F}_{X^0,Y^0}(x_2, y) - \hat{F}_{X^0,Y^0}(x_1, y)\}/\{\hat{S}_{X^0}(x_1) - \hat{S}_{X^0}(x_2)\}$ and the associated variance can be estimated by the bootstrap method. Figure 2.2 presents the estimated survival probability of time to the first infection by the Kaplan-Meier method and the estimated conditional probabilities of recurrent gap times for bacterial (left panel) and viral (right panel) infections. It shows that more patients experienced the first bacterial infection than viral infection. Given the first bacterial infection occurred within the first or second or third week, the probability of having another bacterial infection in the following weeks was similar, which may indicate a weak correlation between the first event time with the following recurrent gap times for the bacterial infections. However, for viral infections, we found that given the first viral infection occurred in the first week, as opposed to the third week after transplant, the probability of having another viral infection within another 3 weeks was twice as high (0.59 [95% CI: 0.37, 0.78] vs. 0.27 [0.20, 0.37]), indicating a positive correlation between time to the first viral infection and the gap times between recurrent viral infections.

The conditional median estimates of the between-infection gap times given different lengths of time from transplant to the first infection are presented for bacterial and viral infections in Figure 2.3. The median between-infection gap time for bacterial infections is relatively constant regardless of the length of time from transplant to the first bacterial infection, which indicates a weak correlation between the time to the first bacterial infection and the gap times between consecutive bacterial infections, consistent with our previous finding. However, for viral infections, the median between-infection gap time shows a slightly increasing trend when the time from transplant to the first infection increases.

Table 2.6: The estimates of the bivariate joint distribution $F_{X^0,Y^0}(x,y)$ (SE $\times 10^3$) of time from transplant to the first infection (x) and gap times between two consecutive infections (y).

x (weeks)	y (weeks)				
	1	2	3	4	5
Bacterial infections					
Overall		$\bar{m}^a$= 0.75; cr^b= 0.52; Trend test $p = 0.09$			
1	0.031 (5)	0.051 (6)	0.067 (8)	0.082 (9)	0.092 (9)
2	0.054 (6)	0.084 (8)	0.112 (10)	0.137 (11)	-[c]
3	0.065 (7)	0.097 (9)	0.131 (11)	-	-
4	0.069 (7)	0.103 (9)	-	-	-
5	0.073 (8)	-	-	-	-
Myeloablative		$\bar{m} = 0.87$; $cr = 0.47$; Trend test $p = 0.18$			
1	0.046 (8)	0.074 (10)	0.090 (12)	0.104 (13)	0.112 (13)
2	0.073 (10)	0.111 (12)	0.143 (14)	0.169 (15)	-
3	0.084 (10)	0.124 (13)	0.164 (15)	-	-
4	0.090 (11)	0.133 (13)	-	-	-
5	0.091 (11)	-	-	-	-
Non-myeloablative		$\bar{m} = 0.59$; $cr = 0.60$; Trend test $p = 0.14$			
1	0.009 (4)	0.018 (6)	0.032 (8)	0.049 (11)	0.064 (12)
2	0.026 (7)	0.046 (10)	0.067 (12)	0.092 (14)	-
3	0.036 (8)	0.059 (11)	0.083 (14)	-	-
4	0.039 (9)	0.061 (11)	-	-	-
5	0.047 (10)	-	-	-	-
Viral infections					
Overall		$\bar{m} = 0.44$; $cr = 0.66$; Trend test $p = 0.14$			
1	0.005 (2)	0.013 (3)	0.015 (4)	0.015 (4)	0.016 (4)
2	0.008 (3)	0.020 (4)	0.026 (5)	0.030 (5)	-
3	0.019 (4)	0.042 (6)	0.049 (7)	-	-
4	0.027 (5)	0.056 (7)	-	-	-
5	0.031 (5)	-	-	-	-
CMV seropositive		$\bar{m} = 0.56$; $cr = 0.59$; Trend test $p = 0.20$			
1	0.007 (3)	0.017 (5)	0.021 (6)	0.022 (6)	0.023 (6)
2	0.011 (4)	0.028 (7)	0.037 (8)	0.042 (9)	-
3	0.027 (7)	0.058 (10)	0.071 (11)	-	-
4	0.036 (8)	0.077 (11)	-	-	-
5	0.044 (8)	-	-	-	-
CMV seronegative		$\bar{m} = 0.28$; $cr = 0.76$; Trend test $p = 0.16$			
1	0.002 (2)	0.007 (4)	0.007 (4)	0.007 (4)	0.007 (4)
2	0.004 (3)	0.011 (5)	0.011 (5)	0.014 (6)	-
3	0.010 (4)	0.020 (7)	0.020 (7)	-	-
4	0.014 (6)	0.028 (8)	-	-	-
5	0.014 (6)	-	-	-	-

[a] Average number of observed infections per subject
[b] Proportion of subjects without any infections
[c] Non-identifiable

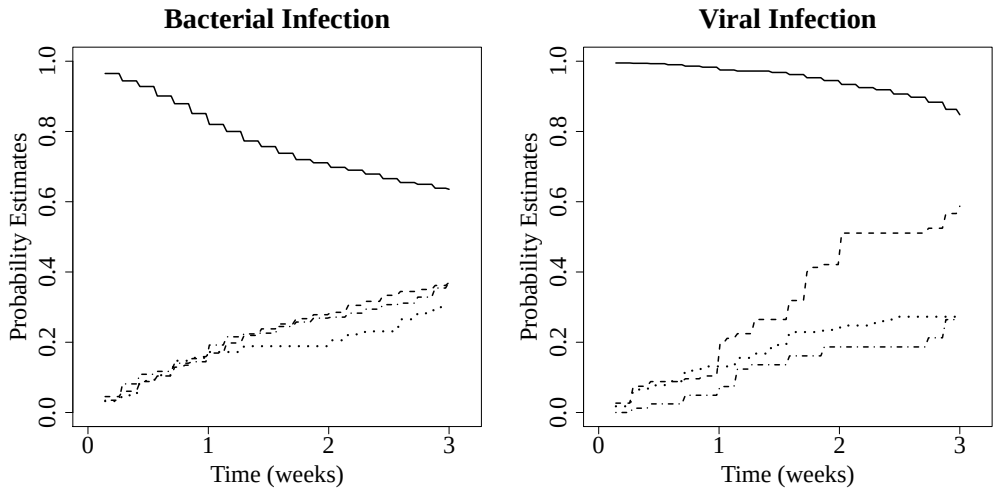

Figure 2.2: The estimated probabilities for the bacterial (left) and viral infection data (right). The solid line (—) is the estimated marginal survival probability of time to the first infection and the dashed (- -), dot-dashed ($-\cdot-$), and dotted ($\cdots$) lines are the conditional cumulative probability estimates of gap times between two consecutive infections given the time from transplant to the first infection in the first, second, and third week post transplant, respectively.

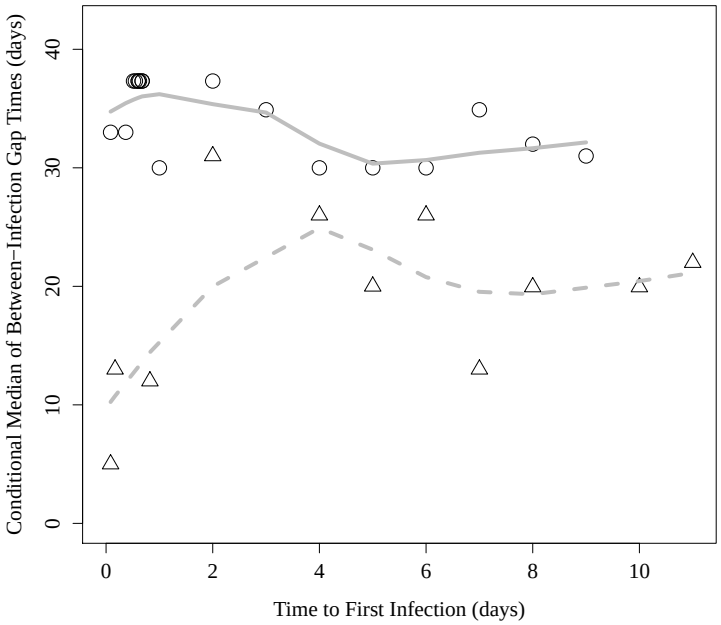

Figure 2.3: The conditional median estimates of the between-infection gap times given different length of time from transplant to the first infection. The circle is for bacterial infections, the triangle is for viral infections, and the solid line and the dashed line are the smooth curves for the two types of infections, respectively.

In the upper panel of Table 2.6, we also present the joint distribution estimates for bacterial infections stratified by pre-transplant regimen. The myeloablative group's probabilities were nearly doubled compared with the non-myeloablative group at all time points. We compared the two groups' joint distribution estimates using a nonparametric permutation test. The null hypothesis was that F_{X^0,Y^0} was the same for the two groups and the supremum test statistic had the form $D = \sup_{(x,y)} |\hat{F}^{(1)}_{X^0,Y^0}(x,y) - \hat{F}^{(2)}_{X^0,Y^0}(x,y)|$ with the superscript $g = 1, 2$ indexing the group assignment. The group index was randomly permuted among the patients 500 times to obtain the sampling distribution of D under the null hypothesis. We found that the difference between the myeloablative and non-myeloablative groups was highly significant ($p < 0.01$). For viral infections, we found that CMV positive serostatus was associated with higher joint probabilities than the CMV negative serostatus (Table 2.6, lower panel) with a significant p-value based on the permutation test ($p < 0.01$).

2.7 Concluding remarks

In this chapter, we developed a method to efficiently estimate the distribution of recurrent gap times while allowing the initial event to be different from the recurrent events. This design is more frequently encountered in prospective studies than the design considered in existing recurrent gap time data methods (e.g., Wang and Chang, 1999, and many others), where patients are enrolled due to an initial event of the same type as the recurrent events. We exploit the exchangeability structure of the gap times between consecutive, same-type events, typically adopted in other recurrent gap time methods (e.g., Wang and Chang, 1999), so that the proposed method could be more efficient than existing methods for bivariate serial event data, where only the first two gap times are used. Because the validity of the proposed method relies on the exchangeability property of the gap times beyond the first recurrent event, we suggest examining the exchangeability condition using the trend test of Wang and Chen (2000)'s before applying the proposed method. When this condition is violated, the method for ordered multivariate gap time data by Schaubel and Cai (2004) can be considered.

Similar to the nonparametric methods by Wang and Chang (1999) and Huang and Wang (2005), our proposed method does not impose distributional assumptions on the

frailty and leaves the within-subject correlation between gap times as a nuisance parameter. When the between-gap association is of interest, one can adopt the nonparametric method by Lakhal-Chaieb et al. (2010) based on Kendall's τ, by using the first two gap times. Alternatively, one can use semiparametric methods such as the bivariate copula model of Lawless and Yilmaz (2011) or the shared-frailty model of Huang and Liu (2007). However, the validity of the inference from these two methods depends on the correct specification of the form of the copula and the parametric distribution of the frailty, respectively.

Other nonparametric methods for bivariate serial event data such as the one proposed by Lin et al. (1999) can also be extended to handle recurrent infection data by using the weighted risk-set technique. For the bivariate case, it has been demonstrated by de Uña-Álvarez and Meira-Machado (2008), through simulation studies, that the Huang-Louis estimator can be more efficient than the Lin-Sun-Ying estimator. More theoretical investigation would be needed to desire general conclusions on their relative efficiency.

Similar to most recurrent event methods, the proposed method assumes the censoring time to be noninformative. However, with a longer follow-up time, death could become a nontrivial part of censoring, hence, the censoring time could be informative about the recurrent event of interest. In this case, modeling the death event jointly with the recurrent gap time data, as in the shared-frailty model proposed by Huang and Liu (2007), may be desired.

The primary focus of this chapter is on nonparametric estimation of the joint distribution of the gap times. Other nonparametric estimators based on the joint distribution estimator such as the conditional distribution and conditional quantile estimators are also provided. The nonparametric estimators proposed in this chapter are not only important in their own right, but also have important statistical applications. As demonstrated in Section 2.6, the proposed estimator $\hat{F}_{X^0,Y^0}$ enables us to make comparisons between subgroups of patients defined by potential risk factors, which can inform a more formal regression analysis.

To assess covariate effects on gap times, one may use Huang (2002)'s regression model for multistate gap time data, where the number of states to be analyzed is prespecified (e.g., two states if only the first two gaps are of interest). This method was originally

proposed for sequential events of different types, hence applying it on recurrent gap times with certain exchangeability property could be inefficient. Alternatively, at the cost of modeling the frailty distribution, one can use the model by Huang and Liu (2007) by including an episode-specific covariate to distinguish the covariate effect on the first gap (transplant to the first infection) from that on the subsequent gap times. We further discuss this in Chapter 3.

A direction of future research could be modeling multivariate recurrent gap time processes, such as the bacterial and viral infection processes jointly. We also note that an infectious event is not a transient event and an infectious episode can last for a certain period of time before being resolved. When both the infection time and infection-free time of an infectious episode are of interest, methods for alternating recurrent event processes by Huang and Wang (2005) and Chang (2004) can be considered.

Chapter 3

Semiparametric regression model for recurrent bacterial infections after hematopoietic stem cell transplantation

3.1 Introduction

Infections after hematopoietic stem cell transplantation (HSCT) are often a major source of mortality and morbidity among transplanted patients. During the early post-transplant period, bacterial infections are predominant among various infection types. Hence, characterizing the underlying early bacterial infection process and identifying risk factors are of primary interest in clinical practice. Our motivating data were from 516 patients who received their first HSCT at the University of Minnesota between 2000 and 2010 using unrelated umbilical cord blood (UCB) as a graft source. Transplanted patients were followed prospectively with infectious events recorded until the occurrence of disease relapse, a second transplant, death, or loss to follow-up. It is well-known that patients who undergo HSCT are at highest risk of infections prior to the engraftment of donor blood cells, which may require as long as 42 days. In this chapter, we focus on bacterial infections observed within 42 days after transplant. The aim of this study

is to identify important risk factors for early phase bacterial infections. Specifically, we are interested in the effect of patient- and transplant-related factors on time from transplant to the first bacterial infection and on the interoccurrence times (i.e., gap times) between one bacterial infection and the next recurrent infection.

As discussed in Wang and Chang (1999), recurrent gap time data have a distinctive structure and impose difficulties in modeling. In particular, the gap times beyond the first recurrent event time are subject to "induced dependent censoring" even when the overall censoring time is assumed to be independent to the recurrent event process. This is due to the correlation between gap times of the same subject. Also, it is noteworthy that the last censored gap times are likely to be longer than uncensored or completely observed gap times due to intercept sampling. Hence, conventional regression methods such as the Cox proportional hazards model or the accelerated failure time (AFT) model for univariate time-to-event data or multivariate clustered survival data are not applicable to recurrent gap time data. In Section 1.2, we discussed several existing regression methods for recurrent gap time data such as Huang and Chen (2003), Sun et al. (2006), Chang (2004, referred as to "Chang's method" hereafter), Lu (2005), and Strawderman (2005). However, these methods commonly assume that all events, including the initial event of the recurrent event process, are of the same type and all gap times including the first event time are identically distributed. Hence, applying these methods to our post-transplant infection data can lead to incorrect inferential results when the time from transplant to the first infection and the gap times between recurrent infections have different clinical significance. Therefore, it is important to account for the difference between the first event time and the following gap times. In Chapter 2, we studied a one-sample method which estimates the joint distribution of the time from transplant to the first infection and the gap times between consecutive, recurrent infections for the same data. However, to the best of our knowledge, there are no regression methods for recurrent gap time data under this setting.

In this chapter, we propose a semiparametric regression model which allows covariates to have different effects on time from transplant to the first infection and on gap times between two consecutive infections beyond the first infection. In particular, we assume that covariate effects are linearly related to the first event time or the gap times on a logarithmic scale. The proposed model is similar in form to the AFT model

for univariate survival data (Kalbfleisch and Prentice, 2002, Chapter 7, and references therein). As recognized in the literature, the AFT model can be more attractive than hazard-based models because of its direct interpretation of covariate effects on time variables. In the proposed model, we assume the existence of subject-specific latent random variables (i.e., frailties) to model within-subject correlation. The distribution of the latent vector is left unspecified, which distinguishes our approach from parametric frailty models such as the one considered by Huang and Liu (2007).

The estimation procedure for our model is motivated by the regression method for multistate gap time data proposed by Huang (2002, referred to as "Huang's method" hereafter). Note that Huang's method can be directly applied to our data by fixing the number of states to two (which then becomes a bivariate gap time method) and restricting the analysis to data up to the second infection. This approach will inevitably lead to loss of information because patients can experience more than two infections during the course of follow-up. In our data, for example, some patients experienced as many as six infections by day 42 after transplant. To make full use of the data, we extend Huang's method by applying the weighted risk-set method discussed in Luo and Huang (2011) to gap times beyond the first infection by taking advantage of the exchangeability property of gap times between two consecutive infections after the first infection.

The remainder of this chapter is organized as follows. In Section 3.2, we describe our proposed model. In Section 3.3, we first briefly review Huang's method for the simplified bivariate serial gap time data and then introduce our proposed model for recurrent infection data. Asymptotic properties of the proposed model are established in Section 3.4. In Section 3.5, we investigate the performance of the proposed method by conducting a series of simulation studies. In Section 3.6, we apply the proposed method to the University of Minnesota post-HSCT bacterial infection data. Concluding remarks are presented in Section 3.7.

3.2 Model setup

We first introduce notation to describe the recurrent infection process after transplant. Let X^0 denote the time from transplant to the first infection and Y_j^0, $j = 1, 2, \ldots$, the

gap times between two consecutive infections. The collection of all gap times of subject i, $i = 1, \ldots, n$ is denoted as $N_i = \{X_i^0, Y_{ij}^0, j = 1, 2, \ldots\}$. Let $\mathbf{A}_i$ denote a $p \times 1$ vector of baseline covariates collected at the time of transplantation. We assume that the log-transformed time from transplant to the first infection and the log-transformed gap times from one infection to the next are linearly related to the covariates, respectively, as

$$\begin{aligned} \log X_i^0 &= \gamma_{i0} + \mathbf{A}_i^T \boldsymbol{\beta}_0 + \epsilon_{i0} \\ \log Y_{ij}^0 &= \gamma_{i1} + \mathbf{A}_i^T \boldsymbol{\beta}_1 + \epsilon_{ij}, j = 1, 2, \ldots, \end{aligned}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are $p \times 1$ vectors of coefficients specific to the first event time and the following gap times, respectively; $(\gamma_{i0}, \gamma_{i1})$ is the subject-specific latent random vector shared by times from the same subject; and $\epsilon_{ij}, i = 1, \ldots, n, j = 0, 1, \ldots$ are mutually independent random errors with mean zero. The random vector $(\gamma_{i0}, \gamma_{i1})$ is used to account for heterogeneity among patients and the correlation between gap times within the same subject. The distribution of both the random vector and the error terms are left unspecified. Hence, the joint distribution of $(X_i^0, Y_{i1}^0, Y_{i2}^0, \ldots)$ is not completely specified, which renders the proposed model a semiparametric rather than a fully parametric model.

We note that in the proposed model, the distinctive $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ allow the same factor to have different effects on the first event time and the following gap times. Also note that under this model, we assume $Y_{ij}^0, j = 1, 2, \ldots$ are independently and identically distributed (i.i.d.) given $\mathbf{A}_i$ and $(\gamma_{i0}, \gamma_{i1})$. Without conditioning on $\mathbf{A}_i$ and $(\gamma_{i0}, \gamma_{i1})$, the Y_{ij}^0's are identically, but not independently distributed (i.e., exchangeable). In our post-HSCT bacterial infection study, we focus on early-stage infections within 42 days post transplant and expect no trend in such a short follow-up period in general, hence the exchangeability condition on the gap times between consecutive infections is a reasonable assumption for our data. In the upcoming sections, we will demonstrate that the exchangeability condition is a key condition for the proposed efficient estimation procedure.

Let C_i be the censoring time from transplant, whose survival function is $G(t) = \Pr(C_i > t)$ with a maximum support τ_C. We then define m_i as the number of observed

infections satisfying $X_i^0 + \sum_{j=1}^{m_i-1} Y_{ij}^0 \leq C_i$ and $X_i^0 + \sum_{j=1}^{m_i} Y_{ij}^0 > C_i$. The censoring time C_i is assumed to be independent of N_i, $(\gamma_{i0}, \gamma_{i1})$, and $\mathbf{A}_i$.

3.3 Estimation methods

3.3.1 Existing method for bivariate serial gap time data

To evaluate the covariate effects on time from transplant to the first infection and on gap times from one infection to the next, one can apply the multistate gap time method proposed by Huang (2002) to data up to the second infection by fixing the number of states (i.e., number of gap times) to two. In what follows, we briefly review this method and demonstrate how it can be applied to post-transplant infection data.

Define $Z_{i0}^0 = X_i^0$ and $Z_{i1}^0 = X_i^0 + Y_{i1}^0$ for times from transplant to the first and the second infections, respectively. For any two subjects indexed by i and i', their difference in covariates is denoted by $\mathbf{A}_{ii'} = \mathbf{A}_{i'} - \mathbf{A}_i$. The transformed times from transplant to the first and the second infections are defined as

$$\begin{aligned} Z_{ii'0}^0(\mathbf{b}_0) &= \exp(\mathbf{A}_{ii'}^T \mathbf{b}_0) X_i^0 \\ Z_{ii'1}^0(\mathbf{b}) &= \exp(\mathbf{A}_{ii'}^T \mathbf{b}_0) X_i^0 + \exp(\mathbf{A}_{ii'}^T \mathbf{b}_1) Y_{i1}^0, \end{aligned}$$

where $\mathbf{b} = (\mathbf{b}_0^T, \mathbf{b}_1^T)^T$, respectively, for $i, i' = 1, \ldots, n$. It follows that $Z_{ii'0}^0(\mathbf{b}_0)$ shares the same distribution with $Z_{i'0}^0$, and $Z_{ii'1}^0(\mathbf{b})$ with $Z_{i'1}^0$ when $\mathbf{b}_0 = \boldsymbol{\beta}_0$ and $\mathbf{b}_1 = \boldsymbol{\beta}_1$. By constructing the transformed time to the second infection as the sum of two transformed gap times (X_i^0 and Y_{i1}^0), the covariate effects on each gap time can be evaluated distinctively. Note that when $\mathbf{A}_i = \mathbf{A}_{i'}$, the transformed times reduce to Z_{i0}^0 and Z_{i1}^0. While the aim is to assess covariate effects on the length of interoccurrence times between events, introduction of time-to-event notation is necessary in order to properly address the problem of induced dependent censoring on gap times after the first infection. Now, consider bivariate vectors $\{Z_{i0}^0, Z_{ii'0}^0(\mathbf{b}_0)\}$ and $\{Z_{i1}^0, Z_{ii'1}^0(\mathbf{b})\}$. It is obvious that given $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, $\{Z_{i0}^0, Z_{ii'0}^0(\boldsymbol{\beta}_0)\}$ has the same distribution as $\{Z_{i'i0}^0(\boldsymbol{\beta}_0), Z_{i'0}^0\}$, denoted by $\{Z_{i0}^0, Z_{ii'0}^0(\boldsymbol{\beta}_0)\} \sim \{Z_{i'i0}^0(\boldsymbol{\beta}_0), Z_{i'0}^0\}$, and also $\{Z_{i1}^0, Z_{ii'1}^0(\boldsymbol{\beta})\} \sim \{Z_{i'i1}^0(\boldsymbol{\beta}), Z_{i'1}^0\}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$. Let $O_L(\cdot, \cdot)$ denote a symmetric and continuous function such that $O_L(t, s) = O_L(s, t)$ which satisfies the condition,

$O_L(t,s) = 0$ if $t \vee s \geq L$, where $a \vee b = \max(a,b)$, for $L < \tau_C$, and defines a functional mapping from two dimensional space to one. Then it naturally follows that, conditional on $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, $O_{L_0}\{Z^0_{i0}, Z^0_{ii'0}(\mathbf{b}_0)\} \sim O_{L_0}\{Z^0_{i'0}, Z^0_{i'i0}(\mathbf{b}_0)\}$, and $O_{L_1}\{Z^0_{i1}, Z^0_{ii'1}(\mathbf{b})\} \sim O_{L_1}\{Z^0_{i'1}, Z^0_{i'i1}(\mathbf{b})\}$ for constants $L_0 < \tau_C$ and $L_1 < \tau_C$ and $\mathbf{b} = \boldsymbol{\beta}$. This implies that when evaluated under the truth, $\mathrm{E}[w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_0)\mathbf{A}_{ii'}O_{L_0}\{Z^0_{i0}, Z^0_{ii'0}(\boldsymbol{\beta}_0)\}] = 0$ and $\mathrm{E}[w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_1)\mathbf{A}_{ii'}O_{L_1}\{Z^0_{i1}, Z^0_{ii'1}(\boldsymbol{\beta})\}] = 0$ where w is a continuous and symmetric scalar weight function satisfying $w(\mathbf{a}_1, \mathbf{a}_2, \mathbf{b}) = w(\mathbf{a}_2, \mathbf{a}_1, \mathbf{b})$ for a fixed $\mathbf{b}$.

Let the observed times from transplant to the first two infections and the corresponding censoring indicators be denoted as $Z_{i0} = Z^0_{i0} \wedge C_i$, $\Delta_{i0} = I(Z^0_{i0} \leq C_i)$, $Z_{i1} = Z^0_{i1} \wedge C_i$, and $\Delta_{i1} = I(Z^0_{i1} \leq C_i)$, where $a \wedge b = \min(a,b)$. The observed gap times are $X_i = Z_{i0}$ and $Y_{i1} = Z_{i1} - Z_{i0}$, respectively. The observed analogs of $Z^0_{ii'0}(\mathbf{b}_0)$ and $Z^0_{ii'1}(\mathbf{b})$ are then defined as

$$\begin{aligned} Z_{ii'0}(\mathbf{b}_0) &= \exp(\mathbf{A}_{ii'}^T\mathbf{b}_0)X_i, \\ Z_{ii'1}(\mathbf{b}) &= \exp(\mathbf{A}_{ii'}^T\mathbf{b}_0)X_i + \exp(\mathbf{A}_{ii'}^T\mathbf{b}_1)Y_{i1}, \end{aligned} \tag{3.1}$$

respectively. Recall that $G(\cdot)$ is the survival function for the censoring time. Then, under the independent censoring assumption, one can easily show that

$$\begin{aligned} \mathrm{E}\left[\left.\frac{\Delta_{i0}O_{L_0}\{Z_{i0}, Z_{ii'0}(\boldsymbol{\beta}_0)\}}{G(Z_{i0} \wedge L_0)}\right| \mathbf{A}_i, \mathbf{A}_{i'}\right] &= \mathrm{E}\left[O_{L_0}\{Z^0_{i0}, Z^0_{ii'0}(\boldsymbol{\beta}_0)\}\middle|\, \mathbf{A}_i, \mathbf{A}_{i'}\right], \text{ and} \\ \mathrm{E}\left[\left.\frac{\Delta_{i1}O_{L_1}\{Z_{i1}, Z_{ii'1}(\boldsymbol{\beta})\}}{G(Z_{i1} \wedge L_1)}\right| \mathbf{A}_i, \mathbf{A}_{i'}\right] &= \mathrm{E}\left[O_{L_1}\{Z^0_{i1}, Z^0_{ii'1}(\boldsymbol{\beta})\}\middle|\, \mathbf{A}_i, \mathbf{A}_{i'}\right]. \end{aligned}$$

It follows that

$$\begin{aligned} \mathrm{E}\left[\mathrm{E}\left[\left. w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_0)\mathbf{A}_{ii'}\frac{\Delta_{i0}O_{L_0}\{Z_{i0}, Z_{ii'0}(\boldsymbol{\beta}_0)\}}{G(Z_{i0} \wedge L_0)}\right| \mathbf{A}_i, \mathbf{A}_{i'}\right]\right] = 0, \text{ and} \\ \mathrm{E}\left[\mathrm{E}\left[\left. w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_1)\mathbf{A}_{ii'}\frac{\Delta_{i1}O_{L_1}\{Z_{i1}, Z_{ii'1}(\boldsymbol{\beta})\}}{G(Z_{i1} \wedge L_1)}\right| \mathbf{A}_i, \mathbf{A}_{i'}\right]\right] = 0. \end{aligned}$$

Then, the following estimating functions, which are U-statistics, are obtained in Huang

(2002):

$$D_0(\mathbf{b}_0) = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_0)\mathbf{A}_{ii'}\frac{\Delta_{i0}O_{L_0}\{Z_{i0}, Z_{ii'0}(\mathbf{b}_0)\}}{\hat{G}_0(Z_{i0}\wedge L_0)}, \tag{3.2}$$

$$D_1(\mathbf{b}) = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_1)\mathbf{A}_{ii'}\frac{\Delta_{i1}O_{L_1}\{Z_{i1}, Z_{ii'1}(\mathbf{b})\}}{\hat{G}_1(Z_{i1}\wedge L_1)}, \tag{3.3}$$

where $\hat{G}_0(t)$ and $\hat{G}_1(t)$ are the Kaplan-Meier estimators of the censoring time's survival function $G(t)$ using the data $\{(Z_{i0}, \Delta_{i0}), i = 1, \ldots, n\}$ and $\{(Z_{i1}, \Delta_{i1}), i = 1, \ldots, n\}$, respectively. Note that the artificial limits L_0 and L_1 are imposed to handle the case in which Z_{i0}^0 and Z_{i1}^0 have larger maximum support than τ_C and that subjects whose first or second infection time is censored only contribute to the denominator in functions (3.2) and (3.3) for the estimation of the censoring time survival function. The estimating equations $D_0(\mathbf{b}_0) = \mathbf{0}$ and $D_1(\mathbf{b}) = \mathbf{0}$ are solved inductively to obtain estimators for $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$.

As discussed in Huang (2002), popular log-rank based estimating equation approaches for the univariate AFT model (see Kalbfleisch and Prentice, 2002) can provide consistent estimates for $\boldsymbol{\beta}_0$, but not for $\boldsymbol{\beta}_1$ unless the time from transplant to the first infection and the gap time between the first and second infections are conditionally independent given the covariates. This strong independence condition is rarely true considering the heterogeneity in patients. The estimating equation based on the U-statistic function in (3.2) has been shown to be an efficient alternative to log-rank based estimating equation approaches by Huang (2002).

3.3.2 Proposed method for post-transplant recurrent infection data

To make better use of information beyond the second infection in our post-HSCT bacterial infection data, we propose to extend Huang's method for bivariate gap time data described in Section 3.3.1 by using the technique discussed in Luo and Huang (2011). It was demonstrated in Luo and Huang (2011) that the weighted risk-set method can be used to pool the exchangeable gap times within a subject to gain efficiency in estimation. This has been used in the one-sample estimation method for post-transplant

recurrent infection data in Section 2.3. We apply the technique to our proposed regression method in a similar fashion. To proceed, define the observed, uncensored gap times beyond the second infection by $Y_{ij} = Y_{ij}^0$ for $j = 2, \ldots, m_i^*$, where $m_i^* = m_i - 1$ and $m_i > 2$. Obviously, we have $\Delta_{ij} = 1$ for $j = 2, \ldots, m_i^*$. Under the assumptions made in Section 3.2, the observed uncensored gap times, $Y_{ij}, j = 1, \ldots, m_i^*$, are i.i.d. conditional on m_i, $(\gamma_{i0}, \gamma_{i1})$, and $\mathbf{A}_i$. It follows that the observed uncensored gap time pairs, $(X_i, Y_{ij}), j = 1, \ldots, m_i^*$, are also conditionally i.i.d. Under this condition, we can replace Z_{i1} with $Z_{ij} = X_i + Y_{ij}, j = 1, \ldots, m_i^*$, and the sum of the transformed gap times, $Z_{ii'1}(\mathbf{b})$ in (3.1) with $Z_{ii'j}(\mathbf{b}) = \exp(\mathbf{A}_{ii'}^T \mathbf{b}_0) X_i + \exp(\mathbf{A}_{ii'}^T \mathbf{b}_1) Y_{ij}$, for $j = 1, \ldots, m_i^*$, and prove that

$$\mathrm{E}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\mathrm{E}\left[\left.\frac{\Delta_{ij} O_{L_1}\{Z_{ij}, Z_{ii'j}(\mathbf{b})\}}{\hat{G}_1(Z_{ij} \wedge L_1)}\right| m_i, (\gamma_{i0}, \gamma_{i1}), \mathbf{A}_i\right]\right] = \mathrm{E}\left[\frac{\Delta_{i1} O_{L_1}\{Z_{i1}, Z_{ii'1}(\mathbf{b})\}}{\hat{G}_1(Z_{i1} \wedge L_1)}\right].$$

Hence, we propose to replace the estimation equation based on (3.3) with the following estimating equation for the estimation of $\boldsymbol{\beta}_1$:

$$D_1^*(\mathbf{b}) = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_1)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij} O_{L_1}\{Z_{ij}, Z_{ii'j}(\mathbf{b})\}}{\hat{G}_1(Z_{ij} \wedge L_1)}\right] = \mathbf{0}. \quad (3.4)$$

The estimator $\hat{\boldsymbol{\beta}}_1$ is derived by solving $D_1^*(\hat{\boldsymbol{\beta}}_0, \mathbf{b}_1) = 0$, where $\hat{\boldsymbol{\beta}}_0$ is the same as the one from the existing method discussed in Section 3.3.1. As discussed earlier, the last censored recurrent gap times are usually longer than uncensored gap times due to intercept sampling. Hence, the censored gap times are not used in Equation (3.4), unless $m_i = 1$, to avoid bias. Compared with $D_1(\mathbf{b})$ in (3.3), additional uncensored recurrent gap times beyond the second infection are used in the construction of (3.4), hence the proposed estimating method is expected to provide more efficient estimation on $\boldsymbol{\beta}_1$ than applying Huang's method on data up to the second infection.

Here, we choose $O_L(t, s) = \log[\{(t \vee s) \wedge L\}] - \log(L)$ and $w = 1$ to achieve numerical stability of the proposed estimation procedure. Specifically, with these functions, the estimating equation become monotone and a unique solution is attainable. Other choices for O_L and w have been also discussed in Huang (2002).

3.4 Asymptotic properties

The asymptotic properties for $\hat{\boldsymbol{\beta}}_0$ corresponding to the time from transplant to the first infection have been established in Huang (2002). Here, we focus on demonstrating the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}$ derived from D_1^*, given that of $\hat{\boldsymbol{\beta}}_0$. We begin by reviewing the large sample study for the bivariate gap time model in Huang (2002).

The estimating function in (3.3) can be rewritten as

$$D_1(\mathbf{b}) = \int_{t,s,\mathbf{a}_1,\mathbf{a}_2} w(\mathbf{a}_1,\mathbf{a}_2,\mathbf{b}_1)(\mathbf{a}_2-\mathbf{a}_1)\frac{O_{L_1}(t,s)}{\hat{G}_1(t\wedge L_1)}\hat{F}(\mathrm{d}t,\mathrm{d}s,\mathrm{d}\mathbf{a}_1;\mathbf{a}_2,\mathbf{b})\hat{H}(\mathrm{d}\mathbf{a}_2), \quad (3.5)$$

where $\hat{F}$ is the empirical estimator of the subdistribution function $F(t,s,\mathbf{a}_1;\mathbf{a}_2,\mathbf{b}) = \Pr[Z_{i1}\le t, \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_0\}X_i + \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_1\}Y_{i1}\le s, \mathbf{A}\le\mathbf{a}_1, \Delta_{i1}=1]$ and $\hat{H}$ is the empirical distribution function of $H(\mathbf{a}_2)=\Pr(\mathbf{A}\le\mathbf{a}_2)$. Through the properties of the components, $\hat{G}_1$, $\hat{F}$, and $\hat{H}$, it is shown that the functional D_1 is continuous and compactly differentiable.

Based on the functional representation of D_1, $D_1^T(\mathbf{b})(\mathbf{b}-\boldsymbol{\beta})$ converges almost surely and uniformly in $\mathbf{b}$ to

$$\mathrm{E}\left[\mathrm{E}\left[w(\mathbf{A}_i,\mathbf{A}_{i'},\mathbf{b}_1)\mathbf{A}_{ii'}^T(\mathbf{b}-\boldsymbol{\beta})O_{L_1}\{Z_{i1}^0, Z_{ii'1}^0(\mathbf{b})\}\right|\mathbf{A}_i,\mathbf{A}_{i'}\right]\right], \quad (3.6)$$

which equals $\mathrm{E}\big[\mathrm{E}\{w(\mathbf{A}_i,\mathbf{A}_{i'},\mathbf{b}_1)\mathbf{A}_{ii'}^T(\mathbf{b}-\boldsymbol{\beta})(O_{L_1}[Z_{i1}^0,\exp\{\mathbf{A}_{ii'}^T(\mathbf{b}-\boldsymbol{\beta})\}Z_{ii'1}^0(\boldsymbol{\beta})]$ $-O_{L_1}\{Z_{i1}^0, Z_{ii'1}^0(\boldsymbol{\beta})\})\big|\mathbf{A}_i,\mathbf{A}_{i'}\}\big]$. Since the proposed estimating function D_1^* in (3.4) converges uniformly to the same limit as D_1 in (3.3), $D_1^{*T}(\mathbf{b})(\mathbf{b}-\boldsymbol{\beta})$ also converges to (3.6). Equation (3.6) equals 0 if $\mathbf{b}=\boldsymbol{\beta}$. Thus, given the consistency of $\hat{\boldsymbol{\beta}}_0$, the consistency of $\hat{\boldsymbol{\beta}}$ follows.

The asymptotic normality of $\hat{\boldsymbol{\beta}}$ can be derived from that of $D(\boldsymbol{\beta})$ and the asymptotic linearity of $D(\mathbf{b})$ at $\mathbf{b}=\boldsymbol{\beta}$, where $D(\mathbf{b})\equiv\{D_0^T(\mathbf{b}_0), D_1^{*T}(\mathbf{b})\}^T$. Using the compact differentiability of (3.5), Huang (2002) shows that by the functional delta method $n^{1/2}\{D_0^T(\boldsymbol{\beta}_0), D_1^T(\boldsymbol{\beta})\}^T$ converges weakly to a normal distribution with mean zero and

variance estimated by $\sum_{i=1}^{n}\left\{\xi_{i0}^{T}(\boldsymbol{\beta}_0),\xi_{i1}^{T}(\boldsymbol{\beta})\right\}^{T}\left\{\xi_{i0}^{T}(\boldsymbol{\beta}_0),\xi_{i1}^{T}(\boldsymbol{\beta})\right\}$ where

$$\begin{aligned}\xi_{i0}(\boldsymbol{\beta}_0) &= n^{-3/2}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_0)\mathbf{A}_{ii'}\left[\frac{\Delta_{i0}O_{L_0}\{Z_{i0},Z_{ii'0}(\boldsymbol{\beta}_0)\}}{\hat{G}_0(Z_{i0}\wedge L_0)}-\frac{\Delta_{i'0}O_{L_0}\{Z_{i'0},Z_{i'i0}(\boldsymbol{\beta}_0)\}}{\hat{G}_0(Z_{i'0}\wedge L_0)}\right]\\ &\quad +n^{-3/2}\int_0^{L_0}\frac{U_0(t,\boldsymbol{\beta}_0)\hat{G}_0(t-)}{Y_0(t)\hat{G}_0(t)}\mathrm{d}\hat{M}_{i0}(t)\end{aligned}$$

and

$$\begin{aligned}\xi_{i1}(\boldsymbol{\beta}) &= n^{-3/2}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_1)\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}O_{L_1}\{Z_{i1},Z_{ii'1}(\boldsymbol{\beta})\}}{\hat{G}_1(Z_{i1}\wedge L_1)}-\frac{\Delta_{i'1}O_{L_1}\{Z_{i'1},Z_{i'i1}(\boldsymbol{\beta})\}}{\hat{G}_1(Z_{i'1}\wedge L_1)}\right]\\ &\quad +n^{-3/2}\int_0^{L_1}\frac{U_1(t,\boldsymbol{\beta})\hat{G}_1(t-)}{Y_1(t)\hat{G}_1(t)}\mathrm{d}\hat{M}_{i1}(t),\end{aligned}$$

in which,

$$\begin{aligned}U_0(t,\boldsymbol{\beta}_0) &= \sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_0)\mathbf{A}_{ii'}\left[\frac{\Delta_{i0}O_{L_0}\{Z_{i0},Z_{ii'0}(\boldsymbol{\beta}_0)\}}{\hat{G}_0(Z_{i0}\wedge L_0)}I(Z_{i0}>t)\right],\\ U_1(t,\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_1)\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}O_{L_1}\{Z_{i1},Z_{ii'1}(\boldsymbol{\beta})\}}{\hat{G}_1(Z_{i1}\wedge L_1)}I(Z_{i1}>t)\right],\end{aligned}$$

$Y_k(t)=\sum_{i=1}^{n}I(Z_{ik}\geq t)$, $\hat{M}_{ik}(t)=I(Z_{ik}\leq t,\Delta_{ik}=0)-\int_0^t I(Z_{ik}\geq s)\mathrm{d}\hat{\Lambda}_k(s)$ for $k=0,1$, and $\hat{\Lambda}_k$ is the Nelson-Aalen estimator corresponding to $\hat{G}_k$. It is easy to observe that D_1^* share the same properties as D_1, i.e., continuous and compactly differentiable. Thus, by the functional delta method, $n^{1/2}D(\boldsymbol{\beta})$ is asymptotically normal with mean zero and variance Ω which can be estimated by applying the weighted risk-set technique to the variance estimate derived in Huang (2002). Define

$$\begin{aligned}\xi_{i1}^{*}(\boldsymbol{\beta}) &= n^{-3/2}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_1)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}O_{L_1}\{Z_{ij},Z_{ii'j}(\boldsymbol{\beta})\}}{\hat{G}_1(Z_{ij}\wedge L_1)}\right.\\ &\quad \left.-\frac{1}{m_{i'}^*}\sum_{l=1}^{m_{i'}^*}\frac{\Delta_{i'l}O_{L_1}\{Z_{i'l},Z_{i'il}(\boldsymbol{\beta})\}}{\hat{G}_1(Z_{i'l}\wedge L_1)}\right]+n^{-3/2}\int_0^{L_1}\frac{U_1^*(t,\boldsymbol{\beta})\hat{G}_1(t-)}{Y_1^*(t)\hat{G}_1(t)}\mathrm{d}\hat{M}_{i1}^*(t),\end{aligned}$$

where

$$U_1^*(t,\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_1)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}O_{L_1}\{Z_{ij},Z_{ii'j}(\boldsymbol{\beta})\}}{\hat{G}_1(Z_{ij}\wedge L_1)}I(Z_{ij}>t)\right],$$

$$Y_1^*(t) = \sum_{i=1}^{n}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij}\geq t),$$

and

$$\hat{M}_{i1}^*(t) = \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij}\leq t,\Delta_{ij}=0) - \int_0^t \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij}\geq s)\mathrm{d}\hat{\Lambda}_1(s).$$

Note that the "weighted average version" of components, ξ_{i1}^*, U_1^*, Y_1^*, and $\hat{M}_{i1}^*$ converges uniformly to the same limit as their counterparts, ξ_{i1}, U_1, Y_1, $\hat{M}_{i1}$ by the exchangeability property of $Y_{ij}^0, j=1,2,\ldots$. Then the variance estimate for $D(\boldsymbol{\beta})$ is established by replacing ξ_{i1} with ξ_{i1}^* as follows,

$$\hat{\Omega} = \sum_{i=1}^{n}\{\xi_{i0}^T(\boldsymbol{\beta}_0),\xi_{i1}^{*T}(\boldsymbol{\beta})\}^T\{\xi_{i0}^T(\boldsymbol{\beta}_0),\xi_{i1}^{*T}(\boldsymbol{\beta})\}.$$

We note that ξ_{i0}, which corresponds to the first event time, is equivalent to the one derived for the bivariate gap time model. Since $\hat{\boldsymbol{\beta}}$ is shown to be consistent for $\boldsymbol{\beta}$, the consistency of the variance estimate $\hat{\Omega}$ follows from the fact that $\sum_{i=1}^n\{\xi_{i0}^T(\mathbf{b}_0),\xi_{i1}^{*T}(\mathbf{b})\}^T$ $\{\xi_{i0}^T(\mathbf{b}_0),\xi_{i1}^{*T}(\mathbf{b})\}$ converges uniformly and almost surely in $\mathbf{b}$ to a limiting function which is continuous at $\mathbf{b}=\boldsymbol{\beta}$ by the Glivenko-Cantelli theorem of Pollard (1984).

We proceed to establish the asymptotic linearity of $D(\mathbf{b})$. For $\mathbf{b}$ converging to $\boldsymbol{\beta}$, it is shown in Huang (2002) that $D_1(\mathbf{b}) = \tilde{D}_1(\mathbf{b}) + o_p(||\mathbf{b}-\boldsymbol{\beta}|| + n^{-1/2})$, where

$$\tilde{D}_1(\mathbf{b}) = \int_{t,s,\mathbf{a}_1,\mathbf{a}_2} w(\mathbf{a}_1,\mathbf{a}_2,\boldsymbol{\beta}_1)(\mathbf{a}_2-\mathbf{a}_1)\frac{O_{L_1}(t,s)}{\hat{G}_1(t\wedge L_1)}\hat{F}(\mathrm{d}t,\mathrm{d}s,\mathrm{d}\mathbf{a}_1;\mathbf{a}_2,\mathbf{b})\hat{H}(\mathrm{d}\mathbf{a}_2). \quad (3.7)$$

We define $\tilde{D}_1^*$ by replacing $\hat{F}(t,s,\mathbf{a}_1;\mathbf{a}_2,\mathbf{b})$ in (3.7) with its weighted average version of

the estimator

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I\left[Z_{ij}\le t, \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_0\}X_i+\exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_1\}Y_{ij}\le s, \mathbf{A}\le\mathbf{a}_1, \Delta_{ij}=1\right].$$

Then, it follows naturally that

$$D_1^*(\mathbf{b}) = \tilde{D}_1^*(\mathbf{b}) + o_p(\parallel \mathbf{b}-\boldsymbol{\beta} \parallel + n^{-1/2})$$

for $\mathbf{b}$ converging to $\boldsymbol{\beta}$. The derivative matrix of $\tilde{D}(\mathbf{b}) = \{\tilde{D}_0^T(\mathbf{b}_0), \tilde{D}_1^{*T}(\mathbf{b})\}^T$, where $\tilde{D}_0$ corresponding to the first event time can be expressed in a similar fashion as $\tilde{D}_1$, is denoted by

$$\hat{\Sigma} = \frac{\partial}{\partial \mathbf{b}}\tilde{D}(\mathbf{b}).$$

Note that nondifferentiable points for $\tilde{D}(\mathbf{b})$ exist in $\mathbf{b}$. To accommodate nondifferentiable functions, Huang (2000) suggests the generalized law of the mean. By applying the generalized law of the mean and by the fact that there exist left and right partial derivatives for $\tilde{D}(\mathbf{b})$ and that both have the same limit Σ (Pollard, 1984), we derive

$$\begin{aligned} D(\mathbf{b}) &= \tilde{D}(\mathbf{b}) + o_p(||\mathbf{b}-\boldsymbol{\beta}|| + n^{-1/2}) \\ &= D(\boldsymbol{\beta}) + \Sigma(\mathbf{b}-\boldsymbol{\beta}) + o_p(||\mathbf{b}-\boldsymbol{\beta}|| + n^{-1/2}) \end{aligned}$$

for $\mathbf{b}$ converging to $\boldsymbol{\beta}$. The asymptotic normality and the linearity of $D(\boldsymbol{\beta})$ naturally lead to the asymptotic normality of $\hat{\boldsymbol{\beta}}$. The asymptotic variance of $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ can be consistently estimated by $\hat{\Sigma}^{-1}\hat{\Omega}(\hat{\Sigma}^{-1})^T$.

3.5 Simulation studies

We conduct a series of simulation studies to evaluate the performance of the proposed method, each with 1000 datasets and $n = 200$ subjects per dataset. We generate time to the first infection and gap times between two consecutive infections for each subject

from the following model

$$\begin{aligned} \log(X_i^0) &= \gamma_{i0} + \mathbf{A}_i^T \boldsymbol{\beta}_0 + \epsilon_{i0} \\ \log(Y_{ij}^0) &= \gamma_{i1} + \mathbf{A}_i^T \boldsymbol{\beta}_1 + \epsilon_{ij}, j = 1, 2, \ldots, \end{aligned}$$

respectively, where $\mathbf{A}_i = (A_{i1}, A_{i2})^T$ and A_{i1} is sampled from Bernoulli(0.5) and A_{i2} is from Uniform[0, 1]. The true covariate effects are set as $\boldsymbol{\beta}_0 = (-0.5, 0.5)^T$ and $\boldsymbol{\beta}_1 = (0.5, 0.5)^T$. We generate the mutually independent error terms ϵ_{ij} from a normal distribution with mean zero and variance equal to 0.1 and the subject-specific latent vector $(\gamma_{i0}, \gamma_{i1})$ from a bivariate normal distribution with mean zero and covariance matrix $\begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$. Note that ρ accounts for the degree of association between the (transformed) time to the first infection, $\log(X_i^0)$, and one of the (transformed) gap times after the first infection, $\log(Y_{ij}^0)$, and σ_1 indicates the level of correlation between two (transformed) gap times after the first infection, $\log(Y_{ij}^0)$ and $\log(Y_{ij'}^0)$. We set $\sigma_0^2 = 0.01$ or 0.1, $\sigma_1^2 = 0.01$ or 0.1, and $\rho = 0$ or 0.5 in different scenarios. The censoring time $C_i, i = 1, \ldots, n$ is sampled from a uniform distribution $[0, U]$, where $U = 10$.

We applied the proposed method to the simulated data with the bootstrap method used for variance estimation. The bootstrap sampling size is set as $B = 200$. In the proposed estimation method, we select constant values smaller than the largest observed follow-up time of Z_{i0} and Z_{i1} for L_0 and L_1, respectively. For comparison, we also applied Huang's method and Chang's method. The variance estimates for these methods were derived by bootstrap sampling with $B = 200$. Table 3.1 summarizes the simulation results. Under all settings, we observe that both the proposed method and Huang's method are virtually unbiased and the Monte-Carlo standard deviations are close to the average bootstrap standard errors (SEs). Note that the two methods share the same estimator for the covariate effects on time to the first infection ($\boldsymbol{\beta}_0$). However, the proposed method yields more efficient results than Huang's method in the estimation of covariate effects on gap times after the first infection ($\boldsymbol{\beta}_1$) in all settings.

As expected, biased results are obtained from Chang's method, which assumes that all gap times, including the time from transplant to the first infection, are identically distributed. Specifically, it fails to capture the different effects of a covariate (A_1) on two

Table 3.1: Summary of simulation results with true coefficients; Monte-Carlo mean of the point estimates (Mean); Monte-Carlo standard deviation (SD$\times 10^3$); and the average bootstrap standard error (SE$\times 10^3$) of the proposed estimator (Proposed) and Huang's method (Huang) for time from transplant to the first infection and gap time(s) after the first infection, and Chang's method (Chang) for all gap times after transplant pooled together.

	Time to 1st infection		Gap times after 1st infection					
	Huang/Proposed		Proposed		Huang		Chang	
Covariate	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
True $\boldsymbol{\beta}$	-0.5	0.5	0.5	0.5	0.5	0.5	-[a]	-
$\sigma_0^2 = 0.01$; $\sigma_1^2 = 0.01$								
$\rho = 0$	$\bar{m}$[b]= 2.89; cr_1[c]= 0.11; cr_2[d]= 0.29							
Mean	-0.501	0.505	0.499	0.496	0.499	0.497	0.148	0.426
SD	60	111	84	149	88	158	49	78
SE	60	108	82	137	88	147	49	78
$\rho = 0.5$	$\bar{m} = 2.89$; $cr_1 = 0.11$; $cr_2 = 0.29$							
Mean	-0.500	0.509	0.494	0.504	0.495	0.505	0.146	0.427
SD	59	108	86	148	90	156	50	78
SE	61	108	83	139	88	148	49	79
$\sigma_0^2 = 0.1$; $\sigma_1^2 = 0.01$								
$\rho = 0$	$\bar{m} = 2.86$; $cr_1 = 0.11$; $cr_2 = 0.29$							
Mean	-0.497	0.497	0.499	0.496	0.501	0.495	0.120	0.437
SD	78	144	89	153	93	158	55	90
SE	79	141	86	145	91	154	55	88
$\rho = 0.5$	$\bar{m} = 2.87$; $cr_1 = 0.11$; $cr_2 = 0.30$							
Mean	-0.502	0.498	0.450	0.493	0.501	0.491	0.114	0.434
SD	78	143	92	153	97	159	57	88
SE	79	140	88	149	93	157	56	89
$\sigma_0^2 = 0.1$; $\sigma_1^2 = 0.1$								
$\rho = 0$	$\bar{m} = 2.98$; $cr_1 = 0.11$; $cr_2 = 0.30$							
Mean	-0.501	0.509	0.499	0.492	0.499	0.490	0.069	0.417
SD	81	147	113	192	117	196	62	107
SE	79	141	108	183	111	190	61	103
$\rho = 0.5$	$\bar{m} = 2.99$; $cr_1 = 0.12$; $cr_2 = 0.30$							
Mean	-0.501	0.505	0.503	0.494	0.504	0.494	0.073	0.420
SD	80	141	118	186	122	193	62	109
SE	79	141	109	185	113	192	62	105

[a]True $\boldsymbol{\beta}$ values do not exist
[b]Average number of observed infections per subject
[c]Average proportion of subjects without any infections
[d]Average proportion of subjects with the first or the second gap times censored

different types of time variables, the time from transplant to the first infection time and the following gap times (-0.5 and 0.5, respectively). Based on our simulation setting, covariate $A_1 = 1$ is associated with shorter time from transplant to the first infection, but prolonged gap time from one infection to the next. By using Chang's method, this distinction is obscured and the overall effect of A_1 is diminished (ranging from 0.06 to 0.15). In our simulated setting, covariate A_2's effect is set to be the same for the two types of time variable (0.5 for both), but the estimated effect of this variable on the pooled gap times based on Chang's method is found to be biased from 0.5 (ranging from 0.41 to 0.44). This suggests that if one of the covariates in Chang's model has a different effect on time to first infection than it has on gap times, the estimate of the effects of other covariates which do not have differential effects would also be affected.

3.6 Application

To illustrate the proposed estimation method, we analyzed the post-HSCT bacterial infection data introduced in Section 3.1. The data are composed of 516 HSCT recipients who used unrelated UCB as a graft source. Since we are interested in the incidence and characteristics of infections after HSCT for both pediatric and adult patients (Saavedra et al., 2002, Barker et al., 2005, Yazaki et al., 2009), we stratify the data to two groups: patients who were younger than 18 (155 patients, 30%) and those 18 or older (361 patients, 70%) at the time of transplant. Table 3.2 summarizes patient- and transplant-related characteristics for the overall group, and for the pediatric and adult cohorts separately.

We focus on early phase bacterial infections experienced within 42 days of transplant. The majority of patients (89%) were censored by the 42 day cut-off, whereas 25 (5%) patients were censored by death, 21 (4%) by relapse, and 10 (2%) by a second transplant before day 42. Among the 25 deaths, only 7 were related to infection, of whom 3 ($< 1\%$ of all patients) were related to bacterial infection. Hence, we do not expect a serious violation of the independent censoring assumption in our data. Infectious episodes were documented in the Blood and Marrow Transplant Database at the University of Minnesota according to the criteria described by Barker et al. (2005). A total of 397 bacterial infectious episodes were observed for all patients, 86 in children and 311 in

Table 3.2: Summary of patient- and transplant-related characteristics.

Variables	No. Patients (%) / Median (Range)[†] All Patients	Children (Age < 18)	Adults (Age ≥ 18)
N	516	155	361
Age at TX	36.9 (0.5−71.4)[†]	9.4 (0.5−17.9)[†]	47.4 (18.1−71.4)[†]
Gender			
Male	304 (59)	100 (65)	204 (57)
Female	212 (41)	55 (35)	157 (43)
Diagnosis			
ALL	131 (25)	67 (43)	64 (18)
AML	217 (42)	63 (41)	154 (43)
CML	19 (4)	1 (1)	18 (5)
Hodgkin's Lymphoma	7 (1)	1 (1)	6 (2)
Multiple Myeloma	1 (0)	0 (0)	1 (0)
Myelodysplastic Syndrome	45 (9)	9 (6)	36 (10)
Myeloproliferative Neoplasm	10 (2)	0 (0)	10 (3)
Neuroblastoma	1 (0)	1 (1)	0 (0)
Non-Hodgkin's Lymphoma	59 (11)	6 (4)	53 (15)
Other Leukemia	21 (4)	7 (5)	14 (4)
Other Malignancy	5 (1)	0 (0)	5 (1)
CMV Serostatus			
Positive	301 (58)	100 (65)	201 (56)
Negative	215 (41)	55 (35)	160 (44)
Type of Transplant			
Double Cord	374 (72)	60 (39)	314 (87)
Single Cord	142 (28)	95 (61)	47 (13)
Conditioning Regimen			
Myeloablative	281 (54)	150 (97)	131 (36)
Non-Myeloablative w ATG	67 (13)	0 (0)	67 (19)
Non-Myeloablative wo ATG	168 (33)	5 (3)	163 (45)
HLA Locus Matching Score			
4/6	262 (51)	44 (28)	218 (60)
5/6	202 (39)	86 (55)	116 (32)
6/6	52 (10)	25 (16)	27 (7)
GVHD Prophylaxis			
CSA/MMF/MTX	449 (87)	104 (67)	344 (95)
Other	67 (13)	51 (33)	16 (4)
CD34+ graft infused ($\times 10^6$/kg)[‡]	0.49 (0.06−27.53)[†]	0.58 (0.06−8.42)[†]	0.47 (0.07−27.53)[†]
Low	130 (25)	35 (23)	95 (26)
High	386 (75)	120 (77)	266 (74)
TNC dose infused ($\times 10^8$/kg)[‡]	0.38 (0.11−4.89)[†]	0.48 (0.15−2.27)[†]	0.36 (0.11−4.89)[†]
Low	139 (27)	29 (19)	110 (30)
High	377 (73)	126 (81)	251 (70)

Abbreviations: TX = transplant; ALL = acute lymphoblastic leukemia; AML = acute myeloblastic leukemia; CML = chronic myeloid leukemia; CMV = cytomegalovirus; ATG = anti-thymocyte globulin; HLA = human leukocyte antigen; GVHD = graft-versus-host disease; CSA = cyclosporin; MMF = mychophenolate mofetil; MTX = methotrexate; TNC = total nucleated cell.
[‡]High: dose $> 1^{st}$ quartile; low: dose $\leq 1^{st}$ quartile.

Table 3.3: Summary of number of patients who experienced k number of bacterial infections within 42 days after transplant, $k = 0, 1, \cdots, 6$.

		No. of infections observed for a patient						
Group	No. patients (%)	0	1	2	3	4	5	6
All Patients	516	266	152	69	15	10	2	2
	(100)	(51.6)	(29.5)	(13.4)	(2.9)	(1.9)	(0.4)	(0.4)
Children	155	92	45	14	3	1	0	0
(Age < 18)	(100)	(59.4)	(29.0)	(9.0)	(1.9)	(0.7)	(0.0)	(0.0)
Adults	361	174	107	55	12	9	2	2
(Age ≥ 18)	(100)	(48.2)	(29.6)	(15.2)	(3.3)	(2.5)	(0.6)	(0.6)

adults during the first 42 days after transplant. On average, each patient experienced 0.77 infections, with children experiencing fewer infections than adults (0.55 vs. 0.86). The detailed summary of the infections can be found in Table 3.3. About 59% of pediatric patients and 48% of adult patients experienced no infections. To assure that the gap times after the first infection were similarly distributed, we carried out the trend test by Wang and Chen (2000) for each patient group. We found no evidence of trend in these gap times (p-value = 1.00 and 0.51 for children and adults, respectively). Hence, the exchangeability condition is a reasonable assumption for the gap times after the first infection in our data.

First, we considered univariate regression models to identify potential risk factors. Results are presented in Table 3.4. We found that older children had significantly shorter time to the first bacterial infection than younger children ($\hat{\beta} = -0.09$, SE = 0.03). Children who received UCB cells from two donors experienced their first bacterial infection earlier than single donor UCB recipients ($\hat{\beta} = -0.81$, SE = 0.28). High total nucleated cell (TNC) dose level ($> 0.34 \times 10^8$/kg, the 1$^{\text{st}}$ quartile among children) was associated with prolonged time to the first bacterial infection for children ($\hat{\beta} = 1.02$, SE = 0.34). No factors were found to be significantly associated with gap times between two consecutive infections for children. For adult patients, older age was associated with longer time to the first bacterial infection ($\hat{\beta} = 0.03$, SE = 0.01). As compared to a myeloablative regimen, a non-myeloablative regimen without anti-thymocyte globulin (ATG) was associated with longer time to the first bacterial infection ($\hat{\beta} = 1.26$, SE = 0.26) and longer gap times from one bacterial infection to the next ($\hat{\beta} = 0.85$, SE = 0.36)

Table 3.4: Summary of univariate regression analysis of risk factors for early bacterial infections for children and adults with estimated regression coefficients (p-values of Wald test based on the bootstrap standard error).

	Children		Adults	
Variables	1^{st} gap	$\geq 2^{nd}$ gap	1^{st} gap	$\geq 2^{nd}$ gap
Age at Transplant (Years)	-0.086 (0.002*)	-0.046 (0.486)	0.031 (0.001*)	0.007 (0.560)
CMV Serostatus				
Positive vs. Negative	-0.488 (0.191)	0.666 (0.139)	-0.099 (0.674)	-0.009 (0.974)
Type of Transplant				
Double vs. Single	-0.806 (0.007*)	-0.078 (0.889)	-0.381 (0.340)	-0.493 (0.360)
Conditioning Regimen (vs. Myeloablative)				
Non-myeloablative w ATG	NI	NI	0.504 (0.139)	0.181 (0.713)
Non-myeloablative wo ATG	NI	NI	1.257 (<0.001*)	0.854 (0.015*)
HLA Match Score				
5-6/6 vs. 4/6	-0.739 (0.060)	-0.536 (0.623)	0.309 (0.205)	0.187 (0.518)
GVHD Prophylaxis				
CSA/MMF/MTX vs. Other	-0.735 (0.068)	0.337 (0.606)	-0.292 (0.664)	-1.321 (0.488)
CD34+ Dose Level				
High vs. Low	-0.097 (0.806)	-1.717 (0.131)	0.050 (0.865)	-0.443 (0.224)
TNC Dose Level				
High vs. Low	1.021 (0.003*)	-0.288 (0.686)	-0.128 (0.659)	-0.185 (0.590)

*P-value < 0.05.
NI: Conditioning regimen was not included in the model for pediatric patients since 97% of children in our data received myeloablative conditioning regimen.

for adult patients. Other factors such as cytomegalovirus serostatus, human leukocyte antigen matching, graft-versus-host disease prophylaxis, and CD34 dose infused were not found to be associated with either type of time variable for either patient cohort.

Multivariable regression analyses were conducted with significant variables identified from the univariate analysis. The results (see Table 3.5) are consistent with the univariate analyses except that the age variable is no longer a significant risk factor for child or adult patients. The loss of significance of age may be due to the confounding of other factors. For example, we found that older adult patients were more likely to receive non-myeloablative conditioning regimen than younger adult patients, while

Table 3.5: Summary of multivariable regression analysis of risk factors for early bacterial infections for children and adults with estimated regression coefficients (p-values of Wald test based on the bootstrap standard error).

	Children		Adults	
Variables	1^{st} gap	$\geq 2^{nd}$ gap	1^{st} gap	$\geq 2^{nd}$ gap
Age at Transplant (Years)	-0.018 (0.644)	-0.069 (0.331)	0.013 (0.386)	-0.015 (0.532)
Type of Transplant				
Double vs. Single	-1.050 (0.014*)	-0.052 (0.958)	NI	NI
Conditioning Regimen (vs. Myeloablative)				
Non-Myeloablative w ATG	NI	NI	0.188 (0.724)	0.564 (0.480)
Non-Myeloablative wo ATG	NI	NI	1.013 (0.011*)	1.163 (0.080)
TNC Dose Level				
High vs. Low	1.250 (0.019*)	-0.408 (0.681)	NI	NI

*P-value < 0.05.
NI: Variables not included in the multivariable model due to the lack of significance in univariate analyses.

non-myeloablative regimen was associated with earlier engraftment and hence fewer bacterial infections. For children, we found that age was associated with both the type of transplant and TNC dose. Specifically, double umbilical cord blood stem cells were used more frequently for older children and older children tended to require higher TNC dose than younger children in our data.

3.7 Concluding remarks

In this chapter, we proposed a semiparametric regression model for recurrent gap time data which allows covariates to have different effects on the first event time and on the following gap times. In our data, a patient's recurrent infection process was initiated by the event of transplant, which is a different type of event than the recurrent events (i.e., infections). Hence, the first event time (i.e., time from transplant to the first infection) and the following gap times (i.e., gap times between two consecutive infections) may have different clinical significance and should be modeled differently. Unlike many existing recurrent gap time regression models (e.g., Huang and Chen, 2003), our proposed model has the flexibility to assess the potentially different covariate effects on the two different

types of gap times. Note that our proposed method still needs the exchangeability condition on the gap times between the same-type recurrent events as many existing recurrent gap time models. Hence, it is advised to examine this condition using the trend test of Wang and Chen (2000) before applying the proposed method as we have demonstrated.

When the exchangeability condition on recurrent gap times beyond the first infection time is not satisfied, one can apply the multistate gap times model of Huang (2002) to the data. The covariate effects on gap times between two consecutive infections are not constrained to be the same in the model. Note that the number of states in Huang's method, which corresponds to the number of infections in our case, needs to be pre-specified. If the number of states is large, however, the events of higher states may become rare, which could result in inefficient estimation.

Our study focuses on early phase infections after transplantation and on the effect of factors which do not vary over time, e.g., patient- and transplant-related characteristics. When a longer follow-up period is of interest, the recurrent gap times' structure may become more complex and in particular it might be affected by time-varying variables. Research in extending the model to handle time-dependent covariates is warranted. In addition, informative censoring may become a nontrivial issue in a study with longer follow-up time. In this case, informative censoring events such as death can be modeled jointly with the recurrent infection process using the method considered by Huang and Liu (2007).

Chapter 4

Semiparametric regression model for bivariate alternating recurrent gap time data

4.1 Introduction

In clinical studies, recurrent event data are frequently encountered where patients experience an event repeatedly over time. These recurrent events often consist of alternating states such as patients' admission to and discharge from hospital. Another example is the relapse and remission of a recurring disease. This type of data is referred to as bivariate alternating recurrent event data. The two states of a recurrence episode could each carry important and distinct information about the underlying health condition of a patient or the severity of a disease. It is of primary interest to develop statistical methods which can make full use of the information in this kind of data.

When analyzing bivariate alternating recurrent event data, the sequential structure of recurrent events imposes challenges such as induced dependent censoring and intercept bias as for other serial event gap time data (Huang and Wang, 2005). In the literature, many studies have been conducted of gap time data analysis that properly accounts for these challenges. Huang and Wang (2005) proposed a nonparametric method for estimating the joint distribution of the durations of the two states of a bivariate

alternating recurrent event process. The nonparametric estimator serves as a basis for understanding the underlying recurrent event process. However, regression methods would be more attractive to researchers who are interested in identifying which risk factors are related to the duration of each state. Some existing regression methods model the interoccurrence time (i.e., gap time) between consecutive recurrent events of the same type, such as the methods considered by Huang and Chen (2003), Chang (2004), Strawderman (2005), Lu (2005), and Luo et al. (2013). Note that these methods are proposed for recurrent events of single type and hence are not directly applicable to a recurrent event process with bivariate alternating states such as hospital admission and discharge. One can apply such methods to durations between admissions while ignoring the time of discharge. However, when evaluating a risk factor on the hospitalization process, researchers may want to study whether it prolongs the in-patient stay as well as whether it shortens the out-of-hospital period. To evaluate the distinctive covariate effects on each of the two states one can apply semiparametric methods for bivariate serial event gap time data or multistate gap time data (when the number of states is two) to the first bivariate gap time pair. Since the time information beyond the first two gap times is not used, these methods are not expected to be efficient.

To assess covariate effects on the length of the two alternating states of bivariate alternating recurrent event data, a rank-based estimating equation approach was introduced by Chang (2004, referred to as Chang's bivariate estimator hereafter, to distinguish it from the estimator for non-alternating recurrent events proposed in the same paper). This model takes the same form as the accelerated failure time (AFT) model (see Kalbfleisch and Prentice, 2002, and reference therein), assuming that the logarithm-transformed gap times of the two alternating states are both linearly related to the covariate effects. To characterize the dependence among recurrent gap times within a subject, frailties, with their distribution left unspecified, are assumed to be shared by that subject's gap times. However, as discussed in Jin et al. (2003), solving rank-based estimating equations is challenging, due to its discontinuity and non-monotonicity in general. In this chapter, we propose a competitive estimating equation approach for bivariate alternating recurrent gap time data under an AFT model of the same form as Chang (2004)'s. The proposed estimating equation approach is based on U-statistics, which is motivated by Huang (2002)'s method for multistate gap time data

(referred to as Huang's method hereafter). The U-statistic based estimating function can be expressed as a continuous and compactly differentiable functional and hence is expected to be more computationally tractable than a rank-based estimating function.

The remainder of the chapter is organized as follows. We introduce the setting and assumptions for our proposed model in Section 4.2. In Section 4.3, we briefly review the estimation method developed in Huang (2002) for multistate gap time data and introduce our proposed method for bivariate alternating recurrent events. Large sample properties of the proposed method are established in Section 4.4. We conduct a series of simulation studies to demonstrate the performance of the proposed method and compare it with Chang's bivariate estimator in Section 4.5. Application of our method to the psychiatric case register (PCR) data is presented in Section 4.6. Some concluding remarks can be found in Section 4.7.

4.2 Model setup

Suppose hospitalization data are collected for a group of patients from the time they are initially admitted to the hospital. We denote the duration of an in-patient stay and an out-of-hospital period as X^0 and Y^0, respectively. Each patient can be readmitted to the hospital repeatedly during follow-up. Then, the underlying alternating hospitalization process for subject i is denoted as $N_i = \{(X_{i1}^0, Y_{i1}^0), (X_{i2}^0, Y_{i2}^0), \ldots\}, i = 1, \ldots, n$. We assume that the two gap times from admission to discharge and from discharge to the next admission (i.e., the durations of each state) are linearly related to covariates on the logarithmic scale, respectively:

$$\log X_{ij}^0 = \gamma_{i1} + \mathbf{A}_i^T \boldsymbol{\beta}_1 + \epsilon_{ij1}, \text{and} \tag{4.1}$$

$$\log Y_{ij}^0 = \gamma_{i2} + \mathbf{A}_i^T \boldsymbol{\beta}_2 + \epsilon_{ij2}, \tag{4.2}$$

where a $p \times 1$ vector $\mathbf{A}_i$ denotes the baseline covariates, $p \times 1$ vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the corresponding regression coefficients for X^0 and Y^0, respectively, $(\gamma_{i1}, \gamma_{i2})$ is the subject-specific latent random vector, and $\epsilon_{ijk}, i = 1, \ldots, n, j = 1, 2, \ldots$, and $k = 1, 2$, are mutually independent random error with mean zero. Note that we use the same covariate vector $\mathbf{A}_i$ for X_{ij}^0 and Y_{ij}^0, $j = 1, 2, \ldots$. However, if some of the covariates are

not of an interest for either X^0 or Y^0, they can be excluded by fixing the corresponding components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to be 0. The distributions of $(\gamma_{i1}, \gamma_{i2})$ and $(\epsilon_{ij1}, \epsilon_{ij2})$ are left unspecified. While patients are independently sampled from a target population, the alternating events within a patient may be correlated. The subject-specific vector $(\gamma_{i1}, \gamma_{i2})$ characterizes the correlation among the recurrence times within a subject. The association between X_{ij}^0 and Y_{ij}^0 is characterized by the correlation of γ_{i1} and γ_{i2}. Variances of γ_{i1} and γ_{i2} account for the level of correlations among X_{ij}^0's and Y_{ij}^0's, $j = 1, 2, \ldots$, respectively. We make the following assumptions on the bivariate recurrent event process:

Assumption 4.1 *Conditioning on* $\mathbf{A}_i$ *and* $(\gamma_{i1}, \gamma_{i2})$*, the bivariate gap time pairs* $\{(X_{ij}^0, Y_{ij}^0), j = 1, 2, \ldots\}$*, are independently and identically distributed (i.i.d) within subject* i*.*

Thus, the bivariate gap time pairs can be described as a renewal process given the baseline covariates and the latent random vector. Note that without conditioning on the latter, the bivariate gap time pairs within a patient are allowed to be correlated. It naturally follows that the bivariate gap time pairs $\{(X_{ij}^0, Y_{ij}^0),\ j = 1, 2, \ldots\}$ are exchangeable.

The alternating bivariate recurrent event process is subject to right censoring. Let C_i denote the censoring time, which has a survival function $G(\cdot)$ with a maximum support τ_C defined by $\tau_C = \sup\{t : G(t) > 0\}$. We adopt the following independent censoring assumption which is common in the literature:

Assumption 4.2 *The censoring time* C_i *is independent of* $N_i = \{(X_{ij}^0, Y_{ij}^0), j = 1, 2, \ldots\}$*,* $\mathbf{A}_i$*, and* $(\gamma_{i1}, \gamma_{i2})$*.*

We denote by m_i the number of observed bivariate pairs, which satisfies

$$\sum_{j=1}^{m_i-1} (X_{ij}^0 + Y_{ij}^0) \leq C_i \quad \text{and} \quad \sum_{j=1}^{m_i} (X_{ij}^0 + Y_{ij}^0) > C_i.$$

The first bivariate pair is possibly censored by C_i. However, the second or higher order bivariate pairs are subject to the induced dependent censoring. The last observed bivariate pair of the i^{th} subject is censored by $C_i^* = C_i - \sum_{j=0}^{m_i-1}(X_{ij}^0 + Y_{ij}^0)$, in which

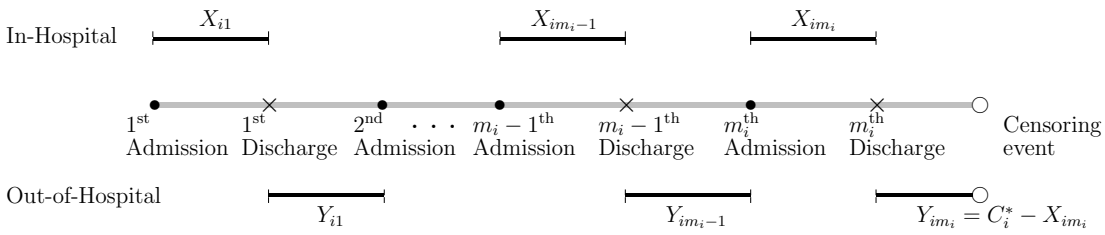

Figure 4.1: Illustration of a typical bivariate alternating recurrent events process.

we set $X_{i0}^0 = Y_{i0}^0 = 0$. Note that both $X_{im_i}^0$ and $Y_{im_i}^0$ or only the latter can be censored. Then, the observed data are $\{(X_{i1}, Y_{i1}), \ldots, (X_{im_{i-1}}, Y_{im_{i-1}}), (X_{im_i}, Y_{im_i}), \Delta_{ij}^X, \Delta_{ij}^Y, j = 1, \ldots, m_i\}$ where $X_{ij} = X_{ij}^0$, $Y_{ij} = Y_{ij}^0$, and $\Delta_{ij}^X = \Delta_{ij}^Y = 1$ for $j < m_i$; and $X_{im_i} = \min(X_{im_i}^0, C_i^*)$, $Y_{im_i} = \min(Y_{im_i}^0, \max(C_i^* - X_{im_i}^0, 0))$, $\Delta_{im_i}^X = I(X_{im_i}^0 < C_i^*)$, and $\Delta_{im_i}^Y = 0$. Figure 4.1 illustrates a typical repeated hospitalization process of a patient.

4.3 Estimation methods

4.3.1 A brief review of Huang's method

In this chapter, we are interested in evaluating the covariate effects on the duration of the two states of a bivariate alternating recurrent event process. We first review a simple case of the multistate gap time model developed in Huang (2002) where the number of states is fixed as two. It is easily seen that applying this simple method to the hospitalization data is valid when the analysis is restricted to data of the first hospitalization episode only, i.e., the time from the first admission to discharge (X_{i1}^0) and the time from discharge to the next admission (Y_{i1}^0). Let the gap time between the first and second admission be denoted as $\tilde{Y}_{i1}^0 = X_{i1}^0 + Y_{i1}^0$. Then, we define the transformed times as

$$\begin{aligned} X_{ii'1}^0(\mathbf{b}_1) &= \exp(\mathbf{A}_{ii'}^T \mathbf{b}_1) X_{i1}^0, \text{and} \\ \tilde{Y}_{ii'1}^0(\mathbf{b}) &= \exp(\mathbf{A}_{ii'}^T \mathbf{b}_1) X_{i1}^0 + \exp(\mathbf{A}_{ii'}^T \mathbf{b}_2) Y_{i1}^0 \end{aligned}$$

corresponding to the duration in hospital and the sum of in-hospital and out-of-hospital times, respectively, where $\mathbf{A}_{ii'} = \mathbf{A}_{i'} - \mathbf{A}_i$ is the difference of the baseline covariates

between subjects i and i', $i, i' = 1, \ldots, n$, and $\mathbf{b} = (\mathbf{b}_1^T, \mathbf{b}_2^T)^T$. While the aim is to evaluate covariate effects on X_{i1}^0 and Y_{i1}^0, we introduce $\tilde{Y}_{i1}^0$ (i.e., time-to-event notation) to properly account for the induced dependent censoring issue. By defining the transformed times for the gap time between the first two consecutive admissions as the sum of individually transformed duration of in-hospital and out-of-hospital times, different effects are allowed to be estimated for same factors (i.e., we allow $\mathbf{b}_1 \neq \mathbf{b}_2$). Furthermore, based on the model, the transformed times $X_{ii'1}^0(\mathbf{b}_1)$ and $\tilde{Y}_{ii'1}^0(\mathbf{b})$ share the same distribution with $X_{i'1}^0$ and $\tilde{Y}_{i'1}^0$, respectively, when evaluated at $\mathbf{b}_1 = \boldsymbol{\beta}_1$ and $\mathbf{b}_2 = \boldsymbol{\beta}_2$. Hence, it is obvious that the bivariate vector $\{X_{i1}^0, X_{ii'1}^0(\boldsymbol{\beta}_1)\}$ has the same distribution as $\{X_{i'i1}^0(\boldsymbol{\beta}_1), X_{i'1}^0\}$, which we denote by $\{X_{i1}^0, X_{ii'1}^0(\boldsymbol{\beta}_1)\} \sim \{X_{i'i1}^0(\boldsymbol{\beta}_1), X_{i'1}^0\}$, and we also have $\{\tilde{Y}_{i1}^0, \tilde{Y}_{ii'1}^0(\boldsymbol{\beta})\} \sim \{\tilde{Y}_{i'i1}^0(\boldsymbol{\beta}), \tilde{Y}_{i'1}^0\}$ given $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. Let L be a number smaller than τ_C. We adopt a symmetric function $O_L(t,s) = O_L(s,t)$ which is continuous on $\{(t,s) : 0 \leq t, s \leq L\}$ and monotonic in t given s and vice versa, and which is a functional mapping from two dimensional space into one. Then, conditional on $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, we have $O_L\{X_{i1}^0, X_{ii'1}^0(\boldsymbol{\beta}_1)\} \sim O_L\{X_{i'1}^0, X_{i'i1}^0(\boldsymbol{\beta}_1)\}$, and $O_L\{\tilde{Y}_{i1}^0, \tilde{Y}_{ii'1}^0(\boldsymbol{\beta})\} \sim O_L\{\tilde{Y}_{i'1}^0, \tilde{Y}_{i'i1}^0(\boldsymbol{\beta})\}$. It follows that

$$\begin{aligned} \mathrm{E}\left[w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_1)\mathbf{A}_{ii'}O_L\{X_{i1}^0, X_{ii'1}^0(\boldsymbol{\beta}_1)\}\right] &= 0, \text{and} \qquad (4.3) \\ \mathrm{E}\left[w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_2)\mathbf{A}_{ii'}O_L\{\tilde{Y}_{i1}^0, \tilde{Y}_{ii'1}^0(\boldsymbol{\beta})\}\right] &= 0, \qquad (4.4) \end{aligned}$$

where the scalar weight function $w(t, s, \mathbf{b}) = w(s, t, \mathbf{b})$ is symmetric and continuous for fixed $\mathbf{b}$.

Note that the times to discharge and to the next admission from the initial admission are both subject to independent right censoring. We denote the observed gap time from the first admission to the next one, which is subject to censoring, as $\tilde{Y}_{i1} = X_{i1} + Y_{i1}$, and the observed transformed times analogous to $X_{ii'1}^0$ and $\tilde{Y}_{ii'1}$ as

$$\begin{aligned} X_{ii'1}(\mathbf{b}_1) &= \exp(\mathbf{A}_{ii'}^T\mathbf{b}_1)X_{i1}, \text{and} \\ \tilde{Y}_{ii'1}(\mathbf{b}) &= \exp(\mathbf{A}_{ii'}^T\mathbf{b}_1)X_{i1} + \exp(\mathbf{A}_{ii'}^T\mathbf{b}_2)Y_{i1}. \end{aligned}$$

Under Assumption 4.2, it follows that $\mathrm{E}\left[\frac{\Delta_{i1}^X O_L\{X_{i1}, X_{ii'1}(\boldsymbol{\beta}_1)\}}{G(X_{i1} \wedge L)} \middle| \mathbf{A}_i, \mathbf{A}_{i'}\right] = \mathrm{E}\left[O_L\{X_{i1}^0, X_{ii'1}^0(\boldsymbol{\beta}_1)\} \mid \mathbf{A}_i, \mathbf{A}_{i'}\right]$ and $\mathrm{E}\left[\frac{\Delta_{i1}^Y O_L\{\tilde{Y}_{i1}, \tilde{Y}_{ii'1}(\boldsymbol{\beta})\}}{G(\tilde{Y}_{i1} \wedge L)} \middle| \mathbf{A}_i, \mathbf{A}_{i'}\right] = \mathrm{E}\left[O_L\{\tilde{Y}_{i1}^0, \tilde{Y}_{ii'1}^0(\boldsymbol{\beta})\} \middle| \mathbf{A}_i, \mathbf{A}_{i'}\right]$,

where $a \wedge b = \min(a, b)$. Following (4.3) and (4.4),

$$\mathrm{E}\left[\mathrm{E}\left[w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_1)\mathbf{A}_{ii'}\frac{\Delta_{i1}^X O_L\{X_{i1}, X_{ii'1}(\boldsymbol{\beta}_1)\}}{G(X_{i1} \wedge L)}\middle| \mathbf{A}_i, \mathbf{A}_{i'}\right]\right] = 0$$

and

$$\mathrm{E}\left[\mathrm{E}\left[w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_2)\mathbf{A}_{ii'}\frac{\Delta_{i1}^Y O_L\{\tilde{Y}_{i1}, \tilde{Y}_{ii'1}(\boldsymbol{\beta})\}}{G(\tilde{Y}_{i1} \wedge L)}\middle| \mathbf{A}_i, \mathbf{A}_{i'}\right]\right] = 0.$$

Then, the following estimating functions can be derived:

$$D_1(\mathbf{b}_1) = n^{-2}\sum_{i=1}^{n}\left[\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_1)\mathbf{A}_{ii'}\frac{\Delta_{i1}^X O_{L_1}\{X_{i1}, X_{ii'1}(\mathbf{b}_1)\}}{\hat{G}_1(X_{i1} \wedge L_1)}\right] \quad (4.5)$$

$$D_2(\mathbf{b}) = n^{-2}\sum_{i=1}^{n}\left[\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_2)\mathbf{A}_{ii'}\frac{\Delta_{i1}^Y O_{L_2}\{\tilde{Y}_{i1}, \tilde{Y}_{ii'1}(\mathbf{b})\}}{\hat{G}_2(\tilde{Y}_{i1} \wedge L_2)}\right] \quad (4.6)$$

where $\hat{G}_1$ and $\hat{G}_2$ are the Kaplan-Meier estimators using data $\{(X_{i1}, 1 - \Delta_{i1}^X), i = 1, \ldots, n\}$ and $\{(\tilde{Y}_{i1}, 1 - \Delta_{i1}^Y), i = 1, \ldots, n\}$, respectively, and L_1 and L_2 are constant values less than τ_C. We impose the artificial limits L_1 and L_2 to address the problem of X_{i1}^0 and $\tilde{Y}_{i1}^0$ having maximum support greater than τ_C. Note that patients who are censored before the first discharge or second admission only contribute in the estimation of the censoring time's survival functions $G(\cdot)$ in the denominator of Equations (4.5) and (4.6). To obtain the estimators for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, we inductively solve estimating equations $D_1(\mathbf{b}_1) = 0$ and $D_2(\mathbf{b}) = 0$.

We note that rank-based estimating equation approaches of the conventional AFT model for univariate survival data (**?**, see) also provide consistent estimates for $\boldsymbol{\beta}_1$ in Model (4.1). However, those approaches cannot be applied to estimate $\boldsymbol{\beta}_2$ in Model (4.2) because of the induced dependent censoring on Y_{i1}^0. Huang (2002) showed that the estimating equation based on the U-statistic function in (4.5) serves as an efficient alternative for rank-based estimating equation approaches on estimating $\boldsymbol{\beta}_1$.

4.3.2 Proposed estimation method

We propose to extend Huang's method introduced in Section 4.3.1 to use more available time information after a patient is admitted to a hospital for the second time. The extension is carried out in a fashion similar to Section 3.3.2.

Let $\tilde{Y}_{ij}^0 = X_{ij}^0 + Y_{ij}^0, j = 1, 2, \ldots$. Denote the observed (censored or uncensored) recurrent gap times between repeated admissions as $\tilde{Y}_{ij} = X_{ij} + Y_{ij}$, for $j = 1, \ldots, m_i^*$, where $m_i^* = m_i - 1$ if $m_i \geq 2$ and $m_i^* = 1$ if $m_i = 1$. The observed transformed times are defined as

$$\begin{aligned} X_{ii'j}(\mathbf{b}_1) &= \exp(\mathbf{A}_{ii'}^T \mathbf{b}_1) X_{ij}, \\ \tilde{Y}_{ii'j}(\mathbf{b}) &= \exp(\mathbf{A}_{ii'}^T \mathbf{b}_1) X_{ij} + \exp(\mathbf{A}_{ii'}^T \mathbf{b}_2) Y_{ij}, j = 1, \ldots, m_i^*. \end{aligned}$$

Under Assumption 4.1, conditioning on m_i, $(\gamma_{i1}, \gamma_{i2})$, and $\mathbf{A}_i$, the observed bivariate gap times $\{(X_{i1}, Y_{i1}), \ldots, (X_{im_i^*}, Y_{im_i^*})\}$ are i.i.d. for $m_i \geq 2$. In the same spirit as the weighted risk-set technique discussed in Luo and Huang (2011), we have

$$\mathrm{E}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} \mathrm{E}\left[\frac{\Delta_{ij}^X O_{L_1}\{X_{ij}, X_{ii'j}(\mathbf{b}_1)\}}{\hat{G}_1(X_{ij} \wedge L_1)} \middle| m_i, (\gamma_{1i}, \gamma_{2i}), \mathbf{A}_i\right]\right] = \mathrm{E}\left[\frac{\Delta_{i1}^X O_{L_1}\{X_{i1}, X_{ii'1}(\mathbf{b}_1)\}}{\hat{G}_1(X_{i1} \wedge L_1)}\right],$$

$$\mathrm{E}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} \mathrm{E}\left[\frac{\Delta_{ij}^Y O_{L_2}\{\tilde{Y}_{ij}, \tilde{Y}_{ii'j}(\mathbf{b})\}}{\hat{G}_2(\tilde{Y}_{ij} \wedge L_2)} \middle| m_i, (\gamma_{1i}, \gamma_{2i}), \mathbf{A}_i\right]\right] = \mathrm{E}\left[\frac{\Delta_{i1}^Y O_{L_2}\{\tilde{Y}_{i1}, \tilde{Y}_{ii'1}(\mathbf{b})\}}{\hat{G}_2(\tilde{Y}_{i1} \wedge L_2)}\right].$$

Then, we propose the following estimating functions:

$$D_1^*(\mathbf{b}_1) = n^{-2}\sum_{i=1}^{n}\left[\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_1)\mathbf{A}_{ii'}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}^X O_{L_1}\{X_{ij}, X_{ii'j}(\mathbf{b}_1)\}}{\hat{G}_1(X_{ij} \wedge L_1)}\right] \quad (4.7)$$

$$D_2^*(\mathbf{b}) = n^{-2}\sum_{i=1}^{n}\left[\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_2)\mathbf{A}_{ii'}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}^Y O_{L_2}\{\tilde{Y}_{ij}, \tilde{Y}_{ii'j}(\mathbf{b})\}}{\hat{G}_2(\tilde{Y}_{ij} \wedge L_2)}\right]. \quad (4.8)$$

We obtain $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ by inductively solving $D_1^*(\mathbf{b}_1) = 0$ and $D_2^*(\hat{\boldsymbol{\beta}}_1, \mathbf{b}_2) = 0$. We choose $O_L(t, s) = \log[\min\{\max(t, s), L\}] - \log(L)$ and $w = 1$ to achieve a monotonic estimating function which guarantees a unique solution. Further discussion on the selection of O_L and w is provided in Huang (2002).

By extending Huang's method to handle bivariate gap times beyond the second hospital admission, the proposed estimation method is expected to gain efficiency. Note that Huang (2002)'s original multistate gap time model can also use data beyond the second admission by fixing the number of states to be greater than two. However, the

model does not properly account for the fact that two states repeatedly alternate since no constraints are made to the covariate effects to be the same for each state. Also, the number of states which needs to be fixed in the multistate gap time model is an unknown random variable rather than a fixed number in our study.

4.4 Asymptotic properties

In this section, we establish the consistency and the asymptotic normality of $\hat{\boldsymbol{\beta}}$. The asymptotic properties of $\hat{\boldsymbol{\beta}}$ are based on the large sample study for Huang's bivariate gap time model (Huang, 2002, in Appendix A). We begin by rewriting the estimating functions (4.5) and (4.6) as

$$D_1(\mathbf{b}_1) = \int_{t,s,\mathbf{a}_1,\mathbf{a}_2} w(\mathbf{a}_1,\mathbf{a}_2,\mathbf{b}_1)(\mathbf{a}_2-\mathbf{a}_1)\frac{O_{L_1}(t,s)}{\hat{G}_1(t\wedge L_1)}\hat{F}_1(\mathrm{d}t,\mathrm{d}s,\mathrm{d}\mathbf{a}_1;\mathbf{a}_2,\mathbf{b}_1)\hat{H}(\mathrm{d}\mathbf{a}_2), \tag{4.9}$$

$$D_2(\mathbf{b}) = \int_{t,s,\mathbf{a}_1,\mathbf{a}_2} w(\mathbf{a}_1,\mathbf{a}_2,\mathbf{b}_2)(\mathbf{a}_2-\mathbf{a}_1)\frac{O_{L_2}(t,s)}{\hat{G}_2(t\wedge L_2)}\hat{F}_2(\mathrm{d}t,\mathrm{d}s,\mathrm{d}\mathbf{a}_1;\mathbf{a}_2,\mathbf{b})\hat{H}(\mathrm{d}\mathbf{a}_2), \tag{4.10}$$

where $\hat{F}_1$, $\hat{F}_2$, and $\hat{H}$ are the empirical estimators of $F_1(t,s,\mathbf{a}_1;\mathbf{a}_2,\mathbf{b}_1) = \Pr\left[X_{i1}\le t, \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_1\}X_{i1}\le s, \mathbf{A}\le\mathbf{a}_1, \Delta_{i1}^X=1\right]$, $F_2(t,s,\mathbf{a}_1;\mathbf{a}_2,\mathbf{b}) = \Pr\left[\tilde{Y}_{i1}\le t, \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_1\}X_{i1} + \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_2\}Y_{i1}\le s, \mathbf{A}\le\mathbf{a}_1, \Delta_{i2}^Y=1\right]$, and $H(\mathbf{a}_2) = \Pr(\mathbf{A}\le\mathbf{a}_2)$, respectively. Huang (2002) showed that D_1 and D_2 are continuous and compactly differentiable functionals through the properties of the components, $\hat{G}_1$, $\hat{G}_2$, $\hat{F}_1$, $\hat{F}_2$, and $\hat{H}$.

Based on the re-expression (4.9) and (4.10), $D_1^T(\mathbf{b}_1)(\mathbf{b}_1-\boldsymbol{\beta}_1)$ and $D_2^T(\mathbf{b})(\mathbf{b}-\boldsymbol{\beta})$ converges almost surely and uniformly in $\mathbf{b}_1$ and in $\mathbf{b}$ to

$$\mathrm{E}\left[\mathrm{E}\left[w(\mathbf{A}_i,\mathbf{A}_{i'},\mathbf{b}_1)\mathbf{A}_{ii'}^T(\mathbf{b}_1-\boldsymbol{\beta}_1)O_{L_1}\{X_{i1}^0, X_{ii'1}^0(\mathbf{b}_1)\}\right|\mathbf{A}_i,\mathbf{A}_{i'}\right]\right] \quad \text{and} \tag{4.11}$$

$$\mathrm{E}\left[\mathrm{E}\left[w(\mathbf{A}_i,\mathbf{A}_{i'},\mathbf{b}_2)\mathbf{A}_{ii'}^T(\mathbf{b}-\boldsymbol{\beta})O_{L_2}\{\tilde{Y}_{i1}^0, \tilde{Y}_{ii'1}^0(\mathbf{b})\}\right|\mathbf{A}_i,\mathbf{A}_{i'}\right]\right], \tag{4.12}$$

respectively. It is easily seen that the estimating functions D_1^* and D_2^* constructed in (4.7) and (4.8) converge uniformly to the same limit as D_1 and D_2. Thus, it follows that $D_1^{*T}(\mathbf{b}_1)(\mathbf{b}_1-\boldsymbol{\beta}_1)$ and $D_2^{*T}(\mathbf{b})(\mathbf{b}-\boldsymbol{\beta})$ also converge almost surely to (4.11) and

(4.12). Since (4.11) equals 0 if $\mathbf{b}_1 = \boldsymbol{\beta}_1$, $\hat{\boldsymbol{\beta}}_1$ is consistent to $\boldsymbol{\beta}_1$. Given the consistency of $\hat{\boldsymbol{\beta}}_1$, the consistency of $\hat{\boldsymbol{\beta}}$ follows from the fact that Equation (4.12) equals 0 if $\mathbf{b} = \boldsymbol{\beta}$.

It is enough to establish the asymptotic normality and linearity of $D(\boldsymbol{\beta}) = \{D_1^T(\boldsymbol{\beta}_1), D_2^T(\boldsymbol{\beta})\}^T$ in order to prove the asymptotic normality of $\hat{\boldsymbol{\beta}}$. Huang (2002) showed that by the functional delta method $n^{1/2}D(\boldsymbol{\beta})$ is asymptotically normal with mean zero and variance Ω using the compact differentiability of (4.9) and (4.10). For the variance estimate we define

$$\begin{aligned}
\xi_{i1}(\boldsymbol{\beta}_1) &= n^{-3/2}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_1)\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}^X O_{L_1}\{X_{i1}, X_{ii'1}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{i1}\wedge L_1)} - \frac{\Delta_{i'1}^X O_{L_1}\{X_{i'1}, X_{i'i1}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{i'1}\wedge L_1)}\right] \\
&\quad + n^{-3/2}\int_0^{L_1} \frac{U_1(t,\boldsymbol{\beta}_1)\hat{G}_1(t-)}{R_1(t)\hat{G}_1(t)}\mathrm{d}\hat{M}_{i1}(t), \text{ and} \\
\xi_{i2}(\boldsymbol{\beta}) &= n^{-3/2}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_2)\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}^Y O_{L_2}\{\tilde{Y}_{i1}, \tilde{Y}_{ii'1}(\boldsymbol{\beta})\}}{\hat{G}_2(\tilde{Y}_{i1}\wedge L_2)} - \frac{\Delta_{i'1}^Y O_{L_2}\{\tilde{Y}_{i'1}, \tilde{Y}_{i'i1}(\boldsymbol{\beta})\}}{\hat{G}_2(\tilde{Y}_{i'1}\wedge L_2)}\right] \\
&\quad + n^{-3/2}\int_0^{L_2} \frac{U_2(t,\boldsymbol{\beta})\hat{G}_2(t-)}{R_2(t)\hat{G}_2(t)}\mathrm{d}\hat{M}_{i2}(t),
\end{aligned}$$

in which

$$\begin{aligned}
U_1(t,\boldsymbol{\beta}_1) &= \sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_1)\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}^X O_{L_1}\{X_{i1}, X_{ii'1}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{i1}\wedge L_1)} I(X_{i1} > t)\right], \\
U_2(t,\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \boldsymbol{\beta}_2)\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}^Y O_{L_2}\{\tilde{Y}_{i1}, \tilde{Y}_{ii'1}(\boldsymbol{\beta})\}}{\hat{G}_2(\tilde{Y}_{i1}\wedge L_2)} I(\tilde{Y}_{i1} > t)\right],
\end{aligned}$$

$R_1(t) = \sum_{i=1}^n I(X_{i1} \geq t)$, $R_2(t) = \sum_{i=1}^n I(\tilde{Y}_{i1} \geq t)$, $\hat{M}_{i1}(t) = I(X_{i1} \leq t, \Delta_{i1}^X = 0) - \int_0^t I(X_{i1} \geq s)\mathrm{d}\hat{\Lambda}_1(s)$, $\hat{M}_{i2}(t) = I(\tilde{Y}_{i1} \leq t, \Delta_{i1}^Y = 0) - \int_0^t I(\tilde{Y}_{i1} \geq s)\mathrm{d}\hat{\Lambda}_2(s)$, and $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ are the Nelson-Aalen estimator corresponding to $\hat{G}_1$ and $\hat{G}_2$, respectively. The variance Ω can be estimated consistently by $\hat{\Omega} = \sum_{i=1}^n \{\xi_{i1}^T(\boldsymbol{\beta}_1), \xi_{i2}^T(\boldsymbol{\beta})\}^T\{\xi_{i1}^T(\boldsymbol{\beta}_1), \xi_{i2}^T(\boldsymbol{\beta})\}$. Now, let $D^*(\boldsymbol{\beta}) = \{D_1^{*T}(\boldsymbol{\beta}_1), D_2^{*T}(\boldsymbol{\beta})\}^T$. It is obvious that D_1^* and D_2^* share the same properties with D_1 and D_2, respectively, and thus, are continuous and compactly differentiable functionals. By applying the functional delta method, we derive that $n^{1/2}D^*(\boldsymbol{\beta})$

converges weakly to a normal distribution with mean zero and variance Ω^*. Define

$$\begin{aligned}\xi^*_{i1}(\boldsymbol{\beta}_1) &= n^{-3/2}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_1)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta^X_{ij}O_{L_1}\{X_{ij},X_{ii'j}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{ij}\wedge L_1)}\right.\\ &\left.-\frac{1}{m_{i'}^*}\sum_{l=1}^{m_{i'}^*}\frac{\Delta^X_{i'l}O_{L_1}\{X_{i'l},X_{i'il}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{i'l}\wedge L_1)}\right]+n^{-3/2}\int_0^{L_1}\frac{U_1^*(t,\boldsymbol{\beta}_1)\hat{G}_1(t-)}{R_1^*(t)\hat{G}_1(t)}\mathrm{d}\hat{M}^*_{i1}(t),\end{aligned}$$

$$\begin{aligned}\xi^*_{i2}(\boldsymbol{\beta}) &= n^{-3/2}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_2)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta^Y_{ij}O_{L_2}\{\tilde{Y}_{ij},\tilde{Y}_{ii'j}(\boldsymbol{\beta})\}}{\hat{G}_2(\tilde{Y}_{ij}\wedge L_2)}\right.\\ &\left.-\frac{1}{m_{i'}^*}\sum_{l=1}^{m_{i'}^*}\frac{\Delta^Y_{i'l}O_{L_2}\{\tilde{Y}_{i'l},\tilde{Y}_{i'il}(\boldsymbol{\beta})\}}{\hat{G}_2(\tilde{Y}_{i'l}\wedge L_2)}\right]+n^{-3/2}\int_0^{L_2}\frac{U_2^*(t,\boldsymbol{\beta})\hat{G}_2(t-)}{R_2^*(t)\hat{G}_2(t)}\mathrm{d}\hat{M}^*_{i2}(t),\end{aligned}$$

in which,

$$\begin{aligned}U_1^*(t,\boldsymbol{\beta}_1) &= \sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_1)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta^X_{ij}O_{L_1}\{X_{ij},X_{ii'j}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{ij}\wedge L_1)}I(X_{ij}>t)\right],\\ U_2^*(t,\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i,\mathbf{A}_{i'},\boldsymbol{\beta}_2)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta^Y_{ij}O_{L_2}\{\tilde{Y}_{ij},\tilde{Y}_{ii'j}(\boldsymbol{\beta})\}}{\hat{G}_2(\tilde{Y}_{ij}\wedge L_2)}I(\tilde{Y}_{ij}>t)\right],\end{aligned}$$

$R_1^*(t) = \sum_{i=1}^n \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(X_{ij}\geq t)$, $R_2^*(t) = \sum_{i=1}^n \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(\tilde{Y}_{ij}\geq t)$, $\hat{M}^*_{i1}(t) = \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(X_{ij}\leq t,\Delta^X_{ij}=0) - \int_0^t \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(X_{ij}\geq s)\mathrm{d}\hat{\Lambda}_1(s)$, and $\hat{M}^*_{i2}(t) = \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(\tilde{Y}_{ij}\leq t,\Delta^Y_{ij}=0) - \int_0^t \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(\tilde{Y}_{ij}\geq s)\mathrm{d}\hat{\Lambda}_2(s)$. Note that by exchangeability, ξ^*_{ik}, U^*_k, R^*_k, and M^*_{ik}, which are the "weighted average versions", converge uniformly to the same limit as their counterparts, ξ_{ik}, U_k, R_k, and M_{ik} for $k=1,2$. Thus, the variance estimate for Ω^* can be estimated by $\hat{\Omega}^* = \sum_{i=1}^n\{\xi^{*T}_{i1}(\boldsymbol{\beta}_1),\xi^{*T}_{i2}(\boldsymbol{\beta})\}^T$ $\{\xi^{*T}_{i1}(\boldsymbol{\beta}_1),\xi^{*T}_{i2}(\boldsymbol{\beta})\}$. By the Glivenko-Cantelli theorem of Pollard (1984), $\sum_{i=1}^n\{\xi^{*T}_{i1}(\mathbf{b}_1),$ $\xi^{*T}_{i2}(\mathbf{b})\}^T\{\xi^{*T}_{i1}(\mathbf{b}_1),\xi^{*T}_{i2}(\mathbf{b})\}$ converges uniformly and almost surely in $\mathbf{b}$ to a limiting function continuous at $\mathbf{b}=\boldsymbol{\beta}$. Hence, the variance estimate $\hat{\Omega}^*$ is consistent for Ω^* given the consistency of $\hat{\boldsymbol{\beta}}$.

Huang (2002) shows that $D_1(\mathbf{b}_1) = \bar{D}_1(\mathbf{b}_1) + o_p(||\mathbf{b}_1-\boldsymbol{\beta}_1|| + n^{-1/2})$ and $D_2(\mathbf{b}) =$

$\bar{D}_2(\mathbf{b}) + o_p(||\mathbf{b}-\boldsymbol{\beta}|| + n^{-1/2})$ for $\mathbf{b}_1$ converging to $\boldsymbol{\beta}_1$, and $\mathbf{b}$ to $\boldsymbol{\beta}$, respectively, where

$$\bar{D}_1(\mathbf{b}_1) = \int_{t,s,\mathbf{a}_1,\mathbf{a}_2} w(\mathbf{a}_1,\mathbf{a}_2,\boldsymbol{\beta}_1)(\mathbf{a}_2-\mathbf{a}_1)\frac{O_{L_1}(t,s)}{\hat{G}_1(t\wedge L_1)}\hat{F}_1(\mathrm{d}t,\mathrm{d}s,\mathrm{d}\mathbf{a}_1;\mathbf{a}_2,\mathbf{b}_1)\hat{H}(\mathrm{d}\mathbf{a}_2) \quad (4.13)$$

$$\bar{D}_2(\mathbf{b}) = \int_{t,s,\mathbf{a}_1,\mathbf{a}_2} w(\mathbf{a}_1,\mathbf{a}_2,\boldsymbol{\beta}_2)(\mathbf{a}_2-\mathbf{a}_1)\frac{O_{L_2}(t,s)}{\hat{G}_2(t\wedge L_2)}\hat{F}_2(\mathrm{d}t,\mathrm{d}s,\mathrm{d}\mathbf{a}_1;\mathbf{a}_2,\mathbf{b})\hat{H}(\mathrm{d}\mathbf{a}_2). \quad (4.14)$$

Then, we derive that $D_1^*(\mathbf{b}_1) = \bar{D}_1^*(\mathbf{b}_1) + o_p(||\mathbf{b}_1-\boldsymbol{\beta}_1|| + n^{-1/2})$ and $D_2^*(\mathbf{b}) = \bar{D}_2^*(\mathbf{b}) + o_p(||\mathbf{b}-\boldsymbol{\beta}|| + n^{-1/2})$, for $\mathbf{b}_1$ converging to $\boldsymbol{\beta}_1$ and for $\mathbf{b}$ to $\boldsymbol{\beta}$, respectively, where $\bar{D}_1^*$ and $\bar{D}_2^*$ are defined by replacing $\hat{F}_1$ and $\hat{F}_2$ in (4.13) and (4.14) with their weighted average version of the estimators

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I\left[X_{ij}\le t, \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_1\}X_{ij}\le s, \mathbf{A}\le\mathbf{a}_1, \Delta_{ij}^X = 1\right] \text{ and}$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I\left[\tilde{Y}_{ij}\le t, \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_1\}X_{ij} + \exp\{(\mathbf{a}_2-\mathbf{A})^T\mathbf{b}_2\}Y_{ij}\le s, \mathbf{A}\le\mathbf{a}_1, \Delta_{ij}^Y = 1\right].$$

Denote the derivative matrix of $\bar{D}^*(\mathbf{b}) = \{\bar{D}_1^{*T}(\mathbf{b}_1), \bar{D}_1^{*T}(\mathbf{b})\}^T$ by $\hat{\Sigma} = \frac{\partial}{\partial\mathbf{b}}\bar{D}^*(\mathbf{b})$. We note that $\bar{D}^*(\mathbf{b})$ is not everywhere differentiable. Huang (2000) proposed the generalized law of mean for accommodating the nondifferentiable functions. By applying the generalized law of the mean, we derive

$$\begin{aligned} D^*(\mathbf{b}) &= \bar{D}^*(\mathbf{b}) + o_p(||\mathbf{b}-\boldsymbol{\beta}|| + n^{-1/2}) \\ &= D^*(\boldsymbol{\beta}) + \Sigma(\mathbf{b}-\boldsymbol{\beta}) + o_p(||\mathbf{b}-\boldsymbol{\beta}|| + n^{-1/2}) \end{aligned}$$

for $\mathbf{b}$ converging to $\boldsymbol{\beta}$. This follows from the fact that left and right partial derivatives exist for $\bar{D}^*(\mathbf{b})$ and have the same limit Σ (Pollard, 1984). The asymptotic normality of $\hat{\boldsymbol{\beta}}$ naturally follows from the asymptotic normality and linearity of $D^*(\boldsymbol{\beta})$. Thus, $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ converges weakly to a normal distribution with mean zero and variance consistently estimated by $\hat{\Sigma}^{-1}\hat{\Omega}^*(\hat{\Sigma}^{-1})^T$.

4.5 Simulation studies

We conduct a series of simulation studies to assess the performance of the proposed method. For each setting, we simulate 1000 datasets with sample size of $n = 150, 300$ from the following model:

$$\begin{aligned} \log X_{ij}^0 &= \gamma_i + A_i\beta_1 + \epsilon_{ij1} \\ \log Y_{ij}^0 &= \gamma_i + A_i\beta_2 + \epsilon_{ij2}, i = 1, \ldots, n, j = 1, 2, \ldots. \end{aligned}$$

The covariate A_i is generated from a Bernoulli distribution with probability 0.5, which indicates group assignment. The error terms ϵ_{ij1} and ϵ_{ij2} are sampled from independent normal distributions with mean zero and variance equal to 0.1. The subject-specific latent variable γ_i is from a standard normal distribution with variance $\sigma^2 = 0, 0.25, 0.5$ for different level of correlation among gap times within each patient. The censoring time C_i is chosen independently for each subject from a uniform distribution $(0, U)$, where $U = 10$. We consider two scenarios to demonstrate the performance of the proposed method relative to Chang's bivariate estimator. We also examine the performance of a single-type recurrent gap time model (Chang, 2004, referrred to as Chang's univariate estimator) when only times between admissions are analyzed. Chang's univariate estimator assumes that gap times between recurrent admissions ($\tilde{Y}_{ij}^0$) are identically distributed and follow the following model:

$$\log \tilde{Y}_{ij}^0 = \gamma_i + A_i\tilde{\beta} + \epsilon_{ij}, i = 1, \ldots, n, j = 1, 2, \ldots.$$

Simulation Scenario 1

In the first scenario, we consider a case where a group of patients has longer gap times between recurrent admissions than the other group, but not necessarily longer out-of-hospital durations. Specifically, we consider the situation in which a group has prolonged time to readmission due to a longer time in hospital during a previous hospitalization episode but has no difference in out-of-hospital duration, by setting $\beta_1 = 1$ and $\beta_2 = 0$. The simulation results are summarized in Table 4.1. We observe that both the proposed method and Chang's bivariate estimator are virtually unbiased, and the averages of the standard errors computed by bootstrap resampling method with sample size $B = 100$

Table 4.1: Summary of the simulation results for Scenario 1 with true coefficients (True); Monte-Carlo mean (Mean); Monte-Carlo SD $\times 10^3$ (SD); average bootstrap SE $\times 10^3$ (SE); and power or size (Power/Size) of the proposed estimator (Proposed); Chang's bivariate estimator (Chang-Biv); and Chang's univariate estimator (Chang).

	$n = 150$					$n = 300$				
	Proposed		Chang-Biv		Chang	Proposed		Chang-Biv		Chang
	β_1	β_2	β_1	β_2	$\tilde{\beta}$	β_1	β_2	β_1	β_2	$\tilde{\beta}$
True	1.0	0.0	1.0	0.0	-	1.0	0.0	1.0	0.0	-
$\sigma^2 = 0$	$\bar{m}^a$= 2.38; cr^b= 0.30					$\bar{m}$= 2.38; cr= 0.30				
Mean	1.002	-0.008	0.939	0.014	0.626	1.001	-0.002	0.981	0.004	0.627
SD	63	150	226	104	44	47	111	141	62	32
SE	62	141	256	118	44	45	105	170	75	31
Power^c/Size^d	1.000	0.084	1.000	0.045	-	1.000	0.084	1.000	0.051	-
$\sigma^2 = 0.25$	$\bar{m}$= 2.63; cr= 0.34					$\bar{m}$= 2.63; cr= 0.33				
Mean	1.001	-0.026	0.991	0.001	0.615	1.002	-0.018	0.998	0.001	0.615
SD	111	233	130	157	109	80	176	81	113	74
SE	115	230	160	177	109	81	168	93	121	77
Power/Size	1.000	0.055	1.000	0.014	-	1.000	0.064	1.000	0.033	-
$\sigma^2 = 0.5$	$\bar{m}$= 2.90; cr= 0.36					$\bar{m}$= 2.91; cr= 0.36				
Mean	1.001	-0.026	0.998	0.009	0.614	1.001	-0.016	0.999	0.006	0.615
SD	146	278	143	205	144	102	197	98	150	99
SE	146	263	182	220	146	104	191	113	154	104
Power/Size	0.999	0.077	1.000	0.005	-	1.000	0.064	1.000	0.028	-

[a] Average number of observed pairs of recurrence times per subject
[b] Average proportion of subjects having the first pair censored
[c] Proportion of 95% confidence interval not including 0 when true coefficient $\neq 0$
[d] Proportion of 95% confidence interval not including 0 when true coefficient $= 0$

are close to the empirical standard deviations under all settings. Both methods have power close to 1 for the estimation of $\boldsymbol{\beta}_1$ when the true $\boldsymbol{\beta}_1 = 1.0$. For $\boldsymbol{\beta}_2 = 0$, the size of both methods are close to the nominal level 0.05. When Chang's univariate estimator is applied, the covariate effects on the in- and out-of-hospital durations are not estimable and the covariate effect on the time between two admissions become obscure. All estimates for $\tilde{\beta}$ are shown to be close to 0.6, which is unsprisingly between the true $\beta_1(= 1)$ and $\beta_2(= 0)$. This simulation scenario illustrates that a group of patients with longer in-hospital stay, hence worse quality of life and more medical cost, could be incorrectly concluded to be in a better position (i.e., prolonged hospital readmission) due to the use of wrong method.

Simulation Scenario 2

In this scenario, we want to present a situation when two groups have a comparable length of gap times between hospital admissions but different length of durations for the in- and out-of-hospital states. Let $\beta_1 = 0.5$ and $\beta_2 = -1$ such that a group has longer in-hospital stay but shorter out-of-hospital duration compared to the other group. Table 4.2 summarizes the results. It can be observed that both the proposed and Chang's bivariate estimators provide virtually unbiased estimates under all settings. While the averages of the bootstrap standard errors with sample size $B = 100$ are close to their empirical standard deviations for the proposed method, those tend to deviate for Chang's bivariate estimator. The proposed method tends to be more powerful in detecting the group effect on both the in- and out-of-hospital periods in this scenario. As expected, the estimated $\tilde{\beta}$ from Chang's univariate method is close to 0. Again, this scenario shows that applying inappropriate method on the bivariate alternating recurrent event data will obscure the real covariate effects on the event process and lead to incorrect scientific conclusion.

In summary, the proposed method is more robust than Chang's bivariate estimator. Our estimator is proven to be applicable under various circumstances.

4.6 Application

For illustration of the proposed method, we analyze a subset of the South Verona PCR data (Tansella, 1991). We studied a total of 336 patients who were diagnosed with schizophrenia or related disorders and contacted the register for the first time between 1981 and 1995. Contacts to South Verona Community Mental Health Services were prospectively collected over time. By the definition provided in Section 1.3.2, the data comprise bivariate alternating recurrent gap times. The two states of interest are the care period and the break period. A total number of 1035 bivariate gap time pairs were observed. On average, each patient experienced about 3.1 episodes of care and break period (range: $1 - 18$). However, the majority of the patients (145, 43.2%) experienced only one care period and a censored break period during follow-up. The longest follow-up time was 5817 days (15.9 years). Among all patients, 47.9% are male; 52.1% female. The age of the patients at onset ranges from 13.7 to 84.0 (median: 37.2). Following

Table 4.2: Summary of the simulation results for Scenario 2 with true coefficients (True); Monte-Carlo mean (Mean); Monte-Carlo SD $\times 10^3$ (SD); average bootstrap SE $\times 10^3$ (SE); and power (Power) of the proposed estimator (Proposed); Chang's bivariate estimator (Chang-Biv); and Chang's univariate estimator (Chang).

	$n = 150$					$n = 300$				
	Proposed		Chang-Biv		Chang	Proposed		Chang-Biv		Chang
	β_1	β_2	β_1	β_2	$\tilde{\beta}$	β_1	β_2	β_1	β_2	$\tilde{\beta}$
True	0.5	-1.0	0.5	-1.0	-	0.5	-1.0	0.5	-1.0	-
$\sigma^2 = 0$	$\bar{m}^a$= 2.91; cr^b= 0.21					$\bar{m}$= 2.92; cr= 0.21				
Mean	0.501	-0.987	0.489	-0.956	0.019	0.499	-0.993	0.499	-0.998	0.018
SD	58	114	70	212	37	43	89	35	70	25
SE	58	108	88	290	38	43	82	54	166	27
Power[c]	1.000	1.000	0.996	0.569	-	1.000	1.000	1.000	0.836	-
$\sigma^2 = 0.25$	$\bar{m}$= 3.23; cr= 0.24					$\bar{m}$= 3.23; cr= 0.24				
Mean	0.499	-0.983	0.492	-0.965	0.005	0.499	-0.990	0.499	-0.997	0.006
SD	119	218	120	222	107	83	164	75	112	74
SE	115	200	133	340	104	83	152	82	182	73
Power	0.977	0.991	0.827	0.452	-	1.000	0.999	0.978	0.913	-
$\sigma^2 = 0.5$	$\bar{m}$= 3.59; cr= 0.26					$\bar{m}$= 3.59; cr= 0.26				
Mean	0.495	-0.987	0.488	-0.931	0.002	0.500	-0.986	0.501	-0.988	0.005
SD	146	239	155	289	134	108	179	98	153	96
SE	147	230	177	409	142	105	173	115	250	100
Power	0.909	0.973	0.614	0.207	-	0.988	1.000	0.889	0.773	-

[a]Average number of observed pairs of recurrence times per subject
[b]Average proportion of subjects having the first pair censored
[c]Proportion of 95% confidence interval not including 0

Table 4.3: Summary of univariate regression analysis of risk factors for being under mental health care. The table shows point estimates (SE) for each variable.

Variables	Care period	Break period
Gender		
Male = 1; Female = 0	0.048 (0.220)	0.013 (0.295)
Age at Onset		
(Years)	-0.009 (0.006)	0.021 (0.008*)
Education level		
Higher = 1; Lower = 0	0.552 (0.198*)	-0.207 (0.306)

*P-value < 0.05.

Huang and Wang (2005), we consider onset age 25 or younger as early onset (64 patients, 19%); and older than 25 as late onset (272 patients, 81%). Ten patients had missing values for education level and were excluded from the analysis. Among all, 131 patients (40.2%) received lower level education and 195 patients (59.8%) were more educated.

We are interested in evaluating the effects of socio-demographic and economic factors on the length of care and break periods. Specifically, it is of interest to identify patient characteristics that are associated with shortened care period and/or prolonged break period. We first apply the proposed estimation method on each factor in univariate analyses. Results are presented in Table 4.3. Education level was found to be a significant factor among various patient-related characteristics. We found that subjects who received secondary or higher education tended to have longer care period than less educated subjects. Table 4.4 presents the results of fitting a multivariable regression model with all three covariates. The results show that a later age of onset of mental health care was significantly associated with extended break period. When other variables are fixed, the length of break period increases by 28% ($= \exp(0.025 \times 10) - 1$) when the age of onset is delayed by 10 years. Also, higher education level is found to be related to an extended care period. Specifically, a patient with higher education tended to have 1.78($= \exp(0.58)$) times longer duration of care period than patients with lower level of education. A previous study conducted on costs of community-based psychiatric care (Amaddeo et al., 1998) has shown that for patient with schizophrenia,

Table 4.4: Summary of multivariable regression analysis of risk factors for being under mental health care. The table shows point estimates (SE) for each variable.

Variables	Care period	Break period
Gender		
Male = 1; Female = 0	-0.126 (0.197)	0.264 (0.235)
Age at Onset		
(Years)	0.000 (0.007)	0.025 (0.008*)
Education level		
Higher = 1; Lower = 0	0.577 (0.238*)	0.162 (0.290)

*P-value < 0.05.

higher education was positively associated with costs of care, which is in line with our finding as extended care period will inevitably trigger more cost.

4.7 Concluding remarks

In this chapter, we proposed a semiparametric regression model to make inference on covariate effects for bivariate alternating event data. The proposed model allows covariates to have different effects on each of the two states, hence leads to better understanding of the underlying recurrent process than models which do not distinguish the two different states. In our estimation procedure, we adopt the weighted risk set method discussed in Luo and Huang (2011) to pool available recurrent gap time pairs together by exploiting the exchangeability structure among the completely observed bivariate gap time pairs. However, when longer follow-up times are studied, the exchangeability property may no longer hold since the bivariate gap times may follow a certain trend due to the improvement or deterioration of health over time. Besides, informative censoring may become an issue when a study has a longer follow-up. Especially for hospitalization due to a disease associated with mortality, death can be a major censoring event informative to the recurrent event process. In this case, the independent censoring assumption required by the proposed method can be violated. Research on regression model which can relax the independent censoring assumption on the bivariate alternating recurrent event data

is certainly warranted.

In practice, it is often of interest to model the dependence structure between the two alternating states. In this chapter, we do not specify the dependence structure between the two alternating states or among different bivariate pairs explicitly, which leads the proposed method to be a robust model compared to models with more specific assumptions on correlation structure. However, when the dependence structure is of primary interest, semiparametric estimation using copula models may be considered.

In this chapter, a comparison of the proposed method with Chang's bivariate estimator was made with the simulation studies. The results show that the proposed method has better power and size closer to the nominal level than Chang's under many settings. More importantly, the proposed method is found to be more computationally tractable than Chang's bivariate method in which covariate effects are estimated by solving estimating equations in form of step functions.

Chapter 5

Conclusion

5.1 Summary of major findings

In this thesis, we have studied statistical methods for recurrent gap time data where the event process consists of different types of gap times, for example, time from transplant to the first infection and the gap times between two consecutive infections in the post-transplant infection data; or the durations of alternating care periods and break periods in the South Verona psychiatric case register data. The commonly adopted assumption for recurrent gap time analysis in the literature, that all gap times are identically distributed, is found to be inappropriate for the data we have investigated. The proposed methods in this thesis were constructed on the basis of appropriate yet inefficient existing estimation methods for bivariate serial gap time data. To improve efficiency, we exploited the exchangeablility property among same-type recurrent gap times, and applied the weighted risk-set technique as discussed in Luo and Huang (2011). This way, we were able to make full use of the available time information in the estimation procedure.

It is worthwhile to emphasize that the data structure or the study design discussed in this thesis are not rare. In many observational studies, subjects are enrolled to a study due to an event which is different from the recurrent events observed during follow-up. Also, if recurrent events are not momentary, they may last for a certain period of time. Then, we easily see an event-period and an event-free-period alternating over time, which is bivariate alternating recurrent event data. Thus, we believe the proposed

nonparametric estimator and semiparametric regression methods will widely serve as useful tools in analyzing recurrent gap time data.

5.2 Future work

There are many interesting research topics to study in recurrent gap time analysis. While challenges arise due to the distinctive sequential structure of recurrent gap time data, researchers are interested in developing statistical methods because they carry useful information. In this section, we briefly introduce some preliminary studies that we conducted which are not included in this thesis. We also present some future research directions.

5.2.1 Dependence structure between gap times

In Chapter 2, we left the dependence structure between gap times unspecified. When the association between the first event time and the subsequent gap times is of primary interest, one can estimate the Kendall's τ between the two types of gap times.

Briefly, we can make use of the proposed bivariate distribution function estimator in Section 2.3 for estimation of the restricted version of Kendall's τ, which is defined as the difference of conditional concordance and discordance probabilities. We denote it as $\tau^* = \Pr\{(X_1^0 - X_2^0)(Y_1^0 - Y_2^0) > 0 \mid X_1^0 + Y_1^0 \leq \tau_C, X_2^0 + Y_2^0 \leq \tau_C\} - \Pr\{(X_1^0 - X_2^0)(Y_1^0 - Y_2^0) < 0 \mid X_1^0 + Y_1^0 \leq \tau_C, X_2^0 + Y_2^0 \leq \tau_C\}$, where (X_1^0, Y_1^0) and (X_2^0, Y_2^0) are independent realizations from (X^0, Y^0). It is straightforward to see that $\tau^* = 4\int\int_{\{(x,y):x+y\leq\tau_C\}} F_{X^0,Y^0}(x-,y-)F_{X^0,Y^0}(\mathrm{d}x,\mathrm{d}y)/\{F_{Z^0,\boldsymbol{V}^0}(\tau_C,\boldsymbol{\infty})\}^2 - 1$ when (X^0, Y^0) are both continuous positive random variables (Wang and Wells, 2000). We can estimate τ^* by $\hat{\tau}^* = 4\int\int_{\{(x,y):x+y\leq\tau_C\}} \hat{F}_{X^0,Y^0}(x-,y-)\hat{F}_{X^0,Y^0}(\mathrm{d}x,\mathrm{d}y)/\{\hat{F}_{Z^0,\boldsymbol{V}^0}(\tau_C,\boldsymbol{\infty})\}^2 - 1$. The weak convergence of $\sqrt{n}(\hat{\tau}^* - \tau^*)$ follows from that of $n^{1/2}(\hat{F}_{X^0,Y^0}(x,y) - F_{X^0,Y^0}(x,y))$ by the functional delta method. Nevertheless, in simulation studies, we found that the measure is influenced substantially by censoring. For example, even when there is no correlation between the first event time and the following gap times, the restricted Kendall's τ presents negative correlation in presence of censoring. We believe this is due to "artificial negative association", namely, when a study has short follow-up time, a long observed first gap time is usually associated with a shorter second gap time. Thus,

even in the presence of positive correlation between true (i.e., uncensored) gap times, the restricted Kendall's τ can be reversed and show an artificial negative association. In general, the restricted Kendall's τ tends to underestimate the true level of positive correlation. Therefore, an alternative nonparametric method to measure association between gap times must be developed.

5.2.2 Joint modeling multiple recurrent event processes with informative censoring

In the post-transplant infection data introduced in Section 1.3.1, multiple types of infections such as bacterial, viral, fungal, and others are documented for transplant recipients. In the models postulated in this thesis, each type of infection process is considered independently, while in practice they may be correlated to one another. It is attractive to researchers to model multiple recurrent event processes jointly.

Recently, several studies have been conducted on modeling multiple recurrent event processes simultaneously in the presence of informative censoring through multivariate frailty. Mazroui et al. (2013) proposed joint models of two types of recurrent events with a terminal event based on the intensity functions. Zeng et al. (2014) studied joint models to analyze multiple recurrent events which are subject to multiple terminal events, where the recurrent events are based on an intensity model and proportional hazards models are postulated for informative censoring times. These methods can handle multiple types of infection processes and also account for multiple informative censoring events such as relapse, a second transplant, or death. However, these models ignore the potential difference of the first event time distribution and the distribution of the recurrent gap times between consecutive infections. Huang and Liu (2007) proposed a joint proportional hazards frailty model where the correlation between the recurrent event process and the terminal event are characterized through frailties following a parametric distribution. In the model, the covariates effects are allowed to be specific to each gap time. That is, effects of the covariates on time to the first event and on gap times between two successive infections can be estimated. The model considers a single recurrent event process with a single terminal event. Extension of the model to accommodate multiple recurrent event processes is warranted. Developing a joint model based on the accelerated failure time model would be another future research direction.

References

Amaddeo, F., Beecham, J., Bonizzato, P., Fenyo, A., Tansella, M., and Knapp, M. (1998). The costs of community-based psychiatric care for first-ever patients: a case register study. *Psychological Medicine*, 28(01):173–183.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer.

Barker, J. N., Hough, R. E., van Burik, J.-A. H., DeFor, T. E., MacMillan, M. L., O'Brien, M. R., and Wagner, J. E. (2005). Serious infections after unrelated donor transplantation in 136 children: impact of stem cell source. *Biology of Blood and Marrow Transplantation*, 11(5):362–370.

Billingsley, P. (1999). *Convergence of probability Measures*. New York: Wiley.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453.

Chang, S.-H. (2004). Estimating marginal effects in accelerated failure time models for serial sojourn times among repeated events. *Lifetime data analysis*, 10(2):175–190.

Chiou, S. H., Kang, S., Kim, J., and Yan, J. (2014). Marginal semiparametric multivariate accelerated failure time model with generalized estimating equations. *Lifetime data analysis*, 20(4):599–618.

Cook, R. J. and Lawless, J. F. (2007). *The statistical analysis of recurrent events*. New York: Springer.

de Uña-Álvarez, J. and Meira-Machado, L. F. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters*, 78(15):2440–2445.

Huang, C.-Y. and Wang, M.-C. (2005). Nonparametric estimation of the bivariate recurrence time distribution. *Biometrics*, 61(2):392–402.

Huang, X. and Liu, L. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics*, 63(2):389–397.

Huang, Y. (2000). Two-sample multistate accelerated sojourn times model. *Journal of the American Statistical Association*, 95(450):619–627.

Huang, Y. (2002). Censored regression with the multistate accelerated sojourn times model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1):17–29.

Huang, Y. and Chen, Y.-Q. (2003). Marginal regression of gaps between recurrent events. *Lifetime data analysis*, 9(3):293–303.

Huang, Y. and Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*, 85(4):785–798.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353.

Johnson, L. M. and Strawderman, R. L. (2009). Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. *Biometrika*, 96(3):577–590.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. Hoboken, NJ: J. Wiley.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. New York: Springer.

Lakhal-Chaieb, L., Cook, R. J., and Lin, X. (2010). Inverse probability of censoring weighted estimates of kendall's τ for gap time analyses. *Biometrics*, 66(4):1145–1152.

Lawless, J. F. and Yilmaz, Y. E. (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, 53(5):779–796.

Lin, D. Y., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*, 86(1):59–70.

Lin, D. Y. and Ying, Z. (2001). Nonparametric tests for the gap time distributions of serial events based on censored data. *Biometrics*, 57(2):369–375.

Lu, W. (2005). Marginal regression of multivariate event times based on linear transformation models. *Lifetime data analysis*, 11(3):389–404.

Luo, X. and Huang, C.-Y. (2011). Analysis of recurrent gap time data using the weighted risk-set method and the modified within-cluster resampling method. *Statistics in medicine*, 30(4):301–311.

Luo, X., Huang, C.-Y., and Wang, L. (2013). Quantile regression for recurrent gap time data. *Biometrics*, 69(2):375–385.

Mazroui, Y., Mathoulin-Pélissier, S., MacGrogan, G., Brouste, V., and Rondeau, V. (2013). Multivariate frailty models for two types of recurrent events with a dependent terminal event: Application to breast cancer data. *Biometrical Journal*, 55(6):866–884.

Pena, E. A., Strawderman, R. L., and Hollander, M. (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association*, 96(456):1299–1315.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Saavedra, S., Sanz, G. F., Jarque, I., Moscardo, F., Jimenez, C., Lorenzo, I., Martin, G., Martinez, J., De la Rubia, J., Andreu, R., et al. (2002). Immune reconstitution/early infections in adult patients undergoing unrelated donor cord blood transplantation. *Bone Marrow Transplantation*, 30(12):937–943.

Schaubel, D. E. and Cai, J. (2004). Non-parametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in medicine*, 23(12):1885–1900.

Strawderman, R. L. (2005). The accelerated gap times model. *Biometrika*, 92(3):647–666.

Sturt, E., Wykes, T., and Creer, C. (1982). Demographic, social and clinical characteristics of the sample. *Psychological Medicine (Monograph Supplement)*, 2:5–14.

Sun, L., Park, D.-H., and Sun, J. (2006). The additive hazards model for recurrent gap times. *Statistica Sinica*, 16(3):919–932.

Tansella, M. (1991). *Community-based psychiatry: long-term patterns of care in South-Verona*, volume 19. Cambridge University Press.

Tansella, M., Micciolo, R., Biggeri, A., Bisoffi, G., and Balestrieri, M. (1995). Episodes of care for first-ever psychiatric patients. a long-term case-register evaluation in a mainly urban area. *The British Journal of Psychiatry*, 167(2):220–227.

van Burik, J.-A. and Weisdorf, D. J. (1999). Infections in recipients of blood and marrow transplantation. *Hematology/oncology clinics of North America*, 13(5):1065–1089.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3. Cambridge University Press.

Visser, M. (1996). Nonparametric estimation of the bivariate survival function with an application to vertically transmitted aids. *Biometrika*, 83(3):507–518.

Wang, M.-C. and Chang, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, 94(445):146–153.

Wang, M.-C. and Chen, Y.-Q. (2000). Nonparametric and semiparametric trend analysis for stratified recurrence times. *Biometrics*, 56(3):789–794.

Wang, W. and Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika*, 85(3):561–572.

Wang, W. and Wells, M. T. (2000). Estimation of kendall's tau under censoring. *Statistica Sinica*, 10(4):1199–1215.

Yazaki, M., Atsuta, Y., Kato, K., Kato, S., Taniguchi, S., Takahashi, S., Ogawa, H., Kouzai, Y., Kobayashi, T., Inoue, M., et al. (2009). Incidence and risk factors of early bacterial infections after unrelated cord blood transplantation. *Biology of Blood and Marrow Transplantation*, 15(4):439–446.

Zeng, D., Ibrahim, J. G., Chen, M.-H., Hu, K., and Jia, C. (2014). Multivariate recurrent events in the presence of multivariate informative censoring with applications to bleeding and transfusion events in myelodysplastic syndrome. *Journal of Biopharmaceutical Statistics*, 24(2):429–442.

Appendix A

Proof of Theorems

A.1 The weak convergence of $\hat{F}^*_{Z,\boldsymbol{V}}$ and $\hat{R}$

Weak convergences of $\hat{F}^*_{Z,\boldsymbol{V}}$ and $\hat{R}$ are required to derive the large sample properties of $\hat{\Lambda}$. Since $\hat{F}^*_{Z,\boldsymbol{V}}$ and $\hat{R}$ are moment type estimators, the stochastic processes $\sqrt{n}\{\hat{F}^*_{Z,\boldsymbol{V}}(t,\boldsymbol{u})-F_{Z,\boldsymbol{V}}(t,\boldsymbol{u})\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{I(m_i\geq 2)}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij}\leq t, \boldsymbol{V}_{ij}\leq \boldsymbol{u}) - F_{Z,\boldsymbol{V}}(t,\boldsymbol{u})\right\}$ and $\sqrt{n}\{\hat{R}(t) - S_Z(t)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij}\geq t) - S_Z(t)\right\}$ converge weakly to Gaussian processes from the Central Limit Theorem. The tightness of the processes is shown by the following inequalities,

$$
\begin{aligned}
&n^2\mathrm{E}[\{\hat{F}^*_{Z,\boldsymbol{V}}(t,\boldsymbol{u}) - F_{Z,\boldsymbol{V}}(t,\boldsymbol{u}) - \hat{F}^*_{Z,\boldsymbol{V}}(t^{'},\boldsymbol{u}^{'}) + F_{Z,\boldsymbol{V}}(t^{'},\boldsymbol{u}^{'})\}^2 \times \\
&\{\hat{F}^*_{Z,\boldsymbol{V}}(t^{''},\boldsymbol{u}^{''}) - F_{Z,\boldsymbol{V}}(t^{''},\boldsymbol{u}^{''}) - \hat{F}^*_{Z,\boldsymbol{V}}(t,\boldsymbol{u}) + F_{Z,\boldsymbol{V}}(t,\boldsymbol{u})\}^2] \\
&\leq \mathit{constant} \times (F_{Z,\boldsymbol{V}}(t,\boldsymbol{u}) - F_{Z,\boldsymbol{V}}(t^{'},\boldsymbol{u}^{'}))(F_{Z,\boldsymbol{V}}(t^{''},\boldsymbol{u}^{''}) - F_{Z,\boldsymbol{V}}(t,\boldsymbol{u})) \text{ and} \\
&n^2\mathrm{E}[\{\hat{R}(t) - S_Z(t) - \hat{R}(t^{'}) + S_Z(t^{'})\}^2\{\hat{R}(t^{''}) - S_Z(t^{''}) - \hat{R}(t) + S_Z(t)\}^2] \\
&\leq \mathit{constant} \times (S_Z(t) - S_Z(t^{'}))(S_Z(t^{''}) - S_Z(t)),
\end{aligned}
$$

for $0 \leq t^{'} \leq t \leq t^{''} \leq L$, $0 \leq u_1^{'} \leq u_1 \leq u_1^{''} \leq L$, and $0 \leq u_2^{'} \leq u_2 \leq u_2^{''} \leq L$ (Billingsley, 1999). Let $\mathcal{S}(\Omega)$ denote the space of bivariate right-continuous functions on Ω with left-hand limits, and $\mathcal{S}^{-}([0,L])$ denote the space of left-continuous functions on $[0,L]$ with right-hand limits. Then, $\sqrt{n}(\hat{F}^*_{Z,\boldsymbol{V}}(t,\boldsymbol{u}) - F_{Z,\boldsymbol{V}}(t,\boldsymbol{u}), \hat{R}(t) - S_Z(t))$ is tight on $\mathcal{S}(\Omega)\times\mathcal{S}^{-}([0,L])$, and converges weakly to a zero mean bivariate Gaussian process.

A.2 Proof of Theorem 2.1

Since $(F_{Z,\boldsymbol{V}}, S_Z)$ satisfies weak convergence and the mapping from $(F_{Z,\boldsymbol{V}}, S_Z)$ to Λ is continuous and compactly differentiable with respect to the supremum norm at a given $(F_{Z,\boldsymbol{V}}, S_Z)$, we apply the functional delta method (van der Vaart, 1998, Chapter 20) to $\sqrt{n}(\hat{\Lambda}(t,\boldsymbol{u}) - \lambda(t,\boldsymbol{u}))$ to obtain the asymptotically i.i.d. representation.

$$
\begin{aligned}
&\sqrt{n}\{\hat{\Lambda}(t,\boldsymbol{u}) - \Lambda(t,\boldsymbol{u})\} \\
&= \sqrt{n}\left\{\int_0^t \frac{\hat{F}^*_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u})}{\hat{R}(s)} - \int_0^t \frac{F_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u})}{S_Z(s)}\right\} \\
&= \int_0^t \frac{\sqrt{n}\{\hat{F}^*_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u}) - F_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u}))\}}{S_Z(s)} - \int_0^t \frac{\sqrt{n}\{\hat{R}(s) - S_Z(s)\}}{S_Z^2(s)} F_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u}) + o_p(1) \\
&= \int_0^t \frac{\sqrt{n}\hat{F}^*_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u})}{S_Z(s)} - \int_0^t \frac{\sqrt{n}\hat{R}(s)}{S_Z^2(s)} F_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u}) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i(t,\boldsymbol{u}) + o_p(1),
\end{aligned}
$$

where $\psi_i(t,\boldsymbol{u}) = \frac{I(m_i \geq 2)}{m_i^*}\sum_{j=1}^{m_i^*} \frac{I(Z_{ij} \leq t, \boldsymbol{V}_{ij} \leq \boldsymbol{u})}{S_Z(Z_{ij})} - \int_{[0,t]} \frac{1}{m_i^*}\sum_{j=1}^{m_i^*} I(Z_{ij} \geq s)\frac{F_{Z,\boldsymbol{V}}(\mathrm{d}s,\boldsymbol{u})}{S_Z^2(s)}$.

By the Central Limit Theorem, weak convergence of $\sqrt{n}\{\hat{\Lambda}(t,\boldsymbol{u}) - \Lambda(t,\boldsymbol{u})\}$ follows. In addition, tightness of the representation can be shown from the weak convergence of $\sqrt{n}\{\hat{F}^*_{Z,\boldsymbol{V}}(t,\boldsymbol{u}) - F_{Z,\boldsymbol{V}}(t,\boldsymbol{u}), \hat{R}(t) - S_Z(t)\}$ adopting the arguments in Breslow and Crowley (1974). Therefore, the asymptotically i.i.d. representation converges weakly to a Gaussian process.

A.3 Proof of Theorem 2.2

We use the asymptotic properties of $\hat{\Lambda}$ in Theorem 2.1 and the compact differentiability of the mapping $\Phi : \Lambda \to F_{Z^0,\boldsymbol{V}^0}$ with respect to the supremum norm at a given Λ to derive the large sample properties. Applying the functional delta method to

$\sqrt{n}\{\Phi(\hat{\Lambda})(t,\boldsymbol{u}) - \Phi(\Lambda)(t,\boldsymbol{u})\}$, we derive

$$
\begin{aligned}
&\sqrt{n}\{\hat{F}^{*}_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u}) - F_{Z^0,\boldsymbol{V}^0}(t,\boldsymbol{u})\} \\
&= \sqrt{n}\{\Phi(\hat{\Lambda}) - \Phi(\Lambda)\} \\
&= \mathrm{d}\Phi_{\Lambda}(\sqrt{n}\{\hat{\Lambda} - \Lambda\})(t,\boldsymbol{u}) + o_p(1) \\
&= \int_0^t \{F_{Z^0,\mathbf{V}^0}(s,\boldsymbol{u}) - F_{Z^0,\mathbf{V}^0}(t,\boldsymbol{u})\}\sqrt{n}\{\hat{\Lambda}(\mathrm{d}s,\boldsymbol{\infty}) - \Lambda(\mathrm{d}s,\boldsymbol{\infty})\} \\
&+ \int_0^t S_{Z^0}(s)\sqrt{n}\{\hat{\Lambda}(\mathrm{d}s,u) - \Lambda(\mathrm{d}s,u)\} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\int_0^t F_{Z^0,\mathbf{V}^0}(s,\boldsymbol{u})\psi_i(\mathrm{d}s,\boldsymbol{\infty}) + \int_0^t S_{Z^0}(s)\psi_i(\mathrm{d}s,\boldsymbol{u}) - F_{Z^0,\mathbf{V}^0}(t,\boldsymbol{u})\psi_i(t,\boldsymbol{\infty})\right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_i(t,\boldsymbol{u}) + o_p(1),
\end{aligned}
$$

where $\phi_i(t,\boldsymbol{u}) = \int_{[0,t]} F_{Z^0,\mathbf{V}^0}(s,\boldsymbol{u})\psi_i(\mathrm{d}s,\boldsymbol{\infty}) + \int_{[0,t]} S_{Z^0}(s)\psi_i(\mathrm{d}s,\boldsymbol{u}) - F_{Z^0,\mathbf{V}^0}(t,\boldsymbol{u})\psi_i(t,\boldsymbol{\infty})$.

The weak convergence of $\sqrt{n}\{\hat{F}^{*}_{Z^0,\mathbf{V}^0}(t,\boldsymbol{u}) - F_{Z^0,\mathbf{V}^0}(t,\boldsymbol{u})\}$ follows from the Central Limit Theorem and the tightness follows from the weak convergence of $\sqrt{n}\{\hat{\Lambda}(t,\boldsymbol{u}) - \Lambda(t,\boldsymbol{u})\}$ by adopting similar arguments in Breslow and Crowley (1974). This completes the proof of Theorem 2.2.

A.4 Proof of Theorem 2.3

We have the asymptotically i.i.d. representation for $\sqrt{n}\{\hat{F}_{Y^0|X^0}(u_2|u_1) - F_{Y^0|X^0}(u_2|u_1)\}$:

$$
\begin{aligned}
&\sqrt{n}\{\hat{F}_{Y^0|X^0}(u_2|u_1) - F_{Y^0|X^0}(u_2|u_1)\} \\
&= \sqrt{n}\left\{\frac{\hat{F}_{X^0,Y^0}(u_1,u_2)}{1-\hat{S}_{X^0}(u_1)} - \frac{F_{X^0,Y^0}(u_1,u_2)}{1-S_{X^0}(u_1)}\right\} \\
&= \frac{1}{1-S_{X^0}(u_1)}\sqrt{n}\left\{\hat{F}_{X^0,Y^0}(u_1,u_2) - F_{X^0,Y^0}(u_1,u_2)\right\} \\
&+ \frac{F_{X^0,Y^0}(u_1,u_2)}{\{1-S_{X^0}(u_1)\}^2}\sqrt{n}\left\{\hat{S}_{X^0}(u_1) - S_{X^0}(u_1)\right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{1}{1-S_{X^0}(u_1)}\phi_i(t,\boldsymbol{u}) - \frac{F_{X^0,Y^0}(u_1,u_2)}{\{1-S_{X^0}(u_1)\}^2}\phi_i'(u_1)\right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i(t,\boldsymbol{u}) + o_p(1),
\end{aligned}
$$

where $\xi_i(t,\boldsymbol{u}) = \phi_i(t,\boldsymbol{u})/\{1-S_{X^0}(u_1)\} - F_{X^0,Y^0}(u_1,u_2)\phi_i'(u_1)/\{1-S_{X^0}(u_1)\}^2$ and $\phi_i'(u_1) = S_{X^0}(u_1)\left\{\frac{I(X_i\le C_i)I(X_i\le u_i)}{S_X(X_i)} - \int_0^{u_1}\frac{I(X_i\ge s)}{S_X^2(s)}\mathrm{d}F_X(s)\right\}$. The weak convergence and the tightness of $\sqrt{n}\{\hat{F}_{Y^0|X^0}(u_2|u_1) - F_{Y^0|X^0}(u_2|u_1)\}$ can be established by the Central Limit Theorem and from the weak convergence of $\sqrt{n}\{\hat{F}_{X^0,Y^0}(u_1,u_2) - F_{X^0,Y^0}(u_1,u_2)\}$ and $\sqrt{n}\{\hat{S}_{X^0}(u_1) - S_{X^0}(u_1)\}$, respectively, by following arguments in Breslow and Crowley (1974).